PARP 4 digital
Grupa PFR

Generatywna AI w badaniach. Praktyczne zastosowania w ewaluacji polityk publicznych

2025

Redakcja:
Karol Olejniczak, Dominik Batorski, Jacek Pokorski

Generatywna AI w badaniach. Praktyczne zastosowania w ewaluacji polityk publicznych

Warszawa 2025

Olejniczak, K., Batorski, D., Pokorski, J. (red.). (2025). *Generatywna AI w badaniach. Praktyczne zastosowania w ewaluacji polityk publicznych [Generative AI in research. Practical applications in public policy evaluation]*. Polska Agencja Rozwoju Przedsiębiorczości. DOI: 10.58142/swps-gr-02/2025

Redakcja merytoryczna:
- Karol Olejniczak, Uniwersytet SWPS
- Dominik Batorski, Uniwersytet Warszawski
- Jacek Pokorski, Polska Agencja Rozwoju Przedsiębiorczości

Autorzy
- Dominik Batorski, Uniwersytet Warszawski
- Kamil Filipek, Uniwersytet Marii Curie-Skłodowskiej
- Andrzej Gołoś, MCM Institute Poland
- Paulina Grabowska, NASDRA, Uniwersytet SWPS
- Andrzej Jędrzejowski, Polska Agencja Rozwoju Przedsiębiorczości
- Paweł Kędzia, Research And Development Laboratory
- Tomasz Kupiec, Uniwersytet Warszawski, Polskie Towarzystwo Ewaluacyjne
- Bartosz Ledzion, EGO – Evaluation for Government Organizations
- Marta Lesiak, Polska Agencja Rozwoju Przedsiębiorczości
- Igor Lyubashenko, Uniwersytet SWPS
- Karol Olejniczak, Uniwersytet SWPS
- Jacek Pokorski, Polska Agencja Rozwoju Przedsiębiorczości
- Grzegorz Rzeźnik, Uniwersytet SWPS
- Jacek Szut, Polska Agencja Rozwoju Przedsiębiorczości
- Dominika Wojtowicz, Akademia Leona Koźmińskiego
- Teresa Wyszyńska, Polska Agencja Rozwoju Przedsiębiorczości

Redakcja językowa
- Renata Włostowska

Redakcja techniczna i skład
- Przemysław Biliczak, Studio IMPRESO

Projekt graficzny
- Polska Agencja Rozwoju Przedsiębiorczości

Publikacja powstała w ramach projektu współfinansowanego z Europejskiego Funduszu Rozwoju Regionalnego w ramach programu Fundusze Europejskie dla Nowoczesnej Gospodarki, 2021-2027.

Spis treści

Słowo wstępne 7

CZĘŚĆ I

Kwestie horyzontalne współpracy z gen AI 9

1. Wprowadzenie do zagadnień zastosowania gen AI w ewaluacji 10
1.1. Zasady działania i wykorzystania gen AI 12
1.2. Spektrum typów gen AI 20
1.3. Spektrum zastosowań gen AI w ewaluacji 23
1.4. Struktura książki 32

2. Promptowanie, czyli jak komunikować się z modelem językowym 35
2.1. Wstęp 35
2.2. Wyzwania po stronie gen AI 36
2.3. Wyzwania po stronie użytkownika 38
2.4. Komunikacja z gen AI 41
2.5. Podsumowanie 62

3. BHP pracy z gen AI – bezpieczeństwo danych i poufność 64
3.1. Wprowadzenie 64
3.2. Udostępnienie przekształconych danych 67
3.3. Identyfikacja danych do przekształcenia 70
3.4. Wykorzystanie narzędzi gen AI jako platformy *low-code* 73
3.5. Wykorzystanie modeli LLM działających na lokalnej infrastrukturze 79
3.6. Podsumowanie 81

4. Integracja modeli gen AI z bazami dokumentów 83
4.1. Wstęp 83
4.2. Jak działa RAG? 84
4.3. Zastosowania RAG w praktyce badawczej 93

4.4. Rozpoczęcie pracy z RAG 95
4.5. Podsumowanie 97

CZĘŚĆ II
Zastosowania gen AI w procesie ewaluacji 99

5. Zamawianie usług ewaluacyjnych – tworzenie Opisu Przedmiotu Zamówienia 100
5.1. Wstęp 100
5.2. Tworzenie OPZ z wykorzystaniem gen AI – ogólne zasady 103
5.3. Podsumowanie 128

6. Analizy jakościowe wspierane gen AI 131
6.1. Wprowadzenie 131
6.2. Specyfika analizy danych jakościowych 132
6.3. Analiza pojedynczego dokumentu 134
6.4. Użycie gen AI w jakościowym kodowaniu indukcyjnym 140
6.5. Użycie gen AI w jakościowym kodowaniu dedukcyjnym 142
6.6. Automatyzacja procesów badawczych z wykorzystaniem gen AI 147
6.7. Podsumowanie 151

7. Przeglądy źródeł wspierane gen AI 154
7.1. Wprowadzenie 154
7.2. Typy przeglądów i rodzaje źródeł w przeglądach 157
7.3. Rodzaje źródeł tekstowych 160
7.4. Dostępność źródeł dla gen AI 165
7.5. Perspektywy zastosowań gen AI w przeglądach źródeł 168
7.6. Wnioski końcowe 180

8. Analizy ilościowe wspierane gen AI 185
8.1. Wprowadzenie: ewolucja dużych modeli językowych 185
8.2. Wybór środowiska analitycznego 186
8.3. Ocena i przygotowanie danych wejściowych 189
8.4. Analiza danych 196
8.5. Ocena wyników 203
8.6. Zakończenie 211

9. Analizy sieciowe wspierane gen AI 213
9.1. Wprowadzenie 213
9.2. Pozyskanie danych 215
9.3. Analiza pozyskanego zbioru danych. 219
9.4. Analizy sieciowe z użyciem modeli gen AI. 221
9.5. Podsumowanie 241

10. Strategie komunikacji i prezentacji wyników. 243
10.1. Wprowadzenie 243
10.2. Wykorzystanie generatywnej AI w dostosowywaniu treści raportów ewaluacyjnych – krok po kroku. 247
10.3. Podsumowanie. 268
10.4. Zakończenie 270

CZĘŚĆ III

Gen AI w administracji publicznej – implikacje systemowe 272

11. Gen AI w organizacjach publicznych – regulacje i *codes of conducts* 273
11.1. Wprowadzenie 273
11.2. Zbadana populacja dokumentów 275
11.3. Zidentyfikowane zasady, wytyczne i wskazówki 279
11.4. Uwagi końcowe 289

12. Perspektywy adaptacji organizacyjnych i systemowych wykorzystania AI w ewaluacji 292
12.1. Wstęp 292
12.2. Czynniki blokujące transformację ewaluacji w kierunku wykorzystania gen AI . . 294
12.3. Czynniki katalityczne dla transformacji ewaluacji w kierunku wykorzystania gen AI 302
12.4. Zakończenie 307

13. Zakończenie 310
13.1. Odkrywanie potencjału gen AI 310
13.2. Architektura współpracy z gen AI 311
13.3. Szersze implikacje płynące z zastosowań gen AI 313

13.4. Refleksja przy eksperymentowaniu – pytania otwarte 315

14. Autorzy 321

15. Aneksy 326

Aneks 1. Informacje o projektach, w których testowaliśmy użycie gen AI 326

Aneks 2. Zaplanowanie przeglądu 327

Aneks 3. Szukanie i selekcja źródeł 329

Aneks 4. Analiza i synteza źródeł 335

Aneks 5. Metodyka wyszukiwania i analizy dokumentów 337

Słowo wstępne

Szanowni Państwo!

Z dużą przyjemnością przekazujemy Państwu publikację pt. **„Generatywna AI w badaniach. Praktyczne zastosowania w ewaluacji polityk publicznych”**, przygotowaną przez ekspertów Polskiej Agencji Rozwoju Przedsiębiorczości i Uniwersytetu SWPS. W publikacji zaprezentowano zagadnienia związane z wykorzystaniem generatywnych modeli sztucznej inteligencji (gen AI) w szeroko rozumianych pracach analityczno-badawczych instytucji. Szczególny nacisk kładziemy przy tym na potencjalne zastosowania gen AI w ewaluacji polityk publicznych, w tym ewaluacji programu Fundusze Europejskie dla Nowoczesnej Gospodarki 2021–2027 (oraz jego następcy na lata 2028–2034).

Publikacja ma formę przewodnika przez świat gen AI – od informacji podstawowych, po konkretne przykłady wykonywania zadań ze wsparciem sztucznej inteligencji, omawiające napotykane wyzwania i sposoby radzenia sobie z nimi. W poszczególnych rozdziałach przedstawiono zasady działania gen AI, opisano metody efektywnej komunikacji z modelami językowymi i przykłady ich zastosowań w analizach ilościowych i jakościowych. Zwrócono również uwagę na kwestie dotyczące bezpieczeństwa danych i rzetelności wyników, a także zasygnalizowano etyczny kontekst wykorzystania sztucznej inteligencji. Ponadto, w publikacji poruszony został temat regulacji wewnętrznych oraz niezbędnych adaptacji organizacyjnych – pilnych do przygotowania i wprowadzenia w życie w najbliższych miesiącach i latach – które według Autorów są kluczowe, aby skutecznie wdrażać rozwiązania oparte na sztucznej inteligencji w instytucjach publicznych.

Choć niniejsze opracowanie skupia się na ewaluacji, z pewnością dostarczy także wartościowych wskazówek osobom, które na co dzień zajmują się innymi badaniami eksploracyjnymi i aplikacyjnymi w instytucjach, w szczególności służącym pokrewnym dziedzinom, takim jak planowanie strategiczne, projektowanie usług, selekcja projektów, monitoring programów czy szeroko pojęta analityka i raportowanie zarządcze.

Liczę, że publikacja stanie się ważnym wkładem w debatę na temat wyzwań – choćby takich jak przejrzystość działań, ochrona danych osobowych czy efektywne zarządzanie

zasobami – które stawia przed instytucjami publicznymi dynamiczny rozwój technologii cyfrowych. Naturalnymi stają się pytania dotyczące m.in. koordynacji wykorzystania sztucznej inteligencji w sektorze publicznym czy określenia odpowiedzialności za wykorzystywane aplikacje i procesy oparte na sztucznej inteligencji. Wyzwaniem jest również wypracowanie efektywnego zarządzania narzędziami AI oraz stworzenie odpowiednich regulacji wewnętrznych, określających zasady korzystania ze sztucznej inteligencji przez pracowników instytucji publicznych i podmioty współpracujące (np. beneficjentów lub wykonawców usług na rzecz sektora publicznego).

Wierzę, że publikacja stanowić będzie inspirującą lekturę oraz wesprze nie tylko specjalistów, ale również osoby decyzyjne, w pracy nad odpowiedzialnym i efektywnym wykorzystaniem gen AI w naszych instytucjach.

Z życzeniami ciekawej lektury,

Joanna Zembaczyńska-Świątek
Zastępca Prezesa Polskiej Agencji Rozwoju Przedsiębiorczości

CZĘŚĆ I

Kwestie horyzontalne współpracy z gen AI

1. Wprowadzenie do zagadnień zastosowania gen AI w ewaluacji

Dominik Batorski
Karol Olejniczak
Jacek Pokorski

Generatywna sztuczna inteligencja (gen AI) jest technologią ogólnego przeznaczenia (ang. *general purpose technology*), co oznacza innowację technologiczną, która ma szeroki, trwały i znaczący wpływ na różne sektory gospodarki oraz społeczeństwo jako całość (Przegalińska, Triantoro, 2024). Jednocześnie innowacja ta rozwija się i podlega upowszechnieniu w tempie bezprecedensowym w historii ludzkości (Mollick, 2024). Wysoce zasadne jest więc szybkie podjęcie krytycznej refleksji nad tym, jak ta nowa technologia może nam pomóc w realizowaniu różnego rodzaju zadań, istotnych w codziennej pracy analityczno-badawczej i ewaluacyjnej.

Ta książka jest pierwszą polską, względnie kompleksową refleksją nad zastosowaniami nowej technologii w praktyce ewaluacji polityk publicznych, osadzoną w krajowym kontekście i opartą na kiełkujących praktykach z gen AI polskiego sektora publicznego.

Główna teza naszej publikacji jest następująca: **Generatywna sztuczna inteligencja otwiera nowe możliwości skutecznego realizowania naszej misji, którą jest doskonalenie rozwiązań publicznych służących społeczeństwu poprzez rzetelne badania ewaluacyjne. Wykorzystanie potencjału gen AI wymaga jednak systematycznego eksperymentowania z jej praktycznymi zastosowaniami oraz świadomej analizy zarówno jej możliwości, jak i ograniczeń.**

Główną grupą docelową, do której kierujemy ten tom, są osoby zlecające i prowadzące badania aplikacyjne polityk publicznych, przede wszystkim przedstawiciele jednostek ewaluacyjnych i osoby realizujące różnego typu prace analityczno-badawcze w instytucjach.

Temat zastosowań sztucznej inteligencji w badaniach społecznych i polityce publicznej jest bardzo szeroki i wielowymiarowy, a nowe praktyki, dyskusje i dylematy pojawiają się praktycznie każdego tygodnia. Na potrzeby tej publikacji dokonaliśmy więc trzech strategicznych wyborów – koncentrujemy się na zagadnieniach możliwych do zorganizowania (zarządzania zmianą, nowym procesem włączającym gen AI) w codziennej pracy analityczno-badawczej instytucji i przyjmujemy praktyczne ramy narracji, a pomijamy nieco niektóre bardziej filozoficzne kwestie współczesnych debat o AI.

W książce skupiamy się przede wszystkim na modelach językowych i wielomodalnych. Modele gen AI obejmują generowanie różnorodnych form treści, takich jak tekst, obraz, film czy głos, ale uznaliśmy, że to modele językowe, które już dziś rewolucjonizują pracę badaczy, stwarzają potencjalnie największe możliwości w interesującym nas obszarze ewaluacji oraz polityk publicznych. Czytelnikom zainteresowanym szerszym spektrum zastosowań AI w politykach publicznych polecamy inne publikacje, o bardziej przekrojowym profilu (Charalabidis et al., 2024; Nielsen et al., 2025).

Do analizy zastosowań gen AI przyjęliśmy ramę jednostek ewaluacyjnych i struktur analityczno-badawczych działających jako „brokerzy wiedzy” (szerszej piszemy o tym w dalszej części tego rozdziału). Taka perspektywa pozwoliła nam przedstawić praktyczne zastosowania gen AI zarówno w zadaniach badawczych, jak i w działaniach związanych z przekazywaniem i upowszechnianiem wyników, w tym skuteczną komunikacją z decydentami.

W pełni dostrzegamy wagę dyskusji o etyce, sprawiedliwości algorytmicznej, wpływie AI na przejrzystość administracji czy na formułowanie sądów wartościujących względem polityk publicznych, jednak w niniejszej publikacji skupiamy się przede wszystkim na wdrażaniu tej technologii do pracy organizacji „uczącej się” (Olejniczak, 2012) i na codziennym wymiarze korzystania z narzędzi gen AI w jednostkach analityczno-badawczych i ewaluacyjnych instytucji. Uważamy, że technologie nie są deterministyczne. Pierwszym krokiem w społecznym kształtowaniu ich użycia jest odkrywanie możliwych zastosowań w kontekście konkretnych zadań i organizacji. Dopiero takie praktyczne doświadczenia są dobrą bazą do identyfikowania dylematów, wyzwań i konstruktywnej dyskusji na temat standardów. Ta publikacja pomaga więc postawić ten pierwszy krok – przybliżamy Czytelnikom technologię gen AI, mając nadzieję, że te doświadczenia zainspirują do stawiania szerszych pytań o rolę i zasady używania gen AI w polityce publicznej oraz jej ewaluacji. Na zakończenie książki

sygnalizujemy kilka kwestii, które naszym zdaniem powinny być przedmiotem refleksji i dyskusji na szerszych forach.

W pozostałej części niniejszego wprowadzenia koncentrujemy się na wyjaśnieniu, czym jest gen AI, na jakich zasadach działa i jakie typy asystentów AI pojawiły się w ostatnim czasie na rynku. Dalej przywołujemy wyniki eksperymentu PARP, który zelektryzował w 2024 r. środowisko ewaluacyjne w Polsce i stanowił również jedną z przesłanek zainicjowania prac nad niniejszą publikacją. Następnie zakreślamy możliwe pola zastosowań gen AI w pracy jednostek ewaluacyjnych i ewaluatorów, osadzając to w szerszej perspektywie brokeringu wiedzy w politykach publicznych. Pod koniec tego rozdziału omawiamy strukturę książki i zawartość poszczególnych rozdziałów.

1.1. Zasady działania i wykorzystania gen AI

Duże modele językowe (LLMs) stanowią fascynujący przykład postępów w nauczaniu maszyn rozumienia i generowania języka. Choć wciąż nie osiągnęły poziomu ogólnej ludzkiej inteligencji, w niektórych dziedzinach zaczynają jej dorównywać, a kolejne przełomy technologiczne znacząco zwiększają ich skuteczność oraz zakres zastosowań. Szybkie tempo innowacji sprawia, że LLM-y nie tylko redefiniują interakcje człowieka z maszyną, ale także poszerzają granice tego, co uznawano za możliwe w przetwarzaniu języka naturalnego (Plaat i inni, 2024; Liu i inni, 2025).

Efektywne korzystanie z dużych modeli językowych wymaga dobrego zrozumienia tego, w jaki sposób są one tworzone (trenowane) i jak działają, ponieważ aspekty te determinują zarówno potencjał, jak i ograniczenia ich praktycznego wykorzystania. Wymóg ten jest związany z niedojrzałością samej technologii, lecz podobnie było z innymi przełomowymi wynalazkami. Choć teraz nie musimy rozumieć technicznych podstaw działania samochodów czy komputerów, aby móc z nich korzystać, to jednak w pierwszych latach od ich powstania, wiedza taka była niezbędna. Znajomość sposobu działania AI pozwala zrozumieć mocne i słabe strony modeli, formułować zapytania w sposób, który maksymalizuje ich potencjał, oraz interpretować otrzymane wyniki z odpowiednim krytycyzmem i z zachowaniem świadomości ich ograniczeń. Znajomość procesu treningu LLM pomaga lepiej zrozumieć źródła potencjalnych błędów czy obciążeń w generowanych odpowiedziach. To jak posiadanie

mapy terenu – im lepiej znamy teren, tym sprawniej możemy się po nim poruszać, unikając pułapek i znajdując optymalne ścieżki dotarcia do celu.

Czym są i jak działają modele językowe

Modele językowe są oparte na złożonych architekturach sieci neuronowych, które „uczą się" wzorców i zależności w ogromnych zbiorach danych tekstowych. Taka sieć przekształca dane, które są „na wejściu" (*input*), na przykład początek tekstu wprowadzony przez użytkownika, w dane „na wyjściu" (*output*), a więc kontynuację tego tekstu lub odpowiedź udzieloną na pytanie użytkownika.

W pierwszej kolejności trzeba wyjaśnić to, jak model „widzi" tekst. Działanie modelu językowego zaczyna się od podzielenia wejściowego tekstu na tokeny – podstawowe jednostki znaczeniowe. Następnie są one przekształcane na język, który rozumieją komputery, czyli język liczb. Jest to tak zwana wektoryzacja. Każde słowo lub jego fragment (token) jest przekształcany w wektor – długą listę cyfr, fachowo nazywaną też embeddingami. Te wektory działają jak DNA słów, kodując ich znaczenie i relacje z innymi słowami. W świecie modelu, wyrazy „kot" i „pies" będą miały podobne wektory, co odzwierciedla ich bliskie semantyczne powiązanie jako zwierząt domowych. Wektory zawierają informacje o ich znaczeniu, ale także o ich relacjach z innymi słowami. Synonimy będą mieć prawie identyczne wartości. Z kolei wektory słów „król" i „królowa" będą podobne, bo oba odnoszą się do władców, ale będą się też nieco różnić, odzwierciedlając różnicę płci. Co ciekawe, jeśli od wektora słowa „król" odejmiemy wektor słowa „mężczyzna" i dodamy wektor słowa „kobieta", otrzymamy wektor bardzo zbliżony do słowa „królowa". To pokazuje, jak wektoryzacja pozwala modelom „rozumieć" i manipulować znaczeniami słów w sposób matematyczny, co jest kluczowe dla ich zdolności do przetwarzania języka naturalnego.

Obecne sukcesy modeli językowych bazują w dużej mierze na wymyślonej i opublikowanej przez Google w 2017 roku architekturze transformerów (Vaswani et al., 2017). Transformery zrewolucjonizowały sposób, w jaki maszyny przetwarzają i „rozumieją" ludzki język, wprowadzając tzw. mechanizm uwagi. Pozwala on modelowi ważyć relacje pomiędzy różnymi słowami w zdaniu podczas generowania reprezentacji dla każdego słowa. Dzięki temu tekst nie jest traktowany jako zbiór słów, ale uwzględniane jest także to, jak poszczególne znaczenia odnoszą się do siebie. Zdolność ta umożliwia modelowi uchwycenie zależności dalekiego

zasięgu i zrozumienie całych sekwencji słów oraz ich kontekstu, co ma kluczowe znaczenie dla zrozumienia większych tekstów. Rozważmy następujące dwa zdania:

> Zdanie A: „Student nie zdał egzaminu, ponieważ był nieprzygotowany".
> Zdanie B: „Student nie zdał egzaminu, ponieważ był trudny".

Pozornie te dwa zdania różnią się tylko jednym słowem – w zdaniu A na końcu mamy „nieprzygotowany", a w zdaniu B – wyraz „trudny". W praktyce jednak różnice dotyczą także tego, do czego odnosi się słowo „był". W pierwszym przypadku to *student był* nieprzygotowany, podczas gdy w drugim to *egzamin był* trudny. Mechanizm uwagi pozwala modelowi wychwycić i uwzględnić takie relacje.

Te wektorowe reprezentacje słów i powiązań między nimi są następnie przetwarzane przez kolejne warstwy sieci neuronowej, której koncepcja i struktura, choć inspirowane ludzkim mózgiem, znacząco się od niego różnią. Sieć neuronowa w LLM składa się z wielu powiązanych ze sobą warstw, z których każda pełni określoną funkcję w procesie „rozumienia" języka. Warstwa wejściowa przyjmuje wektory słów, warstwy ukryte przetwarzają te informacje, wydobywając coraz bardziej złożone i abstrakcyjne cechy, a warstwa wyjściowa generuje końcowy rezultat – przewidując kolejne słowa jedno po drugim, aż do uformowania pełnej odpowiedzi.

Proces przekształcania danych w kolejnych warstwach sieci neuronowej modelu językowego można porównać z pracą zespołu ekspertów analizujących tekst na różnych poziomach. Wyobraźmy sobie, że mamy zdanie: „Kot pije mleko". W pierwszej warstwie, „eksperci" mogą skupić się na pojedynczych słowach, identyfikując słowo „kot" jako rzeczownik, „pije" jako czasownik, a „mleko" jako przedmiot czynności. W kolejnej warstwie inni „eksperci" łączą te informacje, rozpoznając strukturę zdania i relacje między słowami. W jeszcze głębszych warstwach kolejni „specjaliści" mogą wyciągać bardziej abstrakcyjne wnioski – na przykład, że zdanie dotyczy zwyczajów żywieniowych kotów. Każda warstwa przetwarza informacje otrzymane od poprzedniej, dodając swój własny poziom analizy i abstrakcji. To, co zaczyna się jako proste wektory reprezentujące słowa, stopniowo przekształca się w złożone reprezentacje znaczeń, kontekstów i relacji. Na końcu tego procesu ostatnia warstwa jest w stanie wykorzystać wszystkie te przetworzone informacje do generowania odpowiedzi lub przewidywania kolejnych słów w sposób, który uwzględnia nie tylko bezpośredni kontekst, ale także głębsze znaczenia i powiązania.

Generowanie tekstu przez duże modele językowe przypomina nieco grę w zgadywanie następnego słowa, ale na niezwykle zaawansowanym poziomie. Modele te operują na rozkładzie prawdopodobieństwa dla sekwencji słów, co umożliwia im generowanie tekstu poprzez przewidywanie najbardziej prawdopodobnego kolejnego słowa (słów) na podstawie dotychczasowego kontekstu.

Wyobraźmy sobie, że model ma dokończyć zdanie „Kot siedzi na...". W swojej „cyfrowej wyobraźni" model rozważa wszystkie możliwe słowa, które mogłyby się pojawić w tym miejscu – „krześle", „kanapie", „parapecie", „drzewie" i tysiące innych. Dla każdego z tych słów model oblicza prawdopodobieństwo, że właśnie ono powinno znaleźć się w tym miejscu, bazując na całym swoim „doświadczeniu" z treningowych tekstów. To tak, jakby model tworzył listę kandydatów, przypisując każdemu liczbę od 0 do 1, gdzie 1 oznacza absolutną pewność. Słowo „parapecie" może dostać 0,3, „drzewie" 0,1, a „krześle" 0,5.

Model nie wybiera zawsze słowa o najwyższym prawdopodobieństwie, lecz losuje je z rozkładu prawdopodobieństwa, co pozwala na większą różnorodność generowanych tekstów. Proces ten może być modyfikowany przez parametr, który kontroluje stopień losowości wyboru, nazywany temperaturą modelu. Przy niskich wartościach (np. 0,2) model staje się bardziej deterministyczny – wybiera głównie słowa o najwyższym prawdopodobieństwie, co prowadzi do bardziej przewidywalnych i spójnych tekstów. Natomiast wyższa temperatura (np. 1,5) zwiększa rolę losowości, sprawiając, że model częściej wybiera mniej prawdopodobne słowa, co może prowadzić do bardziej kreatywnych, ale też mniej spójnych wyników. Dzięki tej regulacji parametrów modelu można dostosować sposób generowania tekstu w zależności od pożądanego stylu i zastosowania.

Modele językowe są więc podobne do rozwiązań, które znamy z wprowadzania tekstu – różnego rodzaju wirtualnych klawiatur czy autouzupełniania w edytorach tekstu, sugerujących nam dokończenie słowa, które zaczęliśmy wprowadzać, lub podpowiadających, jakie może być kolejne. Zasadnicza różnica polega jednak na tym, że te rozwiązania biorą pod uwagę tylko ostatnie słowo lub kilka słów, natomiast model językowy, przewidując prawdopodobieństwo kolejnego wyrazu, jest w stanie wziąć pod uwagę zarówno sekwencje tysięcy słów, które były wprowadzone wcześniej, jak i dodatkowo zależności między nimi.

Modele językowe nie planują swojej wypowiedzi dużo do przodu. Nie przewidują z góry całych akapitów czy dłuższych fragmentów tekstu. Generują tekst token po tokenie (gdzie

token może być słowem lub jego częścią). Proces generowania ma charakter iteracyjny – jest sekwencyjny i „krótkowzroczny". Model analizuje dotychczasowy kontekst „rozmowy" i ostatnie wprowadzone tokeny. Na tej podstawie przewiduje najbardziej prawdopodobny następny token. Ten token jest dodawany do już wygenerowanego tekstu, a tym samym bierze się go pod uwagę w generowaniu kolejnego tokena. Kroki te są powtarzane aż do zakończenia weryfikacji. Model nie ma więc żadnego „planu" czy „wizji" tego, co chce „powiedzieć". Ma swego rodzaju „wewnętrzny stan", który zawiera informacje o tym, co zostało już powiedziane i jaki jest ogólny kontekst wypowiedzi. Ten stan wpływa na wybór każdego kolejnego tokena. Choć model nie planuje dokładnie całych zdań czy akapitów z wyprzedzeniem, to jego wybory są kształtowane przez ten wewnętrzny stan, co prowadzi zazwyczaj do spójnych i logicznych wypowiedzi. Co więcej, niektóre zaawansowane techniki (np. *beam search*), pozwalają modelowi rozważać kilka możliwych „ścieżek" generowania tekstu jednocześnie i wybrać ostatecznie tę, która wydaje się najlepsza, z uwzględnieniem kilku kolejnych tokenów. Warto również zauważyć, że chociaż model generuje tekst token po tokenie, to jego zrozumienie kontekstu może sięgać daleko wstecz. Dzięki mechanizmowi uwagi, model może odwoływać się do informacji z początku tekstu nawet przy generowaniu jego końcowych fragmentów, co dodatkowo przyczynia się do spójności i logiczności wytworzonego tekstu.

To, co czyni duże modele językowe tak potężnymi narzędziami, to fakt, że są one nauczone języka na podstawie terabajtów treści i tworzą algorytm, który modeluje prawidłowości statystyczne w tych treściach. Można powiedzieć, że model „przeczytał" miliony książek, artykułów, prac naukowych, encyklopedii, forów dyskusyjnych i stron internetowych. Na ich podstawie nauczył się wzorców i relacji pomiędzy milionami słów, dzięki czemu po podaniu początku tekstu jest w stanie sensownie przewidywać ciąg dalszy. „Ucząc się" języka, model zdobywa też wiedzę o świecie, która jest zakodowana w tym, jak języka używamy. Znajomość powiązań między pojęciami i wszystkim, co da się wyrazić za pomocą języka, pozwala modelowi na generowanie tekstu, który wydaje się, a w wielu przypadkach *de facto* jest sensowny. Taki model nie tylko zna słowa i gramatykę, ale także ma szeroką wiedzę ogólną i rozumie subtelności języka. Dlatego właśnie LLM-y potrafią nie tylko tłumaczyć z jednego języka na drugi, ale także pisać poezję, odpowiadać na pytania, czy nawet programować.

Warto jeszcze podkreślić, że trenowanie modeli językowych składa się z dwóch faz. W trakcie szkolenia wstępnego (tzw. *pretraining*) model uczy się na podstawie ogromnego korpusu tekstu przy użyciu nienadzorowanych technik uczenia się, takich jak np. przewidywanie

następnego słowa w zdaniu. Pomaga to modelowi rozwinąć ogólne zrozumienie języka. Następnie model jest dalej dostosowywany (tzw. *fine-tuning*) poprzez uczenie ze wzmocnieniem na podstawie ludzkich informacji zwrotnych (ang. *Reinforcement Learning from Human Feedback*, RLHF). Ten krok dostosowuje parametry modelu, aby dobrze radził sobie z określonymi zadaniami, takimi jak odpowiadanie na pytania lub klasyfikacja tekstu. Modele przechodzą w ten sposób przez setki tysięcy zadań i instrukcji. Analizy pokazują, że tak przygotowane modele znacznie lepiej radzą sobie z typami zadań, które są im znane i których przykłady już „widziały", niż takimi, które są zupełnie nowe.

„Wiedza" modeli językowych

Jednakże gdy mówimy o „uczeniu się" i „wiedzy" modeli językowych, musimy być ostrożni, aby nie przypisywać im ludzkich cech. Wbrew powszechnemu przekonaniu, duże modele językowe nie „zapamiętują" dokładnie treści, na których były trenowane, ani nie posiadają wiedzy w takim sensie, w jakim rozumieją to ludzie. To, co nazywamy „wiedzą" modelu, jest w rzeczywistości skomplikowanym systemem wzorców i statystycznych zależności, zakodowanych w dziesiątkach lub nawet setkach miliardów parametrów sieci neuronowej. Model językowy nie przechowuje dosłownych fragmentów tekstu, na których był trenowany, ale raczej „uczy się" wzorców językowych, relacji między pojęciami i ogólnych zasad generowania tekstu. Dlatego też, choć model może generować tekst, który wydaje się odzwierciedlać głęboką wiedzę na dany temat, nie oznacza to, że „zna" on ten temat w taki sposób, jak znałby go ekspert-człowiek. Model może czasem popełniać błędy, generować niespójne lub nawet nieprawdziwe informacje, szczególnie gdy jest proszony o szczegółowe lub specjalistyczne dane. Najnowszy przykład tych ograniczeń to raport BBC nt. rzetelności faktograficznej modeli streszczających informacje (Elliott, 2025). To dlatego tak ważne jest, aby użytkownicy modeli językowych zawsze weryfikowali generowane przez nie informacje, szczególnie w kontekstach wymagających precyzyjnych informacji i dokładności.

Sposób trenowania i działania modeli językowych ma daleko idące konsekwencje – taki model „nie myśli", lecz tylko generuje tekst najlepiej pasujący do danego kontekstu (tj. wcześniej wprowadzonego tekstu). Dzięki temu jako użytkownicy mamy duży wpływ na działanie modelu. To, jak zdefiniujemy zadanie, jak precyzyjnego kontekstu dostarczymy, będzie wpływało na to, jak model będzie to zadanie wykonywał i jaki będzie efekt. Wiele z technik promptowania, którym poświęcony jest kolejny rozdział, polega właśnie na dostarczaniu lepszego kontekstu modelowi lub skłaniania go do tego, żeby sam sobie ten kontekst

rozszerzał, np. poprzez rozwiązywanie zadania krok po kroku. Model nie tylko pamięta kontekst, ale może się też z niego uczyć, w szczególności w oparciu o przykłady podawane przez użytkownika (tzw. *prompt-based learning*).

Obecne (2025 r.) zastosowania modeli językowych idą jednak o krok dalej, łącząc je z zewnętrznymi źródłami danych, co pozwala przezwyciężyć niektóre z tych ograniczeń. Technika znana jako RAG (ang. *Retrieval-Augmented Generation*) umożliwia integrację modeli z bazami danych, dokumentami czy innymi repozytoriami informacji. Wyobraźmy sobie, że model jest jak niezwykle inteligentny asystent, który nie tylko ma ogólną wiedzę, ale także dostęp do specjalistycznej biblioteki. Gdy otrzymuje pytanie, najpierw przeszukuje tę bibliotekę w poszukiwaniu odpowiednich informacji, a następnie używa ich do sformułowania odpowiedzi. W praktyce działa to tak, że gdy użytkownik zadaje pytanie, system RAG najpierw przeszukuje zewnętrzne źródła danych w poszukiwaniu odpowiednich fragmentów. Te fragmenty są następnie dodawane do zapytania jako element tzw. „prompta" – swoistej podpowiedzi dla modelu. Model wykorzystuje te dodatkowe informacje wraz ze swoją ogólną wiedzą (wzorce językowe, konteksty, relacje) do generowania odpowiedzi. Dzięki temu może dostarczać bardziej precyzyjne, aktualne i kontekstowo trafne odpowiedzi, oparte na konkretnych źródłach. To podejście jest szczególnie cenne w zastosowaniach praktycznych, gdzie modele muszą operować na aktualnych i specyficznych dla danej organizacji danych.

Modele „rozumujące"

Wraz z rozwojem sztucznej inteligencji w 2024 roku pojawiła się nowa klasa modeli – *Large Reasoning Models* (LRMs) – które rozszerzają możliwości tradycyjnych dużych modeli językowych (LLMs) o zdolność do wieloetapowego rozumowania i bardziej złożonego wnioskowania (Plaat et al., 2024, Xu et al., 2025; Open AI, 2025). Klasyczne modele językowe generują odpowiedzi wyraz po wyrazie, na podstawie wyuczonej statystyki języka. Dzięki temu są w stanie robić to niemal natychmiast. W odróżnieniu od nich LRM-y stosują podejście iteracyjne – analizują najpierw problem krok po kroku, a dopiero potem generują odpowiedź na podstawie przeprowadzonej analizy. Charakteryzują się one zmiennym czasem „myślenia", wewnętrznym łańcuchem rozumowania (ang. *Chain-of-Thought*, CoT) oraz zdolnością do samokrytyki, którą wykorzystują w weryfikacji rozumowania, a także eksploracją alternatywnych rozwiązań. Dzięki tym właściwościom mogą okazać się niezwykle użyteczne w badaniach ewaluacyjnych, szczególnie tam, gdzie analiza wymaga syntetycznego

myślenia, porównywania wielu perspektyw (triangulacji) i wyciągania wniosków na podstawie złożonych danych.

Jednym z kluczowych atutów LRM-ów jest ich zdolność do rozwiązywania problemów poprzez wieloetapowe rozumowanie, co ma istotne znaczenie w ewaluacji polityk publicznych, programów społeczno-gospodarczych czy inicjatyw badawczo-rozwojowych. Modele te mogą dekomponować skomplikowane pytania badawcze na mniejsze elementy, oceniać różne scenariusze i uwzględniać kontekstowe zależności w sposób, który tradycyjne LLM-y realizują w ograniczonym zakresie. Ponadto, ich zdolność do samokrytyki i weryfikowania własnych hipotez sprawia, że mogą one identyfikować potencjalne błędy w analizie, co zwiększa rzetelność generowanych ocen i rekomendacji.

Kolejnym ważnym aspektem LRM-ów jest ich mniejsza zależność od nadzoru człowieka w procesie uczenia, co wpływa na większą skalowalność ich zastosowań. W przeciwieństwie do tradycyjnych modeli językowych, które często wymagają intensywnej regulacji poprzez uczenie nadzorowane, LRM-y w dużej mierze opierają się na metodach uczenia poprzez wzmocnienie (*Reinforcement Learning* – RL), gdzie nagrody są określane na podstawie reguł systemowych, a nie wyłącznie ludzkich preferencji. Dzięki temu mogą one lepiej dostosowywać się do dynamicznych warunków ewaluacyjnych, szczególnie w kontekście analizy danych ilościowych i jakościowych oraz modelowania różnych wariantów interwencji.

Pomimo swoich zalet modele „rozumujące" są na razie na bardzo wczesnym etapie rozwoju, a w konsekwencji – narażone są na szereg problemów, które mogą mieć wpływ na ich wykorzystanie, w tym również w badaniach ewaluacyjnych. Wrażliwość na sposób formułowania zapytań (*prompt sensitivity*) sprawia, że ich efektywność może być zależna od precyzji wprowadzonych danych, a długie łańcuchy rozumowania mogą zwiększać ryzyko tzw. „nadmiernego przemyślenia" problemu (*overthinking*) i generowania zbędnych treści (por. Cuadron et al., 2025). Dodatkowo konieczność zapewnienia bezpieczeństwa i kontrolowania jakości generowanych wniosków wymaga dalszych badań nad metodami szkolenia tych modeli. Mimo to należy spodziewać się wzrostu roli LRM-ów w rozwiązywaniu skomplikowanych problemów, a w konsekwencji mogą one stać się wartościowym narzędziem w ewaluacji, zwłaszcza tam, gdzie tradycyjne metody analityczne okazują się niewystarczające.

1.2. Spektrum typów gen AI[1]

W 2024 i na początku 2025 roku na rynku pojawiło się kilka zaawansowanych dużych modeli językowych (LLMs), z których każdy ma unikalne cechy i zastosowania. Wśród nich wyróżniają się zarówno modele o otwartych wagach, takie jak DeepSeek V3 i Llama 3.1, jak i rozwiązania zamknięte, w tym GPT-4o, Claude 3.5 czy Gemini 2.0. Modele te różnią się architekturą, liczbą parametrów, zakresem obsługi multimodalności oraz zastosowaniami w praktycznych zadaniach, takich jak analiza danych, kodowanie czy przetwarzanie języka naturalnego w czasie rzeczywistym. Jednym z kluczowych aspektów porównawczych jest też zakres okna kontekstowego, który wpływa na zdolność modelu do przetwarzania dłuższych fragmentów tekstu. Większość modeli obsługuje już 128 tysięcy tokenów, a Gemini 2.0 nawet 2 mln.

W ostatnich latach Polska dołączyła do grona państw rozwijających własne modele językowe, dostosowane do lokalnych potrzeb i specyfiki języka polskiego. Najbardziej zaawansowane projekty w tym obszarze to Bielik, rozwijany przez Fundację SpeakLeash i AGH, oraz PLLuM, realizowany przez konsorcjum pod kierownictwem Politechniki Wrocławskiej. Bielik koncentruje się na otwartości i szerokim zastosowaniu w różnych dziedzinach, podczas gdy PLLuM ma służyć przede wszystkim administracji publicznej, oferuje rozwiązania wspierające automatyzację procesów urzędowych. Oba modele powstały jako odpowiedź na dominację anglojęzycznych systemów AI, które często nie radzą sobie z polską fleksją, specyfiką prawną czy kontekstem historycznym.

Z technicznego punktu widzenia Bielik 2.3 opiera się na architekturze *decoder-only*, co pozwala na efektywne generowanie tekstu. Model ma 11 miliardów parametrów i obsługuje kontekst do 32 tysięcy tokenów, co umożliwia analizę dłuższych tekstów. W odróżnieniu od niego PLLuM wykorzystuje architekturę *encoder-decoder*, co lepiej sprawdza się w zadaniach wymagających precyzyjnego odwzorowania struktury wejściowych danych, takich jak generowanie odpowiedzi na podstawie dokumentów urzędowych. Model ten został przeszkolony na zbiorze 84 miliardów tokenów, z czego większość stanowią dokumenty prawne i administracyjne, co czyni go wyjątkowo skutecznym w analizie tekstów formalnych. Największy wariant modelu ma 70 miliardów parametrów.

[1] Stan na marzec 2025.

Szczegółowe porównanie najważniejszych dostępnych obecnie (2025) modeli językowych przedstawiono w tabeli 1.1.

Tabela 1.1. Wybrane duże modele językowe

Model	Firma	Liczba parametrów	Okno kontekstu	Dostępność	Modalności	Dodatkowe informacje
GPT	OpenAI	175 miliardów w GPT3.5, prawdopodobnie około biliona w GPT4o	128 tys. tokenów	Model zamknięty, dostępne API	Tekst, audio, obraz	Modele dostępne w ChatGPT
Claude	Anthropic	b.d.	200 tys. tokenów	Model zamknięty, dostępne API		
Gemini	Google DeepMind	b.d.	2 mln. tokenów	Model zamknięty, dostępne API		
LLaMA	Meta AI	415 mld	128 tys. tokenów	Model częściowo otwarty	Tekst, tekst + obraz (wersja 3.3)	
Mixtral	Mistral AI	b.d.	128 tys. tokenów	Model otwarty		
DeepSeek V3	DeepSeek	671 mld	128 tys. tokenów	Model o otwartych wagach		
Bielik	Fundacja SpeakLeash	11 mld	32 tys. tokenów		Tekst	
PLLUM	Konsorcjum jednostek naukowych	70 mld			Tekst	

Źródło: Opracowanie własne na podstawie danych dostawców modeli.

Wśród pierwszych *Large Reasoning Models* wyróżnia się kilka wiodących architektur, które reprezentują różne podejścia do wieloetapowego wnioskowania i rozwiązywania złożonych problemów. Modele o1 i o3 od OpenAI to flagowe rozwiązania koncentrujące się na planowaniu przed generacją odpowiedzi, co pozwala na bardziej spójne i logicznie uzasadnione wyniki. Modele te przeszły zaawansowane szkolenie w zakresie matematyki, kodowania i nauk ścisłych, wykorzystując wzmocnione uczenie przez nagrody (RL) w celu optymalizacji procesu rozwiązywania problemów. Alternatywne podejście reprezentuje DeepSeek-R1, który całkowicie rezygnuje z nadzorowanego dostrajania (SFT), polegając

wyłącznie na uczeniu przez wzmocnienie, co dowodzi, że umiejętności rozumowania mogą wyłaniać się bez bezpośredniego dostrajania na zbiorach ludzkich odpowiedzi. Z kolei seria Qwen, obejmująca Qwen QwQ i Qwen QvQ, kładzie nacisk na specjalizację w konkretnych dziedzinach, takich jak finanse, inżynieria i analiza wizualna.

W przyszłości modele „rozumujące" (LRMs) zostaną zintegrowane z klasycznymi modelami językowymi (LLMs), tworząc hybrydowe systemy zdolne zarówno do płynnej generacji tekstu, jak i głębokiego, wieloetapowego wnioskowania. Zamiast wymagać od użytkownika wyboru odpowiedniego modelu, inteligentne systemy będą automatycznie dostosowywać swoje działanie do charakteru zadania. Jeśli zapytanie użytkownika będzie wymagało precyzyjnej analizy logicznej, model uruchomi wewnętrzny mechanizm wieloetapowego rozumowania, natomiast w przypadku prostszych zadań pozostanie przy standardowej metodzie generacji tekstu. OpenAI zapowiedziało, że GPT-5 będzie pierwszym modelem tej nowej generacji, płynnie łączącym tradycyjne LLM-y z zaawansowanym rozumowaniem w jednym systemie. Taka architektura pozwoli na lepszą efektywność obliczeniową i bardziej naturalną interakcję, eliminując konieczność dostosowywania promptów czy wybierania trybu pracy przez użytkownika.

Teksty to najpopularniejszy, ale nie jedyny rodzaj treści tworzonych przez generatywną sztuczną inteligencję. Drugim obszarem, w którym AI odnosi sukcesy, jest generowanie obrazu. Modele generatywne potrafią tworzyć realistyczne i artystyczne obrazy na podstawie opisów tekstowych. Najbardziej znane przykłady tego typu modeli to DALL-E od OpenAI, Midjourney oraz Stable Diffusion. Modele te znajdują zastosowanie przede wszystkim w projektowaniu, sztuce cyfrowej, tworzeniu grafik na potrzeby marketingu i mediów społecznościowych, dlatego w tej książce poświęcamy im stosunkowo niewiele miejsca.

Względnie nową, ale szybko rozwijającą się dziedziną AI, jest generowanie wideo. Modele takie jak Sora, Synthesia, Pictory, Descript, czy Runway ML potrafią tworzyć i edytować materiały wideo na podstawie tekstu lub innych wskazówek. Umożliwia to szybkie tworzenie animacji, prezentacji czy nawet krótkich filmów bez konieczności angażowania tradycyjnego zespołu produkcyjnego. Również muzyka generowana przez AI staje się coraz bardziej zaawansowana. Modele takie jak MusicGen od Meta, AIVA czy OpenAI Jukebox potrafią komponować utwory muzyczne w różnych stylach i gatunkach. Mogą one być wykorzystywane do tworzenia ścieżek dźwiękowych do filmów, gier czy jako inspiracja dla muzyków.

Kolejnym obszarem, o którym warto wspomnieć, jest generowanie kodu. Modele takie jak GitHub Copilot, Amazon CodeWhisperer, Tabnine czy OpenAI Codex potrafią generować fragmenty kodu na podstawie opisu funkcjonalności lub kontekstu, znacząco wspomagając pracę programistów. Przyspiesza to proces programowania i pomaga w rozwiązywaniu problemów technicznych. Ta klasa modeli ma znaczenie o tyle, że może wspomagać również pracę analityków danych.

1.3. Spektrum zastosowań gen AI w ewaluacji

Impuls do rozważań – historia pierwszego eksperymentu

Jednym z impulsów dla powstania tej książki był eksperyment, zrealizowany na przełomie 2023–2024 r. przez Jednostkę Ewaluacyjną PARP we współpracy z Krajową Jednostką Ewaluacji (Ministerstwo Funduszy i Polityki Regionalnej), która testowała względnie nową technologię gen AI. Celem eksperymentu było sprawdzenie możliwości generatywnej AI w pracy eksperta-ewaluatora i w pracy jednostek ewaluacyjnych oraz zainicjowanie dyskusji w środowisku branżowym na temat wykorzystania tej technologii. Testowano (eksperymentalnie) następującą hipotezę: „Wykorzystanie generatywnej AI może dostarczać równoważnych efektów co praca zespołu ekspertów".

Przyczynkiem do przeprowadzenia eksperymentu i zarazem empiryczną bazą referencyjną było opracowanie tzw. *Issue Paper* dla Polityki Spójności 2021–2027[2], które powstało w rezultacie prac zespołu ekspertów dziedzinowych i uczestników XV Konferencji Ewaluacyjnej w Polsce (Łódź, grudzień 2023)[3], podczas warsztatów poświęconych tematyce społecznej, środowiskowej i cyfryzacji. Praca wówczas wykonana przyczyniła się do powstania eksperckich opracowań na temat kluczowych wyzwań we wskazanych obszarach, z uwzględnieniem roli ewaluacji. Wykorzystując ten materiał, przeprowadzono eksperyment służący weryfikacji możliwości technologii generatywnej sztucznej inteligencji w tego typu procesach badawczo-kreatywnych i analityczno-eksperckich. Za pomocą serwisu ChatGPT (OpenAI) w kilku prostych zapytaniach wygenerowano dokument analogiczny (co do struktury merytorycznej i objętości) względem pierwszego opracowania – stworzonego w procesie

[2] Dokument pt. „Ewaluacja wobec wyzwań przyszłości. Zielona gospodarka, cyfryzacja, spójność społeczna" (opracowanie pokonferencyjne XV Konferencji Ewaluacyjnej)".

[3] Por. XV Międzynarodowa Konferencja Ewaluacyjna.

angażującym ponad 200 osób i wzbogaconego o ekspertyzę dziedzinową. W rezultacie obu procesów uzyskano dwie wersje dokumentów problemowych (ang. *issue papers*):

1. Wersję stworzoną przez zespół ekspertów, we współpracy z uczestnikami konferencji,
2. Wersję eksperymentalną przygotowaną przez pracownika PARP przy wsparciu modelu gen AI, bez angażowania materiałów eksperckich jw.[4].

W następnym kroku ww. opracowania były rozsyłane losowo do zaopiniowania przez uczestników konferencji i członków Zespołu Sterującego Ewaluacją Polityki Spójności – część osób otrzymała materiał przygotowany z udziałem ekspertów, a część materiał wygenerowany z udziałem AI[5].

Żaden z uczestników eksperymentu nie wiedział o istnieniu alternatywnej wersji opracowania. Co do zasady, respondent opiniujący materiał do pewnego stopnia mógł identyfikować się z nim jako współtwórca, jako że uczestniczył w warsztatach kreatywnych, w rezultacie których powstać miał tego typu dokument. Opracowania były oceniane w ramach pięciu kryteriów: 1. Trafność identyfikacji kluczowych wyzwań; 2. Trafność wskazań kluczowych zadań dla polityki spójności; 3. Kompletność materiału; 4. Użyteczność wskazówek dotyczących ewaluacji; 5. Jakość zredagowania materiału – w skali od 1 (ocena bardzo niska) do 5 (ocena bardzo wysoka). Wszyscy uczestnicy został poinformowani o charakterze przedsięwzięcia już po jego zakończeniu. W tym samym czasie udostępniono im wyniki obserwacji.

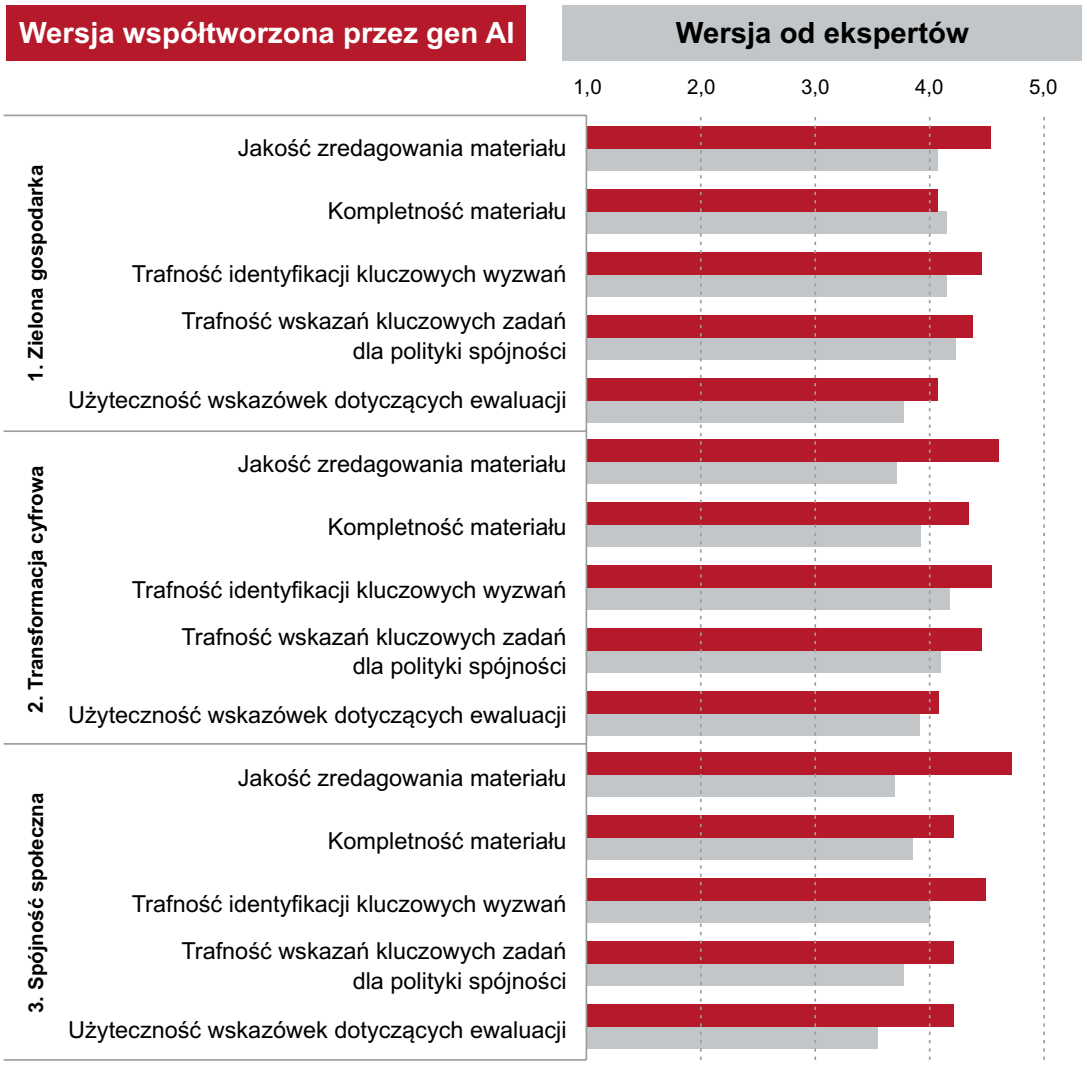
W świetle zebranych danych okazało się, że respondenci średnio w ramach wszystkich kryteriów wyżej ocenili wersję dokumentu opracowaną we współpracy z generatywną AI, niż tę przygotowaną przy zaangażowaniu ich samych oraz ekspertów dziedzinowych. Rozkład ocen był spójny – dokument, którego powstanie wspierała gen AI, oceniono na podobnie wysokim poziomie (średnio 4,3–4,4 pkt), z kolei opracowanie uczestników konferencji

[4] W okresie generowania drugiej wersji dokumentu problemowego (*issue paper),* pierwsza wersja opracowania, która powstała w procesie kreatywnym i eksperckim, nie była jeszcze opublikowana w Internecie, a więc model GPT nie mógł jej wykorzystać.

[5] Do obydwu grup (łącznie N = 169) wysłana została wiadomość zawierająca link do dokumentu wraz z prośbą o zapoznanie się z nim i wypełnienie ankiety CAWI oceniającej dokument. Respondenci otrzymywali link tylko do jednej wersji opracowania (na etapie eksperymentu nie mieli możliwości skonfrontowania dwóch wersji opracowania) – wylosowana połowa badanych otrzymała link do dokumentu opracowanego przez ekspertów, zaś pozostali – do dokumentu opracowanego przy współpracy generatywnej AI. Ankietę wypełniło 18% uczestników (n = 31).

i ekspertów dziedzinowych oceniono niżej, ale także oceny były bardziej zróżnicowane (między 3,8 a 4,1 – w zależności od rozdziału tematycznego).

Rycina 1.1. Wyniki eksperymentu – zestawienie ocen dwóch wersji dokumentu „*Issue Papers* dla Polityki Spójności 2021–2027" (wersja wytworzona we współpracy z gen AI vs. wersja opracowana przez ekspertów)



Źródło: PARP (Jędrzejowski, Pokorski, 2024)

Przytoczone wyniki eksperymentu pozwoliły, po pierwsze, potwierdzić postawioną hipotezę („Wykorzystanie generatywnej AI może dostarczać równoważnych efektów co praca zespołu ekspertów"), po drugie, empirycznie zweryfikować duży potencjał i pola zastosowań AI przy

wsparciu dotychczasowej pracy ekspertów i struktur analityczno-badawczych w instytucjach, i po trzecie, wyzwolić dyskusję w środowisku branżowym o roli gen AI. Dostępność tej technologii oraz jej systematyczne udoskonalanie sygnalizuje bowiem potrzeby reorganizacji dotychczasowych procesów generowania wiedzy w instytucjach i angażowanych zasobów (np. zamawianych usług obcych), z włączeniem asysty gen AI.

Z niniejszego eksperymentu można również wyciągnąć kilka dalszych wniosków. Po pierwsze, mimo że gen AI stosunkowo dobrze syntetyzuje wiedzę, nadal nie rozwiązuje wyzwań związanych z jej przepływem od ewaluacji do decydentów („od producentów do użytkowników").

Po drugie, dokumenty typu *policy papers* generowane przy wsparciu AI (w tym diagnozy problemów i hierarchie celów) nadal wymagają zbudowania społecznego konsensu z uwzględnieniem dynamiki procesów decyzyjnych i demokratycznych.

Po trzecie, włączanie technologii gen AI w procesy analityczno-badawcze i eksperckie instytucji powoduje wzrost znaczenia wewnętrznych struktur odpowiedzialnych za te procesy (np. jednostek ewaluacyjnych) względem niezależnych ewaluatorów (wykonawców, ekspertów zewnętrznych) oraz zwiększenie odpowiedzialności pracowników za wiedzę wytworzoną z wykorzystaniem tej technologii (Pokorski, 2024, 2024).

Spektrum procesów i działań z wykorzystaniem gen AI

Jako ramy naszych rozważań o możliwościach zastosowań gen AI w ewaluacji przejęliśmy perspektywę jednostek ewaluacyjnych (JE) działających jako brokerzy wiedzy w ekosystemie polityk publicznych.

Brokerzy wiedzy są kluczowymi pośrednikami, którzy łączą (lub nawet kształtują) popyt na wiedzę opartą na dowodach z jej podażą. Termin „wiedza" może obejmować całe spektrum dowodów – wyniki badań ewaluacyjnych, ale także analizy, syntezy badań naukowych czy badania eksploracyjne i diagnostyczne. Po stronie popytu jednostki ewaluacyjne (lub inne pokrewne komórki analityczno-badawcze) współpracują z użytkownikami wiedzy, takimi jak decydenci, menedżerowie publiczni projektujący i wdrażający interwencje. Po stronie podaży brokerzy współpracują z producentami wiedzy (badaczami, ekspertami, ewaluatorami),

a także sami produkują wiedzę w formie badań i analiz własnych (Dobbins et al., 2009; Frost et al., 2012; Olejniczak et al., 2016).

Przedstawienie jednostek ewaluacyjnych jako brokerów wiedzy daje trzy zalety w kontekście eksplorowania zastosowań gen AI. Po pierwsze, oferuje całościowe i strategiczne spojrzenie na temat roli jednostek ewaluacyjnych w systemie decyzyjnym polityk publicznych. Przede wszystkim pozwoli nam utrzymać uwagę na nadrzędnej misji JE – wspieraniu decyzji rzetelną wiedzą (wynikami badań, analizami, danymi). Ta misja będzie w najbliższym czasie poddana szczególnej presji, bo łatwość generowania tekstów przez gen AI na pewno przyczyni się do nadprodukcji i inflacji informacji.

Po drugie, perspektywa brokerów wiedzy jest osadzona w istniejącej już praktyce zarówno polskich, jak i zagranicznych jednostek ewaluacyjnych (Olejniczak et al., 2014). Jednocześnie wykracza poza relatywnie wąski świat praktyki ewaluacyjnej. Dzięki temu w naszej książce będziemy mogli odnieść gen AI do całego spektrum zróżnicowanych zadań ewaluacyjnych i uwzględnić różne modele funkcjonowania jednostek ewaluacyjnych. Będziemy też mieli możliwość poszukiwania synergii i odnoszenia się do doświadczeń innych jednostek analityczno-badawczych w administracji publicznej, które także eksperymentują z gen AI (np. działy strategii, zespoły Oceny Skutków Regulacji, zespoły monitoringowe i analityczne).

Po trzecie, myślenie o jednostkach ewaluacyjnych jako łącznikach w systemie wiedzy, a nie tylko producentach ewaluacji, pozwoli nam spojrzeć w przyszłość i zastanowić się nad perspektywami rozwoju i redefinicją ich roli w systemie polityk publicznych. Ułatwienia oferowane przez gen AI mogą już dziś wpłynąć na portfolio działań JE, np. umożliwić jednostkom podejmowanie działań wewnętrznych, których realizacja przy ograniczonych zasobach kadrowych była dotychczas zbyt czasochłonna.

Misją brokerów wiedzy jest tworzenie wartościowej, opartej na badaniach wiedzy i wspieranie w tym zakresie decydentów odpowiedzialnych za projektowanie i wdrażanie polityk publicznych. W praktyce oznacza to prowadzenie trzech procesów:

1. Pomocy w określaniu potrzeb wiedzy danej administracji.
2. Pozyskiwania wiedzy przez badania zamawiane lub własne działania analityczne oraz
3. Wykorzystania wiedzy – przekazywania jej decydentom i wspierania faktycznego użycia wiedzy w konkretnych działaniach publicznych.

Każdy z tych procesów kładzie nacisk na nieco inne kwestie. W ramach każdego z nich jednostki realizują konkretne zadania, a także – w zależności od swojej pozycji i roli w systemie wdrażania danej polityki czy programu – mają całe spektrum możliwych podejść i szczegółowych działań do realizacji.

W przypadku określania potrzeb wiedzy kluczową wartością jest adekwatność, czyli zadbanie, by pytania badawcze przystawały do faktycznych potrzeb wiedzy artykułowanych przez interesariuszy danej polityki i programu. Spektrum działań obejmuje zarówno te systemowe, takie jak budowanie wieloletnich planów ewaluacji (a więc antycypowanie potrzeb wiedzy dla całych interwencji publicznych), jak i monitorowanie oraz ustalanie bieżących potrzeb informacyjnych interesariuszy pojedynczych programów.

Przy pozyskiwaniu wiedzy nasza uwaga skupia się na rzetelności metodycznej badania, bo ona buduje podwaliny pod wiarygodność wyników. Działania jednostek ewaluacyjnych mogą dotyczyć zamawiania i nadzorowania realizacji ewaluacji przed podmioty zewnętrzne lub też prowadzenia własnych działań analityczno-badawczych.

Wreszcie na etapie wykorzystania wiedzy akcent stawiamy na dopasowanie formy i sposobu komunikacji wyników do preferencji odbiorców ewaluacji oraz na właściwy moment czasowy przekazania wyników. Zadania jednostek ewaluacyjnych na tym etapie to cały wachlarz działań dyplomatyczno-taktycznych (udział w różnych spotkaniach, formach konsultacji i wymiany informacji) oraz działania merytoryczno-techniczne – dobór metod, form, technik, które pozwalają skutecznie komunikować wnioski i rekomendacje.

W tabeli przedstawiamy wstępną listę głównych procesów, zadań i dotychczasowych działań z zaznaczeniem, które z nich już dziś mogą być przedmiotem eksperymentowania z gen AI. Warto jednak podkreślić, że z uwagi na szybki postęp technologii mogą się zmieniać zarówno zawartość listy i pola eksperymentowania z AI, jak i ewolucja roli jednostek ewaluacyjnych.

Czytelnicy znajdą w dalszych rozdziałach tej książki właśnie eksperymenty z różnymi zadaniami i działaniami jednostek.

Tabela 1.2. Spektrum działań jednostek ewaluacyjnych

Główne procesy	Zadania Jednostek Ewaluacyjnych	Dotychczasowe działania szczegółowe
1. Określanie potrzeb wiedzy	1.1. Określanie potrzeb systemowych – opracowanie planów ewaluacji	Zebranie potrzeb informacyjnych od interesariuszy
		Opracowanie projektu planu ewaluacji
		Konsultowanie projektu planu ewaluacji i przekazanie planu do zatwierdzenia
		Upublicznienie planu ewaluacji
	1.2. Aktualizacja potrzeb systemowych – aktualizacja planów ewaluacji	Zebranie aktualnych informacji na temat potrzeb informacyjnych od interesariuszy
		Porównanie z planami ewaluacji innych jednostek
		Opracowanie projektu zaktualizowanego planu ewaluacji
		Konsultowanie projektu aktualizacji planu ewaluacji i przekazanie do zatwierdzenia
		Upublicznienie zaktualizowanego planu ewaluacji
	1.3. Określanie bieżących potrzeb informacyjnych interesariuszy	Zebranie szczegółowych informacyjnych dotyczących zidentyfikowanego tematu badania
		Analiza źródeł danych zawierających potencjalne odpowiedzi na zidentyfikowane potrzeby informacyjne
		Podjęcie decyzji o konieczności realizacji badania

Główne procesy	Zadania Jednostek Ewaluacyjnych	Dotychczasowe działania szczegółowe
2A. Pozyskiwanie wiedzy – zamawianie wiedzy	2.1. Przygotowanie zamówienia	Opracowanie szczegółowego zakresu badania (w tym konsultacje z interesariuszami)
		Opracowanie warunków udziału w postępowaniu, w tym opracowanie kryteriów wyboru ofert
	2.2. Ocena ofert i wybór wykonawcy	Ocena merytoryczna ofert
		Przygotowanie i zawarcie umowy z wybranym wykonawcą
	2.3. Ustalenie metodyki badania	Spotkanie otwierające z wykonawcą, przedstawienie wzajemnych oczekiwań
		Opiniowanie raportu metodycznego (a na późniejszym etapie konsultowanie i opiniowanie szczegółowych narzędzi badawczych)
	2.4. Wsparcie w zbieraniu danych wtórnych i pierwotnych	Przygotowanie listu polecającego/informacji o wykonawcy na stronę
		Ustalenie z wykonawcą zakresu niezbędnych do realizacji badania danych z wewnętrznych baz danych
		Opracowanie i przekazanie wykonawcy danych z wewnętrznych zbiorów danych oraz inne wsparcie w pozyskiwaniu zewnętrznych źródeł
	2.5. Odbiór produktów badania	Opiniowanie projektu raportu końcowego (i innych produktów z badania)
		Koordynacja zbierania uwag do projektu raportu końcowego
		Ocena stopnia uwzględnienia uwag
		Przygotowanie protokołu odbioru
		Ocena jakości, rzetelności i użyteczności raportu przy użyciu karty oceny procesu ewaluacji

Główne procesy	Zadania Jednostek Ewaluacyjnych	Dotychczasowe działania szczegółowe
2B. Pozyskiwanie wiedzy – analizy własne	2.1. Analiza QUAL: analizy dokumentów wewnętrznych, badań dostępnych w danym temacie lub zewnętrznych	Określenie celu i pytania/pytań badawczych
		Zdefiniowanie ramy koncepcyjnej
		Zaprojektowanie strategii poszukiwania źródeł
		Poszukiwanie źródeł i filtrowanie wyników
		Analiza źródeł
		Synteza i wnioski praktyczne
	2.2. Analiza QUAN – zbiory danych (własne lub zewnętrzne)	Określenie celu i pytania/pytań badawczych
		Przygotowanie danych
		Analiza danych
		Synteza i wnioski praktyczne
	2.3. Analiza rekomendacji zawartych w Systemie Wdrażania Rekomendacji (SWR)	Określenie celu i pytania/pytań badawczych
		Zebranie dokumentów i integracja zbioru
		Analiza źródeł
		Synteza i wnioski praktyczne
	2.4. Metaewaluacje badań realizowanych w Polsce	Określenie celu i pytania/pytań badawczych
		Zdefiniowanie ramy koncepcyjnej
		Zebranie dokumentów i integracja zbioru
		Analiza źródeł
		Synteza i wnioski praktyczne

Główne procesy	Zadania Jednostek Ewaluacyjnych	Dotychczasowe działania szczegółowe
3. Wykorzystanie wiedzy	3.1. Upowszechnianie wyników wraz z dyskusją nad wynikami badania	Organizacja spotkań z interesariuszami
		Publikacja raportu na stronie ministerstwa/agencji
		Przygotowanie informacji na stronę/fb i inne media cyfrowe
		Przygotowanie informacji dla kierownictwa
	3.2. Koordynacja wdrażania rekomendacji	Wypełnienie tabeli wdrażania rekomendacji
		Uzgodnienie sposobu i zakresu wdrożenia rekomendacji z adresatami
		Wprowadzenie rekomendacji do SWR
		Monitorowanie procesu wdrażania rekomendacji i aktualizacja statusów w SWP
		Przedstawienie postępów wdrażania członkom KM
	3.3. Wykorzystanie wyników badań ewaluacyjnych w bieżących pracach instytucji	Szybkie odpowiedzi na interpelacje
		Uzupełnianie materiałów dla kierownictwa
		Uzupełnianie dokumentów (strategicznych, programowych, operacyjnych)

Źródło: Opracowanie K. Olejniczak, B. Ledzion, I. Lyubashenko w oparciu o dyskusję z przedstawicielami Krajowej Jednostki Ewaluacji – Ministerstwo Funduszy i Polityki Regionalnej.

1.4. Struktura książki

Publikacja ma strukturę składającą się z trzech części. Celem części I (rozdziały 2.–4.) jest przystępne przedstawienie zagadnień horyzontalnych, ważnych w pracy z gen AI w sektorze publicznym. Rozdział 2. wyjaśnia zasady dialogu z AI (promptowania), rozdział 3. omawia kwestie bezpieczeństwa i poufności danych przy zastosowaniu gen AI do działań badawczych, zaś rozdział 4., kończący tę część, wyjaśnia, w jaki sposób można łączyć modele gen AI z wewnętrznymi bazami dokumentów tak, by cyfrowi asystenci mogli produktywnie pracować w oparciu o nasze zbiory publikacji i ekspertyz.

Część II (rozdziały 5.–10.) prezentuje ujęcie procesowe – pokazuje zastosowania gen AI w konkretnych zadaniach badawczo-ewaluacyjnych. Struktura tej części jest oparta na liście

zadań jednostek ewaluacyjnych działających jako brokerzy wiedzy. I tak rozdział 5. zawiera omówienie praktyk projektowania badań i przygotowywania Opisów Przedmiotu Zamówienia ze wsparciem gen AI. Kolejne rozdziały pokazują, jak wykorzystać gen AI przy konkretnych sposobach pozyskiwania wiedzy: prowadzeniu analiz jakościowych (rozdział 6.), przeglądów źródeł (rozdział 7.), przygotowaniu analiz ilościowych (rozdział 8.) i analiz sieciowych (rozdział 9.). Część drugą książki zamyka rozdział 10. pokazujący przykłady procesów i narzędzi wspierających komunikację wyników badań.

W części III (rozdziały 11.–13.) sygnalizujemy szersze implikacje, jakie niesie ze sobą rewolucja gen AI dla praktyki ewaluacji programów i polityk publicznych. Rozdział 11. przybliża pierwsze próby stworzenia standardów i zasad bezpiecznego korzystania z gen AI w organizacjach publicznych. W rozdziale 12. zastanawiamy się nad mechanizmami zmiany i adaptacji do zastosowania gen AI w naszych organizacjach publicznych, w szczególności wewnętrznych strukturach badawczo-analitycznych i ewaluacyjnych instytucji. W zakończeniu książki (rozdział 13.) podsumowujemy praktyczne obserwacje z pierwszych eksperymentów z zastosowaniem gen AI i stawiamy szersze pytania do dalszej dyskusji dla wspólnoty praktyków ewaluacji i badaczy polityk publicznych.

Bibliografia

- Bubeck, S., V. Chandrasekaran, R. Eldan, et al. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv* preprint arXiv:2303.12712.
- Bullock, J.B., Chen, Y.-C., Himmelreich, et al. (red.). (2024). *The Oxford Handbook of AI Governance*. Oxford University Press.
- Charalabidis, Y., Medaglia, R., van Noordt, C. (red.). (2024). *Research Handbook on Public Management and Artificial Intelligence*. Edward Elgar Publishing.
- Christiano, P.F., J. Leike, T. Brown, et ali. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.
- Cuadron, A., Li, D., Ma, W., et al. (2025). The Danger of Overthinking: Examining the Reasoning-Action Dilemma in Agentic Tasks (dostęp: 31.03.2025).
- Dobbins, M., Robeson, P., Ciliska, D., et al. (2009). A description of a knowledge broker role implemented as part of a randomized controlled trial evaluating three knowledge translation strategies. *Implementation Science*, *4*(23).
- European Commission. (2019). *High-Level Expert Group on Artificial Intelligence: Ethics Guidelines for Trustworthy AI*. European Commission.
- Frost, H., Geddes, R., Haw, S., et al. (2012). Experiences of knowledge brokering for evidence-informed public health policy and practice: three years of the Scottish Collaboration for Public Health Research and Policy. *Evidence & Policy*, 8(3), 347–359.
- Gasser, U., Mayer-Schönberger, V. (2024). *Guardrails: Guiding Human Decisions in the Age of AI*. Princeton University Press.

- Jędrzejowski A., Pokorski J. (2024). *Efekty pracy wykonanej przez ekspertów i generatywną AI. Eksperyment przy opracowaniu Issue Paper dla Polityki Spójności 2021–27. Ekspertyza*. Polska Agencja Rozwoju Przedsiębiorczości (dostęp: 31.03.2025).
- Kojima, T., Gu, S.S., Reid, M., et al. (2022). Large language models are zero-shot reasoners. *arXiv* preprint arXiv:2205.11916.
- Mikalef, P. (2024). Responsible AI governance: from ideation to implementation. W: Charalabidis Y., Medaglia R., van Noordt C. (red.), *Research Handbook on Artificial Intelligence and Decision Making in Organizations* (s. 126–142). Edward Elgar Publishing.
- Mollick, E. (2024). *Co-Intelligence: Living and Working with AI*. Penguin.
- Montrosse-Moorhead, B. (2023). Evaluation criteria for artificial intelligence. *New Directions for Evaluation*, *178–179*, 123–134 (dostęp: 31.03.2025).
- Nielsen, S.B., Rinaldi, F.M., Petersson, G.J. (red.). (2025). *Artificial Intelligence and Evaluation: Emerging Technologies and Their Implications for Evaluation*. Taylor & Francis.
- Ociepa, K., Flis, Ł., Wróbel, K., et al. (2024). *Bielik 7B v0.1: A Polish Language Model -- Development, Insights, and Evaluation* (dostęp: 31.03.2025).
- OECD. (2019). *The OECD AI Principles* (dostęp: 31.03.2025).
- Olejniczak, K., Kupiec, T., Raimondo, E. (2014). *Brokerzy wiedzy. Nowe spojrzenie na rolę jednostek ewaluacyjnych*. W: A. Haber, K. Olejniczak (red.), (R)ewaluacja 2. Wiedza w działaniu. Polska Agencja Rozwoju Przedsiębiorczości, s. 67–112.
- Olejniczak, K., Raimondo, E., Kupiec, T. (2016). Evaluation units as knowledge brokers: Testing and calibrating an innovative framework. *Evaluation*, *22*(2), 168–189 (dostęp: 31.03.2025).
- Olejniczak, K. (red.) (2012). *Organizacje uczące się. Model dla administracji publicznej*. Wydawnictwo Naukowe Scholar (dostęp: 31.03.2025).
- Open AI (2025, 8 maja). Reasoning best practices. Learn when to use reasoning models and how they compare to GPT models (dostęp: 31.03.2025).
- Ouyang, L., J. Wu, X. Jiang, et al. (2022). Training language models to follow instructions with human feedback. *arXiv* preprint arXiv:2203.02155.
- Plaat, A., Wong, A., Verberne, S., et al. (2024). Reasoning with Large Language Models, a Survey (No. arXiv:2407.11511). *arXiv* (dostęp: 31.03.2025).
- Pokorski, J. (2024). *Transformacja cyfrowa w systemie ewaluacji wykorzystania funduszy europejskich w Polsce. Praca dyplomowa MBA*. Uniwersytet SWPS, IPI PAN, Woodbury UVU.
- Pokorski, J. (2024). *Co jest potrzebne, aby przeprowadzić ewaluację i w jaki sposób możemy ją realizować?* W: Bienias, S., Kupiec, T., Strzęboszewski, P. (red.). Poradnik ewaluacji dla administracji publicznej, Ministerstwo Funduszy i Polityki Regionalnej.
- Przegalińska, A., Triantoro, T. (2024). *Converging Minds: The Creative Potential of Collaborative AI*. CRC Press.
- Scriven, M. (1991). *Evaluation thesaurus* (4th ed ed.). Sage.
- Urząd M. St. Warszawy (2025). Kierunki odpowiedzialnego wykorzystania generatywnej sztucznej inteligencji w Urzędzie m.st. Warszawy. Poradnik (dostęp: 31.03.2025).
- Vaswani, A., Shazeer, N.M., Parmar, N., et al. (2017). Attention is All you Need. *Neural Information Processing Systems*.
- Wei, J., X. Wang, D. Schuurmans, et al. (2022). Chain of thought prompting elicits reasoning in large language models. *arXiv* preprint arXiv:2201.11903.
- Xu, F., Hao, Q., Zong, Z., et al. (2025). Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models (No. arXiv:2501.09686). *arXiv* (dostęp: 31.03.2025).
- Ziegler, D.M., N. Stiennon, J. Wu, et al. (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

2. Promptowanie, czyli jak komunikować się z modelem językowym

Jacek Szut
Bartosz Ledzion

2.1. Wstęp

Jeżeli rozwój sztucznej inteligencji jeszcze nie ma wpływu na sposób pracy czytelnika niniejszego rozdziału, to niebawem się to zmieni. Szczególną rolę w tym procesie odgrywają modele językowe (LLM), które stają się coraz bardziej dostępnym i powszechnym narzędziem wspierającym pracę w wielu obszarach.

Umiejętność efektywnej pracy z LLM będzie stawała się jedną z kluczowych kompetencji pracownika, analogicznie jak to się stało z „tradycyjnymi" kompetencjami cyfrowymi. Obecnie trudno sobie wyobrazić funkcjonowanie zawodowe bez znajomości podstawowych aplikacji biurowych, podobnie będzie z narzędziami wykorzystującymi sztuczną inteligencję. Zmiana ta dotyczy również środowiska ewaluatorów, zarówno po stronie wykonawców badań ewaluacyjnych (w tym ewaluatorów wewnętrznych), jak i zamawiających ewaluacje.

W rozdziale chcemy się podzielić z czytelnikami wiedzą na temat komunikowania się z modelami językowymi. Czerpiemy ją z kilku opracowań, ale przede wszystkim bazujemy na naszym doświadczeniu. Z gen AI korzystamy od momentu jej udostępnienia dla szerokiego grona użytkowników i wspomagamy się nią m.in. przy zadaniach badawczych, w tym związanych *stricte* z ewaluacją. W naszej codziennej pracy najczęściej jak dotąd sięgaliśmy po modele GPT oraz Claude, chociaż również testowaliśmy możliwości Gemini, Mistral, Copilot, Bielik. Kluczowy wniosek z tych doświadczeń jest jeden: skuteczna komunikacja z modelem jest fundamentem efektywnego wykorzystania jego możliwości. Nie ma jednak skutecznej komunikacji bez zrozumienia ograniczeń modelu, ograniczeń po stronie

użytkownika (człowieka), bez poznania zasad i procesu rozmów z maszyną oraz opanowania podstawowych typów tzw. promptów.

Strukturę rozdziału można nazwać ścieżką poznawania i uczenia się komunikacji z LLM. Naszą podróż rozpoczynamy od omówienia kluczowych wyzwań związanych z korzystaniem z gen AI, tych wynikających z natury samej technologii, jak i tych, które stoją po stronie człowieka (użytkownika). Następnie przechodzimy do omówienia praktycznych zasad i procesu „rozmowy", zaś w końcowej części rozdziału prezentujemy podstawową typologię promptów oraz instrukcje ustrukturyzowane. Niektórzy czytelnicy mogą być zaskoczeni faktem, że w rozdziale nie znajdą rozbudowanych strategii osiągania określonych celów przy wykorzystaniu LLM. Jest to zabieg celowy. Dajemy czytelnikowi podstawy, które pozwolą samodzielnie rozwijać kompetencje i wraz z nabywanym doświadczeniem tworzyć strategie osiągania wyznaczonych celów przy wsparciu gen AI. Można to porównać do przekazania wiedzy dotyczącej gramatyki języka obcego. Gdy już znamy jej podstawy, łatwiej nam formułować wypowiedzi.

Z tego rozdziału czytelnik dowie się, z jakimi wyzwaniami wiąże się korzystanie z gen AI, zapozna z zasadami i procesem komunikacji oraz z podstawowymi typami promptów i technik promptowania. Staraliśmy się położyć nacisk na praktyczne zastosowanie prezentowanych rozwiązań w pracy ewaluatora/ewaluatorki, ilustrując je konkretnymi przykładami i wskazówkami.

2.2. Wyzwania po stronie gen AI

Zanim przystąpimy do pracy z modelem językowym, warto poznać zasady jego działania i potencjalne ograniczenia. Należy pamiętać, że model językowy (niewytrenowany na danych dziedzinowych) może mieć następujące ograniczenia:

- brak specjalistycznej wiedzy,
- przestarzałe informacje,
- zbytnie poleganie na danych treningowych,
- generowanie nieprawdziwych treści,
- trudności w rozumieniu intencji.

W tabeli 2.1. opisujemy główne ograniczenia modeli językowych i metody radzenia sobie z nimi.

Tabela 2.1. Ograniczenia modeli językowych i metody radzenia sobie z nimi

Ograniczenia	Rozwiązanie
Deficyt wiedzy specjalistycznej Modele są trenowane na dużych zbiorach danych i mają deficyty w zakresie wiedzy specjalistycznej lub nie będą skłonne do sięgania po nią, pomimo że nią dysponują. Dzieje się tak dlatego, że taka wiedza stanowi niewielki wycinek ogólnej puli danych znajdujących się w modelu i może on ją pomijać, skupiać się na generowaniu odpowiedzi opartych na danych ogólnych. Będzie prowadziło to do nadmiernego uogólniania odpowiedzi[1]. Jedną z tych dziedzin, z którymi modele stosunkowo słabiej sobie radzą, jest ewaluacja, szczególnie w zakresie odpowiedzi na pytania dotyczące specyficznych zagadnień ewaluacyjnych[2].	Warto dostarczyć modelowi wiedzę i kontekst poprzez wklejenie do prompta fragmentów dokumentacji lub dołączenie jako załączników do konwersacji przykładowych rozwiązań, definicji pojęć oraz oczekiwanego formatu odpowiedzi. Takie działania pozwolą ukierunkować pracę modelu.
Aktualność danych Ten problem ma szczególnie duże znaczenie w przypadku dziedzin, które rozwijają się dynamicznie. Najnowsze odkrycia, teorie itp. mogą nie być uwzględnione w danych treningowych, tym samym nie będą stanowiły źródła wiedzy do sformułowania odpowiedzi na nasze pytania.	Należy udostępnić modelowi wiedzę (np. link do źródła wiedzy, plik CSV z danymi, plik tekstowy) lub dokonać jej aktualizacji, wprowadzając odpowiednie informacje w oknie kontekstowym. Ewentualnie można skorzystać z aplikacji wspomaganych AI, które mają dostęp do Internetu i generując odpowiedź, w pierwszej kolejności szukają informacji w Internecie.

[1] Nadal stanowi to duże wyzwania dla twórców modeli. Z jednej strony starają się oni zapewnić szybką reakcję modelu, co z kolei ogranicza wykorzystywanie wiedzy szczegółowej, analizę niuansów. Z drugiej strony ukierunkowanie modelu na wiedzę specjalistyczną czy analizę niuansów może prowadzić do poważniejszych błędów innego typu. Model będzie wówczas pomijał wiedzę ugruntowaną i uogólnienia (tendencje, wzorce), a skupiał się na niuansach i ostatecznie generował niepoprawne odpowiedzi.

[2] Zalecane jest stworzenie własnego agenta AI, „nakarmienie" go wiedzą w zakresie ewaluacji. Odpowiedzi generowane przez niego są zdecydowanie bardziej trafne od tych wytworzonych przez model ogólny. Przykładowo, po zapytaniu modelu o jego wiedzę na temat programu Fundusze Europejskie dla Nowoczesnej Gospodarki, 2021–2027 (prompt: *Będziemy projektować badanie ewaluacyjne dotyczące programu FENG. Aby to zrobić, musisz mieć wiedzę o tym programie. Napisz mi, co o nim wiesz.*) agent rozwija swoją odpowiedź, w tym uwzględnia strukturę programu i jego działania oraz wymienia sposób realizacji i zarządzania, podczas gdy model ogólny przedstawia jedynie najbardziej ogólne informacje. Odpowiedź agenta składa się z 401 słów, a modelu ogólnego z 215.

Ograniczenia	Rozwiązanie
Nadmierne przywiązanie do danych treningowych Należy również pamiętać o tym, że modele mogą wykazywać tendencję do przypisywania zbyt dużej wagi danym treningowym – preferowania ich kosztem informacji wprowadzanych przez użytkownika[3].	Warto wykorzystać instrukcję (prompt), która ograniczy korzystanie przez model z danych treningowych, tj. zalecać modelowi koncentrowanie się wyłącznie na informacjach dostarczonych przez użytkownika.
Halucynacje W pracy z modelami językowymi mogą wystąpić tzw. halucynacje. Są to odpowiedzi, które są niepoprawne faktograficznie lub wręcz zmyślone[4].	Aby ograniczyć ryzyko takich sytuacji, warto z modelem gen AI pracować iteracyjnie, weryfikować odpowiedzi w innych źródłach (np. wykorzystywać inne modele językowe), modyfikować pytania (np. stosować prompt krokowy, opisany w dalszej części rozdziału).
Problem z rozumieniem intencji Warto mieć na uwadze, że model sam nie domyśli się, co chce uzyskać użytkownik, oraz to, że pomiędzy nim i użytkownikiem może dochodzić do różnic w rozumieniu pojęć.	Pracując z modelem, należy wyjaśniać nasze rozumienie pojęć, sprawdzać różnice w rozumieniu przez drugą stronę i poprawiać model tak, aby uzyskać spójność rozumienia.

Źródło: Opracowanie własne.

2.3. Wyzwania po stronie użytkownika

Po zapoznaniu się z ogólnymi ograniczeniami modeli językowych warto zwrócić uwagę również na wyzwania występujące po stronie użytkownika.

W trakcie pracy z dużymi modelami językowymi użytkownicy często wykazują tendencję do bezkrytycznego akceptowania tworzonych treści. To zachowanie może wyjaśnić koncepcja systemu 1. i systemu 2., wprowadzona przez Daniela Kahnemana, noblistę w dziedzinie

[3] Podczas pracy z danymi GUS dotyczącymi zarobków w gminach model popełnił różne błędy, z których duża część była konsekwencją szumu danych w pliku źródłowym (pozostałości formuł, puste kolumny, ukryte wartości). Jednak jeden błąd zwrócił szczególną uwagę. Model uznał, że gmina X jest tą o najwyższych zarobkach, podczas gdy plasowała się ona daleko poza podium. Okazało się, że gmina ta w kilku poprzednich badaniach rzeczywiście notowała najwyższą wartość wynagrodzeń i informacje o tym znajdowały się danych treningowych modelu.

[4] Przykłady halucynacji: „George Washington podpisał Deklarację Niepodległości 4 lipca 1776 roku jako prezydent Stanów Zjednoczonych” (halucynacja faktograficzna), „tygrysy są większe od lwów, ale lwy są większe od tygrysów” (halucynacja logiczna).

ekonomii behawioralnej. System 1. odpowiedzialny jest za szybkie, niewymagające nakładów energii myślenie intuicyjne i skłania użytkowników do bezkrytycznego akceptowania pierwszych sugestii AI, a jednocześnie sprzyja ignorowaniu błędów oraz nadmiernemu poleganiu na modelach językowych. W rezultacie często pomijamy ważny kontekst i rezygnujemy z własnych pomysłów, polegając zbyt mocno na gotowych odpowiedziach. Wygenerowane przez LLM odpowiedzi nierzadko stają się punktem odniesienia dla całego procesu myślowego użytkownika, utrudniając rozważenie alternatywnych perspektyw.

Rozumiejąc zarówno własne błędy w pracy z AI, jak i ograniczenia modelu, możemy lepiej wykorzystać to narzędzie. Skupmy się teraz na czterech sposobach, które pomogą nam tworzyć bardziej przemyślane zapytania i krytycznie analizować otrzymywane od gen AI odpowiedzi.

Określenie jasnych celów

Precyzowanie jasnych celów przed sformułowaniem zapytania to kluczowy krok w efektywnym korzystaniu z LLM (Bsharat 2024). Zamiast rzucać się na głęboką wodę z mglistym pomysłem, warto poświęcić chwilę na refleksję, czego dokładnie chcemy się dowiedzieć lub co osiągnąć. Określenie konkretnego problemu pomaga ukierunkować nasze myślenie.

Przed przystąpieniem do formułowania zapytania, zadajmy sobie trzy kluczowe pytania:

1. Jaki konkretny rezultat (cel, potrzebę) chcę uzyskać (osiągnąć, spełnić)?
2. Jakie informacje lub kontekst powinienem dostarczyć, aby gen AI lepiej zrozumiała mój cel i moje potrzeby?
3. W jaki sposób zweryfikuję, czy otrzymana odpowiedź jest satysfakcjonująca i wiarygodna?

Ustalenie naszych oczekiwań pozwoli później ocenić, czy odpowiedź gen AI je spełniła. Ta wstępna analiza nie tylko poprawia jakość naszych zapytań, ale także zmusza do głębszego przemyślenia tematu, aktywując tym samym nasze krytyczne myślenie.

Strukturyzacja zapytań

Strukturyzacja promptów to kluczowy element efektywnego korzystania z LLM, który wymaga od użytkownika głębszej refleksji i precyzji w formułowaniu zapytań. Wzorce zapytań

to swoista struktura, która nie tylko kieruje odpowiedziami generowanymi przez AI, ale także zapewnia wysoką jakość wyników. Stosowanie wzorców znacząco rozszerza możliwości konwersacyjne LLM, jednocześnie wymuszając na użytkowniku wpisanie odpowiednich danych i refleksję nad treścią. To z kolei podnosi jakość samego zapytania, a w konsekwencji – otrzymywanych odpowiedzi (White 2023).

Jednym z przykładowych wzorców jest CO-STAR (przedstawiony w dalszej części rozdziału), który służy do otrzymywania spersonalizowanych odpowiedzi, dostosowanych do naszych oczekiwań (Teo 2023).

Iteracyjne podejście

Iteracyjne podejście do tworzenia zapytań to nie tylko technika poprawiająca jakość interakcji z LLM, ale przede wszystkim potężne narzędzie stymulujące krytyczne myślenie użytkownika. Istotą tej metody jest ciągła refleksja (ewaluacja) nad otrzymanymi odpowiedziami i systematyczne udoskonalanie zapytań, co wymaga aktywnego zaangażowania naszych zdolności analitycznych. Kluczowym elementem tej strategii jest praktyka *few-shot prompting*. Wymaga ona od użytkownika przemyślenia i przygotowania kilku przykładowych odpowiedzi przed zadaniem głównego pytania. Ten proces sam w sobie aktywuje krytyczne myślenie, zmuszając nas do głębszego zastanowienia się nad oczekiwanym rezultatem i potencjalnymi ścieżkami rozumowania (Lightman 2023).

Wartościową techniką jest także skłanianie modelu do przedstawienia procesu myślowego krok po kroku, co ułatwia nam krytyczną analizę sposobu „rozumowania" modelu. Warto zauważyć, że choć najnowsze modele często same stosują tę technikę, nasza rola jako krytycznych weryfikatorów jest nadal kluczowa. Musimy być gotowi na zakwestionowanie nawet pozornie logicznego rozumowania AI (OpenAi 2024).

Zadawanie krytycznych pytań w promptach

Umiejętność zadawania krytycznych pytań w promptach to ciekawy sposób na pobudzenie zarówno naszego krytycznego myślenia, jak i zdolności analitycznych modeli językowych. Ta technika prowadzi do pogłębienia odpowiedzi AI, ale również zmusza do refleksyjnego podejścia do formułowania zapytań. Oto kilka przykładowych pytań:

1. **Pytania o perspektywy** – zachęcaj model do rozważenia różnych punktów widzenia, np.: *Jakie są argumenty za i przeciw tej tezie? Rozważ perspektywy co najmniej trzech różnych grup interesariuszy.*
2. **Pytania o konsekwencje** – poproś o analizę potencjalnych skutków, np.: *Jakie mogą być konsekwencje zaproponowanej decyzji? Rozważ aspekty ekonomiczne, społeczne i środowiskowe*.
3. **Pytania o założenia** – zachęć model do identyfikacji ukrytych założeń, np.: *Jakie założenia leżą u podstaw tego rozumowania? Czy są one uzasadnione w tym kontekście?*

Skuteczna praca z modelami językowymi wymaga świadomości własnych ograniczeń i przezwyciężenia tendencji do bezkrytycznego akceptowania odpowiedzi. Możemy to osiągnąć, jak opisano powyżej, poprzez jasne określanie potrzeb i celów, strukturyzację zapytań, iteracyjne podejście i zadawanie krytycznych pytań.

2.4. Komunikacja z gen AI

Uniwersalne zasady i proces komunikacji z gen AI

Skuteczna praca z modelami językowymi opiera się na czterech podstawowych zasadach komunikacji (Phoenix, Taylor, 2023). Są to:

1. **Wyznaczanie celu** – polega na precyzyjnym określeniu kierunku wykonania i kontekstu zadania, aby uzyskać pożądane rezultaty. Może to obejmować nadawanie ról lub dostarczenie modelowi odpowiedniego kontekstu, w tym najlepszych praktyk związanych z danym zadaniem (np. *Jesteś specjalistą do spraw ewaluacji, zajmujesz się ewaluacją programów unijnych adresowanych do przedsiębiorstw i mających na celu poprawę ich innowacyjności … pracujemy nad ewaluacją działania XYZ, które polega na … Twoim zadaniem jest przygotowanie pytań badawczych w następujących obszarach …*).
2. **Określanie formatu** – koncentrowanie się na precyzyjnym zdefiniowaniu oczekiwanej odpowiedzi. Zasada ta może obejmować specyfikację konkretnych formatów danych (np. JSON, YAML, JPG) lub struktur (listy, tabele, obrazy), co ułatwia późniejsze przetwarzanie i integrację z systemami (np. aplikacjami webowymi i mobilnymi, systemami zarządzania bazami danych itp.).

3. **Dostarczanie przykładów** – polega na włączaniu do instrukcji konkretnych wzorców oczekiwanych odpowiedzi. Ta zasada jest szczególnie skuteczna, gdy trudno precyzyjnie wyjaśnić oczekiwania lub gdy użytkownik nie jest ekspertem w danej dziedzinie. Ze względu na ograniczenia długości prompta kluczowe staje się staranne dobieranie różnorodnych przykładów.
4. **Dzielenie zadań** – polega na dzieleniu złożonych instrukcji na mniejsze, mniej skomplikowane części. Może ono obejmować rozbicie zadania na sekwencję powiązanych promptów. To podejście jest szczególnie użyteczne, gdy zadanie jest zbyt skomplikowane dla pojedynczego prompta, lub gdy potrzebujemy większej kontroli nad procesem generowania odpowiedzi przez model.

Zasady promptowania nie są „magicznymi trikami”, lecz ogólnie przyjętymi konwencjami postępowania – użytecznymi w pracy z różnorodnymi systemami generatywnymi, zarówno tekstowymi, jak i graficznymi. Przykład stosowania czterech ww. zasad prezentujemy na rycinie 2.1.

Rycina 2.1. Zastosowanie zasad tworzenia instrukcji do modeli generatywnych

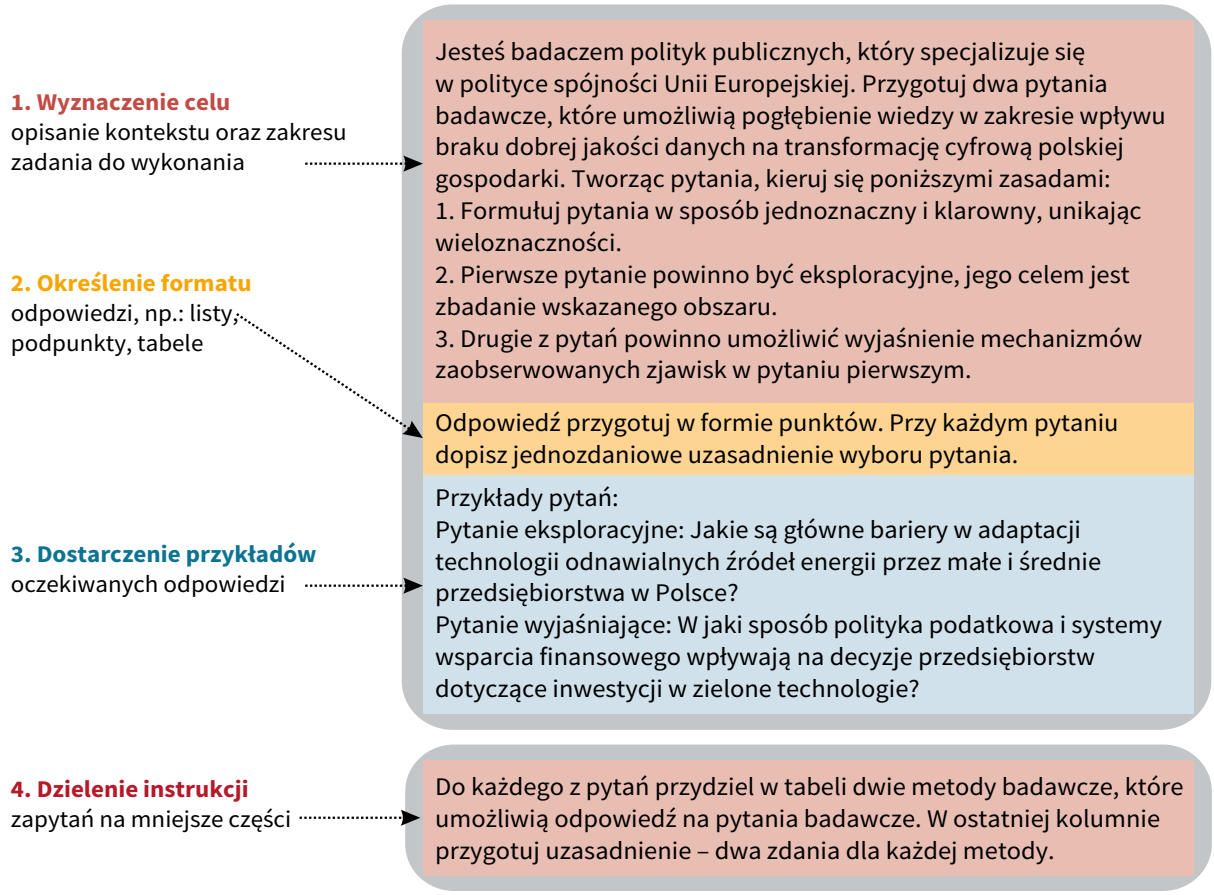


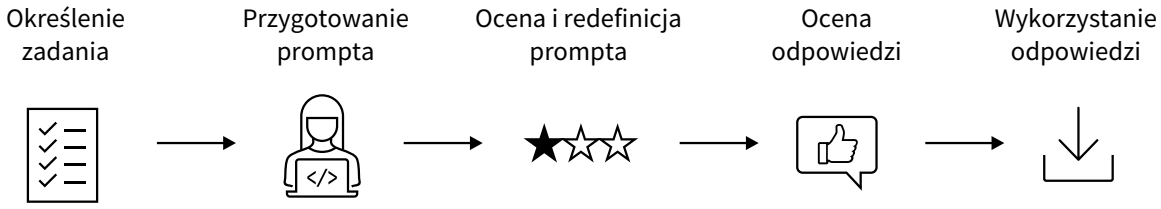
Tabela: Metody badawcze do pytań badawczych dotyczących wpływu braku dobrej jakości danych na transformację cyfrową polskiej gospodarki

Pytanie badawcze	Metoda badawcza	Uzasadnienie
Jakie są kluczowe obszary transformacji cyfrowej polskiej gospodarki, które najbardziej cierpią z powodu braku dobrej jakości danych?	Analiza danych zastanych (*desk research*)	Pozwala na zebranie informacji z raportów, publikacji i istniejących badań, aby zidentyfikować problematyczne sektory i procesy. Zapewnia szeroki obraz sytuacji i umożliwia wstępną identyfikację obszarów do dalszej analizy.
	Wywiady pogłębione (IDI)	Wywiady z ekspertami z kluczowych sektorów dostarczą szczegółowych informacji na temat wpływu niskiej jakości danych na transformację cyfrową. Pozwalają również na uzyskanie kontekstu specyficznego dla różnych branż.

Źródło: Opracowanie własne z wykorzystaniem GPT-4o.

Oprócz stosowania ww. zasad, ważna jest sekwencja kroków, jakie wykonujemy podczas pracy z modelami językowymi. Optymalnie, naszym zdaniem, gdy taki proces składa się z pięciu kroków, przedstawionych na rycinie 2.2.

Rycina 2.2. Proces komunikacji z modelem językowym w pięciu krokach



Źródło: Opracowanie własne

Określenie zadania

Zanim otworzymy aplikację z modelem językowym, warto zatrzymać się na chwilę przy pustej kartce. To moment na przemyślenie naszych potrzeb – co dokładnie chcemy osiągnąć, jaki ma być końcowy efekt. Zapisanie tych założeń pomoże nam później ocenić, czy otrzymana odpowiedź faktycznie spełnia nasze oczekiwania. Jest to również dobry moment na zebranie wszystkich informacji i materiałów, które mogą być potrzebne do przygotowania instrukcji w kolejnym kroku.

Przygotowanie prompta

Gdy mamy jasno określone zadanie, przekształcamy definiowane potrzeby i cele w skuteczny prompt. To jak tłumaczenie naszych zapisków z kartki na język, który model najlepiej zrozumie.

Ocena i udoskonalenie prompta

Ten często pomijany etap może znacząco poprawić jakość otrzymywanych odpowiedzi. Zamiast od razu wysyłać prompt do modelu, warto poświęcić chwilę na jego sprawdzenie. Możemy nawet poprosić sam model o ocenę naszego prompta, np.: ***Zanim udzielisz właściwej odpowiedzi, oceń, proszę, mój prompt pod kątem: jasności celu, określenia formatu, użytych przykładów i podziału na kroki. Co można w nim poprawić?*** Model pomoże nam wychwycić braki i udoskonalić zapytanie, zanim przejdziemy do właściwego zadania.

Ocena otrzymanego wyniku

Po otrzymaniu odpowiedzi od modelu, ale przed jej wykorzystaniem, warto poprosić sam model o jej ocenę. Możemy napisać: ***Zanim przejdę do wykorzystania tej odpowiedzi, oceń ją, proszę, pod kątem: czy trafnie odpowiada na moje pytanie, czy jest merytorycznie poprawna, czy nie ma błędów językowych, czy styl jest odpowiedni, a całość spójna***. Taka samoocena modelu może pomóc nam wychwycić potencjalne problemy i zdecydować, czy potrzebujemy doprecyzować nasze pytanie lub poprosić o korektę odpowiedzi. Jeśli pracujemy nad materiałem, który wymaga od nas wysokiej rzetelności informacji, po otrzymaniu wyników warto zrobić krótką przerwę na refleksję. Zamiast od razu akceptować odpowiedź, należy zweryfikować informację. Pomyślmy o LLM jak o jednym z wielu ekspertów w dyskusji. Tak jak nie polegamy wyłącznie na opinii jednej osoby, nie powinniśmy także „ślepo ufać" gen AI. Traktujmy ją jako jedno ze źródeł informacji. Dobrą praktyką jest sprawdzanie wyniku przynajmniej w dwóch dodatkowych źródłach (Wilson, 2024). W kontekście weryfikacji informacji warto zwrócić też uwagę na zaawansowane techniki, takie jak RAG (*Retrieval-Augmented Generation*). RAG łączy możliwości generatywne modeli językowych z dostępem do zewnętrznych baz wiedzy, co może znacząco zwiększyć dokładność i aktualność odpowiedzi. Jednak korzystając z RAG, pamiętajmy, że to jak używanie zaawansowanej wyszukiwarki – nie zawsze mamy pełną kontrolę nad wyborem materiałów, z których generowane są odpowiedzi. Dlatego kluczowe jest, aby weryfikować wiarygodność i trafność wykorzystanych materiałów w udzielanej odpowiedzi (Polzer, 2024)[5].

Wykorzystanie odpowiedzi

W ostatnim etapie wykorzystujemy zweryfikowaną odpowiedź modelu w praktyce. To moment, gdy wracamy do celów określonych na pustej kartce w pierwszym etapie i sprawdzamy, czy faktycznie uzyskaliśmy to, czego potrzebowaliśmy.

[5] Więcej o pracy modeli gen AI z bazami dokumentów i o technice RAG można znaleźć w rozdziale „Gen AI w organizacjach publicznych – regulacje i codes of conducts".

Skuteczna praca z modelem językowym to przemyślany proces, nie pojedyncze zapytanie. Zaczynamy od kartki i planowania, tworzymy przemyślany prompt, sprawdzamy go z pomocą modelu, oceniamy otrzymaną odpowiedź i dopiero wtedy ją wykorzystujemy. Dodatkowo, na każdym etapie możemy poprosić model gen AI o pomoc w ocenie i usprawnieniu naszej pracy.

Techniki promptowania

W dalszej części opisaliśmy podstawowe typy promptów. Wybraliśmy je dlatego, że uważamy, iż są względnie uniwersalne i nawet w momencie dalszego poprawiania jakości działania LLM zasady, na których się opierają, nadal mogą być przydatne. Uważamy też, że zapoznanie się z prezentowaną typologią będzie stanowiło dobrą bazę dla dalszego, samodzielnego rozwoju czytelnika. Należy zauważyć, że znając podstawowe typy promptów, możemy je łączyć i dostosowywać do różnych sytuacji, budować strategie komunikacji, które po prostu są często kombinacją tych podstawowych typów.

Ponadto, patrząc przez pryzmat potrzeb ewaluacyjnych, lepiej nauczyć się podstawowych zasad, niż bazować na konkretnej strategii podporządkowanej określonemu zadaniu. Dzieje się tak dlatego, że każde badanie ewaluacyjne jest inne i wymaga elastycznego podejścia, a nie bazowania na jednym „przepisie".

Jednocześnie czytelnikowi łatwiej będzie przyswoić podstawowe typy promptów niż wiele szczegółowych strategii, co – również w kontekście dydaktycznej funkcji tej publikacji – ma tę zaletę, że pozwala na stopniowe budowanie kompetencji oraz daje przestrzeń do eksperymentowania i własnych odkryć.

Podzieliliśmy prompty na następujące ich typy:

- proste,
- krokowe,
- warunkowe,
- kreatywne,
- *role playing*,
- ustrukturyzowane.

Instrukcje proste

Technika pozwala na uzyskiwanie szybkich i bezpośrednich odpowiedzi. Tworzenie promptów prostych jest przydatne w momencie, gdy wykonywane przez nas zadanie ma charakter jednorazowy, potrzebujemy szybkiej odpowiedzi i nie zależy nam na jej określonym formacie, oraz nie mamy również potrzeby jej opierania na szerszym kontekście. Tego typu instrukcje mogą się sprawdzać np. w sytuacjach szukania definicji pojęć. Ponadto prompty proste są instrukcjami, które mogą być przydatne również w ramach wątków, w których operujemy bardziej złożonymi promptami. Prompt prosty pozwoli nam np. na sprawdzenie znaczenia określonego terminu, wyjaśnienie jakiejś kwestii, stworzenie prostego przykładu.

Ograniczeniem instrukcji prostych jest uzyskiwanie ogólnych informacji i zasadniczo nie nadają się one do wykonywania bardziej złożonych zadań lub osiągania skonkretyzowanych celów.

Przykładami promptów prostych w obszarze ewaluacji mogą być:

1. *Użyłeś terminu X. Wyjaśnij mi, co on oznacza.*
2. *Napisałeś, że przedsiębiorstwa działają w oparciu o teorię X. Podaj przykład takiego działania.*
3. *Jakie są kluczowe wyzwania zaproponowanej przez ciebie metodyki badania ewaluacyjnego?*

Jeżeli jednak zależy nam na pogłębieniu wiedzy, chcemy tę wiedzę wykorzystać w bardziej konkretnych celach, wówczas warto prompty odpowiednio zmodyfikować, zgodnie z czterema zasadami pisania instrukcji, które przedstawiliśmy wcześniej. Modyfikacja może polegać na dodaniu szczegółowych wskazówek.

1. *Użyłeś terminu X. Wyjaśnij mi, co on oznacza. Zrób to w sposób zrozumiały dla osoby, która nie ma dużej wiedzy na temat statystyki.*
2. *Napisałeś, że przedsiębiorstwa działają w oparciu o teorię X. Podaj przykład takiego działania. Przykład dobierz w taki sposób, aby uwzględniał interwencję publiczną, o której rozmawiamy w tym wątku.*
3. *Jakie są kluczowe wyzwania zaproponowanej przez ciebie metodyki badania ewaluacyjnego? Uzasadnij każde z nich i zaproponuj sposoby radzenia sobie z tymi wyzwaniami.*

Aby pokazać działanie prompta prostego, poprosiliśmy model o stworzenie metodyki badania ewaluacyjnego, sprawdzającego skuteczność wsparcia w zakresie badań, rozwoju i innowacji, przeznaczonego dla małych i średnich przedsiębiorstw (MSP). Model otrzymał również informacje na temat charakterystyki tej interwencji publicznej. W odpowiedzi przygotował złożony opis. Jedną z zaproponowanych technik była tzw. analiza kontrfaktyczna. Prompt prosty posłużył do wyjaśnienia kwestii użycia tej analizy: *Jakie są kluczowe wyzwania zaproponowanej przez ciebie analizy kontrfaktycznej? Uzasadnij każde z nich i zaproponuj sposoby radzenia sobie z tymi wyzwaniami.* Model (Claude 3,5 Sonnet) w odpowiedzi poinformował nas m.in. o tym, że wyzwania to selekcja uczestników, wyjaśnił, na czym polega to wyzwanie, oraz zaproponował rozwiązanie problemu:

> „Analiza kontrfaktyczna w kontekście oceny wsparcia B+R+I dla MSP rzeczywiście wiąże się z szeregiem istotnych wyzwań metodologicznych. Pozwól, że omówię najważniejsze z nich wraz z propozycjami ich rozwiązania:
>
> 1. Selekcja uczestników (*selection bias*)
>
> Wyzwanie:
>
> - Firmy, które aplikowały i otrzymały wsparcie, mogły się systematycznie różnić od pozostałych MSP.
> - Mogły być bardziej innowacyjne, lepiej zarządzane lub mieć większe zasoby już na starcie.
> - To utrudnia znalezienie odpowiedniej grupy kontrolnej.
>
> Rozwiązania:
>
> - Zastosowanie dwustopniowego doboru grupy kontrolnej:
>   - Pierwszy etap: *matching* ze względu na obserwowalne cechy (wielkość, branża, region, historie finansowe).
>   - Drugi etap: uwzględnienie *pre-treatment trends* w działalności B+R (…)”.

Ogólnie rzecz biorąc, można być zadowolonym z uzyskanej odpowiedzi, choć pokazuje ona, że prompt prosty ma ograniczenia. Zwróćmy uwagę, że odpowiedź może nie być zrozumiana przez użytkownika, który nie miał wcześniej do czynienia z analizami kontrfaktycznymi. Przykładowo, nie zostało wyjaśnione pojęcie *pre-treatment trends*, natomiast w całej wypowiedzi, której tutaj nie publikujemy, model dodatkowo użył takich terminów jak: *stagged DiD*, „kontaminacja grupy kontrolnej” itp. Dlatego osoby, które nie mają doświadczenia z analizami kontrfaktycznymi, będą zmuszone do wysłania do modelu

kolejnych zapytań wyjaśniających te kwestie. Można było tego uniknąć poprzez zastosowanie bardziej zaawansowanej instrukcji, np. zawierającej kontekst. *Ad hoc* zmodyfikowaliśmy instrukcję: *Jakie są kluczowe wyzwania zaproponowanej przez ciebie analizy kontrfaktycznej? Uzasadnij każde z nich i zaproponuj sposoby radzenia sobie z tymi wyzwaniami. Pamiętaj, że nie znam się na analizach kontrfaktycznych, nigdy ich nie prowadziłem. Dlatego wyjaśnij w sposób prosty używane przez ciebie terminy, posługuj się też przykładami, abym mógł lepiej zrozumieć ten temat.* Poniżej znajduje się fragment odpowiedzi modelu. Jak widzimy, tym razem postarał się, aby przekaz był łatwy do zrozumienia.

> „Oto główne wyzwania:
>
> 1. Problem «podobnych» firm. Wyzwanie:
> - Firmy, które dostały wsparcie, mogły być już na starcie «lepsze» od innych.
> - To jak porównywanie uczniów, którzy sami zgłosili się na dodatkowe zajęcia, z tymi, którzy się nie zgłosili – ci pierwsi mogli być już bardziej zmotywowani.
> - Przykład: Wyobraźmy sobie dwie firmy produkujące meble:
> - Firma A: już wcześniej inwestowała w innowacje, ma dobrych specjalistów
> - Firma B: nigdy nie prowadziła działań innowacyjnych. Jeśli firma A dostanie wsparcie i osiągnie lepsze wyniki, trudno powiedzieć, czy to dzięki wsparciu, czy dlatego, że już wcześniej była lepsza (…)".

Instrukcje krokowe

Technika ta pozwala na zachowanie kolejności zadań wykonywanych przez model. Umiejętność wykorzystywania tych promptów jest ważna, zwłaszcza w momencie, gdy zadanie jest złożone, zaś wyniki pracy modelu mogą się różnić w zależności od wyników wcześniej wykonanych operacji w ramach danej instrukcji lub wątku, w którym pracujemy (np. wątek poświęcony tworzeniu Opisu Przedmiotu Zamówienia na ewaluację). Bez zastosowania prompta krokowego (lub elementu prompta krokowego) narażamy się na ryzyko, że model pominie jedną z wymaganych przez nas operacji lub wykona ją w sposób niezgodny z oczekiwaną sekwencją. Jednocześnie instrukcje krokowe pozwalają modelowi na tworzenie odpowiedzi na podstawie wcześniej wykonanego zadania, co sprawia, że efekt uzyskanej pracy jest lepszy[6].

[6] Różnica działania modelu wykorzystującego instrukcję prostą (prompt prosty) *versus* prompt krokowy została przedstawiona na schemacie.

Ograniczeniem instrukcji krokowych jest przede wszystkim ich czasochłonność, zarówno na etapie tworzenia prompta, jak również na etapie walidacji. Ponadto musimy pamiętać, że tego typu instrukcja jest narzuceniem modelowi określonej struktury działań, co może zmniejszać jego kreatywność (np. model może „myśleć" linearnie i pomijać alternatywne sposoby rozwiązania problemu). Pewnym rozwiązaniem problemu może być wprowadzenie do instrukcji szerszego kontekstu. Przykładowo model, który ma ogólne informacje na temat naszej interwencji, może projektować metodykę ewaluacji, sięgając do standardowych rozwiązań biznesowych, a tym samym może pominąć specyfikę grupy poddanej oddziaływaniu pomocy publicznej. Może to sprawić, że zaproponowane przez model wskaźniki (np. krótkoterminowe) niekoniecznie będą odpowiednie. Ale jeżeli uzyska bardziej szczegółowy kontekst, wówczas zaproponuje lepsze rozwiązanie, np. skupienie się na oddziaływaniu długoterminowym, które jest bardziej typowe dla specyfiki firm – beneficjentów pomocy na B+R+I.

Pamiętajmy również, że pewne rzeczy stają się dla nas bardziej oczywiste wraz z postępami pracy. Oznacza to, że w instrukcjach tego typu (krokowych) możemy zostać zmuszeni do ponownego wykonania części kroków, zaś zmiana jednego elementu będzie za sobą pociągała konieczność sprawdzenia powiazań z innymi elementami. Ciągłe powracanie do wcześniejszych kroków i ich modyfikacja mogą wydłużyć pracę.

Prompty krokowe przygotowujemy, używając następujących reguł:

1. Staramy się numerować listy lub nagłówki. Dzięki temu wyraźnie oznaczamy każdy krok naszego działania. Ma to znaczenie, bo z jednej strony zmniejsza ryzyko wystąpienia błędu po stronie modelu, z drugiej zaś może być nam to pomocne w rozwijaniu instrukcji – będziemy mogli się w niej odwoływać do konkretnych punktów.
2. Możemy odnosić się do danych plików lub innych materiałów źródłowych.
3. Możemy przedstawiać konkretne wskazówki dotyczące formatu i stylu odpowiedzi.

Poniżej przygotowaliśmy przykładową instrukcję krokową.

> *Zaprojektuj badanie ewaluacyjne dla programu Fundusze Europejskie dla Nowoczesnej Gospodarki (FENG), wykonując następujące kroki:*
> 1. *##Krok 1: Przeanalizuj cele programu FENG i zidentyfikuj 3–5 kluczowych obszarów, które powinny podlegać ewaluacji. Dla każdego obszaru krótko uzasadnij jego wybór.*

2. *##Krok 2: Na podstawie zidentyfikowanych obszarów sformułuj główne pytania badawcze (2–3 dla każdego obszaru). Upewnij się, że pytania są konkretne i mierzalne.*
3. *##Krok 3: Dla każdego pytania badawczego zaproponuj odpowiednie metody i techniki badawcze, uwzględniając zarówno metody ilościowe, jak i jakościowe. Uzasadnij swój wybór.*
4. *##Krok 4: Określ źródła danych potrzebne do odpowiedzi na pytania badawcze. Uwzględnij zarówno dane zastane, jak i te, które trzeba będzie zebrać w trakcie badania.*
5. *##Krok 5: Zaproponuj dobór próby dla badań ilościowych i jakościowych, biorąc pod uwagę specyfikę beneficjentów programu FENG. Określ wielkość próby i kryteria doboru respondentów.*
6. *##Krok 6: Opracuj wstępny harmonogram badania, uwzględniając czas potrzebny na każdy etap (przygotowanie narzędzi, zbieranie danych, analiza, raportowanie). Harmonogram powinien być realistyczny i uwzględniać specyfikę programu FENG.*
7. *##Krok 7: Zidentyfikuj potencjalne ryzyka, które mogą wpłynąć na realizację badania (np. niska responsywność beneficjentów, trudności w dostępie do danych) i zaproponuj sposoby ich minimalizacji.*
8. *##Krok 8: Na podstawie poprzednich kroków przygotuj krótkie (max. 300 słów) streszczenie koncepcji badania, które mogłoby zostać przedstawione instytucji zarządzającej programem FENG.*

Po wykonaniu każdego kroku krótko podsumuj jego wyniki przed przejściem do następnego. Jeśli w trakcie wykonywania któregoś z kroków pojawi się potrzeba dodatkowych informacji, zaznacz to i poproś o doprecyzowanie przed kontynuowaniem. ***Do kolejnego kroku przejdź dopiero po mojej akceptacji poprzedniego.***

Istotna jest część końcowa instrukcji napisana pogrubioną czcionką. Dzięki niej model nie wykona wszystkich kroków naraz, ale zanim przejdzie do następnego kroku, najpierw spyta o naszą akceptację. Takie podejście umożliwi nam utrzymanie kontroli nad przygotowaną treścią. Na rycinie 2.3. znajduje się fragment odpowiedzi przygotowanej przez GPT-4o[7].

[7] Przykład odpowiedzi udzielonej przez Claude. Należy dodać, że jest to połączony wynik dla kroku 1. i 2. po wcześniejszej akceptacji tych kroków przez użytkownika.

Rycina 2.3. Fragment odpowiedzi modelu GPT na prompt krokowy

Krok 1: Analiza celów programu FENG i identyfikacja kluczowych obszarów do ewaluacji

Cele programu FENG:

Program Fundusze Europejskie dla Nowoczesnej Gospodarki (FENG) wspiera innowacyjność i konkurencyjność polskiej gospodarki poprzez finansowanie projektów badawczo-rozwojowych, innowacyjnych technologii oraz transformacji cyfrowej i ekologicznej.

Kluczowe obszary do ewaluacji:

1. Efektywność finansowania innowacji:
   - Uzasadnienie: Ocena wzrostu innowacyjności w przedsiębiorstwach.
2. Wpływ na konkurencyjność gospodarki:
   - Uzasadnienie: Badanie wpływu programu na pozycję polskich firm na rynku międzynarodowym.
3. Transformacja cyfrowa przedsiębiorstw:
   - Uzasadnienie: Ocena adaptacji firm do nowoczesnych technologii.
4. Działania proekologiczne:
   - Uzasadnienie: Ocena wdrażania technologii sprzyjających ochronie środowiska.
5. Efektywność administracyjna i zarządzanie programem:
   - Uzasadnienie: Analiza efektywności zarządzania i przejrzystości procesów aplikacyjnych.

Podsumowanie kroku 1.:

Zidentyfikowano pięć kluczowych obszarów do ewaluacji: efektywność finansowania innowacji, wpływ na konkurencyjność gospodarki, transformacja cyfrowa przedsiębiorstw, działania proekologiczne oraz efektywność administracyjna i zarządzanie programem.

Proszę o akceptację wyników kroku 1. lub sugestie przed przejściem do kroku 2.

Źródło: Opracowanie własne z wykorzystaniem GPT-4o.

Instrukcje warunkowe

Ta technika pozwala nam na zawężenie pracy modelu do określonego przez nas zakresu/obszaru. Stosując ten typ prompta, możemy wprowadzić różne warunki brzegowe,

ograniczenia i reguły. Co do zasady, warto sięgać po ograniczenia (warunki), jeżeli zależy nam na tym, aby model dostarczał bardzo skonkretyzowanych informacji, oraz kiedy chcemy uniknąć niepożądanych treści. Instrukcje warunkowe mogą mieć szerokie zastosowanie w ewaluacji. Dzięki nim możemy ograniczyć liczbę celów badania, pytań badawczych, metod badawczych, pytań ankiety (np. *Zaproponuj nie więcej niż…*), dopasować styl wypowiedzi do naszej grupy docelowej (np. *Pytania ankiety mają być przygotowane w sposób zarozumiały dla przedsiębiorców, nie używaj języka oficjalnego…*), określić ramy czasowe (np. *Uwzględnij wartość wskaźników osiąganą od roku… do roku…*). Pracując zaś z danymi, możemy dokonywać ich selekcji (np. *Przeprowadź analizę załączonego zbioru danych. Zrób to, korzystając z następujących metod analizy statystycznej… i zmiennych na kolumnach… Przedstaw wyniki, ograniczając się wyłącznie do różnic istotnych statystycznie*), możemy również „zmusić model" do pracy w określonym obszarze tematycznym (np. *Szukaj informacji wyłącznie dotyczących efektów interwencji publicznych realizowanych na obszarze Polski*).

Ograniczenia instrukcji warunkowych są podobne do tych występujących przy promptach krokowych, czyli mogą one zbytnio zawęzić obszar działania, a tym samym prowadzić do pominięcia istotnych treści. Tworzenie takich instrukcji również jest czasochłonne, ponieważ wymagają one formułowania bardzo precyzyjnych warunków. Natomiast w przeciwieństwie do instrukcji krokowych, te warunkowe łatwiej jest tworzyć oraz modyfikować (mniejsza czasochłonność pracy).

Poniżej został przedstawiony przykład prompta warunkowego:

> *Przygotuj listę trzech pytań do scenariusza wywiadu pogłębionego z beneficjentem programu rozwoju kompetencji cyfrowych dla osób w wieku 50+. Pytania powinny pomóc w ewaluacji skuteczności programu. Przy tworzeniu pytań przestrzegaj następujących zasad:*
>
> - *Nie używaj żadnych określeń, które mogłyby zostać odebrane jako dyskryminujące ze względu na wiek.*
> - *Unikaj sformułowań sugerujących, że osoby starsze mają problemy z nowymi technologiami.*
> - *Skoncentruj się na praktycznych aspektach wykorzystywania nabytych umiejętności.*
> - *Każde pytanie musi być otwarte (nie może być odpowiedź tak/nie).*
> - *Maksymalna długość pojedynczego pytania to 20 słów.*

Dodatkowo, dla każdego pytania podaj krótkie (max. 15 słów) uzasadnienie, dlaczego jest ono istotne dla ewaluacji programu.

Rycina 2.4. Odpowiedź modelu Claude na prompt warunkowy

Oto lista 3 pytań do scenariusza wywiadu pogłębionego z beneficjentem programu rozwoju kompetencji cyfrowych, zgodnie z podanymi wytycznymi:

1. W jaki sposób umiejętności zdobyte podczas programu wpłynęły na Pana/Pani codzienne aktywności?

Uzasadnienie: Pozwala ocenić praktyczne zastosowanie nabytych kompetencji i ich wpływ na jakość życia uczestnika.

2. Które z narzędzi cyfrowych poznanych w trakcie szkolenia okazały się dla Pana/Pani najbardziej przydatne i dlaczego?

Uzasadnienie: Umożliwia identyfikację najbardziej wartościowych elementów programu z perspektywy uczestnika.

4. Jakie nowe możliwości otworzyły się przed Panem/Panią dzięki umiejętnościom nabytym podczas programu?

Uzasadnienie: Pomaga zrozumieć szerszy wpływ programu na życie zawodowe i społeczne uczestnika.

Źródło: Opracowanie własne z wykorzystaniem Claude 3.5 Sonnet.

Prompty krokowe oraz warunkowe wymagają od użytkownika bardzo precyzyjnego tworzenia instrukcji. Nie oznacza to jednak, że wszystkie typy promptów muszą być tworzone z uwzględnieniem „żelaznych reguł". Kolejny typ prompta stanowi przeciwieństwo dotychczas opisanych. Modele mogą być kreatywne, możemy sprawić, że zaczną myśleć nieszablonowo (*out of box)*.

Instrukcje kreatywne

Ten typ instrukcji nie zawiera ograniczeń i daje modelowi dużą swobodę działania w tworzeniu odpowiedzi.

Dzięki stosowaniu tego typu instrukcji możemy uzyskać kreatywność i większe zróżnicowanie odpowiedzi. Jest to związane z tym, że brak ograniczeń pozwala modelowi bardziej swobodnie zgłębiać temat oraz proponować niestandardowe rozwiązania lub przedstawiać różne punkty widzenia. Ponadto instrukcje te są przydatne, gdy szukamy nowych pomysłów. Model, który nie jest ograniczony instrukcją, może nas pozytywnie zaskoczyć, umożliwiając nam odkrycie np. inspirujących, nieoczywistych koncepcji. Prompty kreatywne mogą być też punktem wyjścia do dalszej pracy i jest to ważne zwłaszcza w momencie, gdy istnieje ryzyko, że wprowadzając ograniczenia, możemy mocno zawęzić pole działania modelu w narzuconych mu ramach. Istotne jest, aby tego typu instrukcja mimo wszystko dotyczyła obszaru, który jest nam względnie dobrze znany. W przeciwnym razie trudniej będzie nam zweryfikować nieszablonowy sposób „myślenia" modelu czy „rozumienia" przez niego poszczególnych pojęć.

Jeżeli szukać zastosowań tych instrukcji w obszarze ewaluacji, to mogą być one przydatne m.in. w opracowaniu rekomendacji (np. *Stwórz niestandardową metodę poradzenia sobie z problemem zidentyfikowanym w badaniu)* i komunikowania wyników (np. *Opracuj pomysł na nieszablonową formę przedstawienia wyników badania…*). Mogą również sprawdzać się w pomocy przy kreśleniu mniej oczywistych efektów interwencji (np. *Rozważ wpływ programu na różne sfery funkcjonowania przedsiębiorstw… Szczególnie interesują mnie nietypowe powiązania pomiędzy różnymi elementami teorii zmiany…)* itp.

Ograniczeniem instrukcji kreatywnych jest przede wszystkim trudność weryfikacji uzyskanych odpowiedzi. Możemy również uzyskiwać odpowiedzi, które są luźno lub nie są powiązane z tematem.

Poniżej znajduje się przykład instrukcji kreatywnej oraz omówienie wyników pracy modelu:

1. *Stwórz kreatywną metaforę lub analogię, która pomoże wyjaśnić złożoność i wielowymiarowość programu FENG oraz proces jego ewaluacji. Możesz czerpać inspirację z dowolnej dziedziny - ekosystemu przyrodniczego, skomplikowanej maszyny, procesu twórczego czy nawet kulinariów. Jak różne elementy FENG (wsparcie innowacji) oraz jego ewaluacji odpowiadałyby elementom w twojej analogii?*

   W odpowiedzi model uznał, że FENG może być przedstawiony na podobieństwo ekosystemu rafy koralowej. Elementom programu FENG zostały przypisane odpowiedniki ekosystemu rafy. Przedsiębiorcy to korale, bo zdaniem modelu ich różnorodność odzwierciedla różnorodność firm, zaś wzrost korali może stanowić symbol

rozwoju innowacyjności, glony symbiotyczne z kolei są utożsamiane z partnerstwem nauka-biznes, prądy morskie to natomiast trendy rynkowe i technologiczne itd. Z kolei proces ewaluacji został opisany jako ekspedycja badawcza. Ewaluatorzy są nurkami, sprzęt badawczy to metody i narzędzie ewaluacyjne, mapa rafy to plan ewaluacji, próbki wody i tkanek to dane ilościowe i jakościowe, obserwacja zachowań odgrywa rolę analizy jakościowej itd.[8]

2. *Zaprojektuj interaktywne doświadczenie lub grę, która mogłaby służyć jako innowacyjne narzędzie do zbierania danych i opinii od beneficjentów programu FENG. Jak to doświadczenie mogłoby angażować uczestników, mierzyć rzeczywisty wpływ programu na ich działalność i jednocześnie edukować o celach FENG?*

W wyniku tego zapytania otrzymaliśmy propozycję zagrania w „FENG Innowator" – interaktywną grę symulacyjną, w której beneficjenci zarządzają wirtualnym przedsiębiorstwem, podejmują decyzje związane z innowacjami, cyfryzacją, zrównoważonym rozwojem i ekspansją międzynarodową. W dalszej części model rozwija opis, przekazuje informacje dotyczące mechaniki gry itd.[9]

Instrukcje typu *role playing*

Ta technika nakazuje modelowi przyjęcie określonej tożsamości lub perspektywy eksperta z danej dziedziny. Daje to możliwości (przybliża nas) do uzyskania odpowiedzi uwzględniającej specyficzny punkt widzenia, w tym odwołujący się do wiedzy specjalistycznej charakterystycznej dla danego obszaru.

Korzystanie z tych promptów jest przydatne w momencie, gdy decydujemy się na pracę w bardziej zawężonym obszarze tematycznym (problemowym) i zależy nam na odpowiedziach silnie powiązanych z badanym tematem (problemem). Ponadto prompt *role-playing* może być wykorzystywany do uzyskania odpowiedzi uwzględniającej perspektywę „ekspercką". Przykładowo, możemy oczekiwać od modelu gen AI, aby rozpatrywał jakieś zagadnienie, wcielając się w rolę specjalisty ds. innowacyjności. Praktycznie nie ma tu ograniczeń – model może się wcielić w rolę badacza, psychoterapeuty, inżyniera, informatyka, ucznia określonej

8 Wynik pracy modelu.
9 Wynik pracy modelu.

klasy szkoły podstawowej itp., choć oczywiście ostateczna skuteczność w przypisanej roli będzie zależała od danych treningowych, na których model był uczony.

Z perspektywy ewaluacji nakazywanie modelowi przyjęcia określonej roli jest przydatnym rozwiązaniem. Przykładowo, możemy od modelu oczekiwać generowania obszarów zainteresowania (potrzeb wiedzy) na etapie projektowania badania, np. obsadzić go w roli jednego z interesariuszy ewaluacji. Innym ciekawym zastosowaniem instrukcji *role playing* może być przygotowywanie rekomendacji lub ich walidacja w momencie, gdy model odgrywa rolę jednego z adresatów. Podczas pracy z modelami testowaliśmy ich działanie w takich rolach jak: specjalista od innowacyjności przedsiębiorstw (praca nad koncepcją ewaluacji), psychoterapeuta i psychiatra (praca nad interpretacją wyników)[10], ekspert-ekonomista zajmujący się pracą akademicką (do oceny wpływu proponowanych koncepcji na różne grupy społeczne, w tym identyfikacji zysków i strat w poszczególnych grupach) itp.

Należy pamiętać, że w przypadku ewaluacji będziemy raczej zmuszeni do nadawania modelowi więcej niż jednej roli lub tworzenia tzw. person, łączących więcej niż jedną rolę[11]. Przyczyną jest to, że często nasze projekty wymagają wiedzy z wielu dziedzin i ograniczenie modelu do jednej funkcji mogłoby zawężać perspektywę, co byłoby niekorzystne dla jakości procesu badawczego, np. pomijałoby koszty społeczne analizowanego rozwiązania, zanadto skupiało się na korzyściach określonej grupy interesariuszy[12]. Ponadto z naszych doświadczeń wynika, że *role-playing* lepiej sprawdzało się w sytuacjach uprzedniego przedstawienia modelowi szczegółowego kontekstu pracy, jaką ma wykonać. Nie wyklucza to jednak możliwości stosowania omawianej instrukcji do zadań o charakterze bardziej ogólnym lub abstrakcyjnym.

Prompty *role-playing* pomagają częściowo rozwiązać problem nadmiernego uogólniania odpowiedzi modelu. Ponadto technika ta może być wykorzystana do testowania różnych

[10] W tym przypadku, ze względu na fakt, że badanie dotyczyło szczególnie wrażliwego obszaru, wyniki pracy modelu zostały poddane ocenie przez specjalistów z tych dziedzin i wypadła ona pozytywnie, choć należało dokonać pewnych zmian. Co ciekawe, model dodatkowo otrzymał instrukcję odwoływania się w swych wypowiedziach do teorii z zakresu zdrowia psychicznego i zrobił to prawidłowo. Nie odnotowaliśmy halucynacji, wnioski były prawidłowo osadzone w teorii.

[11] Więcej o tworzeniu person na potrzeby ewaluacji można znaleźć w rozdziale pt. „Strategie komunikacji i prezentacji wyników".

[12] Przykładowo, proponując integrację usług środowiskowych, model skupiłby się na korzyściach dla klientów (przyjęcie perspektywy eksperta ds. wspierania osób w trudnej sytuacji), zaś pominął kwestie związane z wielośródłowym finansowaniem takich usług (różne budżety i wielość zaangażowanych instytucji itp.).

scenariuszy (weryfikacji hipotetycznych scenariuszy i kontekstów). Co do zasady, technika ta może nam dać:

- **Precyzję i dostosowanie odpowiedzi** – chcemy wykorzystać gen AI do zaprojektowania indywidualnego wywiadu pogłębionego (IDI) z przedstawicielem społeczności lokalnej na temat potrzeb środowiskowych. Zamiast ogólnej instrukcji dla gen AI (*Przeprowadź wywiad na temat…*) możemy precyzyjnie określić rolę (*Wciel się w profesjonalnego moderatora, który ma za zadanie uzyskać szczegółowe informacje na temat wyzwań środowiskowych społeczności lokalnej*). Jak widać, podany przykład będzie poprawiał precyzję oraz ukierunkowanie na konkretny cel.
- **Lepszy przepływ treści** – korzystamy z narzędzi gen AI jako asystentów do transkrypcji i analizy wyników wywiadów. Jeśli zostanie określona rola: *Bądź uważnym i obiektywnym obserwatorem, który szczegółowo dokumentuje przebieg dyskusji,* system będzie podążał za naturalną konwersacją i zachowywał spójność.
- **Zwiększoną kreatywność** – chcemy wspierać działania promujące zrównoważoną gospodarkę. Instrukcja: *Zaproponuj kreatywne strategie komunikacyjne na temat…* jest poprawna i może dać dobre wyniki, lecz standardowe. Określając jednak rolę typu: *Wciel się w awangardową artystkę, której celem jest pobudzenie świadomości społecznej…*, możemy uzyskać bardziej innowacyjne i niekonwencjonalne pomysły.

Ograniczeniem tego typu instrukcji może być brak danych treningowych (lub ich mała ilość) w modelu, co z kolei sprawi, że niektóre role będą odgrywane w sposób bardzo ogólny lub wręcz niezgodny z rzeczywistością – model, chcąc wypełnić swoje zadanie, może zacząć „halucynować". Dlatego szczególnie ważna jest walidacja jego pracy, weryfikacja uzyskanych odpowiedzi. Wcześniej napisaliśmy, że podczas pracy z modelem wcielił się on m.in. w rolę psychoterapeuty i psychiatry oraz że wynik pracy został przeanalizowany przez specjalistów z tych dziedzin. Jednakże trzeba dodać, że odpowiedzi modelu weryfikowaliśmy także za pomocą innego narzędzia opartego na sztucznej inteligencji (Perplexity) oraz sprawdzaliśmy wnioski, korzystając z internetowych baz zawierających opracowania naukowe.

Ponadto zastosowanie omawianej instrukcji jest swego rodzaju warunkowaniem pracy modelu, a tym samym – zawężaniem jego działań. Może zatem dojść do pomijania kwestii istotnych dla problemu, ale uznanych przez model za wykraczające poza przyjętą rolę.

Poniżej przedstawiamy przykład prompta *role-playing*. Tworząc go, staraliśmy się uwzględnić wszystkie atrybuty tego typu instrukcji:

Wciel się w rolę głównego eksperta ds. ewaluacji programów wsparcia innowacyjności i rozwoju gospodarczego, z 20-letnim doświadczeniem w ocenie skuteczności funduszy europejskich. Twoja wiedza obejmuje zarówno aspekty ekonomiczne, jak i technologiczne innowacji, ze szczególnym uwzględnieniem specyfiki polskiego rynku i sektora MŚP. Jesteś znany z holistycznego podejścia do ewaluacji, łączącego analizę ilościową z jakościową oraz z umiejętności identyfikowania długoterminowych efektów interwencji. Twoim zadaniem jest ocena jakości raportu z badania ewaluacyjnego programu Fundusze Europejskie dla Nowoczesnej Gospodarki (FENG), pod tytułem: …

Raport z badania ewaluacyjnego znajduje się w załączniku. W #krok 1# zapoznaj się z raportem. W #kroku 2# oceń raport, uwzględniając dwa kryteria:

- *Kryterium 1. Obiektywność: Czy wnioski są bezstronne i oparte na dowodach?*
- *Kryterium 2. Jakość analizy: Czy dane są analizowane w sposób dogłębny i krytyczny?*

Ocenę badania przygotuj w formie krótkiej notatki, wskazując pozytywne i negatywne oceny raportu dla obu kryteriów.

Rycina 2.5. Fragment odpowiedzi modelu GPT na prompt *role-playing*

Ocena Raportu z Badania Ewaluacyjnego Programu Fundusze Europejskie dla Nowoczesnej Gospodarki (FENG)

Kryterium 1.: Obiektywność

Pozytywne aspekty:

• Raport stara się bazować na szerokim zakresie źródeł, w tym literaturze naukowej, dokumentach rządowych, raportach organizacji międzynarodowych oraz analizach sektora prywatnego. To podejście wskazuje na próbę obiektywnego przedstawienia trendów technologicznych oraz oceny interwencji.

• W raporcie wykazano dbałość o uwzględnienie różnych perspektyw i ryzyka oportunistycznych zachowań beneficjentów, co dodatkowo potwierdza starania o zachowanie obiektywności.

Negatywne aspekty:

• W niektórych częściach raportu brak jest pełnego przedstawienia alternatywnych scenariuszy lub opinii. Na przykład podkreślanie sukcesów interwencji może przesłaniać potencjalne wady lub niepowodzenia.

• Raport może w niektórych miejscach być zbyt optymistyczny w ocenie wpływu programów na beneficjentów, co może sugerować pewną stronniczość na rzecz prezentowania wyników w pozytywnym świetle.

Źródło: Opracowanie własne z wykorzystanie GPT-4o.

Instrukcje ustrukturyzowane

Instrukcje ustrukturyzowane polegają na tworzeniu zapytań kierowanych do modeli językowych na podstawie predefiniowanych szablonów. Zawierają one elementy niezbędne do wygenerowania precyzyjnych i dobrej jakości odpowiedzi. Stosujemy je, aby zapewnić spójność i kompletność informacji w instrukcji, uwzględniając istotne aspekty, takie jak kontekst i cel, czy docelowych odbiorców odpowiedzi.

Przykładowym szablonem stosowanym w pracy z modelami językowymi jest **„CO-STAR”** (Teo, 2023). Został on opracowany przez zespół Data Science & AI w GovTech Singapore. Szablon składa się z sześciu elementów:

- *Context* (Kontekst): opis tła zadania.
- *Objective* (Cel): opis celu zadania.
- *Style* (Styl): opis stylu odpowiedzi.
- *Tone* (Ton): opis tonu odpowiedzi.
- *Audience* (Odbiorcy): opis docelowych odbiorców.
- *Response* (Odpowiedź): opis formatu odpowiedzi.

Poniżej prezentujemy przykład zastosowania tej techniki do stworzenia materiałów służących promocji wyników badania ewaluacyjnego.

> Przykładowa instrukcja CO-START
>
> *#KONTEKST#*
>
> *Wykorzystaj ten kontekst:*
>
> *Nasza organizacja zakończyła badanie skupione na wsparciu przedsiębiorstw w województwie x. W ramach projektów realizowanych z działania X uruchomiono 19 nowych specjalistycznych usług doradczych, z których skorzystało ponad 200 przedsiębiorców. Z usług doradczych świadczonych przez Instytucje Otoczenia Biznesu, akredytowane w systemie X, skorzystało ponad 700 firm. Przedsiębiorstwa te bardzo wysoko oceniły ich jakość (średnia ocena usług na skali od 1 do 5 wahała się od 4,77 do 4,95 w zależności od obszaru tematycznego usługi). 70% firm, dzięki skorzystaniu ze specjalistycznych usług doradczych, wprowadziło innowacje, a prawie 80% zadeklarowało polepszenie swojej pozycji konkurencyjnej.*

#CEL#

Na podstawie wniosków z oceny działania X opracuj trzy posty) na profil urzędu na Facebook. Celem postów jest wzbudzenie zainteresowania wynikami badania. Posty będą ukazywać się w odstępach co 2 dni. Ich treść nie może się powielać. Stosuj emotki.

#STYL#

Kreatywny i inspirujący. Należy naśladować styl kampanii, które skutecznie przekazują wartości i korzyści płynące z innowacji i wsparcia technologicznego."

#TON#

Pozytywny, motywujący, podkreślający możliwości rozwoju. Powinien podkreślać osiągnięcia i realne przykłady zmian, jakie dokonały się dzięki programowi.

#ODBIORCY#

Posty będą skierowane do małych przedsiębiorców zgodnie z profilem:

Profil

Mały przedsiębiorca, zatrudniający do 10 osób, charakteryzuje się dużą elastycznością i zdolnością do szybkiego reagowania na zmiany rynkowe. Działa na lokalnym rynku, często w sektorze usługowym lub handlowym. Dzięki niewielkiej skali działalności może nawiązywać bezpośrednie relacje z klientami, co pozwala lepiej zrozumieć ich potrzeby. Przedsiębiorca pełni wiele funkcji jednocześnie – od zarządzania finansami, przez marketing, po bezpośrednią obsługę klienta, co wymaga wszechstronności i umiejętności wielozadaniowości. Działa lokalnie, angażując się w życie społeczności, co przynosi korzyści w postaci lojalnej bazy klientów. Mały przedsiębiorca wnosi innowacyjność, elastyczność i lokalne zaangażowanie, które są niezbędne dla zrównoważonego rozwoju ekonomicznego.

#ODPOWIEDŹ-FORMAT#

Posty zawierające jedno zdjęcie oraz krótką wiadomość tekstową. Post ma zawierać pytania angażujące społeczność.

Z precyzyjnie określonej instrukcji, opartej na szablonie CO-STAR, model ChatGPT-4o stworzył serię postów odzwierciedlających nasze intencje i wymagania. Na rycinie 2.6. przedstawiamy fragment odpowiedzi wygenerowanej przez model:

Rycina 2.6. Fragment odpowiedzi modelu GPT na prompt ustrukturyzowany

Post 1: Wspieramy Twój rozwój! 🚀

📣 Czy wiesz, że dzięki wsparciu działania X, w województwie x wprowadzono aż **19 nowych specjalistycznych usług doradczych**, które pomogły **ponad 200 przedsiębiorcom** rozwijać skrzydła? 💼

📊 700 firm skorzystało z usług Instytucji Otoczenia Biznesu i oceniło ich jakość na rewelacyjną średnią: od **4,77 do 4,95**!👏

🧐 **Czym są innowacje, które zmieniają Twoją branżę?**
👉 Podziel się swoją historią w komentarzu – chcemy usłyszeć o Twoich sukcesach! 💬

#Innowacje #Rozwój #WsparcieDlaBiznesu

Źródło: Opracowanie własne z wykorzystaniem GPT-4o.

Uzyskany tekst jest dostosowany do specyfiki odbiorców i celów marketingowych, demonstruje skuteczność metody CO-STAR w tworzeniu precyzyjnych instrukcji.

2.5. Podsumowanie

Komunikacja z modelami językowymi to umiejętność, która będzie zyskiwała na znaczeniu wraz z postępującą transformacją cyfrową. Przedstawione w rozdziale ograniczenia po stronie użytkownika i maszyny, zasady i sposób prowadzania „rozmowy" z modelami oraz podstawowe typy promptów – proste, krokowe, warunkowe, kreatywne, *role playing* oraz ustrukturyzowane – dają czytelnikowi uniwersalny zestaw narzędzi, który pozwoli na efektywne wykorzystanie LLM w różnych zadaniach i procesach ewaluacyjnych.

Należy jednak pamiętać o zachowaniu krytycznego spojrzenia na rolę modeli językowych w naszej pracy, pamiętać o ich ograniczeniach, jak również mieć na uwadze te, które stoją po stronie człowieka. Modele mogą powielać błędy, tworzyć błędne informacje czy proponować rozwiązania nieadekwatne do kontekstu. Ponadto bezkrytyczne bazowanie na treściach wytworzonych przez LLM może prowadzić do powierzchowności prowadzonej przez nas pracy (np. wnioski z analiz, rekomendacje), pomijania istotnych kwestii, które jedynie człowiek na podstawie swojej intuicji, wiedzy i doświadczenie byłby w stanie wychwycić.

Mamy jednak nadzieję, że przedstawione w rozdziale wskazówki i przykłady pomogą czytelnikom zbudować swój własny warsztat pracy, oparty na krytycznym myśleniu oraz na znajomości podstaw tworzenia instrukcji dla LLM. Jak wcześniej sygnalizowaliśmy, celem rozdziału nie jest przekazanie wiedzy pozwalającej na rozwiazywanie konkretnych problemów w codziennej pracy ewaluatora, a jedynie dostarczenie informacji umożliwiających samodzielne eksperymentowanie z różnymi typami promptów, do łączenia ich czy adaptowania do własnych projektów oraz do wypracowania różnych strategii postępowania, zgodnie z potrzebami wiedzy.

Przyszłość pracy z LLM niesie ze sobą zarówno szanse, jak i zagrożenia. Od nas, użytkowników, będzie zależało, czy dane nam narzędzia gen AI będą wspierały nasz rozwój i dorobek zawodowy, czy też będą stanowiły źródła uproszczeń i powierzchownych analiz. Liczymy na to, że niniejszy rozdział dostarczył wiedzy, która pozwoli na budowanie nowych kompetencji zawodowych ewaluatorów i minimalizację ryzyk związanych z wykorzystywaniem LLM.

Bibliografia

- Kahneman, D. (2013). *Pułapki Myślenia. O myśleniu szybki i wolnym*. Poznań. Media Rodzina.
- Haber, A., Olejniczak, K. (red.). (2014). *Analizy behawioralne w interwencjach publicznych*. W: (R)ewaluacja: Wiedza w działaniu. Warszawa: Polska Agencja Rozwoju Przedsiębiorczości, s. 17.
- Bsharat, S.M., Myrzakhan, A., Shen, Z. (2024). *Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4*. VILA Lab, Mohamed bin Zayed University of AI, s. 7.
- White, J., Fu, Q., Hays, S., et al. (2023). *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*. Department of Computer Science, Vanderbilt University, Nashville, TN, USA.
- Teo, S. (2023, December 29). *How I Won Singapore's GPT-4 Prompt Engineering Competition: A deep dive into the strategies I learned for harnessing the power of Large Language Models (LLMs)*. Towards Data Science.
- Lightman, H., Kosaraju, V., Burda, Y., et al. (2023). *Let's Verify Step by Step*.
- OpenAI. (2024). *Learning to Reason with LLMs* (dostęp: 31.03.2025).
- Phoenix, J., Taylor, M. (2023). *Prompt Engineering for Generative AI: Future-Proof Inputs for Reliable AI Outputs*. O'Reilly Media.
- Wilson, H.J., Daugherty, P.R. (2024). *Embracing Gen AI at Work: The skills you need to succeed in the era of large language models*. Harvard Business Review, September-October 2024.
- Polzer, D. (2024, June 26). *17 (Advanced) RAG Techniques to Turn Your LLM App Prototype into a Production-Ready Solution: A collection of RAG techniques to help you develop your RAG app into something robust that will last*. Towards Data Science.
- Anthropic. (2024, August 14). *Prompt caching with Claude*.
- *Fundusze Europejskie na rzecz Gospodarki (FENG)*

3. BHP pracy z gen AI – bezpieczeństwo danych i poufność

Andrzej Jędrzejowski

3.1. Wprowadzenie

Celem rozdziału jest przedstawienie podstawowej zasady bezpiecznej współpracy z generatywną sztuczną inteligencją przy realizacji badań, w ramach których przetwarzane są dane osobowe lub inne informacje poufne. W rozdziale zaprezentowano przykładowe rozwiązania umożliwiające wykorzystanie modeli językowych w badaniach ewaluacyjnych, które często mają właśnie taki charakter.

Pojęcie bezpieczeństwa i higieny pracy (BHP) zasadniczo oznacza „przepisy regulujące warunki pracy w zakładach pracy", minimalizujące ryzyka związane z wykonywaniem obowiązków zawodowych. W kontekście pracy z generatywną sztuczną inteligencją można mówić o specyficznym rodzaju BHP, koncentrującym się na ochronie danych oraz poufności informacji. Podobnie jak w tradycyjnym rozumieniu BHP, kluczowe jest tutaj stosowanie odpowiednich procedur i środków ostrożności, które zmniejszają ryzyko nieautoryzowanego dostępu do informacji oraz ich niewłaściwego wykorzystania. Stanowi to szerokie zagadnienie, które na chwilę obecną nie zostało szczegółowo uregulowane. Regulacje, takie jak np. *AI Act*, stanowią dokumenty kierunkowe określające ramy regulacyjne dla rozwoju, wdrażania i użytkowania sztucznej inteligencji na terenie UE.

Szczegółowe zasady, jakimi należy się kierować we współpracy z gen AI w pracy badawczej, wynikają obecnie przede wszystkim z przepisów dotyczących przetwarzania danych osobowych i innych danych poufnych (m.in. RODO). Praca badawcza (w tym badania ewaluacyjne) bazuje głównie na przetwarzaniu informacji (danych). Niezależnie od źródła tych informacji, mogą one zawierać dane osobowe lub poufne, co w praktyce ogranicza możliwości

ich przetwarzania – w szczególności przez podmioty trzecie (niezaangażowane bezpośrednio w proces realizacji badania). Przykładowo, zapisy zawarte w umowach PARP na realizację usług badawczych nakładają na wykonawców (ewaluatorów) m.in. następujące wymogi[1]:

- „Wykonawca dopuści do przetwarzania danych osobowych wyłącznie osoby posiadające stosowne imienne upoważnienia do przetwarzania danych osobowych.
- Wykonawca wyraża zgodę i zobowiązuje się umożliwić kontrolowanie przez Zamawiającego, osoby i podmioty upoważnione przez Zamawiającego oraz inne uprawnione podmioty, czy przetwarzanie powierzonych danych osobowych odbywa się zgodnie z niniejszą umową, przepisami powszechnie obowiązującymi (…)".

Powyższe stanowi znaczące ograniczenie realizacji badań, polegających na przetwarzaniu danych osobowych, we współpracy z narzędziami gen AI dostępnymi *on-line* – bazującymi na infrastrukturze podmiotów trzecich. Narzędzia takie jak np. ChatGPT, Claude, Gemini, Copilot umożliwiają generowanie treści na podstawie zadanego prompta. Ich wykorzystanie nie wymaga instalacji dodatkowego oprogramowania. W wielkim uproszczeniu – proces przetwarzania naszego zapytania dzieje się poza naszym komputerem („w chmurze"), co oznacza, że pomiędzy wprowadzeniem pytania (prompta) a otrzymaniem odpowiedzi (wygenerowanej treści) udostępniona informacja jest wysyłana do zewnętrznej infrastruktury (poza infrastrukturę użytkownika), gdzie następuje jej przetworzenie. W zależności od polityki dostawcy oraz narzędzia i typu konta, z którego korzystamy (w tym warunków, na które się zgadzamy), dostawcy omawianych narzędzi deklarują, że przesyłane przez nas informacje będą dalej wykorzystywane (np. do dalszego trenowania modeli) lub zostaną usunięte. Należy przy tym podkreślić, że niezależnie od tego, czy przesłana przez nas informacja (prompt, pliki źródłowe) będzie dalej przetwarzana, w momencie jej przesłania do narzędzia generatywnej AI trafia ona do zewnętrznej infrastruktury, co do której powinniśmy mieć ograniczone zaufanie. Informacja zostaje udostępniona i nie możemy zagwarantować jej pełnego bezpieczeństwa oraz metody jej wykorzystania. Należy mieć tu na uwadze zarówno przetwarzanie informacji zgodne z zaakceptowanymi warunkami, jak i ryzyko przetwarzania niezgodnego z zaakceptowanymi warunkami (w tym kradzież danych z infrastruktury dostawców poszczególnych narzędzi).

Uwzględniając powyższe, za naczelną zasadę współpracy z narzędziami generatywnej AI warto przyjąć następującą: **Nigdy nie przesyłaj do narzędzi generatywnej AI danych**

[1] Przytoczone zapisy pochodzą z jednej z wybranych umów zawartych przez PARP na realizację usługi badawczej. Przepisy wynikają z powszechnie obowiązujących regulacji prawnych.

osobowych, informacji poufnych ani żadnych treści, których ujawnienie mogłoby naruszyć obowiązujące regulacje prawne lub zobowiązania umowne. Zasada ta dotyczy w szczególności danych osobowych, ale także wszelkiego rodzaju danych wrażliwych, poufnych czy zastrzeżonych ze względów prawnych lub ekonomicznych, które nie powinny być udostępniane podmiotom trzecim bez odpowiednich zabezpieczeń i kontroli. Jeśli organizacja korzysta z zewnętrznej infrastruktury, np. rozwiązań chmurowych, przetwarzanie tych danych powinno odbywać się zgodnie z obowiązującymi regulacjami oraz politykami bezpieczeństwa danej organizacji, zapewniającymi ochronę przed nieautoryzowanym dostępem lub niewłaściwym wykorzystaniem danych.

W praktyce oznacza to, że korzystanie z ogólnodostępnych narzędzi generatywnej AI działających w modelu chmurowym (np. ChatGPT, Claude, Gemini) zazwyczaj nie spełnia wymogów ochrony danych i nie powinno być dopuszczalne w przypadku danych osobowych, informacji poufnych lub objętych innymi regulacjami prawnymi. Ze względu na te ograniczenia konieczne jest poszukiwanie alternatywnych sposobów wykorzystania modeli językowych, które pozwoliłyby na ograniczenie ryzyk związanych z bezpieczeństwem danych.

W dalszej części artykułu przedstawione zostaną trzy przykładowe rozwiązania umożliwiające ograniczenie ryzyk związanych z wykorzystaniem narzędzi generatywnej AI:

1. Udostępnienie przekształconych danych;
2. Wykorzystanie narzędzi generatywnej AI jako platformy *low-code*;
3. Wykorzystanie modeli LLM działających na lokalnej infrastrukturze.

Należy podkreślić, że niezależnie od przedstawionych poniżej rozwiązań możliwość wykorzystania omawianych narzędzi w pracy badawczej jest uzależniona od uwarunkowań prawnych i regulacyjnych, w tym zasad obowiązujących w poszczególnych firmach i organizacjach. Zakres przedstawionych poniżej propozycji koncentruje się na działaniach potencjalnie możliwych do wdrożenia w ramach badań, prowadzonych lub zamawianych przez sektor publiczny, przy założeniu, że w badaniach tych wykorzystywane są m.in. dane osobowe i inne dane poufne, które nie powinny być przetwarzane przez podmioty trzecie (niezaangażowane bezpośrednio w proces realizacji badania).

3.2. Udostępnienie przekształconych danych

Metody przekształcania danych

Zgodnie z proponowaną powyżej zasadą, jeżeli do analizy zadanej informacji chcemy wykorzystać dostępne *on-line* narzędzie generatywnej AI, należy usunąć z tej informacji dane, których nie powinniśmy udostępniać. W tym celu informacje można przekształcić m.in. w procesie maskowania, szyfrowania lub tokenizacji.

Maskowanie

Maskowanie jest metodą anonimizacji danych. W ramach maskowania dane wejściowe są zamieniane w ciąg znaków, który ukrywa ich oryginalną treść. Jest to proces, który jest nieodwracalny – po przekształceniu informacji nie ma możliwości przywrócenia jej oryginalnego brzmienia. Poniżej przedstawiony jest przykład przekształcenia danych poprzez maskowanie wraz z ich wykorzystaniem w ChatGPT[2].

Informacja oryginalna

> Jan Parpowski pracuje w PARP. Krzysztof Unijny pracuje w PARP. Jan Parpowski lubi jeździć na rowerze. Krzysztof Unijny lubi analizować dane.

Informacja przetworzona

> XXX XXX pracuje w XXX. XXX XXX pracuje w XXX. XXX XXX lubi jeździć na rowerze. XXX XXX lubi analizować dane.

Prompt przesłany do ChatGPT

> *Co jest napisane w powyższych zdaniach?*

Odpowiedź od ChatGPT

> W powyższych zdaniach zostały podane informacje o kilku osobach (zastąpionych przez „XXX XXX"), dotyczące ich miejsca pracy oraz zainteresowań. Konkretne treści zdają się sugerować:

[2] W artykule podano szereg przykładów wykorzystania ChatGPT, jednak należy podkreślić, że nie jest celem Autora promowanie tego narzędzia.

- Pierwsze zdanie: „XXX XXX pracuje w XXX.” – osoba (ukryta za „XXX XXX”) pracuje w bliżej nieokreślonym miejscu.
- Drugie zdanie: „XXX XXX pracuje w XXX.” – inna osoba (lub ta sama) również pracuje w nieokreślonym miejscu.
- Trzecie zdanie: „XXX XXX lubi jeździć na rowerze.” – osoba lubi jeździć na rowerze.
- Czwarte zdanie: „XXX XXX lubi analizować dane.” – osoba lubi analizować dane.

Wszystkie zdania dotyczą zainteresowań i miejsc pracy osób, których imiona są ukryte pod „XXX XXX”.

Udostępniona informacja pozostała w pełni anonimowa – ChatGPT przeanalizował ją bez dostępu do danych osobowych. Zamaskowanie wszystkich danych ciągiem znaków „XXX” uniemożliwiło ChatGPT zrozumienie, czy mówimy o jednej czy o wielu osobach (co można uznać zarówno za wadę, jak i zaletę tego rozwiązania).

Szyfrowanie i tokenizacja

Szyfrowanie i tokenizacja są metodami pseudoanonimizacji danych – przekształcenia danych w sposób odwracalny. W ramach szyfrowania i tokenizacji dane wejściowe zamieniane są w ciąg znaków, który przy wykorzystaniu dodatkowych informacji umożliwia odtworzenie pierwotnego brzmienia przekształconej informacji. W przypadku szyfrowania wymaga to dostępu do klucza deszyfrującego; w przypadku tokenizacji – dostępu do bazy danych (klucza kodowego), która zawiera przyporządkowanie tokenów do oryginalnych danych. Poniżej przedstawiony jest przykład przekształcenia danych poprzez tokenizację wraz z ich wykorzystaniem w ChatGPT.

Informacja oryginalna

Jan Parpowski pracuje w PARP. Krzysztof Unijny pracuje w PARP. Jan Parpowski lubi jeździć na rowerze. Krzysztof Unijny lubi analizować dane.

Klucz kodowy (nieudostępniany do ChatGPT)

Token;	Informacja
41dd4080-cd76;	Jan
3f4396d9-908f;	Parpowski
fa0dd52c-9750;	PARP
42b1b4d2-d9d5;	Krzysztof

45ff466b-9869; Unijny

Informacja przetworzona

41dd4080-cd76 3f4396d9-908f pracuje w fa0dd52c-9750. 42b1b4d2-d9d5 45ff466b-9869 pracuje w fa0dd52c-9750. 41dd4080-cd76 3f4396d9-908f lubi jeździć na rowerze. 42b1b4d2-d9d5 45ff466b-9869 lubi analizować dane.

Prompt przesłany do ChatGPT

Co jest napisane w powyższych zdaniach?

Odpowiedź od ChatGPT

Powyższe zdania zawierają informacje o dwóch osobach identyfikowanych przez kody i ich działania lub zainteresowania. Oto co jest napisane:

- Osoba o kodzie 41dd4080-cd76 3f4396d9-908f pracuje w firmie o kodzie fa0dd52c-9750.
- Osoba o kodzie 42b1b4d2-d9d5 45ff466b-9869 również pracuje w firmie o kodzie fa0dd52c-9750.
- Osoba o kodzie 41dd4080-cd76 3f4396d9-908f lubi jeździć na rowerze.
- Osoba o kodzie 42b1b4d2-d9d5 45ff466b-9869 lubi analizować dane.

Zdania opisują osoby poprzez ich miejsce pracy i zainteresowania, wykorzystując unikalne kody do identyfikacji zarówno osób, jak i firmy.

Udostępniona informacja jest anonimowa. ChatGPT poprawnie odczytał, że przekazana informacja dotyczy dwóch osób. Po otrzymaniu odpowiedzi, dysponując kluczem kodowym, istnieje możliwość ponownego odczytania zanonimizowanych danych. W przypadku braku klucza kodowego (lub klucza deszyfrującego), nie będzie takiej możliwości – będziemy wiedzieli, że dane dotyczą dwóch różnych osób, jednak nie będziemy w stanie poznać ich imion i nazwisk. Co należy podkreślić, w sytuacji braku klucza kodowego (lub klucza deszyfrującego) nie ma możliwości powtórzenia procesu przekształcenia danych. Załóżmy sytuację, w której przetworzonych zostało sto zdań. Każde wystąpienie nazwy „PARP" zostało zamienione na token „fa0dd52c-9750". Jeżeli zachowamy klucz kodowy, w sytuacji konieczności przekształcenia kolejnych sto zdań nazwa „PARP" zostanie zamieniona na ten sam token. W przeciwnym przypadku nazwa „PARP" zostanie zamieniona na inny token – co może prowadzić do błędów w analizie danych.

3.3. Identyfikacja danych do przekształcenia

Wykorzystanie NLP

Przekształcenie pojedynczego zdania lub krótkiego tekstu nie stanowi problemu dla ewaluatora – informacje można zanonimizować „ręcznie" (samodzielnie wybrać i przekształcić słowa, które wymagają anonimizacji). W przypadku potrzeby przekształcenia większego zbioru informacji (np. transkrypcja wywiadu lub baza danych) konieczne jest zautomatyzowanie tego procesu.

Rozwiązaniem jest wykorzystanie technik NLP – przetwarzania języka naturalnego. W tym celu w języku Python wykorzystać można *open source'ową* bibliotekę spacy wraz z modelem pl_core_news_lg. Wskazany model języka polskiego umożliwia m.in. tokenizację, rozpoznawanie części mowy, lematyzację[3], analizę składniową oraz rozpoznawanie jednostek nazwanych (NER). Bibliotekę oraz model można zainstalować lokalnie, a przekształcenie danych można przeprowadzić bez udostępniania danych na zewnątrz.

Poniżej przedstawiono kod, który anonimizuje zadany tekst (zmienna *value*). W zadanym tekście wyszukiwane są wszystkie rozpoznane jednostki nazwane, po czym zamieniane zostają ciągiem znaków „XXX"[4].

Narzędzie do anonimizacji – cz.1

```
import spacy
import pl_core_news_lg
nlp = spacy.load("pl_core_news_lg")
doc = nlp(value)
for ent in doc.ents:
    value = value.replace(ent.text, "XXX")
```

Wykorzystanie powyższego kodu umożliwia zamianę zdań: „Jan Parpowski pracuje w PARP. Krzysztof Unijny pracuje w PARP. Jan Parpowski lubi jeździć na rowerze. Krzysztof Unijny lubi analizować dane." na: „XXX pracuje w XXX. XXX pracuje w XXX. XXX lubi jeździć na rowerze.

[3] Sprowadzenie słowa do formy, w jakiej występuje w słowniku.

[4] Niniejszy artykuł nie będzie omawiał szczegółów działania modelu pl_core_news_lg.

XXX lubi analizować dane”. W przypadku potrzeby zanonimizowania wielu jednostek tekstu wskazany kod można zapętlić.

Model pl_core_news_lg rozpoznaje różne kategorie jednostek nazwanych, w tym np.: osoby (persName), organizacje (orgName) i lokalizacje (placeName). W przypadku potrzeby zanonimizowania tylko wybranych kategorii powyższy kod należy uzupełnić o odpowiedni warunek (np. if ent.label_ == "persName").

Przedstawiona powyżej metoda umożliwia maskowanie tekstu – anonimizowane dane zostały zamienione ciągiem znaków „XXX”. W celu przeprowadzenia tokenizacji każdą rozpoznaną nazwę można zamienić na indywidualny numer uuid, do którego wygenerowania można wykorzystać bibliotekę uuid lub shortuuid. Należy przy tym pamiętać, aby generowane numery uuid zachowywać w tabeli – kluczu kodowym (kolumny: uuid/token; anonimizowana informacja). Umożliwi to późniejsze odczytanie zanonimizowanych informacji. W sytuacji przesłania przekształconych informacji do narzędzi gen AI uzyskana odpowiedź będzie zanonimizowana, ale wykorzystanie klucza kodowego umożliwi jej późniejsze poprawne odczytanie (np. ilekroć w odpowiedzi pojawia się token „fa0dd52c-9750”, chodzi o „PARP”).

Wykorzystanie słownika języka polskiego

Przedstawiony powyżej skrypt można rozbudować o anonimizowanie wszystkich nazw własnych (tych, które nie są rozpoznawane przez model pl_core_news_lg). Nazwy własne, co do zasady, są niedopuszczalne w grach słownych (np. scrabble). Uwzględniając powyższe, anonimizowanie tekstu można sprowadzić do usunięcia z niego wszystkich słów, które są niedopuszczalne w grach słownych. Lista wszystkich słów, które są dopuszczalne, jest dostępna na stronie sjp.pl.

Przedstawiony poniżej kod jest kontynuacją kodu zaprezentowanego powyżej: *Narzędzie do anonimizacji – cz. 1*. W pierwszym kroku sprawdzane jest, czy dane słowo występuje w anonimizowanym tekście w formie słownikowej (*token.text.lower() == token.lemma_.lower()*) – jeżeli nie (jest np. odmienione), uznaje się, że słowo nie jest nazwą własną (nie należy go anonimizować). Jeżeli słowo występuje w tekście w formie słownikowej, sprawdzane jest, czy występuje w Słowniku (dic_sjp – słownik należy wcześniej zaimportować) – jeżeli nie, jest ono anonimizowane.

Narzędzie do anonimizacji - cz.2

```
for token in doc:
    if token.text.lower() == token.lemma_.lower():
        word = token.text.strip()
        if word.isprintable() and str(word) != " ":
            if word.lower() not in dic_sjp:
                value = value.replace(word, "XXX")
```

Przedstawiony powyżej kod stanowi przykład możliwego rozwiązania[5] – przed jego wykorzystaniem w praktyce należy zweryfikować jego dokładność na własnym zbiorze danych. W zależności od potrzeb można go samodzielnie modyfikować i optymalizować. Przykładowy test może polegać na wykorzystaniu przedstawionego rozwiązania do zanonimizowania tysiąca zdań (jednostek tekstu) – otrzymany wynik należy samodzielnie przeanalizować pod kątem jego poprawności (czy tekst został poprawnie zanonimizowany). W zależności od rezultatu przeprowadzonego testu należy podjąć samodzielną decyzję o wykorzystaniu opisanego rozwiązania lub jego dalszej modyfikacji.

Automatyzacja procesu anonimizacji danych może umożliwić przetworzenie znaczących ilości danych – np. transkrypcji wywiadów lub dużych baz danych. Należy jednak podkreślić, że poprawność przeprowadzonej anonimizacji może nigdy nie być pełna – algorytm wykona założone przez nas działania, jednak w zależności od przyjętych założeń może się okazać, że: a) część danych, które należałoby zanonimizować, nie została zanonimizowana; b) część danych, które nie powinny zostać zanonimizowane, została zanonimizowana. Warto przy tym zauważyć, że to samo dotyczy procesu anonimizacji dużych zbiorów danych, przeprowadzanej „ręcznie” przez człowieka – należy liczyć się z pewnym prawdopodobieństwem popełnienia błędu. Decyzja o udostępnieniu zanonimizowanych danych do dostępnych *on-line* narzędzi generatywnej AI powinna zostać podjęta ze świadomością braku możliwości zapewnienia pełnej anonimizacji danych (trzeba zakładać pewne ryzyko błędu).

[5] Przedstawione rozwiązanie nie jest gotowym rozwiązaniem – stanowi tylko przykład. Autor nie ponosi odpowiedzialności za skutki jego wykorzystania.

3.4. Wykorzystanie narzędzi gen AI jako platformy *low-code*

Dostępne *on-line* narzędzia gen AI mogą być wykorzystywane jako wsparcie w procesie budowy modeli (w tym np. kodu programistycznego) do analizy danych na lokalnej infrastrukturze (bez konieczności wysyłania danych do zewnętrznej infrastruktury). Narzędzia te mogę pełnić funkcję platform *low-code*[6] poprzez m.in.: budowę skryptów do analizy danych, wyjaśnianie działania skryptów (kodu programistycznego), obsługę błędów – w tym wyjaśnianie przyczyn błędów i propozycje ich rozwiązań[7].

Budowa skryptów do analizy danych

Analiza danych opiera się na wykorzystywaniu języków programistycznych, takich jak Python, SQL, R, M (język wykorzystywany w Power Query). Każde działanie, takie jak np. grupowanie, sumowanie, filtrowanie lub sortowanie, jest zestawem poleceń, które zostają określone w wybranym języku programistycznym. Niektóre narzędzia generatywnej AI można wykorzystać do generowania skryptów – kodu programistycznego, który analizuje określone dane. W przypadku udostępnienia do wybranego narzędzia danych można poprosić zarówno o ich analizę, jak i pokazanie skryptu – kodu, w oparciu o który przeprowadzona została zadana analiza. Co jednak istotne, udostępnienie danych nie jest w tym przypadku niezbędne – narzędzia gen AI mogą opracować skrypt do analizy danych jedynie na podstawie opisu zadanej analizy (wystarczą ogólne informacje, takie jak m.in. język, w którym ma powstać skrypt; nazwa pliku źródłowego; nazwy kolumn oraz opis działań, które chcemy na nich przeprowadzić). Wygenerowany skrypt można następnie przekopiować i uruchomić lokalnie – w tej sytuacji wszystkie dane podlegające analizie pozostają nieudostępnione. W tym rozwiązaniu narzędzia generatywnej AI pełnią funkcję platformy *low-code*, umożliwiającej osobom z mniejszym doświadczeniem programistycznym tworzenie specjalistycznego kodu. Poniżej przedstawiono przykładowy prompt oraz odpowiedź ChatGPT dotyczącą analizy tabeli „Lista projektów realizowanych z Funduszy Europejskich w Polsce w latach 2014–2020”.

[6] Narzędzie generatywnej AI nie są platformami *low-code*, jednak mogą pełnić ich funkcje.

[7] Przykłady takich zastosowań zostały omówione w rozdziałach 8. i 9. niniejszej publikacji.

Prompt przesłany do ChatGPT

Potrzebuję przeanalizować tabelę z pliku Lista_projektow_FE_2014_2020_020624.csv. Tabela zawiera 24 kolumny. Separatorem jest ";"

Nie przesyłam pliku.

Potrzebuję stworzyć tabelę, w której wartości z kolumny nr 12 (Wartość unijnego dofinansowania (w zł, dla projektów EWT w euro)/ Amount of EU co-financing (PLN, for ETC projects EUR) zostaną zsumowane i pogrupowane według kategorii z kolumny nr 6 (Program/Programme) oraz kategorii z kolumny nr 8 (Działanie/Measure). W tabeli chcę uwzględnić tylko wiersze, które w kolumnie nr 15 (Miejsce realizacji projektu/Project location) zawierają tekst "MAZOWIECKIE". Tabelę uszereguj po kolumnach Program/ Programme oraz Działanie/Measure.

Przygotuj mi skrypt w Python, który umożliwi stworzenie takiej tabeli i jej zapisanie do pliku o nazwie "tabela.csv". W skrypcie uwzględnij numery kolumn (nie uwzględniaj nazw kolumn).

Odpowiedź od ChatGPT[8]

```
import pandas as pd
# Wczytanie pliku CSV z separatorem ";"
df = pd.read_csv('Lista_projektow_FE_2014_2020_010924.csv', delimiter=';')
# Filtracja wierszy, które zawierają "MAZOWIECKIE" w kolumnie nr 15
filtered_df = df[df.iloc[:, 14].str.contains('MAZOWIECKIE', na=False)]
# Grupowanie danych według kolumn nr 6 (Program) i nr 8 (Działanie) oraz sumowanie
   wartości z kolumny nr 12
grouped_df = filtered_df.groupby([df.columns[5], df.columns[7]]).agg({df.columns[11]:
   'sum'})
# Resetowanie indeksu dla bardziej przejrzystego zapisu
grouped_df = grouped_df.reset_index()
# Sortowanie tabeli według kolumn Program (nr 6) i Działanie (nr 8)
grouped_df = grouped_df.sort_values(by=[df.columns[5], df.columns[7]])
# Zapisanie wynikowej tabeli do pliku CSV
grouped_df.to_csv('tabela.csv', index=False, sep=';')
print("Tabela została zapisana do pliku 'tabela.csv'.")
```

[8] Z oryginalnej odpowiedzi wygenerowanej przez ChatGPT usunięto fragment niestanowiący kodu Python.

W zależności od poziomu skomplikowania planowanych analiz warto rozważyć udostępnienie dodatkowych informacji, takich jak np.: typ danych znajdujących się w poszczególnych kolumnach (np. dane tekstowe lub dane liczbowe), kodowanie pliku źródłowego (np. kodowanie utf-8 umożliwia poprawne odczytanie polskich liter), separator rozdzielający poszczególne kolumny. Warto rozważyć również udostępnienie pliku zawierającego strukturę danych lub dane przykładowe (testowe, fikcyjne).

Wykorzystanie narzędzi generatywnej AI może stanowić znaczące ułatwienie w budowaniu skryptów do analizy danych. Co jednak należy podkreślić, wykorzystanie tych narzędzi wymaga odpowiednich kompetencji od badacza, który musi być w stanie określić zakres i przebieg planowanej analizy oraz zweryfikować jej efekt. Zgodnie z informacją umieszczoną w ChatGPT (pod polem, w który wpisywany jest prompt): „ChatGPT może popełniać błędy. Sprawdź ważne informacje". Badacz powinien być w stanie zrozumieć wygenerowany kod, aby móc go zmodyfikować lub poprawić.

Wyjaśnianie działania skryptów

Oczywiście o wyjaśnienie i opisanie działania skryptu można również poprosić narzędzia generatywnej AI. Przesłany prompt może dotyczyć zarówno kodu stworzonego przez generatywną AI, jak również stworzonego przez człowieka. Poniżej przedstawiono przykładowy prompt oraz odpowiedź ChatGPT dotyczącą wyjaśnienia działania kodu napisanego w Python.

Prompt przesłany do ChatGPT

Co robi poniższy skrypt:

```
import spacy
import pl_core_news_lg
nlp = spacy.load("pl_core_news_lg")
doc = nlp(value)
for ent in doc.ents:
    value = value.replace(ent.text, "XXX")
```

Odpowiedź od ChatGPT

Ten skrypt używa biblioteki spaCy oraz modelu językowego pl_core_news_lg, aby zanonimizować zidentyfikowane jednostki nazwane (ang. *named entities*) w tekście. Oto jak działa ten kod krok po kroku:

1. Importy i wczytywanie modelu:

```
import spacy

import pl_core_news_lg

nlp = spacy.load("pl_core_news_lg")
```

Skrypt importuje bibliotekę spacy oraz model językowy pl_core_news_lg, który jest przeznaczony do przetwarzania tekstu w języku polskim. Model ten zawiera informacje potrzebne do identyfikacji jednostek nazwanych, takich jak osoby, organizacje, lokalizacje, daty itp.

2. Analiza tekstu:

```
doc = nlp(value)
```

Tekst przechowywany w zmiennej value jest przetwarzany przez model nlp, co skutkuje utworzeniem obiektu doc. Ten obiekt zawiera analizę tekstu, w tym zidentyfikowane jednostki nazwane.

Zastępowanie jednostek nazwanych:

```
for ent in doc.ents:

    value = value.replace(ent.text, "XXX")
```

Skrypt przechodzi przez wszystkie zidentyfikowane jednostki nazwane (ent) w tekście i zastępuje ich tekst (czyli ent.text) ciągiem znaków „XXX”. Dzięki temu wszystkie jednostki nazwane w tekście zostają zanonimizowane.

Podsumowanie: Skrypt ten służy do anonimizacji tekstu poprzez zamianę wszystkich zidentyfikowanych jednostek nazwanych na „XXX”. Dzięki temu informacje takie jak imiona, nazwiska, nazwy miejsc czy organizacji zostają ukryte, co może być użyteczne w kontekście ochrony danych osobowych lub przygotowywania tekstów do dalszego przetwarzania bez ujawniania wrażliwych informacji.

Wykorzystanie gen AI do wyjaśniania działania skryptów umożliwia zrozumienie narzędzi badawczych – dostępnych publicznie, stworzonych przez innych uczestników procesu

badawczego lub przez inne modele gen AI na podstawie zadanych promptów. Zrozumienie kodu programistycznego, tworzącego dane narzędzie badawcze, umożliwia jego dalszy rozwój i dostosowanie do innych zadań i potrzeb.

Obsługa błędów

Narzędzia generatywnej AI mogą zostać również wykorzystane do obsługi błędów pojawiających się w trakcie pracy skryptu. Po przesłaniu prompta zawierającego rodzaj i tekst błędu otrzymamy wyjaśnienie błędu oraz opis jego prawdopodobnych przyczyn. Po dołączeniu kodu, który ten błąd spowodował, otrzymamy dodatkowo poprawioną wersję kodu. Umożliwia to samodzielną edycję kodu programistycznego osobom z mniejszym doświadczeniem programistycznym. Poniżej przedstawiono przykładowy prompt oraz odpowiedź ChatGPT dotyczącą obsługi błędu w kodzie napisanym w Python.

Prompt przesłany do ChatGPT

Poniższy kod zwraca błąd "AttributeError: 'spacy.tokens.doc.Doc' object has no attribute 'drzewce'". Wyjaśnij go i zaproponuj rozwiązanie

```
import spacy
import pl_core_news_lg
nlp = spacy.load("pl_core_news_lg")
doc = nlp(value)
for ent in doc.drzewce:
    value = value.replace(ent.text, "XXX")
```

Odpowiedź od ChatGPT

Błąd AttributeError: 'spacy.tokens.doc.Doc' object has no attribute 'drzewce' wskazuje, że próbujesz uzyskać dostęp do atrybutu o nazwie drzewce, który nie istnieje w obiekcie typu Doc (spacy.tokens.doc.Doc).

Wyjaśnienie błędu

W Twoim kodzie:

import spacy

import pl_core_news_lg

nlp = spacy.load("pl_core_news_lg")

doc = nlp(value)

for ent in doc.drzewce:

value = value.replace(ent.text, "XXX")

Problem tkwi w linii for ent in doc.drzewce:. Atrybut drzewce nie jest prawidłowym atrybutem obiektu Doc. Możliwe, że przez przypadek użyłeś nieprawidłowej nazwy atrybutu.

Prawidłowy atrybut

Dla obiektu Doc w spaCy, aby uzyskać dostęp do jednostek nazwanych, używa się atrybutu ents. Oto poprawiony kod:

import spacy

import pl_core_news_lg

nlp = spacy.load("pl_core_news_lg")

doc = nlp(value)

for ent in doc.ents: # Poprawiona linia - użycie prawidłowego atrybutu 'ents''

value = value.replace(ent.text, "XXX "")

Poprawki:

Zmieniono doc.drzewce na doc.ents, ponieważ ents to właściwy atrybut, który zawiera listę wykrytych jednostek nazwanych w analizowanym tekście.

Podsumowanie

Błąd wystąpił, ponieważ użyto nieprawidłowego atrybutu drzewce zamiast ents. Atrybut ents przechowuje informacje o jednostkach nazwanych, takich jak osoby, miejsca, daty itp., co jest kluczowe dla funkcji anonimizacji w kodzie. Po wprowadzeniu tej poprawki skrypt powinien działać poprawnie i zamieniać wykryte jednostki nazwane na "XXX".

3.5. Wykorzystanie modeli LLM działających na lokalnej infrastrukturze

Korzystanie z modeli LLM nie musi ograniczać się do wykorzystywania narzędzi dostępnych w chmurze (wymagających udostępniania danych na zewnątrz). Rozwiązania tego typu są stosowane m.in. z powodu braku konieczności ponoszenia kosztów rozwoju i utrzymania infrastruktury – sprawna praca modeli LLM wymaga znacznej mocy obliczeniowej, która nie jest dostępna dla większości komputerów. Wykorzystanie modeli LLM (opensource'owych) lokalnie (na własnej infrastrukturze) jest jednak możliwe. Przykładowym rozwiązaniem, które to umożliwia, jest program Ollama[9].

Ollama jest platformą opensource'ową, która umożliwia pobieranie i wykorzystywanie szeregu modeli LLM lokalnie – bez dostępu do Internetu. Oprogramowanie jest darmowe – plik instalacyjny jest dostępny na stronie projektu. Ollama jest dostępna na systemy operacyjne macOS, Linux oraz na Windows. Omawiany program umożliwia lokalną obsługę ponad stu modeli, w tym m.in.: Llama, Mistral, Gemma (pełna lista dostępnych modeli). Modele instalowane są niezależnie od instalacji Ollama – w zależności od potrzeb można zainstalować lub odinstalować dany model.

Zapytania do wybranego modelu są wysyłane przez terminal (wbudowany w system operacyjny program umożliwiający komunikację z modelem LLM za pomocą poleceń tekstowych) lub poprzez skrypt napisany w języku Python. Wykorzystanie Ollama poprzez Python jest możliwe dzięki bibliotece Ollama. Wykorzystanie Pythona daje w tym przypadku możliwość tworzenia zapętlonych poleceń analizujących zadany zbiór danych.

Przykładowy skrypt Python wykorzystujący bibliotekę Ollama oraz model Llama3.1

```
import ollama
response = ollama.chat(model='llama3.1', messages=[
  {
    'role': 'user',
```

[9] Artykuł nie ma na celu promowania programu Ollama ani wybranych modeli LLM – poniższy tekst jest opisem przykładowych dostępnych rozwiązań.

```
        'content': 'Why is the sky blue?',
    },
])
print(response['message']['content'])
```

Odpowiedź od modelu Llama 3.1

The sky appears blue to us during the day due to a phenomenon called Rayleigh scattering. Here's a simplified explanation:

- Sunlight and Atmosphere: Sunlight, or solar radiation, is made up of a spectrum of colors that, combined, look white to our eyes. When sunlight enters Earth's atmosphere, it encounters molecules of nitrogen and oxygen, among others.
- Scattering of Light: These molecules scatter the light in all directions, but they scatter shorter (blue) wavelengths more than longer (red) wavelengths. This is known as Rayleigh scattering, named after the British physicist Lord Rayleigh who first described it in the 19th century.
- Color Perception: The scattered blue light is distributed throughout the atmosphere, reaching our eyes from all directions. This is why the sky appears blue, especially in the daytime when the sun is overhead.
- Variations: The color of the sky can change under different conditions. During sunrise and sunset, the sky can appear red or orange because the sunlight has to travel through more of the Earth's atmosphere, scattering the shorter wavelengths and allowing the longer wavelengths (reds and oranges) to reach our eyes.

In essence, the blue appearance of the sky is a result of the scattering of sunlight by the atmosphere, with blue light being scattered more than the other colors due to its shorter wavelength.

Ollama pozwala na wykorzystanie modeli LLM bez konieczności udostępniania analizowanych danych na zewnątrz. Rozwiązanie to umożliwia wykorzystanie narzędzi gen AI do analizy danych bez konieczności ich wcześniejszego zanonimizowania oraz bez ryzyk wiążących się z udostępnieniem danych do zewnętrznej infrastruktury (ryzyka zostały opisane w pierwszej części rozdziału). Należy jednak podkreślić, że korzystanie z modeli LLM lokalnie wymaga odpowiedniej infrastruktury – każdy model potrzebuje innych zasobów do poprawnego działania. Dla przykładu, uwzględniając wyłącznie rozmiar modeli: model Moondream 2 ma rozmiar 829 MB; model Llama 3.1 w wersji 405B (składający się z 405 miliardów parametrów) – 231 GB.

3.6. Podsumowanie

„**Nigdy nie przesyłaj do narzędzi generatywnej AI danych osobowych, informacji poufnych ani żadnych treści, których ujawnienie mogłoby naruszyć obowiązujące regulacje prawne lub zobowiązania umowne**”. Jest to podstawowa zasada, którą należy się kierować, aby bezpiecznie korzystać z narzędzi gen AI. Zasada ta jest szczególnie istotna w kontekście badań i ewaluacji, w ramach których przetwarzane mogą być dane osobowe lub inne dane poufne, które nie powinny być przetwarzane przez podmioty trzecie (niezaangażowane bezpośrednio w proces realizacji badania). To ogranicza możliwość wykorzystania narzędzi gen AI dostępnych *on-line* („w chmurze”), opierających się na przetwarzaniu danych na infrastrukturze podmiotu, który dostarcza przedmiotowe narzędzie. Ograniczenie to nie oznacza jednak braku możliwości wykorzystania gen AI w pracy badawczej.

Pierwszym możliwym rozwiązaniem jest poprzedzenie przetwarzania danych na infrastrukturze zewnętrznej przekształceniem danych na infrastrukturze lokalnej. Udostępniane dane nie powinny zawierać danych osobowych, pozostałych danych poufnych oraz innych, których z określonych powodów nie należy udostępniać. Dane te można usunąć w procesie anonimizacji (np. maskowania) lub pseudonimizacji (np. szyfrowania lub tokenizacji) przeprowadzonym na własnej infrastrukturze. Rozwiązanie to umożliwia udostępnienie do narzędzi gen AI danych, które wówczas już nie zawierają danych chronionych – dane osobowe i inne poufne nie zostaną udostępnione i nie będą przetwarzana na infrastrukturze zewnętrznej.

Drugim rozwiązaniem jest wykorzystanie narzędzi generatywnej AI jako wsparcia w budowaniu i rozwijaniu skryptów do analizy danych. Dostępne *on-line* narzędzia generatywnej AI mogą wygenerować lub zmodyfikować kod programistyczny służący analizie danych – bez konieczności dostępu do danych. Na podstawie opisu naszych potrzeb (np. opisu tabel i zawartych w nich danych) możemy – bez konieczności udostępniania danych – wygenerować model (algorytm, skrypt), który wykorzystamy do ich analizy. Wówczas przetwarzanie danych będzie prowadzone wyłącznie na infrastrukturze lokalnej.

Trzecim rozwiązaniem może być wykorzystanie modeli LLM zainstalowanych na lokalnej infrastrukturze. Rozwiązanie to wymaga nakładów na rozwój i utrzymanie infrastruktury komputerowej, jednak umożliwia bezpieczne wykorzystanie narzędzi gen AI do analizy danych, bez konieczności udostępniania danych poza lokalną infrastrukturę.

Konieczność zapewnienia bezpieczeństwa danych ogranicza możliwość wykorzystywania narzędzi gen AI w badaniach i ewaluacji, w których przetwarzane są dane osobowe lub inne informacje poufne. Nie oznacza to jednak całkowitego wykluczenia ich zastosowania. Należy dążyć do wypracowania rozwiązań, które umożliwią współpracę z narzędziami gen AI w sposób zgodny z regulacjami dotyczącymi ochrony danych, szczególnie poprzez unikanie przetwarzania na zewnętrznych serwerach treści (w szczególności danych osobowych i innych danych poufnych), które nie powinny być udostępnione poza organizację.

4. Integracja modeli gen AI z bazami dokumentów

Paweł Kędzia
Bartosz Ledzion

4.1. Wstęp

W dobie cyfryzacji i rosnącej ilości informacji analiza polityk publicznych staje przed wyzwaniem efektywnego przetwarzania ogromnych zbiorów dokumentów. Tradycyjne, czasochłonne metody analityczne nie są już wystarczające w obliczu rosnących potrzeb szybkiego podejmowania decyzji dotyczących złożonych zagadnień strategicznych.

W niniejszym rozdziale prezentujemy podejście do tego problemu, które skupia się na wykorzystaniu modeli językowych w połączeniu z bazami wiedzy i dokumentów.

Jak wspomniano we wcześniejszych rozdziałach[1], przegląd źródeł tekstowych (np. opracowań analitycznych, raportów z badań) to zadania trwające nierzadko wiele roboczodni. Obecnie procesy te można usprawnić dzięki zastosowaniu RAG (*Retrieval-Augmented Generation*), czyli rozwiązania, który łączy dwie kluczowe funkcjonalności: wyszukiwanie (*retrieval*) informacji w bazie dokumentów oraz generowanie (*generation*) odpowiedzi przez model gen AI.

W praktyce działa to tak, że system RAG najpierw przeszukuje bazę danych w poszukiwaniu istotnych informacji, a następnie model generatywny tworzy odpowiedź wyłącznie na podstawie znalezionych treści.

[1] Por. rozdziały „Analizy jakościowe wspierane gen AI” oraz „Przeglądy źródeł wspierane gen AI”.

Dzięki zastosowaniu RAG użytkownik zyskuje trzy kluczowe korzyści (Zucchini, 2024):

- większą dokładność działania modelu językowego poprzez ograniczenie zjawiska konfabulacji (system opiera odpowiedzi na rzeczywistych danych ze sprawdzonych źródeł (Lee & Kim, 2024)),
- transparentność, dzięki możliwości weryfikacji źródeł informacji,
- personalizację, ponieważ system wykorzystuje specyficzne dane organizacji, dostosowując odpowiedzi do konkretnego kontekstu.

Zastosowanie gen AI, w tym technik RAG, wspiera analizę dokumentów i znacząco zmienia sposób pracy badaczy i specjalistów zajmujących się przetwarzaniem informacji. Badania pokazują, że może ono trzykrotnie zwiększyć produktywność w zadaniach tekstowych, co obejmuje m.in. analizę i przetwarzanie dużych zbiorów dokumentów (Hartley et al., 2024). Dzięki tym narzędziom zadania związane z identyfikacją kluczowych treści i podejmowaniem decyzji opartych na danych, które wcześniej zajmowały tygodnie, można teraz wykonać w ciągu kilku godzin – dużo szybciej i łatwiej.

W niniejszym rozdziale przedstawimy sposób działania systemów RAG, ich praktyczne zastosowanie w badaniach ewaluacyjnych oraz omówimy różne możliwości rozpoczęcia pracy z tą technologią – od gotowych rozwiązań dla początkujących po zaawansowane systemy wymagające umiejętności programistycznych.

Rozdział ten pozwoli czytelnikom nie tylko lepiej zrozumieć potencjał wykorzystania RAG, ale także wskaże kierunki dalszego rozwoju i doskonalenia tej metody pracy z dokumentami w procesie badawczym.

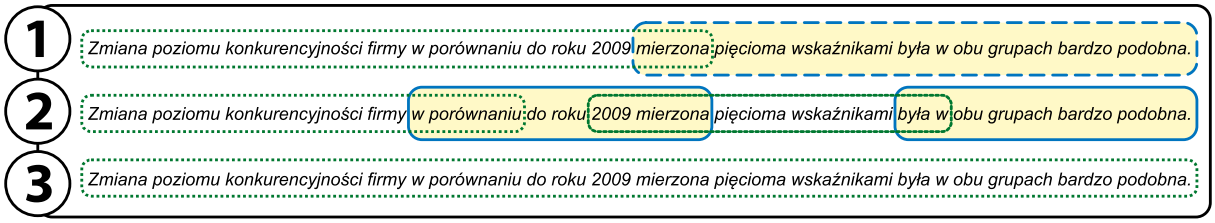
4.2. Jak działa RAG?

RAG – inteligentna biblioteka wiedzy

Na początku, dla lepszego zrozumienia działania systemu RAG, sięgniemy po analogię do biblioteki i bibliotekarza. Wyobraźmy sobie nowoczesną bibliotekę, gdzie zamiast tradycyjnego bibliotekarza mamy zaawansowany system sztucznej inteligencji. W tradycyjnej bibliotece, gdy czytelnik poszukuje informacji, bibliotekarz wykorzystuje system katalogowania, w tym indeksy i zakładki, aby szybko odnaleźć potrzebne treści. W świecie

cyfrowym te funkcjonalności realizuje właśnie RAG. Załóżmy, że system RAG przetwarza stustronicowy raport ewaluacyjny. Na początku system dzieli tekst na mniejsze fragmenty – podobnie jak książka podzielona jest na akapity, strony, rozdziały i sekcje. W przypadku raportu możne podzielić go na fragmenty, np. o stałej długości po 500–1000 znaków (tzw. chunki). Na rycinie 4.1. przedstawiliśmy trzy sposoby podziału tego samego tekstu na chunki o różnej długości[2].

Rycina 4.1. Podział tekstu na chunki o różnych długościach



Źródło: Opracowanie własne.

W pierwszym przypadku długość chunka ustawiona została na dziesięć słów oraz jedno z nich wykorzystane zostało jako część wspólna sąsiadujących chunków (ostatnie słowo z pierwszego chunku jest pierwszym słowem w drugim). Przy takim podziale otrzymujemy dwa chunki (wyróżnione zostało słowo, które przenika między sąsiadującymi chunkami):

- Zmiana poziomu konkurencyjności firmy w porównaniu z rokiem 2009 **mierzona,**
- **mierzona** pięcioma wskaźnikami była w obu grupach bardzo podobna.

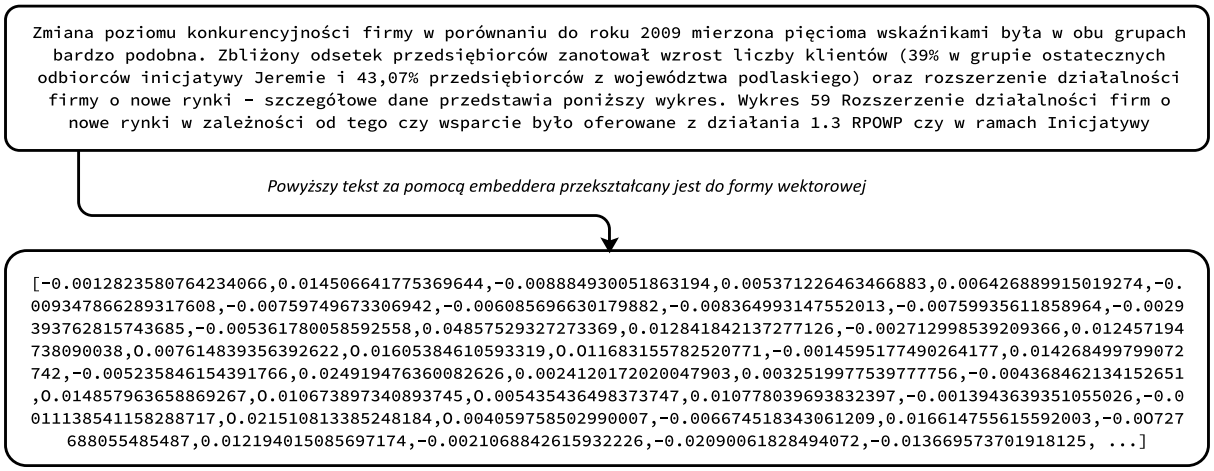
W drugim przypadku długość jednego chunka została ustawiona na sześć słów, dwa z nich wykorzystane są jako część wspólna chunków sąsiadujących. Przy takiej konfiguracji otrzymamy nie dwa, a cztery chunki. Z kolei w trzecim przypadku długość chunka ustawiona została na 30 słów, a pięć słów jako część wspólna sąsiadujących chunków. Przy takiej konfiguracji tekst jest za krótki, dlatego w całości trafia do jednego chunka. W praktyce długość pojedynczego chunka dostosowywana jest do rozwiązywanego problemu.

Następnie system RAG dla każdego chunka tekstu tworzy tzw. embeddingi, czyli matematyczne reprezentacje znaczenia tekstu w przestrzeni wielowymiarowej. Aktualnie najbardziej popularne są bi-encodery (Reimers, 2019), czyli metody, które modelują zależności semantyczne między dwoma tekstami. Tworzenie embeddingu można

[2] Przytoczony przykład jedynie ilustruje proces chunkowania, w rzeczywistości długości chunków są większe, np. 150, 300 słów. Wielkość ta ustalana jest w zależności od typu danych.

porównać do tworzenia szczegółowego indeksu tematycznego, w którym każde hasło ma swoje umiejscowienie w systemie katalogowym biblioteki. Dla przykładowego raportu ewaluacyjnego oznacza to, że każdy fragment zawartego w nim tekstu zostaje przekształcony w wektor liczbowy, który przechowuje jego znaczenie semantyczne. Na rycinie 4.2. przedstawiliśmy przykładowy tekst oraz jego reprezentację w postaci embeddingu.
W przytaczanym przykładzie wykorzystaliśmy ogólnodostępny embedder silver-retriever-base (Rybak, 2024).

Rycina 4.2. Przykład zmiany tekstu na postać wektorową



Źródło: Opracowanie własne.

Systemy tego typu tworzone są w taki sposób, aby były w stanie dopasować do siebie nie tylko teksty na poziomie słów, ale przede wszystkim na poziomie znaczenia tych tekstów, dlatego mówi się, że posługują się one **semantycznymi wektorami**. Przy czym o ile tekst można stosunkowo łatwo przekształcić w tzw. embeddingi (zmienić do postaci wektorowej), o tyle informacje zawarte na wykresach i obrazach nie zawsze da się prawidłowo przechwycić i odzwierciedlić w embeddingach. Przykładowo, system RAG odczyta ogólny kształt wykresu, ale może nie uchwycić wartości liczbowych czy zależności przedstawionych na nim, co może prowadzić do niepełnej lub niedokładnej ekstrakcji informacji.

W przygotowanej bazie dokumentów (wirtualnej bibliotece) system organizuje ich przetworzone fragmenty według semantycznego podobieństwa. Ta czynność przypomina układanie książek na półkach, nie tylko wg kolejności alfabetycznej, ale także z uwzględnieniem ich powiązań tematycznych, np.: grupowania raportów

ewaluacyjnych odnoszących się tylko do wyzwań na rynku pracy, na jakie napotykają osoby z niepełnosprawnościami. W systemie RAG takim odzwierciedleniem półek jak w bibliotece jest **semantyczna baza danych** (np. Milvus).

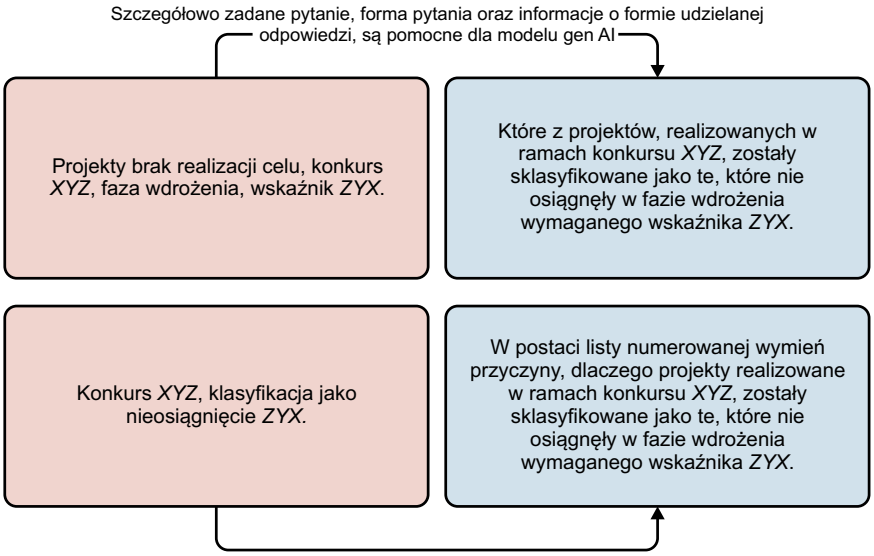
Semantyczna baza danych przechowuje embeddingi i umożliwia ich efektywne przeszukiwanie. Można by to porównać do pracy bibliotekarza, który przypina do stron w poszczególnych książkach karteczki z dodatkowymi informacjami, takimi jak słowa kluczowe, kategorie tematyczne czy krótkie streszczenia danej strony. Te adnotacje miałyby później pomagać w szybkim odnalezieniu potrzebnych fragmentów, bez konieczności czytania całej książki. Podobnie wygląda to w systemach RAG, gdzie do fragmentów tekstu możemy przypisać **metadane (np. informacje o tematyce danego fragmentu), które później ułatwić mają wyszukiwanie odpowiednich treści.**

Do przechowywania tych wszystkich informacji z dokumentów stosuje się bazy relacyjne (np. PostgreSQL) albo grafowe (Neo4j). Z kolei proces zapisywania dokumentów w bazach nazywa się **indeksowaniem.**

Wyszukiwanie informacji w systemie – w roli cyfrowego bibliotekarza

Jeżeli użytkownik zadał pytanie do systemu RAG: *Jakie były efekty programu Cyfrowa Szkoła w latach 2021–2024?,* to system – podobnie jak bibliotekarz – musi przede wszystkim odpowiednio zinterpretować intencję pytającego, identyfikując kluczowe elementy pytania, w tym przypadku: nazwę programu, zakres czasowy oraz fakt, że interesują go skutki tego przedsięwzięcia. Warto w tym miejscu przypomnieć o konieczności precyzyjnego zadawania pytań. Jeżeli użytkownik zada do systemu mało konkretne pytanie, które pozostawi dużą swobodę jego interpretacji, wtedy odpowiedź modelu generatywnego może nie być zgodna z oczekiwaniami. Model potrzebuje bardzo konkretnych instrukcji – na rycinie 4.3. przedstawione zostały przykłady mało konkretnych, hasłowych zapytań oraz ich „konkretniejsze” odpowiedniki.

Rycina 4.3. Przykład mało konkretnych oraz bardziej precyzyjnych zapytań



Źródło: Opracowanie własne.

Po rozszyfrowaniu zapytania w kolejnym kroku system RAG przeszukuje swoją bazę wektorową, wykorzystując algorytmy podobieństwa semantycznego. Przeważnie obliczany jest kosinus kąta[3] między reprezentacją wektorową zapytania użytkownika a reprezentacjami zapisanymi w bazie wektorowej. Na tej podstawie dobierane są takie fragmenty tekstów, których kąt w stosunku do reprezentacji pytania jest jak najmniejszy. Ten proces przypominać mógłby pracę bibliotekarza, który zamiast ponownie czytać po kolei każdą książkę, potrafiłby błyskawicznie wskazać najbardziej odpowiednie fragmenty książek, którymi dysponuje, na podstawie ich tematycznego powiązania z pytaniem czytelnika odwiedzającego bibliotekę. Co więcej, system RAG wykracza poza proste dopasowanie słów kluczowych, a potrafi identyfikować powiązane koncepcje i zbliżone struktury znaczeniowe (synonimy), czyli podobnie jak doświadczony bibliotekarz „rozumie" szerszy kontekst zapytania.

Na tym etapie pracy z systemami RAG warto wprowadzić tzw. modele cross-encoders (CE) – (Reimers, N. (2019)). Są to specjalne modele sztucznej inteligencji, które działają jak eksperci oceniający powiązania między dwoma fragmentami tekstu. Wyobraźmy sobie, że mamy do czynienia z pytaniem i kilkoma potencjalnymi odpowiedziami. Cross-encoder działa w tym przypadku jak wprawny analityk jakościowy, który czyta zarówno pytanie (kod), jak

[3] *Cosine similarity*.

i każdą z odpowiedzi (treści), a następnie przyznaje punkty (koduje zgodnie z ustalonym kluczem, nakłada wagi) za to, jak dobrze odpowiedź pasuje do pytania. W systemach wyszukiwania informacji cross-encoders często używane są do ulepszania wyników. Najpierw szybsze, ale mniej dokładne modele (zwane bi-encoders) znajdują potencjalne odpowiedzi, a potem CE dokładniej oceniają i szeregują te wyniki, zapewniając lepszą jakość końcowych odpowiedzi. Dlatego tworząc system RAG, warto wprowadzić zarówno modele bi-encoder, jak i cross-encoder. Na etapie indeksacji i wyszukiwania, gdzie ten proces musi być szybki, stosuje się modele bi-encoder, zaś w procesie optymalizacji wyników – model CE, który jest dokładniejszy (np. może ułożyć wyniki w ranking, poczynając od najbardziej trafnych).

Generowanie odpowiedzi – synteza informacji

W systemie RAG to model generatywny (np. PLLUM, Bielik, GPT, Claude, Gemini) odpowiada za analizę zebranych fragmentów dokumentów i tworzenie z nich spójnej odpowiedzi. Model działa tu podobnie jak ekspert, który najpierw zapoznaje się z dostarczonymi materiałami, a następnie formułuje odpowiedź. Jeżeli model generatywny nie otrzyma precyzyjnych instrukcji oraz odpowiednio dobranych fragmentów tekstów od użytkownika, może zacząć uzupełniać informacje własnymi „przypuszczeniami” (halucynować, konfabulować).

Niezwykle istotne jest precyzyjne podawanie źródeł informacji. System powinien działać jak skrupulatny naukowiec, oznaczając każdą informację odpowiednim odnośnikiem (przypisem) do dokumentu źródłowego. To nie tylko kwestia przejrzystości, ale także zabezpieczenie się przed niekontrolowaną kreatywnością modeli generatywnych. Gdy system jednoznacznie oznaczy źródłami pochodzenia treść generowanej odpowiedzi, łatwiej jest wówczas użytkownikowi wychwycić miejsca, gdzie model mógł halucynować. Warto pamiętać, że końcowa odpowiedź powinna być nie tylko precyzyjna, ale także zrozumiała dla użytkownika. System musi znaleźć równowagę między dokładnością cytowania źródeł a płynnością udzielanej odpowiedzi bez uszczerbku dla jej merytoryki – tak jak ekspert, który potrafi przedstawić skomplikowane informacje w przystępny sposób.

Podczas doboru modeli generatywnych do systemów RAG warto uwzględnić kilka kluczowych aspektów. Przede wszystkim należy zdecydować, czy model będzie uruchamiany lokalnie na serwerach użytkownika, czy w chmurze. Jest to istotne pod względem wydajności, kosztów oraz bezpieczeństwa danych. Modele działające w chmurze, takie jak OpenAI GPT czy Google Gemini, oferują elastyczność skalowania i stały dostęp do najnowszych aktualizacji, ale

mogą wiązać się z większymi kosztami i wymogami dotyczącymi ochrony danych. Z kolei lokalne modele, takie jak Meta-Llama czy dla języka polskiego pLLama i Bielik, umożliwiają uruchamiane ich na własnych serwerach. Daje to większą kontrolę nad samym modelem (dotrenowywanie, rozwój) oraz pełną kontrolę nad przechowywanymi danymi. Jednak takie podejście wymaga odpowiednich zasobów sprzętowych i specjalistycznej wiedzy do zarządzania infrastrukturą[4].

Ograniczenia i wyzwania dla systemów RAG

Ograniczenia techniczne

Jedną z najistotniejszych ograniczeń modeli generatywnych jest kwestia ilości danych, które dostarcza się do modelu. Każdy model ma określony limit tokenów, które może przetworzyć w jednym czasie (na wejściu do modelu). Tę właściwość modelu nazywamy jego **kontekstem.** Gdy mamy do czynienia z bardzo długimi dokumentami, należy je odpowiednio dzielić na mniejsze części, tak aby odpowiednio wypełniały kontekst modelu. Im ten kontekst jest mniejszy, tym bardziej przypomina czytanie książki po jednej stronie – tracimy wtedy szerszy obraz całości. Szerszy kontekst umożliwia modelowi analizę większej ilości danych w jednym czasie, co przekłada się na precyzję udzielanej odpowiedzi. Dla przykładu, długość kontekstu modelu Llama-3.3 to 128k tokenów, z kolei Mistral posiada kontekst długości 32k tokenów.

Kolejnym istotnym wyzwaniem jest analiza danych osadzonych w dokumentach, takich jak tabele i wykresy. System może **przekształcić obraz** *w* embedding, ale nie potrafi w pełni zrozumieć złożonych relacji przedstawionych w nim danych. To jak próba odczytania mapy przez kogoś, kto widzi tylko kolory, ale nie rozumie oznaczeń symbolami. Szczególnie problematyczne są dokumenty skanowane, gdzie jakość tekstu może być bardzo zniekształcona. Technologia OCR (*Optical Character Recognition*) z jednej strony nie zawsze radzi sobie z różnymi czcionkami, przekrzywionymi stronami czy adnotacjami odręcznymi, a z drugiej – jest niedostępna we wszystkich popularnych LLM.

Kwestia języków wprowadza dodatkowy poziom złożoności. Choć systemy RAG mogą obsługiwać wiele języków, ich skuteczność może się różnić w zależności od języka.

[4] Przykładowo, MetaAI do trenowania modelu LLama3 wykorzystała dwa klastry obliczeniowe, w każdej z nich znajdowało się po 24 576 kart graficznych Nvidia H100.

Szczególnie trudne są dokumenty zawierające mieszankę języków – na przykład raport techniczny po polsku z angielskimi terminami specjalistycznymi. System musi wtedy kojarzyć fakty z różnych języków, co w przypadku gdy model gen AI nie jest odpowiednio douczony w tym zakresie, może prowadzić do nieścisłości i błędów w interpretacji.

Wyzwania jakościowe

W przypadku aspektów jakościowych stajemy przed bardziej subtelnymi wyzwaniami. Dokładność systemu, choć generalnie wysoka, nie jest doskonała. Istnieje realne ryzyko pominięcia istotnych informacji, szczególnie gdy są one wyrażone w sposób nieoczywisty lub wymagają głębszego zrozumienia kontekstu. To jak próba złożenia skomplikowanej układanki, gdzie brak jednego elementu może znacząco zmienić odbiór całego obrazu. Problemy z kontekstem są szczególnie widoczne przy złożonych zagadnieniach. System może mieć trudność z uchwyceniem niuansów, ironii czy kontekstu kulturowego. Na przykład przy analizie dokumentów historycznych może nie uwzględnić ówczesnych realiów społecznych, co prowadziłoby do błędnej interpretacji faktów.

Wiarygodność generowanych odpowiedzi stanowi osobny obszar wyzwań. System musi nie tylko znaleźć informacje, ale także ocenić ich wiarygodność i spójność. Szczególnie niebezpieczne jest zjawisko **halucynacji** – gdy model językowy uzupełnia luki w wiedzy własnymi stwierdzeniami. To jak świadek, który, zeznając na przesłuchaniu, nieświadomie uzupełnia szczegóły wydarzenia niewidzianego przez siebie w całości. Weryfikacja źródeł staje się kluczowa, szczególnie gdy mamy do czynienia z informacjami sprzecznymi lub niejednoznacznymi. System musi rozpoznać, które źródła są bardziej wiarygodne, oraz połączyć informacje z różnych dokumentów w całość. Spójność odpowiedzi jest szczególnie trudna do utrzymania przy złożonych zapytaniach. System musi zachować logiczną strukturę wywodu, jednocześnie nie wpadając w pułapkę nadmiernego upraszczania lub ignorowania ważnych niuansów. Można to porównać do streszczania skomplikowanej teorii naukowej, gdzie trzeba zachować precyzję, nie tracąc przy tym zrozumiałości dla odbiorcy. Do automatycznej oceny systemów RAG stosuje się takie miary jak te przestawione w tabeli 4.1.

Tabela 4.1. Miary oceny wyszukiwania i wygenerowanej odpowiedzi

Nazwa miary	Opis
Retrieval Precision	Miara ta pokazuje, jak duży procent dokumentów zwróconych przez system jest istotny i odpowiada na zadane pytanie, co pozwala ocenić, jak trafne są wyniki wyszukiwania (AutoRAG 2024). Miara *precison* = 1 oznacza, że wszystkie zwrócone dokumenty są istotne. Dla systemów wymagających dużej dokładności (np. analizy prawne) wartość powinna być wyższa niż 0,9. Dla bardziej elastycznych zastosowań 0,7 może być wystarczającą wartością.
Retrieval Recall	Miara ta skupia się na tym, czy system potrafi znaleźć wszystkie istotne wyniki (fragmenty informacji związane z zapytaniem) wśród pierwszych K miejsc na liście wyników. Na przykład jeśli w zbiorze testowym jest 20 ważnych fragmentów, a system znalazł 10 z nich w pierwszych 15 wynikach, oznacza to, że znalazł połowę wszystkich istotnych dokumentów (AutoRAG 2024). Idealna wartość miary Recall wynosi 1, co oznacza, że system odzyskał wszystkie istotne dokumenty. W zależności od zastosowania w systemach, gdzie kompletność informacji jest kluczowa (np. badania naukowe), minimalna wartość powinna wynosić co najmniej 0,9, z kolei w zadaniach, gdzie szybkość i zwięzłość są ważniejsze, wartości powyżej 0,7 mogą być wystarczające.
MRR	Za pomocą MRR (Craswell N., 2009) ocenia się, jak szybko system potrafi znaleźć pierwszą istotną odpowiedź. Im bliżej początku listy znajduje się pierwszy trafny wynik, tym wyższa wartość MRR. Akceptowalna wartość bardzo zależy od kontekstu. Dla wyszukiwarek internetowych MRR może wynosić powyżej 0,8, jednak podczas bardziej złożonych zadań (np. interpretacja wyników badań) akceptowalne mogą być niższe wartości (np. 0,5–0,7).
NDCG	Miara NDCG (*ang. Normalized Discounted Cumulative* Gain) ocenia zarówno trafność wyników, jak i ich pozycję na liście wyszukiwania. NDCG bierze pod uwagę fakt, że trafniejsze wyniki, które pojawiają się wyżej na liście, są bardziej wartościowe (Dhinakaran, A., 2024). NDCG przyjmuje wartości od 0 do 1, gdzie 1 oznacza idealne uporządkowanie wyników według ich relewantności, a 0 wskazuje na całkowicie nieadekwatny ranking, w którym najważniejsze wyniki są na końcu listy.
BLEU	Miara ta porównuje wygenerowaną odpowiedź z zestawem referencyjnych odpowiedzi, koncentrując się na zgodności występujących w nich n-gramów (np. ciągów słów o zadanej długości). BLEU przyjmuje wartości od 0 do 1, gdzie 1 oznacza idealną zgodność wygenerowanego tekstu z tekstem referencyjnym, a 0 – brak zgodności (BLEU 2024).
ROUGE	To miara, która ocenia wygenerowany tekst przez porównanie go z zestawem odpowiedzi referencyjnych, koncentrując się na liczbie wspólnych sekwencji. ROUGE przyjmuje wartości od 0 do 1, gdzie 1 oznacza idealne pokrycie wygenerowanego tekstu z tekstem referencyjnym, a 0 – brak pokrycia (ROUGE 2023).

Źródło: Opracowanie własne.

Miary z powyższej tabeli wykorzystywane są do automatycznej oceny odpowiedzi wygenerowanej przez wybrany model gen AI. Miary te nie są widoczne dla końcowego użytkownika systemu RAG, niemniej na podstawie tych miar mogą być podejmowane decyzje o kierunkach trenowania modelu. Miary takie jak Retrieval Precision, Recall, MRR, NDCG i BLEU można weryfikować za pomocą oznaczonych zbiorów testowych oraz narzędzi

analitycznych (LlamaIndex, Pytrec_eval), które porównują wyniki systemu z oczekiwanymi odpowiedziami.

Niezależnie jednak od automatycznej oceny systemów RAG, najistotniejszą rolę odgrywa tutaj człowiek, który ocenia wygenerowaną odpowiedź pod kątem jej trafności, zrozumiałości, użyteczności oraz zgodności z kontekstem zapytania, co dostarcza cennych informacji o rzeczywistej jakości generowanych treści i ich przydatności.

4.3. Zastosowania RAG w praktyce badawczej

Możemy wyodrębnić dwie główne kategorie zastosowań systemów RAG w praktyce badań aplikacyjnych.

Po pierwsze, RAG może być doskonałym narzędziem do szybkiej analizy jakościowego materiału badawczego, np. z wywiadów pogłębionych lub z pytań otwartych pochodzących z ankiet[5]. Takie systemy mogą umożliwić szybkie wstępne kodowanie transkrypcji wywiadów (odpowiedzi na pytania otwarte) według zdefiniowanych kategorii. Pozwolają one zidentyfikować główne tematy i podtematy oraz porównywać odpowiedzi między różnymi respondentami. Systemy RAG pozwalają także na identyfikację cytatów reprezentatywnych dla danych kategorii czy sugerowanie powiązań między różnymi wątkami oraz generowanie podsumowań.

Podczas ewaluacji dużego programu, w którym zebrano materiał pochodzący ze stu wywiadów pogłębionych, RAG może pomóc w pierwszym etapie analizy poprzez identyfikację powtarzających się tematów i sugerowanie wstępnych kategorii analitycznych, na których powinna się skupić uwaga badaczy.

Wykorzystanie systemu RAG w analizie wywiadów może znacząco usprawnić proces badawczy, jednak należy pamiętać o jego właściwym zastosowaniu. System świetnie sprawdzi się jako narzędzie do szybkiego przeglądu materiału i do zapewnienia wstępnej orientacji w zakresie głównych tematów, szczególnie gdy mamy do czynienia z obszernym materiałem badawczym. Jednak w przypadku, gdy zależy nam na głębokiej i szczegółowej

[5] Więcej na ten temat w rozdziale pt. „Analizy jakościowe wspierane gen AI”.

analizie, rekomendowane jest najpierw przeprowadzenie tradycyjnego kodowania (Friese, 2023), a następnie umieszczenie zakodowanego materiału w systemie RAG do efektywnego przeszukiwania i analizy.

W praktyce najlepsze efekty przynosi korzystanie z dedykowanych dla badaczy aplikacji analitycznych lub ze spersonalizowanych rozwiązań tworzonych w podejściu *code* lub *low/no-code* i dostosowanych do specyfiki prowadzonych badań. Z drugiej strony systemy RAG są także doskonałymi narzędziami do integracji i analizy danych pochodzących z wielu badań, artykułów naukowych i opracowań analitycznych. Systemy te pozwalają na efektywne łączenie różnorodnych źródeł informacji, zwiększają dostęp do rozproszonej wiedzy i ułatwiają szybkie pozyskiwanie kluczowych informacji do opracowań analitycznych. RAG nie tylko generuje zwięzłe podsumowania złożonych zbiorów danych, ale również aktywnie wskazuje potencjalne luki w materiale badawczym oraz niespójności wymagające dodatkowej weryfikacji. Ciekawym przykładem takiego rozwiązania jest AIDA (*Artificial Intelligence for Development Analytics*) – platforma stworzona przez UNDP, która umożliwia inteligentne przeszukiwanie ponad 6000 raportów ewaluacyjnych, błyskawicznie dostarczając poszukiwanych informacji i dowodów.

W praktyce system RAG znajduje zastosowanie w różnych kontekstach. Decydenci mogą wykorzystać go do szybkiego pozyskiwania dodatkowych argumentów i dowodów wspierających ich decyzje[6], uzyskują bowiem niemal natychmiastowy dostęp do odpowiednich badań i analiz. Z kolei komórki analityczne, które planują nowe badania, mogą wykorzystać RAG do sprawnego przeglądu dostępnych badań i źródeł zewnętrznych, co pozwoli im zweryfikować aktualny stan wiedzy w danym obszarze (dorobek naukowo-badawczy) i precyzyjniej określić zakres planowanego badania.

Sami badacze mogą także wykorzystywać systemy RAG do pogłębienia wiedzy na temat kontekstu eksplorowanych problemów badawczych. Dobrym przykładem jest zastosowanie takich rozwiązań jak Perpelxity czy Scite, które umożliwiają przeprowadzenie szybkiej eksploracji wybranych zagadnień na podstawie danych dostępnych w Internecie, źródeł naukowych oraz treści z social mediów.

[6] Niestety, może to również służyć niebezpiecznym praktykom poszukiwania argumentów pod z góry ustaloną tezę.

Systemy RAG są coraz częściej integrowane z agentami AI, co rozszerza ich możliwości – nie tylko potrafią zidentyfikować właściwe źródła i wyodrębnić z nich poszukiwane treści, ale także mogą przygotować kompleksowe opracowania analityczne na zadane tematy. Należy jednak zauważyć, że w tych rozwiązaniach mamy ograniczony wpływ na proces selekcji źródeł, gdyż wyspecjalizowani agenci AI samodzielnie decydują, które materiały będą najbardziej odpowiednie do udzielenia odpowiedzi na nasze zapytanie. Wśród wiodących rozwiązań w tym obszarze znajdują się Google's Deep Research oraz funkcje Deep Research oferowane przez OpenAI i Perplexity.

W praktyce badawczej systemy RAG oferują badaczom dwie komplementarne ścieżki wykorzystania: jako narzędzie wspomagające analizę pierwotnego materiału badawczego oraz jako system wspierający syntezę i integrację wiedzy z różnorodnych źródeł zewnętrznych i wewnętrznych. W obu przypadkach kluczowe jest zachowanie kontroli nad procesem analitycznym oraz krytyczna weryfikacja wykorzystywanych źródeł, szczególnie w kontekście rosnącej integracji systemów RAG z agentami AI, które mogą autonomicznie selekcjonować materiały do analizy.

4.4. Rozpoczęcie pracy z RAG

Rozpoczęcie pracy z systemami RAG może przebiegać różnymi ścieżkami, w zależności od poziomu zaawansowania technicznego użytkownika i specyficznych potrzeb organizacji. W praktyce możemy wyróżnić trzy główne podejścia: korzystanie z gotowych rozwiązań dla użytkowników początkujących, platformy *low/no-code* dla średniozaawansowanych oraz budowa systemu od podstaw dla użytkowników z doświadczeniem programistycznym. Przyjrzyjmy się bliżej każdej z tych opcji.

Gotowe rozwiązania (dla początkujących)

Gotowe rozwiązania to idealna opcja dla osób, które dopiero zaczynają swoją przygodę z systemami RAG i nie mają doświadczenia w programowaniu. Te platformy oferują intuicyjne interfejsy, które pozwalają na szybkie rozpoczęcie pracy z naszą biblioteką dokumentów, bez potrzeby posiadania zaawansowanej wiedzy technicznej.

Obecnie dostępnych jest kilka narzędzi wykorzystujących RAG do pracy z dokumentami naukowymi. Google oferuje Notebook LM – popularne narzędzie do analizy dokumentów z funkcją generowania podsumowań i notatek. Jego kluczową zaletą jest śledzenie cytatów, co zapewnia wiarygodność źródeł, oraz obsługa różnych formatów plików.

Podobne możliwości oferują wiodące modele językowe. Zarówno ChatGPT jak i Claude posiadają funkcję ‘projects’, która pozwala na wgranie własnych dokumentów w różnych formatach i prowadzenie analizy poprzez zadawanie pytań dotyczących zgromadzonych materiałów.

Z kolei dla naukowców przydatne mogą być platformy Elicit/Scite, wyspecjalizowane w procesie badawczym. Wykorzystują one RAG do szybkiego przeszukiwania oraz syntezy publikacji i artykułów naukowych.

Rozwiązania *no-code* (dla średniozaawansowanych)

Rozwiązania *low/no-code* to narzędzia, które umożliwiają budowanie systemów RAG bez potrzeby pisania kodu. Dzięki graficznym interfejsom użytkownicy mogą łatwo tworzyć aplikacje oraz automatyzować różne procesy, co czyni je idealnymi dla średniozaawansowanych użytkowników, którzy chcą wykorzystać moc technologii bez konieczności posiadania umiejętności programistycznych. Przykładami takich aplikacji mogą być: **Relevance AI**, **n8n** czy **Make**, które umożliwiają tworzenie i wdrażanie narzędzi AI, w tym RAG, opartych na dużych modelach językowych (LLM), bez potrzeby programowania. Oferują one wizualny interfejs do budowy rozwiązań, obsługują wielu dostawców LLM i zapewniają wbudowane magazyny wektorowe do przetwarzania danych.

Budowa od zera (dla zaawansowanych)

Najbardziej zaawansowaną opcją jest budowa własnego systemu RAG od podstaw. To podejście, choć wymaga doświadczenia w programowaniu i znajomości przetwarzania języka naturalnego (NLP), daje pełną kontrolę nad systemem i możliwość jego precyzyjnego dostosowania do konkretnych potrzeb organizacji.

Twórcy tego typu systemów mają do wyboru dwie główne ścieżki techniczne. Pierwsza to wykorzystanie biblioteki *Hugging Face Transformers* – otwartego narzędzia do pracy

z modelami językowymi. Pozwala ona na trenowanie własnych modeli RAG oraz ich dokładne dostosowanie do specyfiki posiadanych danych i potrzeb organizacji.

Druga opcja to integracja poprzez API z najnowszymi, potężnymi modelami językowymi (np.: Claude 3.5 Sonnet, GPT o1, Bielik). Choć wymaga ona umiejętności programistycznych, daje dostęp do najbardziej zaawansowanych możliwości tych modeli i pozwala wykorzystać je w tworzonych aplikacjach. Takie rozwiązanie, mimo że wymaga więcej czasu i zasobów na wdrożenie, może być optymalne dla organizacji o bardzo specyficznych wymaganiach lub potrzebie pełnej kontroli nad procesem przetwarzania danych.

4.5. Podsumowanie

Wdrożenie technologii RAG w badaniach polityk publicznych otwiera nowe możliwości, stawiając jednocześnie przed badaczami, ewaluatorami i analitykami istotne wyzwania techniczne. Podstawę sprawnego działania tego rozwiązania stanowi właściwe przygotowanie danych źródłowych. Obejmuje to odpowiednią segmentację dokumentów, ich organizację oraz indeksowanie, przy czym szczególną uwagę należy zwrócić na ograniczenia w przetwarzaniu elementów graficznych, takich jak wykresy czy tabele.

RAG znajduje zastosowanie w dwóch głównych obszarach badań. Pierwszy to sprawna analiza materiału jakościowego takiego jak odpowiedzi na pytania otwarte. System przyspiesza proces kodowania i analizy, choć przy pogłębionych analizach powinien być łączony z tradycyjnymi metodami. Drugi obszar to synteza wiedzy z różnorodnych źródeł zastanych, szczególnie przydatna w planowaniu zakresu badań i wsparciu procesu decyzyjnego.

Szczególnie ważne jest takie skonfigurowanie instrukcji, by system tworzył odpowiedzi wyłącznie na podstawie dostarczonych materiałów, a nie swojej wbudowanej wiedzy. Chociaż RAG zmienia sposób pracy badaczy, nie powinni oni nadmiernie polegać na zautomatyzowanych analizach, lecz zachować krytyczne podejście do generowanych wyników. Jest to zgodne z najnowszymi wynikami badań naukowych, które pokazują, że chociaż narzędzia gen AI zwiększają efektywność pracy, kluczowe pozostaje zachowanie krytycznego myślenia w ocenie ich wyników. Jest to szczególnie istotne, gdyż badania te także dowodzą, że większe zaufanie do AI często wiąże się z mniejszym zaangażowaniem

w krytyczną analizę, podczas gdy wyższe zaufanie do własnych kompetencji prowadzi do bardziej wnikliwej oceny generowanych rezultatów (Lee et al., 2025).

Dostępność różnych wariantów wdrożenia – od gotowych platform dla początkujących, przez rozwiązania *low/no-code*, po zaawansowane rozwiązania – pozwala dostosować poziom techniczny do możliwości organizacji i użytkownika. Niezależnie od wybranej ścieżki, ich wdrożenie wymaga przemyślanego podejścia do aspektów prawnych, organizacyjnych i etycznych. W tej nowej rzeczywistości uwzględniającej AI rola badaczy ewoluuje – stają się oni ekspertami łączącymi wiedzę dziedzinową ze świadomością technologiczną.

Bibliografia

- AutoRAG. (2024). *Tips to Understand RAG Retrieval Metrics* (dostęp: 31.03.2025).
- BLEU - *Bilingual evaluation understudy*. (2024) (dostęp: 31.03.2025).
- Craswell, N. (2009). *Mean Reciprocal Rank*. W: Liu, L., Özsu, M.T. (red.). Encyclopedia of Database Systems. Springer, Boston, MA.
- Dhinakaran, A. (2023). Demystifying NDCG. How to best use this important metric for monitoring ranking models (dostęp: 31.03.2025).
- Friese, S. (2023, April 22). *Two case studies on the application of generative AI in qualitative data analysis: Success and Disenchantment* (dostęp: 31.03.2025).
- Hartley, J.S., Jolevski, F., Melo, V., Moore, B. (2024). *The labor market effects of generative artificial intelligence*. SSRN (dostęp: 31.03.2025).
- Lee, H., Kim, S. (2024). *Bring Retrieval Augmented Generation to Google Gemini via External API: An Evaluation with BIG-Bench Dataset* (dostęp: 31.03.2025).
- Lee, H.-P., Sarkar, A., Tankelevitch, L., et al. (2025). The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. W: CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan. ACM (dostęp: 31.03.2025).
- Reimers, N. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084*.
- ROUGE. (2024). *Recall-Oriented Understudy for Gisting Evaluation* (dostęp: 31.03.2025).
- Rybak P., Ogrodniczuk M. (2024). *Silver Retriever: Advancing Neural Passage Retrieval for Polish Question Answering*. W: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 14826–14831, Torino, Italia. ELRA and ICCL.
- Zucchini, L. (2024, February 7). *Introduction to RAG — GenAI Systems for Knowledge*. Curiosity (dostęp: 31.03.2025).

CZĘŚĆ II

Zastosowania gen AI w procesie ewaluacji

5. Zamawianie usług ewaluacyjnych – tworzenie Opisu Przedmiotu Zamówienia

Jacek Szut

5.1. Wstęp

Niniejszy rozdział ma na celu pokazanie możliwości, jakie daje wykorzystanie gen AI w pracy jednostki ewaluacyjnej (JE), ulokowanej w instytucji sektora finansów publicznych i zamawiającej usługi ewaluacyjne dotyczące określonych programów lub polityk. Ze względu na ograniczenie objętości rozdziału skupiono się głównie na opisie działań, które wykonano przy wsparciu LLM (przede wszystkim model Claude/Anthropic, niekiedy GPT-4o/ OpenAI)[1] w zadaniach związanych z przygotowaniem Opisu Przedmiotu Zamówienia (OPZ) na ww. usługi. Wybór tego dokumentu był celowy. Stanowi on bowiem jeden z ważniejszych wytworów w pracy JE. Jednocześnie jest opracowaniem złożonym i wymagającym od jego autora znajomości metodyki badań ewaluacyjnych, formułowania celów badawczych oraz towarzyszących im pytań badawczych. Istotne jest również posiadanie wiedzy w przedmiocie badania (w tym kontekstu i dotychczasowych ustaleń naukowo-badawczych) na temat potrzeb interesariuszy oraz znajomość zagadnień z zakresu prawa zamówień publicznych. To ostatnie wynika z faktu, że JE, tworząc OPZ, pozostaje z reguły także odpowiedzialna za przygotowanie kryteriów oceny ofert oraz warunków umożliwiających wykonawcom ubieganie się o zamówienie.

W trakcie przygotowywania tekstu staraliśmy się działać modelowo, tzn. prowadzić pracę zgodnie z procesem, który uznaliśmy za optymalny dla powstania dobrej jakości OPZ. Jednak zdajemy sobie sprawę, że przedstawione tu przykłady nie znajdą zastosowania we wszystkich JE, a ich dobór nie musi być odzwierciedleniem pełnowartościowego OPZ –

[1] Był to wybór czysto subiektywny. Zdaniem autora, w momencie przygotowywania artykułu wskazane modele najlepiej radziły sobie z zadaniami wykonywanymi przez JE.

stanowią jedynie elementy owego modelu. Ponadto podczas pracy nad rozdziałem stanęliśmy przed wyzwaniem wymyślenia fikcyjnego OPZ. Nie był to zatem dokument powiązany z którymkolwiek formalnie zaplanowanym badaniem czy realizowanym programem (np. ujętym w Planie Ewaluacji programu Fundusze Europejskie dla Nowoczesnej Gospodarki 2021–2027 lub innej interwencji publicznej).

Uważamy jednak, że przedstawiony proces, w którym gen AI pomaga w przygotowaniu OPZ, dobrze ilustruje potencjał wykorzystania tego modelu pracy oraz pokazuje, w jaki sposób gen AI może pomóc JE. Przedstawione przykłady pozwalają zobaczyć, jak można zintegrować sprawność narzędzi sztucznej inteligencji z doświadczeniem i wiedzą specjalistyczną pracowników JE, aby osiągnąć satysfakcjonujący (gdzie było to możliwe) efekt.

Należy dodać, że naszym założeniem było zgromadzenie informacji, które będą pomocne w odpowiedzi na pytania: 1. Czy w pracach JE warto korzystać ze wsparcia LLM? 2. Czy wsparcie LLM może sprawić, że efekty (produkty) prac JE będą wyższej jakości, a jeżeli nie, to jakie ewentualne inne korzyści niesie ze sobą sięganie po LLM?

Zanim przejdziemy dalej, wyjaśnimy, czym jest OPZ oraz dlaczego praca nad tym dokumentem może być czasochłonna i wymagająca dużej wiedzy. Jest to niezbędne dlatego, że nie każdy z czytelników musi znać specyfikę pracy w JE.

Opis Przedmiotu Zamówienia to dokument wymagany przez ustawę Prawo Zamówień Publicznych (uPzp). W art. 134 ustawodawca określa zawartość Specyfikacji Warunków Zamówienia (SWZ), która zgodnie z ust. 1 tego artykułu powinna zawierać m.in. co najmniej opis przedmiotu zamówienia. Jednocześnie art. 99 ustawy zawiera zasady opisywania przedmiotu zamówienia. Możemy przeczytać m.in., że OPZ opisuje się w sposób jednoznaczny i wyczerpujący, za pomocą dostatecznie dokładnych i zrozumiałych określeń, z uwzględnieniem wymagań i okoliczności mogących mieć wpływ na sporządzenie oferty; określa się (w OPZ) wymagane cechy dostaw, usług lub robót budowlanych.

Zasadniczo zapisy uPzp nakazują zamawiającemu przygotowanie OPZ tak, aby ten zawierał informacje, które są niezbędne do pełnego zrozumienia wymagań zamawiającego dotyczących przedmiotu zamówienia oraz sposobu jego realizacji. OPZ musi być sporządzony w sposób umożliwiający każdemu z wykonawców opracowanie odpowiedniej oferty, tj. adekwatnej do przedmiotu zamówienia, sposobu wykonania oraz kryteriów oceny oferty.

Z drugiej strony oferta taka musi być sporządzona w sposób pozwalający zamawiającemu na jej rzetelną ocenę, na co również ma wpływ przygotowany OPZ[2]. W zależności od specyfiki zamówienia omawiany dokument powinien zawierać dokładny opis zamówienia, wymogi dotyczące standardów realizacji usługi/dostaw/robót budowalnych itp.

W przypadku zamawiania usług ewaluacyjnych ukształtowała się praktyka umieszczania w OPZ informacji dotyczących: 1. kontekstu realizacji badania, 2. jego celów oraz 3. pytań badawczych, 4. wymagań metodologicznych i 5. oczekiwanych produktów zamówienia. Ponadto w OPZ możemy znaleźć informacje na temat współpracy pomiędzy zamawiającym i wykonawcą, np. w zakresie odbioru i oceny produktów. Takie podejście cechuje zdecydowaną większość JE w Polsce. Można je uznać za standard.

Natomiast nie wykształciła się jednolita praktyka dotycząca stopnia szczegółowości zapisów OPZ. Dotyczy to zwłaszcza metodyki badania. Są JE, które przygotowują dokument zawierający zapisy względnie ogólne, np. wymagają zastosowania metod ilościowych, a nie określają ich parametrów (technika, sposób doboru próby, zasady stosowania wag postrealizacyjnych itd.). Są również JE, które opisują to w sposób szczegółowy, np. określają technikę, preferowaną metodę doboru próby itd.

Oba podejścia mają swoje plusy i minusy. W przypadku zapisów o charakterze ogólnym daje się wykonawcom możliwość wykazania się np. innowacyjnością na etapie przygotowania oferty i ostatecznie uzyskania wysokiej jakości produktu dzięki temu, że wykonawca podszedł do projektu w sposób nieszablonowy. Jednak z drugiej strony narażamy się wówczas na ryzyko wpłynięcia ofert cechujących się niską jakością (ale spełniających wymagane minimum) i ostatecznie otrzymania produktów końcowych o relatywnie niewielkiej wartości merytorycznej[3]. Z odwrotną sytuacją mamy do czynienia, gdy nasz OPZ zawiera szczegółowe wytyczne dotyczące np. metodyki badania. Ograniczamy tym wykonawcy możliwość stosowania innowacyjnego lub po prostu lepiej przemyślanego podejścia badawczego.

[2] Przy czym wytyczne dotyczące sposobu przygotowania oferty oraz zastosowanych kryteriów oceny ofert są zawarte w innym dokumencie – Specyfikacja Warunków Zamówienia (SWZ). OPZ stanowi jeden z załączników do SWZ.

[3] Pamiętajmy, że zamawiający musi stosować określony sposób oceny oferty i jest to szczególnie trudne w przypadku zamawiania usług opartych na wiedzy (a do takich zaliczamy badania ewaluacyjne), gdzie jakość oferty (jej poszczególnych elementów merytorycznych) nie może być jednoznacznie ujęta w kryteria zerojedynkowe (spełnia/nie spełnia). Zamawiającego zawsze mogą spotkać zarzuty, że jego ocena była nazbyt subiektywna (ekspercka).

Jednakże jeżeli dobrze znamy się na metodyce badań ewaluacyjnych, mamy jednocześnie dużą pewność, że badanie będzie zrealizowane poprawnie.

5.2. Tworzenie OPZ z wykorzystaniem gen AI – ogólne zasady

W trakcie przygotowywania OPZ możemy korzystać z gen AI od samego początku, tj. zaangażować ją w momencie pracy nad strukturą dokumentu i jego poszczególnych części lub wykorzystywać w wybranych przez siebie obszarach. Innymi słowy, sztuczna inteligencja może nam towarzyszyć w pracy nad każdą częścią OPZ i wówczas budujemy w trakcie rozmowy z nią „długi wątek" dotyczący naszego OPZ. Możemy również z niej korzystać w celu dopracowania wcześniej przygotowanego przez nas dokumentu.

Jeżeli decydujemy się na „długi wątek", musimy uwzględnić ograniczenia kontekstowe. Model może zapomnieć wcześniejsze ustalenia. Dlatego na początku każdego nowego etapu pracy warto nakazać modelowi dokonanie podsumowania dotychczasowych jej efektów. Ponadto w dalszej pracy w takim „wątku", tworząc prompty, lepiej odwoływać się do wcześniejszych wyników, niż tworzyć instrukcje ogólne. Zamiast *zaproponuj metodykę*, lepiej napisać *biorąc pod uwagę, że cele badania to…, a grupy objęte badaniem to…, przygotuj propozycję metodyki* (tutaj możemy dodać dodatkowe wytyczne).

Niezależnie od tego, na jaki wariant się zdecydujemy („długi wątek" vs. „praca w wybranych obszarach"), autor zdecydowanie rekomenduje pracę zgodną z zasadą od ogółu do szczegółu. Pozwala ona na stopniowe rozwijanie wiedzy gen AI na temat wykonywanego zadania i prowadzi do otrzymywania lepszych odpowiedzi. Z drugiej strony użytkownik ma wówczas lepszą kontrolę nad pracą gen AI, łatwiej prowadzić mu jej walidację. Lepiej zatem nakazać modelowi, aby przygotował ogólny zarys metodyki badania, niż od razu wymagać od niego przedstawiania opisu szczegółowego.

Ponadto istotne jest, abyśmy przekazali modelowi dokładny kontekst zadania, niezależnie od tego, czy decydujemy się na wariant pracy nad całym OPZ, czy nad jego wybranymi elementami. Może to być opis interwencji, która ma być poddana ewaluacji itp. Zasadniczo to, co przekażemy gen AI, musi korespondować z wykonywanym przez nią zadaniem.

Przykładowo, jeżeli decydujemy się na doskonalenie wytworzonych przez nas fragmentów OPZ i jednym z nich będzie zaproponowanie metodyki badania, wówczas niezbędne jest dostarczenie modelowi informacji ważnych dla tego etapu pracy. Będą to więc cele badania i pytania badawcze, choć oczywiście warto mu również przekazać informacje na temat grup poddanych interwencji, ich specyfiki itd.

W trakcie pracy z modelem możemy się zdecydować na trzy ścieżki postępowania: 1. model jako inicjator działań, 2. człowiek jako inicjator działań lub 3. hybrydowa – raz model podejmuje pierwszy krok, innym razem robi to człowiek. W tabeli 5.1. przedstawiono porównanie tych trzech podejść.

Tabela 5.1. Porównanie trzech podejść do pracy z gen AI przy tworzeniu OPZ

Aspekt	Mocne strony	Słabe strony	Kiedy stosować	Dla kogo
Model inicjuje	• Szybki start • Szeroki przegląd możliwości •Ustandaryzowana struktura • Efektywne czasowo	• Ryzyko powierzchowności • Możliwe pominięcie specyfiki • Pasywna rola człowieka • Mniejsze zaangażowanie poznawcze	• Nowe obszary tematyczne • Potrzeba szybkiego prototypu • Blokada twórcza • Wysoka presja czasu	• Osoby w całkiem nowych obszarach • Przy wysokiej presji czasowej • Koordynatorzy projektów • W sytuacjach nietypowych
Człowiek inicjuje	• Rozwój własnych kompetencji • Głębsze zrozumienie tematu • Uwzględnienie specyfiki • Pełna kontrola procesu	• Wolniejszy start • Większy nakład czasu • Możliwe luki w wiedzy • Wymaga doświadczenia	• Znane obszary • Projekty specjalistyczne • Standardowe sytuacje • Czas nie jest krytyczny	• Doświadczeni pracownicy • Eksperci dziedzinowi • Osoby z wiedzą w temacie • Przy typowych projektach
Podejście hybrydowe	• Elastyczność działania • Łączy zalety obu podejść • Optymalizacja procesu • Dostosowanie do sytuacji	• Wymaga doświadczenia w pracy z gen AI • Może być niespójne • Trudniejsze zarządzanie • Wymaga planowania	• Złożone projekty • Różnorodne elementy OPZ • Zróżnicowany poziom wiedzy • Mieszane priorytety	• Osoby znające pracę z gen AI • Zespoły projektowe • Przy zróżnicowanych zadaniach • Doświadczeni koordynatorzy

Źródło: Opracowanie własne z wykorzystaniem Claude 3.5 Sonnet.

Potrzeby wiedzy

Ważnym etapem pracy nad OPZ jest określenie celów badania. Możemy je przygotować, kierując się doświadczeniem (z wcześniejszych ewaluacji) oraz wiedzą w obszarze badania. Jednakże każdy projekt ma swoich odbiorców, a ich potrzeby wiedzy nie zawsze muszą w pełni korespondować z tym, co zaprojektowaliśmy, bazując na naszej wiedzy. Gen AI może być pomocna w określaniu hipotetycznych potrzeb wiedzy interesariuszy. Nie zwalnia nas to od późniejszych konsultacji z interesariuszami, ale wówczas taki proces będzie prowadzony na podstawie wcześniej przygotowanego zestawu celów, który przypuszczalnie będzie spójny z ich potrzebami. Jednocześnie wygenerowany (i zweryfikowany przez nas) materiał będzie stanowił dobrą podstawę do pogłębienia rozmowy na temat celów badania (obszarów zainteresowania odbiorców ewaluacji). Takie podejście jest wartościowe zwłaszcza wtedy, gdy przygotowywany przez nas projekt stanowi swego rodzaju nowość. Niemniej, również w przypadku projektów powtarzalnych warto przynajmniej przetestować możliwości gen AI. Efektem tego może być ujawnianie nowych potrzeb, których wcześniej nie zauważaliśmy.

W jaki sposób przeprowadzić taki proces? Nie ma jednej „złotej reguły". Poniżej zaprezentowano trzy rozwiązania.

1. Nakazanie gen AI wyszukania luk informacyjnych w dotychczasowej wiedzy[4]:

> *W bazie wiedzy masz projekt dotyczący modelowania makroekonomicznego efektów POIR i POPW. Chcę wykonać podobne badanie, ale dotyczące innych programów. Wykonaj następujące kroki: 1) Sprawdź, jakie były cele tego projektu 2) Zastanów się, czy te cele były kompletne, czy też można było zaproponować dodatkowe cele 3) Jeżeli można zaproponować dodatkowe cele, wówczas zrób to i uzasadnij.*

[4] Można spróbować nakazania gen AI, aby szukała luk wiedzy poprzez odwołanie się do ogólnej wiedzy modelu, ale prawdopodobnie w przypadku badań ewaluacyjnych wynik pracy modelu nie będzie satysfakcjonujący. Innym rozwiązaniem, być może o relatywnie wysokiej efektywności, będzie skorzystanie z agentów powiązanych z bazami naukowymi. Aczkolwiek w tym przypadku dobre wyniki będą dotyczyły ogółu światowego dorobku badawczego, zaś kontekst krajowy, w tym zwłaszcza ewaluacyjny, nie zostanie uwzględniony lub będzie uwzględniony w minimalnym stopniu.

Rycina 5.1. Fragment odpowiedzi

2. Kompletność celów

Analizując cele tego projektu, można zauważyć, że były one dobrze dobrane do głównego zadania, czyli oceny makroekonomicznych efektów programów. Cele obejmowały kluczowe obszary, takie jak skuteczność, komplementarność, efektywność i identyfikację barier, co pozwalało na kompleksową ocenę wpływu programów na rozwój gospodarczy.

3. Propozycja dodatkowych celów

Dodatkowe cele, które mogłyby być uwzględnione w badaniu, to:

1. Ocena długoterminowej trwałości efektów programów: Warto byłoby przeanalizować, czy osiągnięte rezultaty mają charakter długoterminowy, czy są jedynie krótkotrwałe. To mogłoby pomóc w ocenie trwałości wpływu interwencji i ich rzeczywistego wkładu w zrównoważony rozwój gospodarczy.
   - Uzasadnienie: Trwałość efektów jest kluczowym aspektem każdego programu

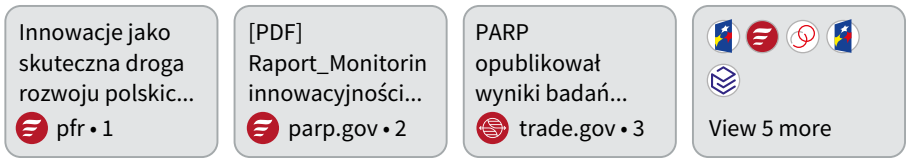
Źródło: Opracowanie własne z wykorzystaniem EvalDesign (agent oparty na GPT-4o, stworzony przez autora rozdziału do pracy w obszarze ewaluacji).

Przedstawiony prompt stanowi jedynie przykład. W zależności od naszych potrzeb na etapie szukania luk informacyjnych możemy oczekiwać od gen AI wykonania innego zadania, np. zaproponowania nowych celów i jednocześnie ewentualnej rezygnacji z niektórych, które były przedmiotem wcześniejszego badania, podjęcia próby bardziej syntetycznego ujęcia celów itp.

2. Nakazanie gen AI przygotowania celów badania:

> *Przygotowuję badanie dotyczące wspierania innowacyjności polskich przedsiębiorstw. Wsparcie to odbywa się w ramach programów unijnych. Jakie mogą być cele takiego badania?*

Rycina 5.2. Fragment odpowiedzi



Cele badania wspierania innowacyjności polskich przedsiębiorstw w ramach programów unijnych

Badanie dotyczące wspierania innowacyjności polskich przedsiębiorstw w ramach programów unijnych może mieć

następujące cele:

1. Ocena skuteczności programów unijnych we wspieraniu innowacyjności

Celem badania może być ocena, na ile programy unijne faktycznie przyczyniają się do zwiększenia innowacyjności polskich firm. Analiza może obejmować m.in.:

- Wpływ dofinansowania z UE na wdrażanie innowacji w przedsiębiorstwach (1,3)
- Ocenę efektywności kosztowej wsparcia innowacyjności (4)
- Identyfikację kluczowych barier i czynników sukcesu we wdrażaniu innowacji przez firmy korzystające ze środków UE (1,3)

Źródło: Opracowanie własne z wykorzystaniem Perplexity.

3. Nakazanie gen AI odegrania roli określonego interesariusza:

W Polsce jest realizowany program współfinansowany ze środków unijnych. Jest to program FENG. Jedno z działań programu, umownie nazwijmy je działaniem X, ma na celu wzrost innowacyjności oraz stopnia prowadzenia i wykorzystania prac B+R. Działanie to jest adresowane do przedsiębiorstw mikro oraz małych. Przedsiębiorcy składają wniosek o wsparcie, w którym opisują swój projekt, jego cele oraz różne parametry finansowe itp. Następnie wniosek jest oceniany przez ekspertów z instytucji pośredniczącej, w oparciu o kryteria formalne (np. zdolność finansowa przedsiębiorcy do realizacji projektu) i merytoryczne (np. zgodność projektu przedsiębiorcy z celami działania X). Jeżeli wniosek uzyska pozytywną ocenę, wówczas przedsiębiorca podpisuje umowę i przystępuje do realizacji projektu, część kosztów poniesionych przez niego wydatków zostanie mu zwrócona. Programem zarządza instytucja zarządzająca (IZ), zaś

jest on wdrażany przez instytucje pośredniczące (IP). IZ jest ministerstwo, a IP to z reguły agencja, np. rządowa.

Mamy rok 2024 i wdrażanie działania X jest na początkowym etapie, uruchomiono pierwsze konkursy i podpisano pierwsze umowy. Chcemy zaprojektować badanie ewaluacyjne tego działania. W związku z tym chcemy określić potrzeby informacyjne interesariuszy. Wciel się w przedstawiciela IZ i określ, jakie mogą być obszary twojego zainteresowania, które powinny zostać uwzględnione w badaniu ewaluacyjnym. Uzasadnij swoją odpowiedź.

Rycina 5.3. Fragment odpowiedzi

Jako przedstawiciel Instytucji Zarządzającej (IZ) programem FENG jestem szczególnie zainteresowany kilkoma kluczowymi obszarami, które powinny zostać uwzględnione w badaniu ewaluacyjnym działania X. Oto moje główne obszary zainteresowania wraz z uzasadnieniem:

1. Efektywność procesu wyboru projektów: Uzasadnienie: Jako IZ musimy upewnić się, że proces selekcji projektów jest skuteczny w wyłanianiu inicjatyw o najwyższym potencjale innowacyjnym. Chcemy zrozumieć, czy kryteria oceny są odpowiednio skalibrowane i czy pozwalają na identyfikację projektów, które rzeczywiście przyczynią się do wzrostu innowacyjności i wykorzystania prac B+R w sektorze mikro i małych przedsiębiorstw.
2. Adekwatność wsparcia do potrzeb beneficjentów: Uzasadnienie: Interesuje nas, czy oferowane wsparcie odpowiada rzeczywistym potrzebom mikro i małych przedsiębiorstw w zakresie innowacji i B+R. Chcemy zidentyfikować ewentualne luki lub bariery, które mogą ograniczać skuteczność działania X.
3. Wstępna ocena wpływu na innowacyjność i B+R: Uzasadnienie: Mimo wczesnego etapu wdrażania, chcemy uzyskać pierwsze sygnały dotyczące wpływu działania X na zwiększenie innowacyjności i stopnia prowadzenia prac B+ R wśród beneficjentów. Pomoże to w ewentualnym dostosowaniu programu w przyszłości.
4. Efektywność współpracy między IZ a IP: Uzasadnienie: Jako IZ chcemy ocenić, czy nasza współpraca z Instytucją Pośredniczącą przebiega sprawnie i czy nie ma obszarów wymagających usprawnienia, co mogłoby wpłynąć na lepszą realizację działania X.

5. Ocena systemu monitorowania: Uzasadnienie: Chcemy upewnić się, że nasz system monitorowania dostarcza odpowiednich i aktualnych danych, niezbędnych do skutecznego zarządzania działaniem X i całym programem FENG.
6. Komplementarność z innymi działaniami FENG i programami wsparcia: Uzasadnienie: Interesuje nas, jak działanie X wpisuje się w szerszy kontekst wsparcia innowacyjności i B+R w Polsce. Chcemy zrozumieć, czy nie występują nakładania się lub luki w stosunku do innych instrumentów wsparcia.
7. Pierwsze wnioski dotyczące trwałości efektów: Uzasadnienie: Mimo wczesnego etapu, chcemy zbadać perspektywy trwałości efektów projektów po zakończeniu finansowania. To pomoże nam w planowaniu długoterminowego wpływu programu.
8. Ocena obciążeń administracyjnych: Uzasadnienie: Chcemy zrozumieć, czy procedury administracyjne nie stanowią nadmiernego obciążenia dla beneficjentów, szczególnie biorąc pod uwagę specyfikę mikro i małych przedsiębiorstw.

Uwzględnienie tych obszarów w badaniu ewaluacyjnym pozwoli nam jako IZ na kompleksową ocenę początkowego etapu wdrażania działania X, umożliwiając ewentualne korekty i usprawnienia, a tym samym zwiększając skuteczność i efektywność całego programu FENG w kontekście wspierania innowacyjności i B+R wśród mikro i małych przedsiębiorstw w Polsce.

Źródło: Opracowanie własne z wykorzystaniem Claude 3.5 Sonnet.

W zależności od poziomu naszej satysfakcji z odpowiedzi wygenerowanej przez model możemy modyfikować lub rozwijać pracę. Takim rozwinięciem będzie nakazanie gen AI uszczegółowienia, rozwinięcia określonego elementu czy obszaru zainteresowania. Przykładowo: *Zaproponowałeś ocenę obciążeń administracyjnych. Opisz, jakie konkretne elementy tego zagadnienia cię interesują, wyjaśnij, dlaczego to jest ważne.* Pamiętajmy również o tym, aby nie przyjmować bezkrytycznie uzyskanej odpowiedzi. To, że wśród obszarów zainteresowań wygenerowanych przez gen AI nie występują te, których my byśmy oczekiwali od interesariusza, nie oznacza, że nasze myślenie jest błędne. Może być tak, że to gen AI z nieznanych nam przyczyn nie uwzględniła ich w swojej pracy. Dlatego możemy ją o to dopytać, np. *Myślałem, że w przedstawionych przez ciebie propozycjach znajdzie się również X i Y. Proszę, oceń, czy X i Y nie powinny stanowić zainteresowania interesariusza. Uzasadnij swoją odpowiedź.*

Analiza przedstawionych powyżej przykładów pozwala zobaczyć, że każdorazowo modele pracowały, wykorzystując przedstawiony przez użytkownika kontekst, który jednak różnił się pod względem zakresu udostępnianych modelowi informacji. GPT uzyskał dostęp do raportu ewaluacyjnego, Perplexity[5] działała w oparciu o bardzo uproszczony opis, ale trafnie kierujący ją w stronę problematyki innowacyjności w programach UE, z kolei Claude uzyskał podstawowe informacje poprzez wprowadzanie opisu do okna tekstowego.

Poprawne wprowadzenie kontekstu ma kluczowe znacznie dla wyniku pracy modelu. Dotyczy to całości pracy nad OPZ (i innymi dokumentami). Pamiętajmy, że model może mieć stosunkowo mało informacji na temat ewaluacji. Dlatego pracując z gen AI nad taką tematyką, szczególnie ważne jest przekazanie mu wiedzy zawierającej specyfikę obszaru prowadzonej analizy. **Modele, co do zasady, potrafią pracować nad różnorodnymi tematami, ale wymagają dokładnego ukierunkowania, zwłaszcza jeżeli wykonywane zadanie dotyczy mniej popularnych dziedzin. Konsekwencją naszego zaniedbania w tym zakresie może być ryzyko wystąpienia zbyt ogólnych odpowiedzi lub takich, które bardziej odpowiadają innym dziedzinom, np. badaniom marketingowym i społecznym, a nie ewaluacji interwencji publicznej.**

Model Claude 3.5 Sonnet, wcielając się w rolę przedstawiciela IZ, wymienił osiem obszarów zainteresowania. Jest to duża liczba i uwzględnienie ich wszystkich może być niewykonalne. Gen AI pomoże w dokonaniu selekcji, wybrania takich opcji, które mogą być najbardziej istotne w danym momencie. Na potrzeby rozdziału postanowiono kontynuować pracę z obszarami potrzeb zgłoszonymi przez „przedstawiciela IZ". Poniżej zaprezentowano wyniki dla dwóch opcji: 1. reakcja na informację, że po prostu nie możemy w badaniu uwzględnić niektórych obszarów, 2. reakcja na informację, że samo badanie ma również określony cel (wyniki badania będą omawiane z przedstawicielami instytucji odpowiadających za inne działania).

[5] W trakcie przygotowań do realizacji badania, aby określić potrzeby badawcze, cele badania itd., możemy również korzystać z innych narzędzi, podobnych w swej funkcjonalności do Perplexity. Przykładowo może być to Elicit lub Consensus. Aplikacje te zostały stworzone z myślą o przeszukiwaniu baz naukowych, choć w ich przypadku kontekst krajowy (PL) będzie uwzględniony w niewielkim stopniu, a przynajmniej tak to wygląda z doświadczenia autora.

Rycina 5.4. Fragment odpowiedzi będącej reakcją na instrukcję, że część obszarów nie może zostać uwzględniona

Rozumiem potrzebę hierarchizacji obszarów badawczych. Jako przedstawiciel IZ dokonam priorytetyzacji wcześniej wymienionych obszarów, dzieląc je na trzy kategorie: bardzo ważne, ważne i mniej ważne. Oto moja propozycja wraz z uzasadnieniem:

Bardzo ważne:

1. Efektywność procesu wyboru projektów
2. Wstępna ocena wpływu na innowacyjność i B+ R
3. Adekwatność wsparcia do potrzeb beneficjentów

Uzasadnienie: Te trzy obszary są kluczowe dla oceny skuteczności działania X w realizacji jego głównych celów. Efektywny proces wyboru projektów jest fundamentem sukcesu całego działania. Wstępna ocena wpływu, nawet na tym wczesnym etapie, pozwoli nam szybko zidentyfikować potencjalne problemy i dostosować program. Adekwatność wsparcia jest niezbędna, aby zapewnić, że działanie X faktycznie odpowiada na realne potrzeby mikro i małych przedsiębiorstw w zakresie innowacji i B+ R.

Źródło: Opracowanie własne z wykorzystaniem Claude 3.5 Sonnet.

Rycina 5.5. Fragment odpowiedzi będącej reakcją na instrukcję, że wyniki badania będą omawiane z przedstawicielami instytucji odpowiadających za inne działania

Rozumiem, że kontekst prezentacji wyników badania przed przedstawicielami instytucji odpowiadających za inne działania zmienia nieco priorytety. W takim razie proponuję następującą hierarchizację:

Bardzo ważne:

1. Wstępna ocena wpływu na innowacyjność i B+ R
2. Komplementarność z innymi działaniami FENG i programami wsparcia
3. Adekwatność wsparcia do potrzeb beneficjentów

Uzasadnienie: Te obszary są kluczowe w kontekście spotkania z przedstawicielami innych działań. Wstępna ocena wpływu pozwoli pokazać pierwsze efekty naszego działania. Komplementarność jest istotna dla zrozumienia, jak nasze działanie wpisuje się w szerszy obraz wsparcia innowacyjności. Adekwatność wsparcia pokaże, jak dobrze odpowiadamy na potrzeby beneficjentów w porównaniu z innymi działaniami.

Źródło: Opracowanie własne z wykorzystaniem Claude 3.5 Sonnet.

Model w obu przypadkach dokonał trafnej hierarchizacji. W pierwszej odpowiedzi skupił się na efektywności systemu wyboru projektów, wpływie na B+R oraz dopasowaniu do potrzeb beneficjentów. W drugiej natomiast, mając informację o dodatkowym parametrze, który musi być uwzględniony (omawianie wyników z przedstawicielami innych instytucji), zrezygnował z oceny efektywności systemu i wprowadził w to miejsce komplementarność z działaniami innych instytucji. **Nakazywanie modelowi dokonania selekcji celów, pytań badawczych, technik badawczych itd. jest przydatnym rozwiązaniem. Możemy je stosować zarówno w przypadku treści wygenerowanych przez model, jak i stworzonych przez nas samodzielnie.** Funkcja ta nie tylko sprawi, że przygotowane przez nas założenia badania będą wykonalne (osiągalne), ale również może być przydatna podczas rozmów z naszymi interesariuszami[6], którzy, nie znając metodyki badań, nierzadko starają się zrealizować w ramach jednego badania wiele celów.

W związku z powyższym model został poproszony o analizę problemów badawczych, ocenę ich liczby, ewentualną selekcję oraz przygotowanie uzasadniania dla kierownika instytucji, w której działa JE. Celem zadania było uzyskanie przekonujących argumentów za rezygnacją z niektórych problemów badawczych. Claude wykonał to zadanie. Po pierwsze, przeanalizował listę problemów badawczych, pod drugie, ocenił je pod kątem wykonalności w ramach jednego badania ewaluacyjnego, po trzecie, przeprowadził selekcję oraz – po czwarte – uzasadnił ją w postaci propozycji listu dla dyrektora. Oczywiście to od nas zależy, czy zaakceptujemy ten wynik. Procedurę możemy powtarzać lub już samodzielnie dokonać zmian[7].

Zanim przejdziemy dalej, warto zasygnalizować ważną kwestię – tj. jedno z niebezpieczeństw związanych z korzystaniem z gen AI. Wcześniej postulowano wykorzystanie modelu do dokonania selekcji celów badania itp. Generatywna sztuczna inteligencja względnie dobrze radzie sobie z przygotowywaniem celów, pytań badawczych i potrafi je racjonalnie uzasadnić. Niemniej, może to skutkować coraz częstszym publikowaniem zamówień na usługi ewaluacyjne, w których OPZ zawierają bardzo obszerne listy celów badania lub pytań badawczych. Powinniśmy unikać (nawet jeżeli pytania są dobrze opracowane) wprowadzania do OPZ tego typu zestawu celów/pytań. Może to prowadzić to tego, że nasze projekty będą *de facto* niewykonalne i zniechęcimy wykonawców do ubiegania się o zamówienie. Ewentualnie

[6] Możemy nakazać modelowi gen AI przeprowadzenie hierarchizacji w określonym kontekście, w tym z przygotowaniem uzasadnienia dla interesariuszy odwołującym się np. do metodyki badań społeczno-ekonomicznych.

[7] Pełna odpowiedź.

będziemy przyciągać wykonawców, którzy będą starali się ominąć problem długiej listy pytań, prowadząc badanie niekoniecznie w sposób zgodny z najlepszymi praktykami.

Cele badania i pytania badawcze

Kiedy mamy zidentyfikowane potrzeby badawcze, możemy przystąpić do sformułowania celów badania. Wykorzystano obszary zainteresowań wskazane przez „przedstawiciela IZ". Natomiast w rzeczywistości często obszary te są wynikiem konsultacji z więcej niż jedną grupą interesariuszy. Jeżeli liczba obszarów jest duża, możemy korzystać z gen AI do analizy różnic i podobieństw oraz stworzenia wspólnej listy zagadnień, tak aby ich liczba była możliwie mała, a przynajmniej adekwatna do pojedynczego projektu badawczego. Postępujemy w sposób analogiczny do przedstawionego wcześniej. Wprowadzamy do modelu gen AI listę potrzeb wszystkich grup interesariuszy, a następnie prosimy model, aby dokonał ich weryfikacji, znalazł różnice i podobieństwa. Kolejnym krokiem będzie dokonanie selekcji, ale ważne jest poinstruowanie gen AI co do kryterium lub kryteriów, jakimi powinna się kierować. Możemy również nakazać jej dokonanie modyfikacji obszarów poprzez ich konsolidacje, tak aby możliwie szeroko uwzględniała potrzeby różnych grup. Ostatecznie potrzeby te sprowadzamy do poziomu celów badania, w czym również pomoże nam gen AI.

Poproszono gen AI o przygotowanie celów badania ewaluacyjnego. Podstawę stanowiły obszary badawcze wygenerowane przez model postawiony w roli „przedstawiciela IZ". Uzyskany produkt nie był zadowalający. Model jedynie odtworzył obszary badania, nieznacznie je modyfikując[8].

Ze względu na to, że autorowi zależało na ograniczeniu liczby celów, została zmieniona treść prompta. Po pierwsze, przekazano informację, że zależy tu na konsolidacji (grupowaniu) oraz że działanie ma prowadzić do wyodrębnienia nie więcej niż czterech celów. Tym razem wynik był wyraźnie lepszy[9], choć kwestią dyskusyjną jest to, czy mógłby on zostać zaakceptowany. Wydaje się, że wymaga on dalszego dopracowania, zwłaszcza że wyraźnie widać problemy pojęciowe (logiczne). Zwróćmy uwagę, że w punkcie pierwszym model zaproponował ocenę skuteczności i adekwatności, a jednocześnie w jednym z podpunktów odwołuje się do efektywności itp. Zdecydowano się na dalszą pracę z modelem. Została mu

8 Wynik tego działania.

9 Wynik tego działania.

przekazana instrukcja zawierająca przykłady niespójności jego propozycji[10] i na tej podstawie miał dokonać zmian, co też uczynił[11]. W rezultacie uzyskaliśmy dużo lepszy wynik[12]. Również na tym etapie nie jest to wersja, która bezpośrednio nadaje się do uwzględniania w naszym OPZ. Aczkolwiek stanowi bazę dobrej jakości, którą możemy łatwo poprawić, zmodyfikować adekwatnie do naszych potrzeb.

Przyjmijmy jednak, że zaakceptowaliśmy propozycję wygenerowaną przez model i chcemy przejść do dalszego kroku – przygotowania pytań badawczych. Co do zasady, najbardziej efektywnym podejściem jest tzw. iteracja[13]. Natomiast musimy się zdecydować na to, czy model zacznie pracę od pytań przez nas wytworzonych i następnie będzie dokonywał ich analizy, czy pójdziemy w innym kierunku – model tworzy wstępną listę pytań, a następnie my je oceniamy i przekazujemy instrukcje co do kierunku zmian. Również mamy do wyboru pracę, przy asyście gen AI, wykorzystującą zestaw celów badania – model od razu przygotowuje listę pytań do wszystkich celów, o ile nie zdecydujemy się pracować z nim osobno nad każdym celem i z odpowiednimi do niego pytaniami badawczymi. Pierwsze rozwiązanie może generować niezbyt zadowalające odpowiedzi w momencie, gdy mamy dużo celów badawczych. Drugie z kolei jest czasochłonne. W trakcie pracy nad pytaniami badawczymi możemy stosować prostą instrukcję lub rozbudowaną, która będzie zawierać np. liczbę oczekiwanych przez nas pytań badawczych („nie więcej niż"), informację o konieczności bezpośredniego odwoływania się pytań badawczych do celów badania, rodzaju odpowiedzi (np. pytania umożliwiające uzyskanie mierzalnych odpowiedzi), przygotowaniu „podpytań" do każdego pytania, uwzględnieniu w części pytań możliwości zastosowania metod ilościowych lub jakościowych itp.

Ponadto, zgodnie z tym co podkreślano wcześniej, należy wprowadzić ograniczenie liczby pytań badawczych lub na kolejnych etapach pracy przeprowadzić ocenę znaczenia pytań

[10] Pokazywanie modelowi jego błędów, poprawienie treści wygenerowanych przez model jest działaniem, które zwiększa jego wiedzę i pozwala na zmniejszanie ryzyka powtórzeń tego rodzaju zdarzeń na kolejnych etapach pracy. Jednakże dotyczy to tylko działania wyłącznie w ramach jednego wątku/projektu (np. model Claude, model GPT).

[11] Wynik tego działania.

[12] Co prawda, niekoniecznie jest on idealny, np. zawiera niejasność zawarta w propozycji „badania wpływu procesu wyboru projektów na skuteczność działania". Z pewnością wymaga to przeformułowania w celu uzyskania większej precyzji przekazu. Nie jest jasne, dlaczego w pkt. 4. model zaproponował opracowanie rekomendacji, co samo w sobie nie jest złe, ale jednocześnie jest niespójne z pozostałymi punktami.

[13] Możemy sukcesywnie poprawiać pytania badawcze, wprowadzać nowe spostrzeżenia, co sprawia, że model „uczy się". Jednocześnie zmniejszamy ryzyko pominięcie ważnych kwestii.

dla celów badania. Możemy również wymagać od modelu przedstawienia odpowiedzi w formie tabeli, której kolumny odpowiadają np. za obszar badania, cel badania, pytania badawcze oraz uzasadnienie pytań w kontekście celów badania. Możemy także oczekiwać (wymagana oczywiście jest odpowiednia instrukcja i funkcjonalność modelu) przygotowania podsumowania wykonanego zadania, np. w postaci schematu, na którym będą widoczne powiązania pomiędzy obszarami, celami i pytaniami badawczymi. Jest to ciekawe rozwiązanie, pomaga w szybszej walidacji propozycji przedstawionej przez gen AI i dyscyplinuje wykonywanie zadania pod względem spójności logicznej i przejrzystości.

Prompt prosty: *Na podstawie podanego celu badania przygotuj listę pytań badawczych do ewaluacji skuteczności wsparcia w ramach działania X.*

Wynik (dostęp: 14.11.2024).

Rycina 5.6. Fragment odpowiedzi będącej reakcją na powyżej przedstawiony prompt złożony

Lista pytań badawczych do ewaluacji skuteczności wsparcia w ramach działania X

1. W jakim stopniu zostały zrealizowane założone cele działania X w zakresie zwiększenia innowacyjności wśród beneficjentów?
   - Jak zmieniła się liczba wprowadzonych innowacji przez beneficjentów po otrzymaniu wsparcia?
   - Jaki jest charakter wprowadzonych innowacji (produktowe, procesowe, organizacyjne, marketingowe)?
2. W jakim stopniu działanie X przyczyniło się do intensyfikacji prac B+R wśród beneficjentów?
   - Jak zmieniły się nakłady na B+R wśród beneficjentów po otrzymaniu wsparcia?
   - Czy wzrosła liczba pracowników zaangażowanych w prace B+ R?
   - Czy beneficjenci nawiązali nowe współprace w zakresie B+R z innymi podmiotami (np. uczelniami, instytutami badawczymi)?
3. Jakie czynniki miały największy wpływ na skuteczność działania X?
   - Które elementy wsparcia beneficjenci oceniają jako najbardziej przydatne?
   - Jakie wewnętrzne czynniki w firmach beneficjentów sprzyjały skutecznemu wykorzystaniu wsparcia?

Źródło: Opracowanie własne z wykorzystaniem Claude 3.5 Sonnet.

Prompt złożony: *Na podstawie podanego celu badania przygotuj listę pytań badawczych do ewaluacji skuteczności wsparcia w ramach działania X, kierując się następującymi wytycznymi:*

- *Sformułuj 8–10 pytań badawczych, upewniając się, że każdy aspekt celu jest pokryty co najmniej jednym pytaniem.*
- *Rozpocznij każde pytanie od odpowiedniego słowa pytającego (np. „W jakim stopniu…”, „Jakie…”, „Jak…”, „Czy…”).*
- *Uwzględnij zarówno pytania dotyczące wyników ilościowych (np. stopień realizacji celów), jak i aspektów jakościowych (np. czynniki wpływające na skuteczność).*

Wynik (dostęp: 14.11.2024).

Rycina 5.7. Fragment odpowiedzi będącej reakcją na powyżej przedstawiony prompt złożony

Wyjaśnienie zgodności z wytycznymi
Lista pytań badawczych została opracowana zgodnie z podanymi wytycznymi:

1. **Liczba pytań**: Lista zawiera 10 pytań, co mieści się w zalecanym przedziale 8–10 pytań.
2. **Rozpoczęcie pytań**: Każde pytanie rozpoczyna się od odpowiedniego słowa pytającego:
   - „W jakim stopniu…” (pytania 1, 8)
   - „Jak…” (pytania 2, 7)
   - „Które…” (pytanie 3)
   - „Jakie…” (pytania 4, 9)
   - „W jaki sposób…” (pytanie 5)
   - „Czy…” (pytania 6, 10)
3. **Aspekty ilościowe i jakościowe**: Lista obejmuje zarówno pytania dotyczące wyników ilościowych (np. pytania 1, 2, 6), jak i aspektów jakościowych (np. pytania 3, 4, 7).
4. **Pokrycie aspektów celu badania**:
   - Analiza stopnia realizacji celów: pytania 1, 2, 6
   - Zwiększenie innowacyjności i intensyfikacja prac B + R: pytania 1, 2, 7, 8

- Identyfikacja czynników wpływających na skuteczność: pytania 3, 4
- Badanie wpływu procesu wyboru projektów: pytanie 5

Dodatkowo uwzględniono pytania o trwałość efektów (8), niezamierzone skutki (9) i wpływ na współpracę (10), co poszerza zakres ewaluacji i zapewnia kompleksowe podejście do oceny skuteczności wsparcia.

Źródło: Opracowanie własne z wykorzystaniem Claude 3.5 Sonnet.

Po analizie wyżej zamieszczonych wyników możemy stwierdzić, że w obu przypadkach model wygenerował odpowiedź, którą można uznać za solidną podstawę do dalszej pracy. Część pytań wydaje się być na tyle poprawna, że mogłyby zostać wprowadzone do OPZ, choć drugi prompt sprawdził się lepiej. Pytania są bardziej zwięzłe i mają jasno określony aspekt. Ponadto, zgodnie z instrukcją zawartą w prompcie, są one łatwiejsze w przełożeniu na narzędzia badawcze, a propozycja przedstawiona przez model jest skoncentrowana na kluczowych zagadnieniach/priorytetach, co powinno ułatwić realizację badania.

Konceptualizacja i operacjonalizacja

Zanim przejdziemy dalej, chcemy wyjaśnić, że w niniejszym rozdziale pojęcie „konceptualizacja" ograniczyliśmy do budowania ram definicyjnych. Natomiast nie zajęliśmy się w nim wykorzystaniem gen AI do doboru teorii czy konstruowaniu ogólnych ram analitycznych.

Gen AI pomoże nam również w przygotowaniu konceptualizacji i operacjonalizacji. Proces ten jest prosty i powinien przebiegać analogicznie jak wcześniej opisane działania dotyczące obszarów, celów i pytań badawczych. Możemy wprowadzić instrukcję, w której opiszemy szczegółowo, jak ma przebiegać praca gen AI, na których elementach nam najbardziej zależy itp., a później zmodyfikować wyniki pracy modelu. Należy pamiętać, że powinniśmy pracować z modelem, który wcześniej został przez nas „nakarmiony" wiedzą na temat przedmiotu badania oraz jego celów, pytań badawczych. Bez tego model będzie „zgadywał" i wytwarzał treść niskiej jakości.

Poniżej przedstawiono wyniki pracy modelu w tym zakresie. Celowo zastosowano prompt prosty: *Wkleję ci listę pytań badawczych. Na jej podstawie przygotuj konceptualizację.*

Rycina 5.8. Fragment odpowiedzi w reakcji na instrukcję dotyczącą przygotowania konceptualizacji

Na podstawie podanej listy pytań badawczych mogę zaproponować konceptualizację dla ewaluacji skuteczności wsparcia w ramach działania X. Poniżej znajduje się analiza kluczowych pojęć, zmiennych oraz sposób ich operacjonalizacji.

1. Cele działania X
   - Definicja: Cele działania X to założone rezultaty, które program wsparcia miał osiągnąć, w szczególności w zakresie zwiększenia innowacyjności.
   - Zmienne: Stopień realizacji założonych celów.
   - Operacjonalizacja: Ocena stopnia realizacji celów za pomocą wskaźników takich jak liczba nowych innowacji, wzrost nakładów na B+R, liczba nawiązanych współprac itp.
2. Innowacyjność
   - Definicja: Wprowadzanie nowych lub znacząco ulepszonych produktów, procesów, metod organizacyjnych czy marketingowych.
   - Zmienne:
     - Liczba wprowadzonych innowacji.
     - Rodzaj innowacji (produktowe, procesowe, organizacyjne, marketingowe).
     - Zmiana w liczbie innowacji po otrzymaniu wsparcia.
   - Operacjonalizacja:
     - Liczba nowych produktów, procesów itp. wprowadzonych na rynek.
     - Zmiana w liczbie i charakterze innowacji w porównaniu z okresem sprzed otrzymania wsparcia.

Źródło: Opracowanie własne z wykorzystaniem GPT-4o.

Zwraca uwagę, że model nie ograniczył się wyłącznie do zdefiniowania pojęć. Zaproponował nam również zmienne oraz dokonał operacjonalizacji. Wykroczył więc poza swoje zadanie. Przypuszczalnie jest to konsekwencja: 1. danych treningowych modelu lub 2. zachowania w pamięci przez model preferencji użytkownika[14]. Nieco inny wynik uzyskaliśmy, prosząc

[14] W czasie pracy z GPT mam aktywną funkcję personalizacji. Sprawia ona, że model „uczy się" użytkownika. W tym przypadku zapamiętał, że podczas pracy nad konceptualizacją kolejnym moim krokiem jest przygotowanie listy zmiennych oraz operacjonalizacja.

o wykonanie tego samego zadania model Claude[15]. Niemniej, niezależnie od modelu, wynik pracy można uznać za daleki od ideału. Z jednej strony przygotowane definicje są wąskie, z drugiej zaś modele wygenerowały również treść, której od nich nie wymagaliśmy. Stało się tak dlatego, że nasza instrukcja nie była precyzyjna i dała modelowi zbyt dużą autonomię. Zobaczmy, jak zmienił się wynik pracy po wprowadzeniu bardziej skonkretyzowanego prompta: *Konceptualizacja to proces tworzenia pojęć, definiowania obiektów, zjawisk itd. Chcę wyłącznie konceptualizację i nic poza nią. Nie chcę wymiarów analizy, wskaźników oraz relacji*[16]*. Chcę tylko rozbudowaną konceptualizację.* Tym razem odpowiedź obu modeli była wyraźnie lepsza, pojęcia zostały opisane w sposób zwięzły, ale jednocześnie trudno mówić o nadmiernym ich zawężeniu. Wyniki znajduje się poniżej:

- Claude (dostęp: 14.11.2024),
- GPT – patrz rycina 5.9.

Rycina 5.9. Fragment odpowiedzi w reakcji na instrukcję dotyczącą przygotowania konceptualizacji

1. Realizacja celów działania X
    - Cel działania X: Obejmuje ogólną ideę i zamierzenia programu wsparcia, które miały na celu zwiększenie innowacyjności wśród beneficjentów. Realizacja celów oznacza osiągnięcie zakładanych rezultatów w postaci zmian w działalności beneficjentów, takich jak wprowadzenie nowych produktów, usług, metod, oraz zwiększenie aktywności w obszarze badań i rozwoju (B+R).
2. Innowacyjność
    - Innowacyjność: Zjawisko wprowadzania nowych lub znacząco ulepszonych produktów, procesów, usług, modeli biznesowych lub sposobów organizacji pracy, które przynoszą wartość dodaną w postaci efektywności, wzrostu zysków, poprawy jakości produktów lub usług. W kontekście działania X innowacyjność odnosi się do zdolności beneficjentów do adaptacji i implementacji nowych rozwiązań, które mogą przynieść przewagę konkurencyjną.

Źródło: Opracowanie własne z wykorzystaniem GPT-4o.

[15] Wynik tego działania.

[16] W przypadku GPT instrukcja nakazywała pominięcie takich elementów jak zmienne i operacjonalizacja.

Uznano, że wygenerowany materiał nadaje się do dalszej pracy, a będzie nią operacjonalizacja. GPT oraz Claude miały ją wykonać, kierując się prostą instrukcją: *Przygotowałeś konceptualizację. Przygotuj w oparciu o nią operacjonalizację* i poradziły sobie z tym zadaniem:

- Claude (dostęp: 14.11.2024),
- GPT – patrz rycina 5.10.

Rycina 5.10. Fragment odpowiedzi w reakcji na instrukcję dotyczącą przygotowania konceptualizacji

1. Realizacja celów działania X
    - Cel działania X:
        - Operacjonalizacja: Ocena realizacji celu poprzez porównanie liczby zaplanowanych i zrealizowanych projektów innowacyjnych, analiza raportów końcowych, wskaźników sukcesu (np. liczba wprowadzonych innowacji, wzrost nakładów na B+R) oraz badania ankietowe wśród beneficjentów.
2. Innowacyjność
    - Innowacyjność:
        - Operacjonalizacja:
            - Pomiar liczby innowacji wprowadzonych przez beneficjentów przed i po otrzymaniu wsparcia, np. poprzez analizę dokumentacji projektowej.
            - Przeprowadzenie ankiet lub wywiadów z beneficjentami w celu zbadania, czy wsparcie przyczyniło się do wprowadzenia nowych rozwiązań.
            - Analiza wzrostu liczby patentów, nowych produktów na rynku czy procesów zarejestrowanych po wsparciu.

Źródło: Opracowanie własne z wykorzystaniem GPT-4o.

Można uznać, że co do zasady, etap konceptualizacji i operacjonalizacji został przez gen AI wykonane dobrze, choć również w tym przypadku ich działania były poddawane korekcie – modyfikacji instrukcji.

Metodyka

Poproszono model Claude, na potrzeby niniejszego rozdziału, o rozpisanie zarysu metodyki, która miała się odnosić do wcześniej wygenerowanych pytań badawczych. Uzyskany wynik charakteryzuje się wysokim poziomem trafności i nadaje się do dalszego procedowania[17]. Model zaproponował analizę danych zastanych, badanie ilościowe, indywidualne wywiady pogłębione, badania fokusowe, studia przypadku oraz analizę kontrfaktyczną. Przedstawił dla każdej metody zarys sposobu jej realizacji (np. zaproponował grupy respondentów oraz wielkość prób), powiązał swoją propozycję z pytaniami badawczymi i – co ważne – zadbał o możliwość odpowiedzi na te pytania na bazie wiedzy pochodzącej z więcej niż jednego źródła (z zachowaniem zasady triangulacji). Oczywiście w ramach dalszej pracy można uszczegółowiać wszystkie elementy wygenerowane przez model, np. nakazać mu rozwinięcie pomysłu na analizę kontrfaktyczną itp.

Ciekawym rozwiązaniem jest wykorzystanie modelu do testowania pomysłów dotyczących metodyki[18]. Gen AI pomoże w podjęciu decyzji, np. dokonując krytycznej analizy przedstawionych mu założeń, wskazania ich mocnych i słabych stron, jak również może zaproponować alternatywne rozwiązania. Dotyczy to także pracy z narzędziami badawczymi. Przy czym doświadczenia autora pokazują, że LLM stosunkowo słabo radzi sobie z ich tworzeniem, zwłaszcza gdy praca przechodzi z poziomu ogólnego na szczegółowy. Prawdopodobnie ta krytyczna obserwacja bierze się z tego, że w pracy ewaluatora narzędzia badawcze są zazwyczaj złożone, a ich poprawne przygotowanie wymaga nie tylko wiedzy na temat metodyki tworzenia narzędzi, ale również szczegółowych informacji o przedmiocie badania i niuansach kontekstowych. Takiego narzędzia gen AI jak na razie nie potrafiła przygotować. Natomiast testowana w stosukowo prostym działaniu, polegającym na stworzeniu kwestionariusza do badania satysfakcji użytkownika czata GPT, poradziła sobie dobrze (choć nie ustrzegła się błędów, np. poruszania dwóch kwestii w jednym pytaniu, posługiwania się środkiem skali przy odpowiedziach dychotomicznych, skala niesymetryczna). Fragment ankiety znajduje się w tabeli 5.2.

[17] Wynik tego działania.

[18] Oczywiście w taki sam sposób możemy testować nasze pomysły dotyczące celów badania, pytań badawczych itp.

Tabela 5.2. Fragment ankiety wygenerowanej przez GPT-4o

1. Ogólne wrażenia	
Jakie ogólne wrażenie miałeś/miałaś podczas korzystania z agenta GPTs?	Bardzo pozytywne / Pozytywne / Neutralne / Negatywne / Bardzo negatywne
Czy interakcja była dla ciebie intuicyjna i przyjemna?	Zdecydowanie tak / Raczej tak / Neutralnie / Raczej nie / Zdecydowanie nie
Czy rozmawiając z agentem GPTs, czułeś/czułaś, że odpowiedzi są naturalne, czy może sprawiały wrażenie sztuczności?	Bardzo naturalne / Raczej naturalne / Neutralne / Raczej sztuczne / Bardzo sztuczne
2. Jakość odpowiedzi	
Czy odpowiedzi agenta GPTs były trafne i zgodne z twoimi oczekiwaniami?	Zdecydowanie tak / Raczej tak / Neutralnie / Raczej nie / Zdecydowanie nie
Czy odpowiedzi były wystarczająco szczegółowe? Czy może brakowało istotnych informacji?	Wystarczająco szczegółowe / Raczej wystarczająco / Neutralnie / Brakowało szczegółów

Źródło: Opracowanie własne z wykorzystaniem GPT-4o.

Gen AI może być również wartościowym wsparciem na etapie optymalizacji narzędzi badawczych i metodyki. Poniżej znajdują się dwa przykłady. Pierwszy pokazuje jej wsparcie przy rozwiązaniu problemu doboru techniki badania, drugi zaś przedstawia pracę z narzędziem badawczym.

1. Problem doboru próby[19],
2. Narzędzie badawcze[20].

Zatrzymajmy się przez chwilę na narzędziu badawczym. Model, co prawda, poprawnie wykonał zadanie, ale jednocześnie wygenerowana odpowiedź jest ogólna i nie daje nam

[19] Zastosowany prompt: *Populację generalną stanowią beneficjenci działania X. Są to przedsiębiorcy z sektora MSP. Ich liczba wynosi 3000. Planuję przeprowadzić ankietowanie CAWI i zbadać co najmniej 500 firm. Chcę to zrobić, wysyłając email do wszystkich beneficjentów. Czy moja propozycja jest właściwa metodologicznie? Jakie są jej słabe i mocne strony? Czy mogę zastosować alternatywne podejście, które jest lepsze metodologicznie od mojej propozycji?*

[20] Zastosowano prompt: *Sprawdź, czy załączone narzędzie badawcze (scenariusz IDI): 1) jest zgodny z celami badania, 2) ma spójny logicznie układ bloków pytań, 3) pytania zawarte w poszczególnych blokach pytań pasują do tych bloków, 4) pytania zawarte w blokach pytań wyczerpują problematykę bloku, 4) pytania są sformułowane w sposób prosty, 5) pytanie są sformułowane w sposób poprawny metodologicznie. (od nowej linii). Zależy mi na tym, aby twoja odpowiedź była oparta na wiedzy merytorycznej (…) Jeżeli nie jesteś pewien jakiejś kwestii, wówczas napisz, że nie możesz tego ocenić.*

pewności, czy rzeczywiście zostały zidentyfikowane np. wszystkie niedoskonałości. Dlatego kolejnym krokiem powinno być przejście na poziom szczegółowy. Autor postanowił sprawdzić, czy problem zidentyfikowany w pytaniach P2 i P6 nie występuje w innych częściach scenariusza. Okazało się, że działanie było słuszne, wspomniane pytania były przykładem i modyfikację można wprowadzić w innych częściach scenariusza IDI[21]. Należy zaznaczyć, że praca z narzędziem badawczym odbywała się w ramach specjalnie utworzonego projektu[22] przeznaczonego do działań JE PARP w ramach ewaluacji systemu wyboru projektów FENG.

Kryteria oceny ofert

Przygotowanie kryteriów oceny ofert bywa jednym z trudniejszych etapów (być może najtrudniejszym) pracy nad dokumentacją zamówienia publicznego. Należy przygotowywać system oceny w taki sposób, aby preferował oferty wysokiej jakości merytorycznej w odniesieniu do przedmiotu zamówienia. Z drugiej strony kryteria nie mogą nakładać na wykonawcę zbyt dużych obciążeń podczas pracy nad ofertą oraz muszą być zgodne z uPzp.

Gen AI pomaga w tym zadaniu, ale wymaga to wcześniejszego „nakarmienia" jej nie tylko informacjami o badaniu, ale również przekazania informacji na temat uPzp. Warto także wskazać jej przykłady stosowanych przez nas kryteriów. Ponadto istotne jest wprowadzenie treści ukierunkowującej działanie sztucznej inteligencji, tzn. kierunku ich tworzenia, określenia parametrów kryteriów, na jakich nam zależy. W przeciwnym razie dostaniemy produkt daleki od oczekiwań.

[21] Wynik tego działania.

[22] Modele Claude i GPT umożliwiają tworzenie projektów. To stworzona przez użytkownika przestrzeń pracy (*workspace*) z jej własną historią oraz bazą wiedzy. W ramach projektów użytkownik może wprowadzić swoje oczekiwania względem modelu (np. *będziesz ewaluatorem zajmującym się…*), dokumenty lub inne pliki. Jednocześnie GPT umożliwia stworzenie własnego „agenta", któremu również możemy przekazać instrukcję działania oraz dokumenty i inne pliki. Możemy np. przygotować „agenta", którego celem będzie wspomaganie użytkownika przy tworzeniu OPZ. Powinniśmy dać mu instrukcję (czego od niego oczekujemy, w jaki sposób ma pracować itp.) oraz w zależności od naszych potrzeb dokumenty, które będą tworzyć bazę wiedzy „agenta".

Rycina 5.11. Fragment odpowiedzi w reakcji na instrukcję dotyczącą przygotowania kryteriów oceny[23]

1. Cena (C): Kryterium ceny jest standardowym i nieodzownym elementem oceny. Powinno stanowić np. 40% całkowitej oceny. Oferta z najniższą ceną otrzymuje maksymalną liczbę punktów, a pozostałe proporcjonalnie mniej.

2. Jakość realizacji usługi (J): Ocena jakości powinna być oparta na doświadczeniu wykonawcy, które odnosi się do realizacji podobnych zamówień. Weryfikowane mogą być takie aspekty jak liczba zrealizowanych projektów badawczych, stopień skomplikowania projektów oraz oceny otrzymane od poprzednich zamawiających. Proponowane waga: 30%.

3. Termin realizacji {T): Kryterium to pozwala na premiowanie ofert zakładających krótszy czas realizacji zamówienia. Waga: 10%.

4. Zespół ekspercki (Z): Ocena powinna być dokonana na podstawie doświadczenia i kompetencji osób wchodzących w skład zespołu badawczego. W tym kontekście warto uwzględnić liczbę ekspertów oraz ich doświadczenie w prowadzeniu badań ewaluacyjnych. Waga: 10%.

5. Innowacyjność metodologii (I): Kryterium to dotyczy oceny proponowanej metodologii badawczej. Oferty mogą być oceniane pod kątem zastosowania nowych, innowacyjnych metod, które zwiększą efektywność badania i wpłyną na jakość wyników. Waga: 10%.

Źródło: Opracowanie własne z wykorzystaniem GPT-4o.

Po poprawieniu instrukcji, dodaniu informacji o tym, czego w kryteriach nie może być oraz podkreśleniu, że zależy nam na ocenie umiejętności Wykonawcy, wynik był lepszy, ale nadal dalece niedoskonały. Różnicą było jednak to, że otrzymaliśmy odpowiedź, w której część elementów nadaje się do wykorzystania w dalszych pracach.

[23] Model uzyskał dostęp do OPZ oraz opis art. 240 uPzp i zadanie *Przygotuj propozycję kryteriów*.

Rycina 5.12. Fragment odpowiedzi w reakcji na instrukcję dotyczącą przygotowania kryteriów oceny

1. **Metodologia badania** (30%)
   - **Opis:** Ocenie będzie podlegać spójność i trafność zaproponowanej metodologii badania w odniesieniu do celów i zakresu badania. Metodologia musi umożliwiać szczegółową analizę systemu wyboru projektów, z uwzględnieniem jakościowych i ilościowych aspektów badawczych oraz triangulacji danych.
   - **Sposób oceny:** Punkty będą przyznawane na podstawie oceny szczegółowości, jasności oraz adekwatności zaproponowanych metod badawczych. Metodologia powinna uwzględniać analizę danych zastanych, badania jakościowe, badania ilościowe (CAWI) oraz warsztaty rekomendacyjne.
2. **Zdolność do osiągnięcia celów badania** (25%)
   - **Opis:** Ocenie będzie podlegać zdolność Wykonawcy do przeprowadzenia badania, które przyczyni się do realizacji celów, w tym oceny przygotowania i funkcjonowania systemu wyboru projektów FENG.
   - **Sposób oceny:** Punkty będą przyznawane za adekwatność i kompletność działań proponowanych przez Wykonawcę, w tym za umiejętność identyfikacji czynników wpływających na skuteczność wyboru projektów oraz propozycje usprawnień systemowych.

Źródło: Opracowanie własne z wykorzystaniem GPT-4o.

Autor postanowił pogłębić propozycję dotyczącą metodologii. Choć była ona ogólna, to ograniczenie modelu do „myślenia” wyłącznie w tym kontekście wyraźnie zmieniło efekt pracy. W rezultacie uzyskano produkt, który przybrał formę zbliżoną do oczekiwanej.

Rycina 5.13. Fragment odpowiedzi w reakcji na instrukcję dotyczącą przygotowania kryteriów oceny[24]

1. Kompleksowość i spójność metodologii badania (15%)

Opis: Ocenie podlega szczegółowość i spójność zaproponowanej metodologii badawczej. W tym kryterium kluczowe jest, aby Wykonawca zaproponował metody badawcze, które pozwolą na realizację pełnego zakresu badania, zgodnie z jego celami i zakresem opisanym w OPZ.

Elementy, które będą oceniane:

- Dobór technik badawczych: Ocenie podlega, czy Wykonawca zaproponował odpowiednie techniki badawcze, np. analiza danych zastanych, wywiady pogłębione (IDI), badania ilościowe (CAWI), warsztaty rekomendacyjne oraz inne uzupełniające metody, które wspierają analizę jakościową i ilościową.
- Zgodność metodologii z celami badania: Czy zaproponowane metody badawcze umożliwiają ocenę kluczowych obszarów badania, takich jak trafność, skuteczność, użyteczność oraz efektywność systemu wyboru projektów?
- Jasność opisu procesu badawczego: Ocena, czy Wykonawca w sposób jednoznaczny i zrozumiały opisał proces realizacji badania, co pozwala na łatwe zrozumienie, jak będzie przebiegać analiza poszczególnych obszarów i tematów badawczych.
- Innowacyjność podejścia badawczego: Ocenie podlega, czy Wykonawca zaproponował nowatorskie podejścia do realizacji badania, np. zastosowanie nowych narzędzi analizy danych, sposobów organizacji warsztatów czy formułowania rekomendacji.

Źródło: Opracowanie własne z wykorzystaniem GPT-4o.

Następnym krokiem było ponowne skoncentrowanie się na jednym z zaproponowanych przez gen AI elementów oceny, zaproszenie chata do uszczegółowienia z uwzględnieniem warunków: 1. wykonawca musi wiedzieć, w jaki sposób przygotować opis, 2. wykonawca musi wiedzieć, w jaki sposób ten opis będzie oceniany. Ponownie dostaliśmy wynik, który trudno bezpośrednio wprowadzić od dokumentacji, ale dający dobrą podstawę do wprowadzania przez użytkownika zmian i modyfikacji.

[24] Prompt: *Podoba mi się propozycja oceny metodologii badania. Chciałbym, abyś przygotował na tej podstawie dwa kryteria. Zastanów się, jakie elementy metodologii możemy oceniać.*

Rycina 5.14. Fragment odpowiedzi w reakcji na instrukcję dotyczącą przygotowania kryteriów oceny

1. Opis technik badawczych:

- Wykonawca powinien opisać, jakie techniki badawcze zamierza zastosować w każdym z obszarów badania (np. analiza danych zastanych, badania jakościowe, badania ilościowe, warsztaty rekomendacyjne).
- Dla każdej techniki Wykonawca powinien podać szczegóły, w tym narzędzia, jakie zamierza zastosować, oraz przewidywane korzyści płynące z wyboru danej techniki dla osiągnięcia celów badania.

2. Uzasadnienie doboru technik badawczych:

- Wykonawca musi przedstawić uzasadnienie, dlaczego wybrane techniki są odpowiednie do realizacji celów badania. W szczególności powinien odnieść się do specyfiki badanych obszarów i celów ewaluacji programu FENG.
- W uzasadnieniu Wykonawca powinien wyjaśnić, jak zastosowanie danych technik pozwoli na uzyskanie wiarygodnych, rzetelnych i przydatnych wyników.

B) Sposób oceny ofert – przyznawanie punktów za dobór technik badawczych

Propozycja oceny ofert powinna być obiektywna i precyzyjna, a jednocześnie opierać się na kilku kluczowych aspektach, które pozwolą porównać oferty pod kątem jakości doboru technik badawczych. Poniżej znajduje się propozycja przyznawania punktów:

1. Adekwatność technik do celów badania (0–10 pkt)
   - 0–3 pkt: Techniki badawcze nie są dostosowane do celów badania, lub ich opis jest niekompletny. Brak uzasadnienia wyboru technik lub nieadekwatność wybranych metod do złożoności badania.
   - 4–6 pkt: Techniki badawcze są częściowo adekwatne, jednak niektóre z proponowanych metod mogą nie pozwalać na pełne osiągnięcie założonych celów badania. Uzasadnienie doboru metod jest ogólnikowe.
   - 7–10 pkt: Dobór technik badawczych jest w pełni adekwatny do założonych celów badania.

Każda technika jest dobrze uzasadniona i odpowiada specyfice obszaru badawczego oraz wymagań zamówienia.

Źródło: Opracowanie własne z wykorzystaniem GPT-4o.

Autor uważa, że dalsza praca nad kryteriami powinna być prowadzona przez człowieka. Co prawda, przypuszczalnie istnieje możliwość ich doskonalenia z wykorzystaniem gen AI, jednak ten proces byłby czasochłonny i obarczony ryzykiem porażki. Na tym etapie poprawa wyniku końcowego wymaga bowiem wiedzy specjalistycznej oraz doświadczenia, a tego model nie posiada. Natomiast istotny jest fakt, że czas potrzebny do uzyskania wyników widocznych na ryc. 5.14. wyniósł raptem kilka minut. Można więc mówić o wyraźnym wpływie gen AI na produktywność pracy nad tego typu elementami dokumentacji zamówienia.

5.3. Podsumowanie

Generatywna sztuczna inteligencja może być z powodzeniem wykorzystywana do różnych prac prowadzonych przez JE, ale warunkiem powodzenia i uniknięcia frustracji jest dokładne zapoznanie się z ograniczeniami po stronie człowieka i modelu oraz nauczenie się poprawnego formułowania promptów[25]. Kluczowe jest też zrozumienie, że praca z gen AI to proces iteracyjny, wymagający stałej wymiany informacji, udoskonalania otrzymywanych wyników, ich weryfikacji, dokumentowania skutecznych sposobów formułowania promptów, dzielenia się doświadczeniami (w zespole JE) i systematycznego rozwijania kompetencji w zakresie pracy ze sztuczną inteligencją. Należy również pamiętać, że technologia gen AI rozwija się dynamicznie, co oznacza, że sposób jej wykorzystania w pracy JE będzie ewoluował.

Gen AI nie jest jednak cudownym narzędziem, które będzie wykonywało za nas zadania. Po przeanalizowaniu wyników pracy nad poszczególnymi fragmentami OPZ widać, że wartość gen AI nie leży w tworzeniu gotowych produktów, lecz we wspieraniu procesu twórczego. Innymi słowy, obecna generacja gen AI nie zastąpi ewaluatora, lecz poszerzy jego możliwości. Przedstawione w tym rozdziale przykłady jednoznacznie pokazują, że praktycznie ani razu nie uzyskaliśmy produktu, który byłby gotowy do umieszczenia w OPZ. Każdorazowo musieliśmy dokonywać korekty treści wygenerowanej przez modele i dzięki temu zbliżaliśmy się do efektu, który można nazwać względnie satysfakcjonującym.

Z pewnością gen AI jest dobrym narzędziem do inspirowania, przełamywania twórczej niemocy, weryfikacji wytworzonych przez nas treści, w tym proponowania usprawnień.

[25] Więcej na ten temat w rozdziale pt. „Promptowanie, czyli jak komunikować się z modelem językowym”.

Ma tym samym zastosowanie – co zostało wykazane w tym rozdziale – na każdym etapie tworzenia OPZ. Mimo ujawniających się niedoskonałości rezultatów pracy, gen AI zawsze dostarczała cennych informacji.

Na koniec warto zadać następujące pytania dotyczące szerszego zastosowania gen AI w pracy jednostki ewaluacyjnej:

1. **Czy rzeczywiście warto korzystać z gen AI w działaniach JE?** Na podstawie doświadczeń autora, polegających na codziennym wspomaganiu się gen AI w pracy zawodowej, należy stwierdzić, że sztuczna inteligencja przyczynia się do:
   a) Poprawy szybkości wykonywania zadań – nawet jeżeli nie jesteśmy zadowoleni z uzyskanego rezultatu, to mamy szybki dostęp do propozycji, które możemy udoskonalać samodzielnie lub z pomocą gen AI;
   b) Poprawy jakości wykonywanych zadań – nawet jeżeli jesteśmy ekspertami w swojej dziedzinie, to gen AI pomoże nam zidentyfikować ewentualne błędy popełnione przez nas i zaproponować rozwiązania. W tym przypadku cenne jest także to, że gen AI daje nam możliwości zwiększenia różnorodności perspektyw, w tym może zaproponować alternatywne podejścia lub rozwiązania, które mogły nam nie przyjść na myśl, co wzbogaca proces analityczny i twórczy;
   c) Pozyskiwania nowej wiedzy – dzięki gen AI możemy szybko uzyskać odpowiedź na wiele pytań. Oczywiście odpowiedzi należy weryfikować, niemniej taki proces powinien dotyczyć każdego źródła, z jakiego korzystamy. Gen AI nie tyle wprowadziła w tym zakresie coś nowego, co bardziej uczula nas na tym punkcie.
2. **Czy praca z LLM może sprawić, że efekty (produkty) pracy JE będą wyższej jakości?**? Produkty pracy JE będą wyższej jakości, ale wyłącznie w momencie zachowania odpowiedniego standardu pracy. Tym samym efekty będą się różniły w zależności od tego, kto z gen AI będzie korzystał. W przypadku pracowników z niewielkim doświadczeniem należy się spodziewać poprawy jakości pracy, przy jednocześnie relatywnie wysokim niebezpieczeństwie mniejszej wrażliwości na błędy generowane przez sztuczną inteligencję. Ponadto mniej doświadczeni pracownicy będą zmuszeni do większej liczby iteracji, o ile będzie im zależało na otrzymaniu produktu wolnego od błędów. Z kolei w przypadku osób doświadczonych proces iteracji będzie zachodził szybciej, będzie wymagał mniej kroków, chociażby ze względu na fakt łatwiejszej identyfikacji błędów gen AI i wprowadzania bardziej skonkretyzowanych instrukcji.

Kluczem do nabycia umiejętności w obsłudze gen AI jest podejmowanie prób, codzienne eksperymentowanie (także w odniesieniu do różnych, niekiedy wręcz banalnych zadań) oraz cierpliwość. Dlatego nie bójmy się z niej korzystać i nie zrażajmy się niepowodzeniami. Pozytywne efekty powinny być widoczne bardzo szybko.

Bibliografia

- Ustawa Prawo zamówień publicznych, Dz. U. z 2024 r. poz. 1320.
- *Program FENG* (dostęp: 14.11.2024).
- *Specyfikacja Warunków Zamówienia: Przeprowadzenie badania pn. „Ewaluacja systemu (w tym kryteriów) wyboru projektów FENG”* (dostęp: 14.11.2024).
- *Analiza wybranych działań POIR oraz POPW na poziomie sektorowym i makroekonomicznym za pomocą modelu makroekonomicznego* (dostęp: 14.11.2024).
- *Prompt Library* (dostęp: 14.11.2024).
- Olah, C., Jermyn, A. (2024). *Reflections on Qualitative Research* (dostęp: 14.11.2024).

6. Analizy jakościowe wspierane gen AI

Igor Lyubashenko

6.1. Wprowadzenie

Zanim zagłębimy się w praktyczne zastosowania generatywnej sztucznej inteligencji (gen AI) w badaniach jakościowych, warto przypomnieć podstawy funkcjonowania dużych modeli językowych (LLM), które będą odgrywać kluczową rolę w omawianych poniżej procesach. Jak wspomniano w rozdziale 1., modele językowe nie mają wiedzy w tradycyjnym rozumieniu tego słowa. Zamiast tego wykorzystują skomplikowany system wzorców i statystycznych zależności pomiędzy koncepcjami i pojęciami, z którymi „zapoznały się" w trakcie procesu trenowania.

Ta charakterystyka modeli językowych przekłada się na szereg „umiejętności", które mogą znaleźć praktyczne zastosowanie w analizach jakościowych. Wyróżnić możemy dwie kluczowe „zdolności":

1. Zdolność analizowania tekstów poprzez wyszukiwanie semantyczne, wykraczające poza proste wyszukiwanie oparte na słowach kluczowych. Modele językowe zdolne są do wyszukania interesujących nas fragmentów w oparciu o ich znaczenie.
2. Zdolność do kategoryzowania tekstów na podstawie ich semantyki, co umożliwia zaawansowane i precyzyjne grupowanie informacji.

Te „umiejętności" modeli językowych otwierają drogę do ich zastosowania w analizie danych jakościowych, co wiąże się zarówno z nowymi możliwościami, jak i wyzwaniami dla badaczy zajmujących się ewaluacją polityk publicznych. W kolejnych podrozdziałach szczegółowo omówiono zastosowania gen AI w analizach jakościowych, prezentując konkretne metody i przykłady, które mogą pomóc w efektywnym wykorzystaniu tej innowacyjnej technologii w pracy ewaluatorów.

Na początku konieczne jest podkreślenie dwóch istotnych kwestii. Po pierwsze, warto przytoczyć trafną obserwację Ethana Mollicka (2024), który stwierdza, że każde nasze doświadczenie z AI to kontakt z „najgorszą wersją", z jaką się zetkniemy. To spostrzeżenie celnie oddaje niezwykle dynamiczny charakter rozwoju tej technologii. Modele gen AI są regularnie aktualizowane, a każda nowa wersja przynosi znaczące rozszerzenie ich możliwości.

Po drugie, choć rozumiemy ogólne zasady działania tej technologii, nasze możliwości wyjaśnienia mechaniki (przyczynowo-skutkowo) procesów zachodzących między wprowadzeniem zapytania a uzyskaniem odpowiedzi są ograniczone, co przyznają sami twórcy zaawansowanych modeli językowych. Paradoksalnie, głównym źródłem wiedzy o rzeczywistych możliwościach modeli gen AI staje się eksperymentowanie i próby wykonania z ich pomocą określonych zadań.

Te cechy gen AI mają istotne implikacje dla dalszych rozważań. W przeciwieństwie do tradycyjnego oprogramowania modele gen AI nie mają zamkniętego zestawu funkcji. Co więcej, dostępna literatura naukowa w tym zakresie składa się głównie z opisów różnorodnych eksperymentów z wykorzystaniem gen AI. W rezultacie, przynajmniej w momencie powstawania tego rozdziału, nie istnieje jeszcze żaden „kanoniczny" standard stosowania gen AI w badaniach jakościowych, do którego można byłoby się odwołać, porównywalny do ugruntowanych metod jakościowych czy ilościowych. Zamiast tego do gen AI podchodzić musimy w kategoriach „strategii" prowadzenia dialogu z modelem językowym, aby osiągnąć pożądane rezultaty. Kluczowe staje się umiejętne wykorzystanie tych zdolności gen AI, co do których mamy pewność, pamiętając, że ewentualne ulepszenia mogą jedynie poprawić jakość procesu pracy z danymi jakościowymi.

W kolejnych częściach rozdziału będziemy odwoływać się do najbardziej reprezentatywnych przykładów tekstów opisujących eksperymenty z gen AI. Należy jednak pamiętać o dużej dynamice rozwoju tej dziedziny. W momencie kiedy tekst trafi do rąk Czytelnika, liczba dostępnych przykładów może być znacznie większa.

6.2. Specyfika analizy danych jakościowych

Zanim przedstawione zostaną specyficzne strategie wykorzystania gen AI w analizach jakościowych, warto podkreślić specyfikę tego rodzaju danych. Choć kategoria „dane

jakościowe" jest niezwykle szeroka i zróżnicowana, w niniejszym rozdziale skoncentrowano się wyłącznie na danych tekstowych, pomijając inne formy, takie jak obrazy czy nagrania audio (choć transkrypcje nagrań audio jak najbardziej mieszczą się w kategorii danych tekstowych).

Opierając się na fundamentalnej pracy Gary'ego Goertza i Jamesa Mahoney'a (2006), możemy wyróżnić kilka kluczowych cech, charakteryzujących dane jakościowe. Przede wszystkim dane te często określane są jako „bogate" (*rich*) lub „gęste" (*thick*). Oznacza to, że zawierają one wiele warstw znaczeniowych i cechują się wysokim stopniem złożoności. Ta wielowymiarowość sprawia, że analiza danych jakościowych jest procesem wymagającym i często niejednoznacznym.

Złożoność i wieloznaczność danych jakościowych oznaczają, że ich analiza nieuchronnie wiąże się z koniecznością interpretacji. W przeciwieństwie do danych ilościowych, które można analizować przy użyciu standardowych metod statystycznych, dane jakościowe wymagają od badacza głębszego zrozumienia kontekstu i umiejętności wydobycia ukrytych znaczeń.

Analiza danych jakościowych w swojej istocie jest niczym innym jak systematycznym podejściem do interpretacji i strukturyzacji danych tekstowych, mającym na celu odkrycie znaczeń, wzorców i relacji w badanym materiale, co zazwyczaj jest osiągane za pomocą przypisywania fragmentów badanego materiału do odpowiednich etykiet (por. Kuckartz, 2019; Pope et al., 2006; Puppis, 2019). Proces ten jest sednem tzw. kodowania jakościowego. Johnny Saldaña (2016) podkreśla kilka kluczowych aspektów tego procesu.

1. Kodowanie to aktywna interpretacja danych, a nie tylko ich etykietowanie. Każde przypisanie kodu wiąże się z decyzją analityczną dotyczącą znaczenia danych.
2. Poprzez kodowanie badacze zaczynają dostrzegać konceptualne i teoretyczne powiązania w swoich danych, co pomaga w identyfikacji wzorców i relacji, które mogą nie być od razu oczywiste.
3. Kodowanie zachęca badaczy do głębokiej refleksji nad znaczeniem danych.
4. Kodowanie służy zarówno redukcji danych (poprzez podsumowywanie), jak i ich komplikacji (poprzez rozszerzanie i rekonceptualizację).
5. Sam wybór metod kodowania jest decyzją analityczną, która kształtuje sposób, w jaki badacz wchodzi w interakcję z danymi i je rozumie.

Kodowanie jakościowe materiału badawczego pozwala w kolejnych krokach analizy wyszukać pewne powtarzające się wzory, które można pogrupować w odpowiednie kategorie

w obrębie szerszych tematów, w ten sposób wydobywając z danych określone znaczenia. Należy podkreślić, że w naukach społecznych istnieje cały szereg różnych podejść do kodowania jakościowego, których omówienie wykraczałoby poza zakres tematyczny tego rozdziału. Z punktu widzenia celowości niniejszego opracowania ważne jest rozróżnienie na dwa rodzaje kodowania jakościowego. Pierwszy to **kodowanie indukcyjne**. W tym przypadku badacz wychodzi od „pustej kartki" i koduje materiał „oddolnie", na bieżąco się z nim zapoznając. W tym podejściu niezbędne jest przeprowadzenie co najmniej dwóch tur kodowania – w drugiej następuje uporządkowanie wytworzonych w ten sposób kodów. Takie podejście jest charakterystyczne przede wszystkim dla badań o charakterze eksploracyjnym. Drugi rodzaj kodowania jakościowego to **kodowanie dedukcyjne**. W tym przypadku badacz podchodzi do materiału z gotową listą kodów. Innymi słowy, wie, czego szuka, zaś proces kodowania sprowadza się – w pewnym uproszczeniu – do wyszukania w materiale odpowiednich fragmentów.

Jak już zostało podkreślone we wstępie, co najmniej dwa argumenty sugerują, że gen AI może być skutecznie użyta w tym procesie. Po pierwsze, w obecnym stanie rozwoju modele językowe są niezwykle efektywnym narzędziem do ekstrakcji, organizacji, przetwarzania i podsumowywania danych jakościowych (Hitch, 2023). Po drugie, jak zauważają niektórzy badacze, gen AI może zapewnić swego rodzaju obiektywną walidacje dla procesu kodowania jakościowego, który nieuchronnie ma subiektywny charakter (Schmitt, 2024). Ten ostatni argument jest poniekąd kontrintuicyjny, ze względu na dość powszechne w dyskursie publicznym obawy wynikające z możliwych uprzedzeń zaszytych w danych treningowych gen AI. Chodzi jednak o to, że wspomniane we wstępie wzorce i statystyczne zależności ujęte w modelach mogą dostarczać ocenę bardziej „obiektywną" niż ocena badacza, uwzględniając nową perspektywę i nieoczywiste spostrzeżenia.

6.3. Analiza pojedynczego dokumentu

Zanim przejdziemy do bardziej zaawansowanych form analizy jakościowej z wykorzystaniem gen AI, warto przyjrzeć się podstawowej formie pracy z tą technologią przy analizie pojedynczego dokumentu. Jest to kluczowy etap pozwalający zrozumieć możliwości i ograniczenia tej technologii, a także wypracować strategie jej wykorzystania w bardziej zaawansowanych zastosowaniach. Jest to szczególnie istotne w kontekście ewaluacji polityk publicznych, gdzie często pracujemy z obszernymi dokumentami źródłowymi różnego typu.

Doświadczenia zebrane podczas pracy z pojedynczym dokumentem będą stanowić podstawę do rozwoju bardziej złożonych metod kodowania indukcyjnego i dedukcyjnego, które omówimy w kolejnych sekcjach.

Tradycyjnie praca badacza z jednostkowym dokumentem polega na jego uważnej lekturze, sporządzaniu notatek i wyciąganiu wniosków. Generatywna AI wprowadza nową jakość do tego procesu poprzez możliwość prowadzenia swoistego „dialogu" badacza z tekstem. Zamiast biernego przyswajania treści możemy aktywnie zadawać pytania dotyczące zawartości dokumentu, prosić o wyjaśnienia czy pogłębione analizy konkretnych fragmentów. Model językowy najpierw przetwarza cały dokument, tworząc jego reprezentację w swojej pamięci kontekstowej, co pozwala mu na precyzyjne odnoszenie się do konkretnych fragmentów tekstu i ich wzajemnych powiązań. Co istotne, w przeciwieństwie do sytuacji, gdy model odpowiada na pytania ogólne, bazując na swoim treningu, przy pracy z konkretnym dokumentem ryzyko tzw. halucynacji znacząco spada – model może bezpośrednio odwoływać się do treści dokumentu jako podstawowego źródła informacji. Ta interaktywna forma pracy z dokumentem przypomina rozmowę z bardzo uważnym czytelnikiem, który dokładnie zapoznał się z analizowanym tekstem, lub przypomina wywiad pogłębiony.

W praktyce takie „czatowanie z dokumentem" sprowadza się do prostej sekwencji kroków. W pierwszym kroku przekazujemy modelowi językowemu interesujący nas dokument – może to nastąpić poprzez skopiowanie tekstu bezpośrednio do okna dialogowego odpowiedniego modelu lub – w przypadku niektórych interfejsów – załączenie pliku. Model zapoznaje się wówczas z treścią dokumentu, co pozwala mu odpowiadać na nasze pytania w kontekście przekazanego materiału. Jest to o tyle istotne, że w kolejnych pytaniach nie musimy już powtarzać treści dokumentu – model będzie się odnosił do przekazanego wcześniej tekstu. Warto przy tym pamiętać, że w przypadku bardzo obszernych dokumentów może być konieczne ich podzielenie na mniejsze części, które model powinien analizować osobno[1].

Przyjrzyjmy się, jak może wyglądać proces analizy transkrypcji wywiadu przy wsparciu gen AI. Jako przykład załóżmy, że analizujemy wywiad z przedstawicielem urzędu miasta,

[1] Możliwość analizy dokumentu zależy od tzw. wielkości okna kontekstowego modelu językowego, czyli liczby tokenów (elementów tekstu), które model może przetworzyć w ramach jednej „rozmowy". W momencie przygotowania tego rozdziału dla najpopularniejszych modeli było to od kilkudziesięciu tysięcy do nawet 2 milionów tokenów. Podzielenie dokumentu na części może być niezbędne, jeśli objętość tekstu jest zbliżona do wartości okna kontekstowego modelu, z którym pracujemy.

przeprowadzony w ramach badania reakcji miast Unii Metropolii Polskich na napływ uchodźców wojennych na początku 2022 r.[2].

Analiza takiego materiału może przebiegać dwuetapowo. W pierwszym etapie warto zadać modelowi pytania otwarte, które pozwolą na całościowe spojrzenie na zawartość transkrypcji wywiadu.

- *Jakie główne wyzwania związane z napływem uchodźców wskazuje rozmówca?*
- *Jak rozmówca opisuje proces organizacji pomocy w mieście?*
- *Jakie podmioty/instytucje są wymieniane w kontekście pomocy uchodźcom?*
- *Jak ewoluowało podejście miasta do sytuacji w czasie?*

Ten pierwszy, eksploracyjny etap analizy często prowadzi do odkrycia wątków i aspektów, które mogą nie być oczywiste przy pierwszym czytaniu wywiadu. Model językowy może zwrócić uwagę na powtarzające się tematy lub powiązania między różnymi elementami wypowiedzi rozmówcy. Jest to szczególnie cenne w przypadku wywiadów dotyczących złożonych procesów organizacyjnych, gdzie często różne wątki są ze sobą powiązane i przeplatają się w toku rozmowy.

Po tym wstępnym rozpoznaniu możemy przejść do bardziej szczegółowych pytań, koncentrujących się na konkretnych aspektach badanego zjawiska.

- *Jakie konkretne rozwiązania zostały opracowane i wdrożone przez miasto w kontekście napływu uchodźców?*
- *Jak wygląda współpraca z organizacjami pozarządowymi w kontekście opracowania i wdrażania rozwiązań mających na celu wsparcie uchodźców?*
- *Jakie są konkretne wyzwania, przed którymi stanęło miasto w pierwszych miesiącach kryzysu?*

Takie dwuetapowe podejście do analizy pojedynczego wywiadu pozwala nie tylko na wydobycie kluczowych informacji, ale także na ich systematyczne uporządkowanie. Może ono być szczególnie wartościowe, gdy planujemy analizę większej liczby podobnych dokumentów. Praca z pojedynczym dokumentem służy wtedy jako swoisty pilotaż,

[2] Dane są dostępne w repozytorium ScienceShare Repository.

pozwalający na dopracowanie strategii zadawania pytań i weryfikację jakości otrzymywanych odpowiedzi. Jest to szczególnie pomocne przy opracowywaniu systemu kodów do późniejszej systematycznej analizy jakościowej. Struktura zadawanych pytań może stanowić podstawę do sformułowania precyzyjnych definicji kodów. Pytania ogólne mogą pomóc w identyfikacji głównych kategorii kodowych, podczas gdy pytania szczegółowe pozwalają przetestować różne sposoby definiowania kodów i subkodów oraz sprawdzić, jak model radzi sobie z ich identyfikacją. Można eksperymentować z różnymi poziomami szczegółowości pytań i weryfikować, czy są one wystarczająco precyzyjne i wzajemnie rozłączne. Wracając do przykładu „czatowania" z transkrypcją wywiadu, wynikający z niego system kodów może wyglądać następująco:

- Rozwiązania instytucjonalne: kod obejmuje wypowiedzi dotyczące formalnych struktur i mechanizmów stworzonych przez miasto w odpowiedzi na kryzys uchodźczy. Zawiera odniesienia do powołanych jednostek koordynujących, wprowadzonych procedur, systemów obiegu informacji oraz procesów decyzyjnych. Nie obejmuje działań nieformalnych ani tymczasowych rozwiązań z pierwszych dni kryzysu. Przykłady: utworzenie centrum koordynacyjnego, wdrożenie systemu rejestracji, ustanowienie oficjalnych kanałów komunikacji.
- Współpraca międzysektorowa: kod odnosi się do wypowiedzi opisujących interakcje między administracją miejską a podmiotami spoza struktur samorządowych w kontekście pomocy uchodźcom. Obejmuje zarówno sformalizowaną współpracę (umowy, porozumienia), jak i działania nieformalne. Szczególną uwagę zwraca się na mechanizmy koordynacji, podział zadań oraz przepływ zasobów w ramach współpracy. Przykłady: wspólne projekty z NGO, współdziałanie z wolontariuszami, partnerstwa z biznesem.
- Bariery implementacyjne: kod obejmuje wskazane przez rozmówcę przeszkody i trudności w realizacji działań pomocowych. Uwzględnia zarówno problemy wewnętrzne (organizacyjne, kadrowe, finansowe), jak i zewnętrzne (prawne, systemowe). Nie obejmuje ogólnych refleksji na temat trudności sytuacji, skupia się na konkretnych barierach w realizacji zaplanowanych działań. Przykłady: problemy z interpretacją przepisów, niedobory zasobów, konflikty kompetencyjne.

Oczywiście w praktyce ewaluacyjnej spotykamy się z różnymi typami jakościowych źródeł tekstowych, które pozwalają na wydobycie z nich wartościowych informacji. Transkrypcje wywiadów czy spotkań fokusowych mają zazwyczaj bardziej swobodną strukturę i zawierają treści pozwalające na wychwycenie:

- wyrażanych opinii i ich kontekstu (jak w przedstawionym wyżej przykładzie),
- przykładów i interesujących przypadków,
- propozycji i sugestii zmian,
- emocjonalnego ładunku wypowiedzi.

Innym powszechnie spotykanym w ewaluacji typem jakościowego źródła danych są dokumenty programowe (lub projektowe), stanowiące podstawę interwencji publicznej. Zazwyczaj mają one wystandaryzowaną strukturę, zawierającą diagnozę sytuacji, cele, planowane działania i spodziewane rezultaty. Z takich dokumentów podczas tego typu analizy, omówionej powyżej, można wyodrębniać:

- założenia dotyczące mechanizmów zmiany (teoria zmiany),
- hierarchię celów i powiązania między nimi,
- planowane działania i ich uzasadnienie,
- system wskaźników i ich wartości docelowe,
- zidentyfikowane ryzyka i sposoby ich minimalizacji.

Z kolei źródła w postaci raportów badawczych, które również mają określoną strukturę, pozwalają na wydobycie na potrzeby ewaluacji m.in. informacji o:

- zastosowanej metodologii i jej ograniczeniach,
- kluczowych ustaleniach badawczych,
- dowodach wspierających wnioski,
- rekomendacjach i ich uzasadnieniach,
- zidentyfikowanych dobrych praktykach.

Niezależnie od typu źródła danych (dokument, transkrypcja), sposób formułowania pytań do modelu językowego ma kluczowe znaczenie dla jakości otrzymywanych odpowiedzi. Pytanie ogólne, np. *O czym jest ten wywiad?*, prawdopodobnie przyniesie powierzchowną odpowiedź modelu. Natomiast pytanie ukierunkowane na konkretny aspekt, np. *Jakie konkretne trudności w rozliczaniu dotacji opisuje rozmówca?*, pozwoli badaczowi uzyskać bardziej precyzyjną odpowiedź. Podobnie kontekst zawarty w pytaniu wpływa na interpretację tekstu przez model, np. pytanie: *Biorąc pod uwagę obecną trudną sytuację rynkową, jakie problemy w prowadzeniu działalności wskazuje rozmówca?* może przynieść inną odpowiedź niż pytanie proste: *Jakie problemy w prowadzeniu działalności wskazuje rozmówca?*

Struktura pytania również ma znaczenie. Pytania otwarte sprzyjają eksploracji (*Jak rozmówca opisuje proces założenia firmy?*), podczas gdy pytania zamknięte są lepsze do weryfikacji konkretnych hipotez (*Czy rozmówca wykorzystał całą kwotę dotacji, zgodnie z biznesplanem?*). Wreszcie, użyte słownictwo powinno odpowiadać charakterowi dokumentu. W przypadku dokumentów programowych warto używać języka formalnego z zakresu polityk publicznych, przy raportach badawczych – terminologii metodologicznej, a przy wywiadach – języka bardziej potocznego.

Taka interaktywna praca z dokumentem ma kilka istotnych ograniczeń. Model może nie uchwycić szerszego kontekstu, szczególnie jeśli wymaga on wiedzy spoza dokumentu. Przykładowo, gen AI może nie powiązać wspomnianych przez rozmówcę trudności ze zmianami w przepisach, które zaszły w danym okresie. Istnieje też ryzyko błędnych interpretacji, zwłaszcza gdy rozmówca używa ironii lub gdy wypowiedź jest wieloznaczna. Model może również przeoczyć ważne niuanse emocjonalne – ton wypowiedzi, wahanie czy szczególne podkreślanie pewnych kwestii przez rozmówcę.

Kolejnym wyzwaniem jest tendencja modeli do nadmiernego uogólniania lub „uzupełniania" brakujących informacji na podstawie typowych schematów zarejestrowanych w danych treningowych modelu. Na przykład jeśli rozmówca wspomina o problemach z pozyskaniem klientów, model może „założyć", że chodziło o brak środków na marketing, nawet jeśli podczas wywiadu nie było o tym mowy. W przypadku dokumentów strategicznych może automatycznie przypisywać standardowe cele czy wskaźniki, które nie zostały faktycznie określone, a przy analizie raportów ewaluacyjnych – domniemywać istnienie typowych barier wdrożeniowych nieobecnych w oryginalnym tekście. Dlatego kluczowe jest weryfikowanie przez badawcza odpowiedzi wygenerowanych przez model (w szczególności poprzez odniesienia do oryginalnego tekstu), a także stosowanie poleceń, w których nakazuje modelowi, aby w swojej odpowiedzi ograniczył się wyłącznie do zawartości udostępnionego dokumentu, a jeśli nie znajduje adekwatnej treści do zadanego pytania, jednoznacznie to komunikował (np. źródło nie zawiera treści pozwalających na udzielenie odpowiedzi na pytanie)[3].

Ten sposób wykorzystania gen AI stanowi fundament dla bardziej zaawansowanych form analizy jakościowej, które zostaną omówione w kolejnych podrozdziałach.

[3] Więcej na temat prowadzania dialogu z modelem gen AI można znaleźć w rozdziale pt. „Promptowanie, czyli jak komunikować się z modelem językowym".

6.4. Użycie gen AI w jakościowym kodowaniu indukcyjnym

Opisane w poprzednim podrozdziale podejście do pracy z pojedynczym dokumentem stanowi punkt wyjścia do bardziej systematycznej analizy większej liczby materiałów badawczych. Kodowanie indukcyjne z wykorzystaniem gen AI polega w istocie na rozszerzeniu strategii „czatowania z dokumentem" na cały korpus danych. Zamiast analizy pojedynczej transkrypcji zadajemy te same, starannie przemyślane pytania eksploracyjne do serii dokumentów.

Przykładowo, jeśli dysponujemy serią wywiadów z przedstawicielami różnych miast na temat reakcji na kryzys uchodźczy, możemy systematycznie wydobyć opinie urzędników magistratów – np. o tym, jak ewaluowało podejście miast do zarządzania tą sytuacją – zadając to samo pytanie w odniesieniu do poszczególnych transkrypcji. Przykładowy prompt może brzmieć następująco:

> *Przeanalizuj, jak w wypowiedziach rozmówcy opisana jest zmiana w podejściu miasta do zarządzania sytuacją uchodźczą w czasie. Zwróć szczególną uwagę na:*
> - *rozróżnienie między działaniami początkowymi a późniejszymi;*
> - *momenty przełomowe wymuszające zmianę podejścia;*
> - *zmiany w strukturach organizacyjnych i procesach.*
>
> *Przedstaw chronologicznie opisane zmiany, cytując odpowiednie fragmenty wypowiedzi.*

Zadając takie pytanie do wywiadów w naszym korpusie danych, otrzymujemy serię uporządkowanych chronologicznie zwięzłych opisów zmian w podejściu poszczególnych miast do sytuacji kryzysowej. Taki materiał stanowi bogaty zbiór informacji o procesach adaptacyjnych w różnych kontekstach miejskich. Co istotne, w wypowiedziach rozmówców mogą pojawiać się zarówno podobne wzorce (na przykład przejście od działań *ad hoc* do rozwiązań systemowych), jak i różnice wynikające ze specyfiki lokalnej. Analiza tych odpowiedzi pozwala na indukcyjne wyodrębnienie kluczowych etapów reakcji na kryzys, typowych punktów zwrotnych czy powszechnie napotykanych wyzwań, wymuszających zmianę podejścia. Na tej podstawie możemy zacząć wyciągać wnioski nt. procesów adaptacyjnych samorządów miejskich w sytuacji kryzysowej, które będą zakorzenione w faktycznych doświadczeniach badanych przedstawicieli miast.

Zastosowanie gen AI w tym kontekście wiąże się z tymi samymi wyzwaniami, o których była mowa w poprzednim podrozdziale. Przede wszystkim aby model mógł efektywnie pracować nad materiałem, konieczne jest przyznanie mu znacznego zakresu autonomii (to model „podejmuje decyzję" odnośnie do tego, jaki element analizowanego tekstu odpowiada na nasze pytanie). To z kolei może obniżać efektywność procesu badawczego. Głównym problemem może się stać konieczność dokładnej weryfikacji pracy gen AI przez człowieka. W praktyce oznacza to, że badacz musiałby najpierw samodzielnie zapoznać się z analizowanym materiałem, co absolutnie wnosi wartość dodaną dla ogólnego rozeznania się w poruszanej problematyce (zwłaszcza jeśli badacz sam nie realizował wszystkich wywiadów), ale częściowo niweluje potencjalne korzyści czasowe wynikające z automatyzacji. Co więcej, ewentualne konieczne poprawki, wymuszone błędami wygenerowanymi przez AI dostrzeżonymi przez badacza znającego materiał, mogą docelowo wymagać ponownego, ręcznego zakodowania materiału, co oznaczałoby wykonanie „kroku wstecz" i dodatkowo wydłużenie procesu badawczego.

Warto zauważyć, że w Internecie dostępne są narzędzia umożliwiające automatyczne kodowanie za pomocą prostych skryptów (QualiGPT, 2024; Zhang et al., 2023). Jednakże brakuje publikacji naukowych, które rzetelnie oceniałyby skuteczność i wiarygodność takiego podejścia. To ograniczenie utrudnia ocenę rzeczywistej wartości tych rozwiązań w kontekście wykorzystania w badaniach naukowych czy ewaluacyjnych.

Podsumowując, obecny stan rozwoju technologii gen AI teoretycznie umożliwia przeprowadzenie jakościowego kodowania indukcyjnego. Jednak proces ten obciążony jest znaczną niepewnością, co do kompletności i trafności analizy przeprowadzonej przez model językowy. Istnieje ryzyko, że gen AI może pominąć istotne fragmenty lub błędnie zinterpretować kontekst. Ponadto korzystanie z takiej analizy wymaga akceptacji wyniku pracy gen AI w formie, w jakiej został przedstawiony, co może być problematyczne z punktu widzenia rzetelności badawczej.

W związku z tymi ograniczeniami warto traktować modele językowe raczej jako narzędzia wspomagające proces kodowania indukcyjnego, a nie jako samodzielne rozwiązanie. Gen AI może pełnić rolę asystenta, który:

a) uzupełnia analizę przeprowadzoną przez badacza, co może potwierdzić wstępne wnioski lub zwrócić uwagę na brakujące/pominięte aspekty,
b) pomaga w uporządkowaniu systemu kodowego wypracowanego w trakcie analizy,

c) wspiera badacza w identyfikacji potencjalnych wzorców lub tematów w analizowanym materiale badawczym.

Takie podejście pozwala wykorzystać potencjał gen AI przy jednoczesnym zachowaniu kontroli człowieka nad procesem badawczym i zapewnieniu odpowiedniego poziomu krytycznej analizy. Badacz pozostaje kluczowym elementem procesu, który wykorzystuje gen AI jako narzędzie wspomagające, a nie zastępujące ludzką interpretację i wnioskowanie.

6.5. Użycie gen AI w jakościowym kodowaniu dedukcyjnym

Nieco inaczej wygląda możliwość wykorzystania gen AI w zakresie jakościowego kodowania dedukcyjnego, gdy dysponujemy precyzyjnie określonym systemem kodów (innymi słowy: wiemy, czego szukamy w analizowanym materiale badawczym). W takim przypadku możemy zarysować dwa różne podejścia do zastosowania gen AI, z których każde opiera się na różnych zdolnościach modeli językowych. Pierwsze podejście koncentruje się na identyfikacji i wydobyciu konkretnych fragmentów tekstu odpowiadających zdefiniowanym kodom, z wykorzystaniem zdolności modeli językowych do wyszukiwania semantycznego. Drugie podejście służy do automatycznej kategoryzacji jednostek analizy (np. pojedynczych wypowiedzi czy fragmentów dokumentów) według określonego schematu kodowego, bazuje się przy tym na zdolności modeli do klasyfikacji tekstu. Te dwa podejścia różnią się nie tylko sposobem wykorzystania możliwości modeli językowych, ale przede wszystkim charakterem i szczegółowością uzyskiwanych wyników. Przyjrzyjmy się dokładniej każdemu z nich.

Pierwsze podejście opiera się na zdolności modeli językowych do wyszukiwania semantycznego. Proces ten można podzielić na następujące etapy:

1. Przygotowanie materiału badawczego i przekształcenie go do postaci wektorowej[4].
2. Opracowanie systemu kodów jakościowych z precyzyjnymi definicjami.

[4] Przekształcenie tekstu do postaci wektorowej (nazywane też wektoryzacją lub embeddingiem) polega na reprezentacji słów i fragmentów tekstu w postaci ciągów liczb. Taka matematyczna reprezentacja tekstu umożliwia modelom językowym efektywne przetwarzanie i analizę treści, przy zachowaniu informacji o znaczeniu i kontekście słów. Jest to standardowa procedura przygotowania danych tekstowych do analizy z wykorzystaniem sztucznej inteligencji.

3. Zadawanie modelowi serii pytań opartych na przygotowanym systemie kodów. Celem jest odnalezienie fragmentów tekstu odpowiadających poszczególnym kodom.
4. Zapisanie wyników analizy w dogodnym formacie.

Ta metoda wykorzystuje zdolność modelu gen AI do zrozumienia kontekstu i znaczenia tekstu, co pozwala na wyszukiwanie fragmentów odpowiadających interesującym nas kodom jakościowym.

Dla przykładu, wróćmy do wspomnianych wyżej wywiadów z urzędnikami miejskimi i przykładowego systemu kodów, przedstawionego w podrozdziale poświęconym pracy z pojedynczym dokumentem. Aby przeprowadzić opisaną procedurę, korpus naszych danych należy dodać do używanego przez nas modelu jako wektorową bazę wiedzy. W momencie przygotowania tego rozdziału (wrzesień – listopad 2024) najbardziej zaawansowaną i ogólnodostępną możliwość w tym zakresie proponował ChatGPT (OpenAI) poprzez funkcję tzw. dostosowanych GPT (*custom GPT*). Podobne rozwiązanie było dostępne także w tzw. projektach Claude'a (Anthropic). Ten krok umożliwia podjęcie próby wydobycia interesującej nas wiedzy z całego korpusu danych. Przykładowy prompt:

> Przeanalizuj wywiady i zidentyfikuj wszystkie fragmenty odpowiadające następującemu kodowi jakościowemu:
> "Rozwiązania instytucjonalne: formalne struktury i mechanizmy stworzone przez miasto w odpowiedzi na kryzys uchodźczy. Szukaj odniesień do:
> *a) powołanych jednostek koordynujących,*
> *b) wprowadzonych procedur,*
> *c) systemów obiegu informacji, procesów decyzyjnych."*
> *d) Nie uwzględniaj działań nieformalnych ani tymczasowych rozwiązań z pierwszych dni kryzysu.*
>
> *Dla każdego znalezionego fragmentu podaj:*
> - *dokładny cytat,*
> - *przypisany kod,*
> - *krótkie uzasadnienie.*

Odpowiednie zapytanie należy zadać w odniesieniu do wszystkich kodów znajdujących się w naszym systemie kodowym.

Drugie podejście bazuje na zdolności modeli językowych do kategoryzacji. Proces ten składa się z następujących kroków:

1. Przygotowanie materiału badawczego do analizy, na przykład w postaci zbioru indywidualnych wypowiedzi uczestników programu szkoleniowego czy beneficjentów interwencji publicznej. Może to być jakościowy feedback zebrany w ramach ewaluacji (opinie o szkoleniu, sugestie zmian w programie, opisy doświadczeń z realizacji projektów), odpowiedzi na pytania otwarte z ankiet ewaluacyjnych czy komentarze uczestników zebrane podczas wywiadów grupowych. W takich przypadkach każda pojedyncza wypowiedź stanowi odrębną jednostkę analizy, którą można przypisać do odpowiednich kategorii w ramach naszego systemu kodowego.
2. Przygotowanie systemu kodów jakościowych z precyzyjnymi definicjami.
3. Zadawanie modelowi pytań mających na celu ocenę, czy dany fragment materiału badawczego odpowiada któremuś z wcześniej zdefiniowanych kodów jakościowych.
4. Zapisanie wyników analizy w dogodnym formacie.

Zastosowanie tego procesu do analizy wspomnianych już wywiadów sprowadza się do podziału tekstu na pojedyncze wypowiedzi (dotyczące określonego zagadnienia, np. poruszanego w scenariuszu wywiadu) – co zazwyczaj jest wykonywane przy transkrypcji – a następnie wielokrotnego użycia prompta o strukturze i logice układu podobnej do poniższego:

> *Przeanalizuj poniższy fragment wywiadu i określ, czy odpowiada on któremuś z następujących kodów.*
>
> *[fragment wywiadu]*
>
> *Kody jakościowe:*
> 1. *ROZWIĄZANIA INSTYTUCJONALNE: formalne struktury i mechanizmy stworzone przez miasto (jednostki koordynujące, procedury, systemy obiegu informacji, procesy decyzyjne). Nie obejmuje działań nieformalnych ani tymczasowych z pierwszych dni kryzysu.*
> 2. *WSPÓŁPRACA MIĘDZYSEKTOROWA: interakcje między administracją miejską a podmiotami zewnętrznymi (formalne i nieformalne), mechanizmy koordynacji, podział zadań, przepływ zasobów.*

3. *BARIERY IMPLEMENTACYJNE: konkretne przeszkody w realizacji działań (wewnętrzne: organizacyjne, kadrowe, finansowe; zewnętrzne: prawne, systemowe). Nie obejmuje ogólnych refleksji o trudnościach.*

Zwróć odpowiedź w następującym formacie:

- *Jeśli fragment odpowiada jednemu kodowi: [NAZWA_KODU]*
- *Jeśli fragment odpowiada dwóm kodom: [NAZWA_KODU_1, NAZWA_KODU_2]*
- *Jeśli fragment nie odpowiada żadnemu kodowi: INNE*

Nie dodawaj żadnych dodatkowych wyjaśnień ani uzasadnień.

Przykłady właściwego kodowania:

1. *Fragment: "Powołaliśmy specjalny zespół koordynacyjny, który spotykał się codziennie o 9.00 rano. Wprowadziliśmy też elektroniczny system raportowania dla wszystkich wydziałów". Kod: [ROZWIĄZANIA_INSTYTUCJONALNE]*
2. *Fragment: "Na początku było chaotycznie, wolontariusze działali na własną rękę. Dopiero po dwóch tygodniach udało nam się wypracować system współpracy z NGOsami i przydzielić im konkretne zadania". Kod: [WSPÓŁPRACA_MIĘDZYSEKTOROWA]*
3. *Fragment: "Na początku było chaotycznie, wolontariusze działali na własną rękę. Dopiero po dwóch tygodniach udało nam się wypracować system współpracy z NGOsami i przydzielić im konkretne zadania". Kod: [WSPÓŁPRACA_MIĘDZYSEKTOROWA]*
4. *Fragment: "Największym problemem był brak odpowiednich procedur i niejasne przepisy. Do tego brakowało nam ludzi mówiących po ukraińsku". Kod: [BARIERY_IMPLEMENTACYJNE]*
5. *Fragment: "Stworzyliśmy centrum informacyjne we współpracy z organizacjami pozarządowymi. My zapewniliśmy lokal i sprzęt, oni dostarczyli tłumaczy i wolontariuszy". Kod: [ROZWIĄZANIA_INSTYTUCJONALNE, WSPÓŁPRACA_MIĘDZYSEKTOROWA]*

To podejście daje badaczowi większą kontrolę nad procesem kodowania i umożliwia łatwiejsze przeprowadzenie wyrywkowej kontroli jakości, szczególnie w przypadku kodowania dużych zbiorów danych.

Podstawowa różnica pomiędzy przedstawionymi podejściami użycia gen AI do dedukcyjnego kodowania jakościowego dotyczy stopnia autonomii przyznanej modelowi językowemu.

Podejście 1. (oparte na wyszukiwaniu semantycznym) daje modelowi większą swobodę w wyborze relewantnych fragmentów, co czyni ją jednocześnie bardziej wrażliwą na jakość procesu wektoryzacji materiału badawczego. Z drugiej strony, jest ona bardziej efektywna, jeśli chodzi o wykorzystanie dostępnej mocy obliczeniowej.

Podejście 2. (oparte na kategoryzacji) ma bardziej „mechaniczny" charakter, oferuje badaczowi większą kontrolę nad procesem kodowania. Jest ono już powszechnie stosowana w analizie sentymentu krótkich wypowiedzi, takich jak opinie konsumentów czy wpisy w mediach społecznościowych (Kheiri & Karimi, 2023). Jednak jej zastosowanie może być znacznie szersze i obejmować różnorodne systemy kodów, m.in. takie jak zawarte w powyższym przykładzie.

Warto zauważyć, że każde z wspomnianych podejść służy nieco odmiennym celom analitycznym. Podejście oparte na wyszukiwaniu semantycznym koncentruje się na identyfikacji różnorodnych przypadków i przykładów odpowiadających zdefiniowanym kodom, pozwalając na wychwycenie także pośrednich odniesień i kontekstowych wskazówek. Z kolei podejście oparte na kategoryzacji ma na celu systematyczne przyporządkowanie każdego analizowanego fragmentu tekstu do określonej kategorii w systemie kodowym, skupia się na bezpośrednich i jednoznacznych odniesieniach do zdefiniowanych elementów. Ta różnica w celach przekłada się na zastosowanie obu podejść. Pierwsze może być bardziej użyteczne w sytuacjach, gdy chcemy zidentyfikować i zgromadzić przykłady ilustrujące poszczególne kody. Drugie sprawdza się lepiej, gdy potrzebujemy kompleksowej kategoryzacji całego materiału badawczego według ściśle określonych kryteriów. Świadomość tych różnic jest istotna przy wyborze podejścia odpowiedniego dla naszego badania – pierwsze może być bardziej odpowiednie do eksploracyjnych aspektów badania, natomiast drugie sprawdzi się lepiej w sytuacjach wymagających bardziej rygorystycznego wyszukania określonych wątków w materiale badawczym.

W obu przypadkach kluczowa jest jakość i precyzja definicji kodów jakościowych. Należy stwierdzić, że dobrze zdefiniowane kody znacznie zwiększają prawdopodobieństwo wykonania przez gen AI dokładnej i rzetelnej analizy jakościowej.

Obie strategie wiążą się także z innym ograniczeniem. Wynikiem takiej analizy będzie baza danych zawierająca znalezione w materiale badawczym fragmenty odpowiadające wcześniej zdefiniowanym kodom (czy to w wyniku wyszukiwania semantycznego, czy to w wyniku kategoryzacji). Istnieje zatem ryzyko pominięcia innych, potencjalnie istotnych wątków, które nie zostały uwzględnione w systemie kodów, a które zazwyczaj są odnotowywane w procesie kodowania ręcznego. Problem ten można rozwiązać, uzupełniając kodowanie dedukcyjne opisanym w poprzednim podrozdziale kodowaniem indukcyjnym (przykładowo, prosząc model o podsumowanie głównych wątków tematycznych zawartych we fragmentach skategoryzowanych jako „inne").

Podsumowując, wykorzystanie gen AI w jakościowym kodowaniu dedukcyjnym może wiązać się ze znaczącymi korzyściami w zakresie efektywności i skalowalności analizy. Jednakże jak w przypadku każdego narzędzia badawczego, kluczowe jest świadome jego stosowanie z uwzględnieniem zarówno potencjału, jak i ograniczeń.

6.6. Automatyzacja procesów badawczych z wykorzystaniem gen AI

Przedstawione w niniejszym rozdziale rozważania wyraźnie pokazują możliwości automatyzacji procesów badawczych, polegające na oddelegowaniu modelowi językowemu zadań o charakterze analitycznym. Prawdziwy potencjał automatyzacji ujawnia się jednak dopiero przy wykorzystaniu modeli językowych poprzez interfejs programistyczny aplikacji (API). Otwiera to drogę do przetwarzania dużych ilości danych jakościowych.

Warto podkreślić, że wszystkie omówione w tym rozdziale podejścia i strategie można realizować poprzez bezpośrednią interakcję z modelem językowym w interfejsie konwersacyjnym (np. ChatGPT, Claude). Taki sposób pracy polega na ręcznym wykonywaniu kolejnych kroków analizy – badacz musi każdorazowo wkleić tekst do analizy, sformułować odpowiednie polecenie dla modelu, a następnie skopiować i zapisać otrzymaną odpowiedź. Taka bezpośrednia interakcja z modelem jest szczególnie wartościowa na początkowym etapie analizy jakościowej, gdy badacz dopracowuje sposób formułowania poleceń (promptów) i weryfikuje, czy model rozumie zadanie zgodnie z oczekiwaniami. Jednak przy analizie większej liczby dokumentów ręczne powtarzanie tych samych czynności dla

każdego tekstu czy fragmentu staje się czasochłonne i nieefektywne, przekształcając się w mechaniczne przepisywanie poleceń i kopiowanie odpowiedzi.

Natomiast dostęp do modeli językowych poprzez API umożliwia tworzenie stosunkowo prostych skryptów, które pozwalają na wielokrotne, automatyczne wykonywanie opisanych powyżej sekwencji czynności.

Korzystanie z modeli językowych poprzez API, w przeciwieństwie do interfejsu czatu, umożliwia także znacznie większą kontrolę nad ich zachowaniem. Jest to możliwe dzięki wykorzystaniu tzw. promptów systemowych (*system prompts*), które definiują podstawowe ramy działania modelu. Prompt systemowy można porównać do szczegółowej instrukcji określającej „zasady gry", jest to sposób, w jaki model ma interpretować dane wejściowe i formatować swoje odpowiedzi. Przykładowo, dla omówionego wcześniej podejścia do kodowania opartego na kategoryzacji możemy zdefiniować *system prompt*, który będzie zawierać:

- pełną definicję roli, jaką nadajemy modelowi,
- szczegółowe definicje wszystkich dopuszczalnych kodów,
- precyzyjne określenie formatu odpowiedzi (np. przez zastosowanie struktury JSON),
- wyraźne ograniczenie, że model może używać wyłącznie predefiniowanych kodów.

> *Jesteś asystentem wspierającym proces kodowania jakościowego. Twoim zadaniem jest klasyfikacja fragmentów tekstu zgodnie z poniższym systemem kodów.*
>
> *DOSTĘPNE KODY I ICH DEFINICJE:*
>
> *1. ROZWIĄZANIA_INSTYTUCJONALNE*
> *Definicja: Formalne struktury i mechanizmy stworzone przez miasto w odpowiedzi na kryzys uchodźczy.*
> *Obejmuje: powołane jednostki koordynujące, wprowadzone procedury, systemy obiegu informacji, procesy decyzyjne.*
> *Nie obejmuje: działań nieformalnych ani tymczasowych rozwiązań z pierwszych dni kryzysu.*
>
> *2. WSPÓŁPRACA_MIĘDZYSEKTOROWA*
> *Definicja: Interakcje między administracją miejską a podmiotami zewnętrznymi.*

Obejmuje: formalne i nieformalne współdziałanie, mechanizmy koordynacji, podział zadań, przepływ zasobów.

3. BARIERY_IMPLEMENTACYJNE
Definicja: Konkretne przeszkody w realizacji działań pomocowych.
Obejmuje: problemy wewnętrzne (organizacyjne, kadrowe, finansowe) i zewnętrzne (prawne, systemowe).
Nie obejmuje: ogólnych refleksji o trudnościach sytuacji.

INSTRUKCJE:

1. Analizuj każdy fragment wyłącznie pod kątem powyższych kodów.

2. Zawsze zwracaj odpowiedź w następującym formacie JSON:

{
"kod_główny": [jeden z: "ROZWIĄZANIA_INSTYTUCJONALNE", "WSPÓŁPRACA_MIĘDZYSEKTOROWA", "BARIERY_IMPLEMENTACYJNE"],
"kod_dodatkowy": [jeden z powyższych kodów lub null],
"pewność": [liczba od 0 do 1]
}

3. Jeśli fragment nie odpowiada żadnemu z kodów, użyj:

{
"kod_główny": null,
"kod_dodatkowy": null,
"pewność": 1
}

Taki prompt systemowy defininiuje zachowanie modelu, natomiast *user prompt* (czyli zapytanie, które jako użytkownicy wysyłamy do modelu) będzie wyglądał w takim przypadku bardzo prosto:

Fragment transkrypcji do analizy: [tekst automatycznie dodany przez skrypt]

Istotnym aspektem korzystania z API modeli językowych są koszty, które naliczane są zazwyczaj przez serwis usługodawcy oferującego model za każdy token (podstawową jednostkę tekstu) w prompcie i w odpowiedzi. W praktyce oznacza to, że koszt analizy rośnie wraz z długością analizowanych fragmentów tekstu, ich liczbą, złożonością prompta systemowego oraz długością generowanych odpowiedzi. By zobrazować skalę kosztów, rozważmy przykład analizy stu wywiadów, każdy o długości około dziesięciu tysięcy słów, z wykorzystaniem systemu kodowania zawierającego dziesięć kodów. Przy typowych stawkach rzędu $0,01 za tysiąc tokenów (stan na koniec 2024 r.) koszt pojedynczego przejścia przez taki materiał może wynieść kilkanaście dolarów. W przypadku konieczności wielokrotnej analizy tego samego materiału lub wykorzystania droższych modeli koszty mogą szybko wzrosnąć do znaczących kwot.

W związku z tym efektywna automatyzacja wymaga jednocześnie przemyślanej strategii optymalizacji kosztów. Można to osiągnąć poprzez wstępne przetwarzanie tekstu i usuwanie nieistotnych fragmentów, dzielenie materiału na mniejsze, sensowne jednostki analizy czy wykorzystanie mechanizmów buforowania, by unikać powtórnej analizy tych samych fragmentów[5]. Skutecznym podejściem może być również wykorzystanie mniejszych, tańszych modeli do wstępnej selekcji materiału. Wyselekcjonowane fragmenty mogą być następnie użyte na dwa sposoby: albo do szczegółowej analizy przy pomocy droższych, bardziej zaawansowanych modeli, albo jako zbiór treningowy do dostosowania (*fine-tuning*) wybranego modelu do specyfiki naszego zadania. To ostatnie podejście może być szczególnie korzystne w długoterminowych projektach badawczych, gdzie mamy do przeanalizowania duże ilości podobnych tekstów – jednorazowa inwestycja w dostosowanie modelu może znacząco obniżyć koszty dalszych analiz.

Alternatywnym podejściem do optymalizacji kosztów jest wykorzystanie otwartych (*open-source*) modeli językowych, które można uruchomić na własnej infrastrukturze obliczeniowej. Modele takie jak Llama czy Mistral mogą oferować możliwości porównywalne z modelami

[5] Buforowanie polega na zapisywaniu wyników już przeprowadzonych analiz w sposób umożliwiający ich ponowne wykorzystanie – dzięki temu unikamy sytuacji, w której model musiałby analizować te same fragmenty tekstu wielokrotnie. Na przykład jeśli ten sam cytat pojawia się w kilku dokumentach, jego interpretacja przez model może zostać zapisana i wykorzystana ponownie bez konieczności powtórnego przeprowadzania (i opłacania) tej samej analizy.

komercyjnymi, szczególnie w przypadku wyspecjalizowanych zadań (do takich należy zaliczyć kodowanie jakościowe). Wiąże się to z koniecznością poniesienia początkowych nakładów na sprzęt oraz jego utrzymanie, ponieważ praca takich modeli wymaga m.in. zaawansowanych kart graficznych. Jednakże w dłuższej perspektywie może to być rozwiązanie bardziej ekonomiczne, szczególnie przy dużej skali prowadzonych badań. Dodatkową zaletą tego podejścia jest możliwość dostosowania (*fine-tuning*) modelu do specyfiki konkretnych zadań badawczych. Przykładowo, model można dotrenować, douczyć na zbiorze wcześniej zakodowanych wywiadów, co potencjalnie zwiększa trafność kodowania w ramach określonego projektu badawczego czy dziedziny.

Podsumowując, warto także podkreślić, że automatyzacja nie oznacza wyeliminowania roli badacza, ale sygnalizuje znaczącą jej zmianę. Automatyzacja wymaga starannego przemyślenia procesu, który powinien odpowiadać specyfice badanego materiału. Konieczny jest także nadzór i okresowa weryfikacja wyników, szczególnie w początkowych fazach projektu. Innymi słowy, badacz musi nie tylko zaprojektować cały proces, ale także monitorować jego przebieg i weryfikować otrzymywane rezultaty.

6.7. Podsumowanie

Generatywna sztuczna inteligencja otwiera nowe możliwości w dziedzinie badań jakościowych, oferując narzędzia, które mogą znacząco zwiększyć efektywność i zakres analizy danych tekstowych. Kluczowe zastosowania gen AI w tym obszarze obejmują wspomaganie kodowania indukcyjnego oraz automatyzację kodowania dedukcyjnego. Dzięki gen AI otwierają się także nowe możliwości w zakresie przetwarzanie dużych zbiorów danych tekstowych.

Intencją autora niniejszego rozdziału było sformułowanie wskazówek, które zachowają swoją aktualność możliwie jak najdłużej, stanowiąc solidną podstawę do pracy z gen AI w kontekście analiz jakościowych. Strategie przedstawione w tym rozdziale służyć mogą jako praktyczna podstawa do formułowania efektywnych promptów, niezbędnych do realizacji analizy jakościowej z wykorzystaniem gen AI. Warto podkreślić, że te wskazówki, w połączeniu z szerszą wiedzą na temat zasad promptowania, stanowić mogą solidną podstawę do eksperymentowania z dostępnymi dużymi modelami językowymi. Postęp w wykorzystaniu gen AI w badaniach jakościowych wymaga systematycznej pracy badaczy. Warto zacząć od

testowania omówionych tutaj metod na małych, znanych badaczom zbiorach danych (lub też na niewielkiej próbie z dużego korpusu danych). Takie podejście umożliwia dokładną kontrolę procesu kodowania jakościowego poprzez systematyczne porównywanie wyników uzyskanych przez model z kodowaniem wykonanym przez doświadczonych badaczy. Pozwala to nie tylko na ocenę jakości pracy wykonanej przez tę technologię, ale także udoskonalenie procedur wykorzystania gen AI w konkretnym kontekście badawczym.

Dla badaczy korzystających z gen AI w badaniach jakościowych kluczowe jest stopniowe wprowadzanie tej technologii do procesu badawczego, łączenie tradycyjnych metod z nowymi możliwościami modeli językowych oraz ciągłe doskonalenie umiejętności formułowania promptów. Warto podkreślić, że gen AI nie zastępuje krytycznego myślenia i eksperckiej wiedzy badacza, ale stanowi raczej potężne narzędzie wspomagające.

Przyszłość badań jakościowych z wykorzystaniem gen AI wiąże się z potrzebą dalszych badań nad wiarygodnością i trafnością analiz prowadzonych z użyciem tej technologii. Już teraz odbywa się proces integracji gen AI z istniejącymi narzędziami do analizy jakościowej (takimi jak MaxQDA, Atlas.ti). Niesie to szereg korzyści. Po pierwsze, pozwala na wykorzystanie sprawdzonych interfejsów i procesów pracy, do których badacze są już przyzwyczajeni. Po drugie, umożliwia łączenie tradycyjnych metod analizy jakościowej z możliwościami gen AI w ramach jednego środowiska pracy. Po trzecie, ułatwia weryfikację i korektę wyników pracy wykonanej przez gen AI. Pozwala to badaczom płynnie przełączać się między „ręczną" a wspomaganą przez gen AI analizą, przy zachowaniu pełnej kontroli nad procesem badawczym. Podobnie jak w przypadku modeli językowych wyspecjalizowanych w tworzeniu kodu komputerowego, można oczekiwać rozwoju specjalistycznych modeli gen AI, dostosowanych do specyfiki badań jakościowych. Warto testować tego typu rozwiązania, ponieważ najpewniej staną się one w najbliższej przyszłości nieodłącznym elementem naszego warsztatu badawczego.

Rola badacza jakościowego w „erze AI" również ulega transformacji. Zamiast skupiać się na mechanicznym kodowaniu, badacze będą coraz bardziej koncentrować się na: projektowaniu procesu analitycznego, z uwzględnieniem w nim odpowiedniej roli dla tej czy innej formy gen AI, nadzorze nad procesem analizy oraz na interpretacji wyników. Wymaga to niewątpliwie rozwoju nowych kompetencji.

Podsumowując, generatywna sztuczna inteligencja oferuje znaczący potencjał w zakresie wspomagania i rozszerzania możliwości analiz jakościowych, w tym również na potrzeby ewaluacji polityk publicznych. Jednocześnie jej wykorzystanie wymaga ostrożnego podejścia, świadomości ograniczeń oraz ciągłego doskonalenia metodologii. Kluczowe będzie znalezienie równowagi między efektywnością oferowaną przez narzędzia gen AI a głębokim, kontekstowym zrozumieniem analizowanych zjawisk społecznych.

Bibliografia

- Hitch, D. (2023). Artificial Intelligence Augmented Qualitative Analysis: The Way of the Future? *Qualitative Health Research*, 10497323231217392 (dostęp: 31.03.2025).
- Kheiri, K., Karimi, H. (2023). *SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning* (arXiv:2307.10234) (dostęp: 31.03.2025).
- Kuckartz, U. (2019). Qualitative Text Analysis: A Systematic Approach. W: Kaiser, G., Presmeg, N. (red.), *Compendium for Early Career Researchers in Mathematics Education*, 181–197. Springer International Publishing (dostęp: 31.03.2025).
- Mahoney, J., Goertz, G. (2006). A Tale of Two Cultures: Contrasting Quantitative and Qualitative Research. *COMPASSS Working Papers*, *2006*(37) (dostęp: 31.03.2025).
- Mollick, E. (2024). *Co-intelligence: Living and working with AI*. Portfolio/Penguin.
- Pope, C., Ziebland, S., Mays, N. (2006). Analysing Qualitative Data. W: Pope, C., Mays, N. (red.), *Qualitative Research in Health Care*, wyd. 1, s. 63–81. Wiley (dostęp: 31.03.2025).
- Puppis, M. (2019). Analyzing Talk and Text I: Qualitative Content Analysis. W: Van Den Bulck, H., Puppis, M., Donders, K., et al. (red.), *The Palgrave Handbook of Methods for Media Policy Research* (s. 367–384). Springer International Publishing. Saldaña, J. (2016). *The coding manual for qualitative researchers* (3E [Third edition]). SAGE (dostęp: 31.03.2025).
- Schmitt, B. (2024). Transforming qualitative research in phygital settings: The role of generative AI. *Qualitative Market Research: An International Journal*, *27*(3), 523–526 (dostęp: 31.03.2025).
- Zhang, H., Wu, C., Xie, J., et al. (2023). *Redefining Qualitative Analysis in the AI Era: Utilizing ChatGPT for Efficient Thematic Analysis* (arXiv:2309.10771). arXiv (dostęp: 31.03.2025).

7. Przeglądy źródeł wspierane gen AI

Karol Olejniczak
Tomasz Kupiec
Dominika Wojtowicz

7.1. Wprowadzenie

Cel rozdziału

W tym rozdziale koncentrujemy się na pracy z tekstowymi źródłami wtórnymi – publikacjami wyników badań, różnego rodzaju dokumentami, literaturą przedmiotu, jak również innymi materiałami tekstowymi dostępnymi w formie cyfrowej.

Przeglądy źródeł tekstowych (ang. *literature reviews*) to jedna z najbardziej powszechnych i popularnych metod pozyskiwania wiedzy przez administrację publiczną. Ten termin obejmuje w praktyce całe spektrum podejść – od syntez wewnętrznych raportów danej organizacji, przez przeglądy źródeł internetowych, przeglądy literatury, po przeglądy systematyczne. W krajowej praktyce terminy te są często mieszane, a niektóre z nich używane nieco na wyrost.

Z perspektywy praktyki polityk publicznych kluczowa jest siła dowodów, czyli wiarygodność wniosków płynących z danego badania. W końcu podejmowanie decyzji w oparciu o nierzetelną wiedzę może być równie szkodliwe jak podejmowanie decyzji bez jakichkolwiek podstaw merytorycznych. Dlatego w literaturze i praktyce *Evidence-Based Policy* przykłada się wysoką wagę do jakości badań. W przypadku przeglądów źródeł zastanych ich jakość opiera się na dwóch filarach:

a) puli informacji, na których budowane są wnioski (rozległość i jakość źródeł) oraz
b) rzetelności przeprowadzonego procesu – jego przejrzystej, uporządkowanej i możliwej do replikacji procedurze.

Prowadzenie przeglądów wysokiej jakości, szczególnie przeglądów systematycznych, jest jednak zasobochłonne. Generatywna sztuczna inteligencja (gen AI) stwarza zupełnie nowe możliwości syntezy wiedzy naukowej (Heidt, 2023). W ostatnich latach pojawił się szereg publikacji opisujących eksperymentalne zastosowania gen AI do przeglądów źródeł (Glickman & Zhang, 2024; van Dijk et al., 2023; Wagner et al., 2021; Ye et al., 2024). Powstają też konkretne rozwiązania cyfrowe kuszące możliwościami automatyzacji kolejnych etapów badania, w tym szybkości wyszukiwania i syntezy źródeł (Matthews, 2021). Kluczową pozostaje jednak kwestia, jak stosowanie tych potencjalnych ułatwień wpływa na jakość wyników badań gabinetowych (w szczególności tekstowych źródeł danych wtórnych), a w konsekwencji – ich użyteczność w procesach decyzyjnych.

W tym rozdziale próbujemy więc odpowiedzieć na następujące pytanie: **Jak możemy usprawnić za pomocą gen AI przegląd tekstowych źródeł danych, dbając jednocześnie o ich wysoką jakość, a co za tym idzie – użyteczność w procesach decyzyjnych administracji publicznej?**

Źródła obserwacji

Nasze rozważania prowadzimy na podstawie materiałów zebranych podczas dwóch quasi-systematycznych przeglądów literatury poświęconych zastosowaniom AI w polityce publicznej (University of Washington, Seattle: wrzesień – grudzień 2023, Temple University, Filadelfia: czerwiec – sierpień 2024) oraz własnych doświadczeń i testów w ramach projektów badawczo-aplikacyjnych. Listę i krótkie opisy tych projektów zamieściliśmy w Aneksie 1.

W naszych testach skoncentrowaliśmy się na powszechnie dostępnych narzędziach AI, które nie wymagają od użytkowników umiejętności programistycznych czy zaawansowanych komend. Testowaliśmy następujące trzy grupy gen AI:

- LLMy: ChatGPT 4.o, Claude 3.5, Meta - Llama,
- RAGi: Perplexity, Consensus, elicit, SCISPACE, Scite,
- wyspecjalizowanych asystentów gen AI: ASReview, Connected Papers, Open Knowledge Maps, Petal AI.

Struktura i adresaci rozdziału

W kolejnej części rozdziału omawiamy trzy kwestie fundamentalne dla rzetelności przeglądów źródeł tekstowych i zastosowań gen AI w tych zadaniach. Po pierwsze, porządkujemy terminologię typów przeglądów występujących w literaturze i praktyce naukowej, proponując typologię dostosowaną do zadań ewaluacji. Wyjaśniamy, że poziom „systematyczności" determinuje rzetelność wyników.

Po drugie, omawiamy całe spektrum rodzajów źródeł tekstowych, które można wykorzystywać przy przeglądach, wskazujemy też ich funkcjonalne zalety i ograniczenia. Tu również zwracamy uwagę na relację, w tym wypadku funkcjonalną, między doborem źródeł a rzetelnością budowanych na ich podstawie wniosków i konkluzji.

Wreszcie po trzecie, wyjaśniamy kwestię dostępności i widoczności różnych źródeł cyfrowych dla asystentów gen AI. Oprogramowania gen AI często stwarzają iluzję wszechstronności i totalności źródeł. W rzeczywistości różne rozwiązania komercyjne opierają się na różnych, często dość wybiórczych, fragmentarycznych zbiorach źródeł. Wiedza na podstawie jakiego zbioru zasobów tekstowych gen AI buduje swoje odpowiedzi, jest kluczowa dla efektywnego korzystania z jej pomocy.

W głównej części rozdziału prezentujemy nasze testy zastosowań gen AI do konkretnych zadań w ramach trzech etapów przeglądów: w planowaniu badania, w szukaniu i selekcji źródeł, w analizie i syntezie. W rozdziale omawiamy pierwsze wnioski i obserwacje, w aneksach zamieściliśmy dla zainteresowanych Czytelników szczegóły wykonanych procedur.

W konkluzjach rozdziału podsumowujemy nasze obserwacje z zastosowań gen AI i podejmujemy szerszą refleksję nad możliwymi kierunkami zmian. Szczególną uwagę zwróciliśmy na warunki związane z wykorzystaniem AI w poszczególnych etapach przeglądu, co pozwoliło na określenie obszarów, gdzie technologia ta może być przydatna, a gdzie aktualnie może stanowić ryzyko dla rzetelności wyników.

Rozdział skierowany jest zarówno do badaczy prowadzących ewaluacje, jak i do osób odpowiedzialnych za zlecanie i odbiór badań. Wykonawcy badań znajdą tu informacje pomocne w prawidłowym przeprowadzeniu przeglądu tekstowych źródeł zastanych, a także

wskazówki, jak w pełni wykorzystać potencjał gen AI, jednocześnie zachowując ostrożność przy ocenie wyników generowanych przez gen AI. Zlecający i odbierający badania natomiast dowiedzą się, jak określić w szczegółowych opisach przedmiotu zamówienia dopuszczalne warunki wykorzystania AI w ramach przeglądów oraz jak ocenić wpływ tej technologii na wiarygodność otrzymanych wyników i rekomendacji.

7.2. Typy przeglądów i rodzaje źródeł w przeglądach

Spektrum przeglądów źródeł

Jak wskazano we wprowadzeniu, przeglądy tekstowych źródeł zastanych są jedną z najpopularniejszych metod pozyskiwania wiedzy. Co równie ważne, według wielu autorów, stanowią one źródło najbardziej wiarygodnych dowodów naukowych, na których decydenci powinni opierać swoje decyzje (np. Petticrew, Roberts, 2003).

Każdy, kto przeprowadzał przeglądy źródeł zastanych i zapoznawał się z ich wynikami, musiał zauważyć, że ich jakość, zakres, cel, sposób raportowania są bardzo zróżnicowane. Oczywistym jest, że nie każdy przegląd jest wystarczająco rzetelny i wiarygodny, by być uznanym za źródło dowodów oraz wystarczającą podstawę kluczowych decyzji.

Zrozumienie roli i potencjalnego znaczenia przeglądów źródeł zastanych jako metody badawczej oraz możliwości wykorzystania gen AI w procesie ich realizacji wymaga uporządkowania podstawowych pojęć i przedstawienia głównych typów przeglądów. Właśnie temu poświęcony jest ten podrozdział.

Początki systematycznego podejścia do przeglądów źródeł zastanych sięgają połowy XVIII wieku i Jamesa Linda, który próbował usystematyzować wyniki dużej liczby raportów na temat zapobiegania i leczenia szkorbutu (Chalmers et al, 2002). Za początek rozwoju współczesnego naukowego podejścia do przeglądów przyjmuje się pracę Karla Pearsona z 1904 r., będącą przeglądem dowodów na skuteczność szczepionki przeciwko durowi brzusznemu (Cooper et al, 2019). Z pola medycyny praktyka przeglądów rozprzestrzeniła się na inne i obecnie na gruncie różnych dyscyplin naukowych rozpoznaje się nawet 50 typów

przeglądów (Kastner et al, 2016; Tricco et al, 2016; Sutton et al, 2019). Uporządkowanie tej różnorodnej gamy podejść zaproponował Booth et al (2021), identyfikując siedem szerokich „rodzin" przeglądów: tradycyjne przeglądy, przeglądy systematyczne, przeglądy przeglądów, szybkie przeglądy, przeglądy jakościowe, przeglądy mieszanych metod oraz przeglądy celowe.

Poniżej (tabela 7.1.) prezentujemy własną propozycję możliwie uproszczonej typologii, dostosowanej do celu tego rozdziału – przedstawienia perspektywy zastosowań gen AI w procesie przeglądu źródeł tekstowych. Składają się na nią: przegląd systematyczny, quasi-systematyczny, tradycyjny i pobieżny[1].

Tabela 7.1. Proponowana typologia przeglądów na potrzeby ewaluacji polityk publicznych

Typ przeglądu	Funkcja	Charakterystyka	Zastosowanie w ewaluacji
Przegląd systematyczny	Synteza i podsumowanie istniejącej wiedzy ze wszystkich dostępnych, wiarygodnych źródeł tekstowych, potrzebnej do odpowiedzi na zadane pytanie.	Podstawowe pojęcie w kontekście przeglądu źródeł tekstowych jako metody badawczej. „Systematyczny" oznacza prawidłowo przeprowadzony, uwzględniający wszystkie istotne dowody (wyczerpujący, obejmuje wszystkie dostępne źródła tekstowe w przyjętym zakresie) oraz umożliwiający wiarygodne wnioskowanie (Hammersley, 2002). Atrybuty przeglądu systematycznego to (Booth et al, 2021): 1. szczegółowa metodyka, zwykle powszechnie zaakceptowana przez określone środowisko badaczy i praktyków, opisana w wytycznych instytucji (np. Cochrane Collaboration, Campbell Collaboration) albo artykułach naukowych; 2. protokół przeglądu, chroniący przed zmianami zakresu w trakcie badania, umożliwiający weryfikację i replikację innym badaczom i odbiorcom; 3. przejrzyste raportowanie przebiegu przeglądu, zwykle zgodnie z zaakceptowanymi w danym środowisku wytycznymi[2]; (4) ocena jakości zgromadzonych źródeł tekstowych według przyjętych wytycznych, check-listy.	Podsumowanie wiedzy o udokumentowanych już efektach interwencji, przewidywania co do skuteczności interwencji w innych kontekstach. Zastosowanie pożądane, ale ograniczone czasochłonnością i potrzebnymi kompetencjami (Mazur, Orłowska, 2018), w związku z czym ten typ przeglądu nie jest stosowany w praktyce badań ewaluacyjnych w Polsce.

[1] Przyjmując powszechnie stosowaną konwencję w tzw. hierarchiach dowodów (np. Petticrew, Roberts, 2003), prezentujemy typy przeglądów od najbardziej rygorystycznego metodycznie i oferującego najbardziej wiarygodne dowody, do najmniej rzetelnego.

[2] Najbardziej rozpowszechnionym sposobem raportowania przebiegu przeglądu systematycznego, który w związku z tym stał się niejako wyróżnikiem „systematyczności" przeglądu, jest PRISMA (Page et al., 2021).

Typ przeglądu	Funkcja	Charakterystyka	Zastosowanie w ewaluacji
Przegląd quasi-systematyczny	Dostarczenie wiedzy „wystarczająco wiarygodnej” w odpowiednim czasie.	Każdy przegląd, w ramach którego następuje świadome odstąpienie od założeń przeglądu systematycznego, wymuszone ograniczeniem zasobów. Najbardziej typowym przykładem jest szybki przegląd, który w zamian za uproszczenia metodyki oferuje znacznie krótszy czas realizacji przy nadal satysfakcjonujących rezultatach. Definiującą cechą tego typu przeglądów jest większa elastyczność oraz kształtowanie zakresu i sposobu realizacji nie w oparciu o uniwersalne wytyczne, a negocjacje między odbiorcą i wykonawcą. Ustalenia dotyczące ograniczeń i uproszczeń na etapach poszukiwania i oceny źródeł tekstowych, analizy i syntezy oraz tego, jak te uproszczenia wpłyną na wiarygodność wnioskowania, powinny być przejrzyście dokumentowane (Sutton et al, 2019).	Podsumowanie wiedzy o efektach interwencji, przewidywania co do skuteczności interwencji w innych kontekstach. Praktyczny ze względu na ramy czasowe i elastyczność podejścia.
Przegląd tradycyjny	Podsumowanie literatury (tzw. dorobku naukowo-badawczego) zademonstrowanie, że posiada się wiedzę w określonym obszarze, wstęp do właściwych badań.	Tradycyjny przegląd jest niesystematyczny w wyżej przedstawionym znaczeniu – zwykle odbywa się bez protokołu, nie ma jasno określonego zakresu, metodyki gromadzenia źródeł tekstowych, oceny ich jakości i analizy, przez co jest niereplikowalny i niewery-fikowalny. Jeśli badacze zakładają, że ma spełniać funkcję przeglądu systematycznego, to taki przegląd jest nieprawidłowy. Istnieją dopuszczalne, „właściwe” rodzaje przeglądu tradycyjnego, np. krytyczny, który w rezultacie przeglądu literatury określonego tematu może prowadzić do postawienia hipotezy lub stworzenia modelu, stanowiących wstęp do zasadniczego badania.	Gromadzenie wiedzy kontekstowej, przygotowanie do gromadzenia danych pierwotnych. Nie powinien być wykorzystywany do podsumowywania wiedzy o efektach interwencji.
Przegląd pobieżny	Pozorowanie, że dokonało się przeglądu i podsumowało wiedzę w określonym obszarze.	Wykonane samodzielnie lub za pomocą coraz powszechniejszych narzędzi AI wyszukiwanie i podsumowanie kilku, kilkunastu przypadkowych pozycji na dany temat, bez weryfikacji ich jakości, wiarygodności, reprezentatywności.	Nie powinien być wykorzystywany w ewaluacji.

Źródło: Opracowanie własne.

7.3. Rodzaje źródeł tekstowych

Powszechny dostęp do Internetu otwiera możliwość korzystania z różnego rodzaju materiałów – od artykułów naukowych, przez raporty i blogi, aż po fora dyskusyjne i *social media*. Mnogość i różnorodność źródeł, z których możemy czerpać wiedzę, sprawia, że przy ich doborze musimy pamiętać o dwóch fundamentalnych kwestiach.

Po pierwsze, użyteczność źródeł tekstowych zależy bezpośrednio od celu naszych poszukiwań, a w konsekwencji dobór źródeł musi być ściśle dostosowany do rodzaju pytania badawczego. Różne typy źródeł spełniają różne funkcje i odpowiadają na odmienne potrzeby badawcze. Na przykład, jeśli celem jest zrozumienie powszechnych opinii, to blogi, media społecznościowe i fora internetowe mogą dostarczyć użytecznych informacji. Jeżeli natomiast celem badań jest określenie skuteczności pewnych typów interwencji publicznych, powinniśmy sięgnąć do treści z publikacji recenzowanych (podręczniki, artykuły naukowe) lub uznanych instytucji (np. OECD, Bank Światowy).

Po drugie, funkcjonuje hierarchia jakości źródeł. Nawet w ramach tej samej kategorii materiałów istnieją różnice w jakości i wiarygodności informacji. Na przykład książki wydawane przez uznane wydawnictwa akademickie oferują bardziej sprawdzone i zaufane treści niż te wydawane przez nieznanych wydawców. Podobnie artykuły publikowane w renomowanych i uznanych periodykach naukowych są bardziej wiarygodne niż publikacje w czasopismach, które często publikują treści bez należytej weryfikacji. Treści umieszczone na stronach uznanych instytucji, takich jak m.in. think thanki czy firmy consultingowe, będą bardziej rzetelne niż podmiotów tworzonych *ad hoc* przez pasjonatów, a nie specjalistów w danym obszarze. I w końcu blogi i podcasty – je również cechuje różna wiarygodność, a dodatkowo, często są obarczone subiektywnością lub nawet stronniczością (co nie znaczy, że nie istnieją materiały zamieszczane w Internecie, które są tworzone przez osoby kompetentne i mogą być wykorzystywane w ewaluacjach). Ta hierarchia jest istotna, ponieważ wpływa bezpośrednio na jakość danych, które wykorzystujemy w badaniach.

W tabeli 7.2. przedstawiono zestawienie różnych typów źródeł wraz z informacją o zaletach i ograniczeniach ich wykorzystania, jak również o ich dostępności.

Tabela 7.2. Rodzaje źródeł i ich charakterystyka

Typ źródła	Krótka charakterystyka	Zalety i ograniczenia źródła	Dostęp do źródła
Encyklopedie tematyczne	Są jednym z podstawowych źródeł wiedzy, szczególnie na etapie wstępnego rozpoznania tematu i mogą stanowić punkt wyjścia do dalszych badań. Stanowią zbiory zwięzłych, zorganizowanych alfabetycznie haseł, które dostarczają ogólnych informacji na temat szerokiego zakresu zagadnień.	+ Umożliwiają szybkie wyszukiwanie haseł; + Jako tworzone przez specjalistów stanowią zaufane źródło wiarygodnych i sprawdzonych informacji; + W wersjach cyfrowych ich treści często poddawane są systematycznym aktualizacjom; – Nie dostarczają szczegółowej, dogłębnej wiedzy, ich rola ogranicza się głównie do dostarczania podstawowych definicji i ogólnych informacji.	Zwykle w zbiorach zamkniętych (płatnych).
Podręczniki specjalistyczne (ang. *handbooks*)	Oferują szczegółowe i zaawansowane informacje, łącząc większość prac dotyczących jednego tematu (dyscypliny, tematu, specjalizacji). *Handbooks* mogą służyć jako źródło wiedzy o kluczowych teoriach, koncepcjach oraz terminologii, które są niezbędne do zrozumienia danego tematu.	+ Mogą dostarczyć rzetelnego przeglądu aktualnego (na moment publikacji) stanu wiedzy na dany temat (najnowsze osiągnięcia, główne nurty literatury); + Oferują praktyczne podejście do tematu, często wzbogacone o przykłady zastosowania teorii w realnych sytuacjach; + Mogą służyć jako referencyjne źródło wiedzy, zwłaszcza w rozwiązywaniu specyficznych problemów; – Nie zawsze dostarczają aktualnych danych, zwłaszcza w dziedzinach, które szybko się rozwijają; – Mogą być bardzo szczegółowe, fragmentaryczne, ograniczać się do konkretnego kontekstu lub zastosowania, co sprawia, że mogą nie obejmować innych perspektyw czy szerszego kontekstu teoretycznego; – Często są napisane specjalistycznym językiem.	Zwykle w zbiorach zamkniętych.

Typ źródła	Krótka charakterystyka	Zalety i ograniczenia źródła	Dostęp do źródła
Podręczniki (ang. *textbooks*)	Podręczniki to ustrukturyzowane opracowania służące do systematycznego przedstawienia fundamentalnej wiedzy w danej dziedzinie. Ich celem jest wprowadzenie czytelnika w podstawowe pojęcia, teorie i procesy, często w sposób chronologiczny lub tematycznie zorganizowany. Często służą jako kompendium wiedzy z danej dziedziny. W odróżnieniu od poradników, które koncentrują się na praktycznych zastosowaniach i rozwiązywaniu konkretnych problemów, podręczniki dostarczają podstawowej wiedzy teoretycznej w sposób usystematyzowany.	+ Dostarcza bardzo dobrego podsumowania definicji, głównych nurtów literatury itp.; + Oferują syntetyczne i uporządkowane podejście do zagadnień; – Jeden autor relacjonuje pracę innych badaczy, co sprawia, że wiedza jest filtrowana przez osobiste preferencje, mogą więc nie uwzględniać alternatywnych teorii lub innowacyjnych metod; – Mogą być mniej aktualne, ponieważ proces ich tworzenia i publikacji zajmuje czas, co sprawia, że rzadziej odzwierciedlają najnowsze badania lub zmieniające się trendy.	Zwykle w zbiorach zamkniętych.
Artykuły w czasopismach recenzowanych	Artykuły w czasopismach recenzowanych to prace naukowe, które przed publikacją przechodzą przez proces recenzji eksperckiej (ang. *peer review*). Oznacza to, że ich jakość, metodologia oraz rzetelność danych są oceniane przez niezależnych ekspertów w danej dziedzinie.	+ Jakość publikacji oceniana jest przez niezależnych recenzentów, co zwiększa wiarygodność i rzetelność zawartych w nich informacji; – Są głównie adresowane do naukowców, co sprawia, że mogą być trudne do zrozumienia dla osób spoza danej dziedziny.	Zwykle w zbiorach zamkniętych, część artykułów w wolnym dostępie.
Prace dyplomowe / rozprawy doktorskie	Obszerniejsze opracowania naukowe, których celem jest zaprezentowanie przez studentów lub doktorantów oryginalnych wyników badań, często na wysoce specjalistyczny temat, które mogą wnieść nowe informacje do danej dziedziny. Prace nadzorowane są przez promotorów, którzy dbają o poprawność merytoryczną i metodologiczną badań.	+ Zawierają oryginalne badania, które mogą dostarczyć nowych informacji i perspektyw w badanej dziedzinie; – Są zazwyczaj szczegółowe i dogłębnie analizują wybrane zagadnienie; – Jakość może być nierówna (pisane przez osoby z mniejszym doświadczeniem badawczym); – Dostęp do prac jest selektywny – nie wszystkie prace dyplomowe są publikowane lub łatwo dostępne w publicznych bazach danych.	Zwykle w wolnym dostępie.

Typ źródła	Krótka charakterystyka	Zalety i ograniczenia źródła	Dostęp do źródła
Badania / raporty techniczne	Badania i raporty techniczne to dokumenty, które prezentują wyniki badań, analiz, ewaluacji lub eksperymentów prowadzonych w ramach projektów, często finansowanych przez rządy, samorządy, przemysł, firmy konsultingowe, instytucje akademickie lub organizacje międzynarodowe. Zazwyczaj przedstawiają wyniki analiz, badań i ewaluacji prowadzonych na szeroką skalę. Są tworzone przez zespoły ekspertów w danej dziedzinie.	+ Tworzone przez prestiżowe instytucje, takie jak Komisja Europejska, Parlament Europejski, OECD, Bank Światowy, Ministerstwa, agencje rządowe, co zwiększa ich wiarygodność; + Najczęściej prowadzone przy zaangażowaniu znacznych środków finansowych, co sprawia, że badania są przekrojowe, łączą różne źródła danych, mogą dotyczyć większej liczby państw itp.; + Dostarczają praktycznych rekomendacji i wniosków; – Rzadko są poddawane recenzji naukowej (ang. *peer review*), a jakość i rzetelność danych mogą się różnić w zależności od źródła finansowania, celów raportu i zespołu ekspertów.	W zbiorach zamkniętych lub w wolnym dostępie.
Artykuły w czasopismach branżowych	Artykuły w czasopismach branżowych są zwykle tworzone z myślą o praktykach zawodowych i osobach działających w określonej branży, a ich głównym celem jest dostarczenie informacji użytecznych w codziennej pracy o nowych trendach, technologiach i rozwiązaniach praktycznych. W przeciwieństwie do artykułów naukowych są one oceniane przede wszystkim pod kątem praktycznej przydatności i aktualności, a nie pod względem rygoru akademickiego.	+ Często mają bezpośrednie, praktyczne znaczenie dla specjalistów i pracowników danej branży, co czyni je cennym źródłem bieżących informacji; + Mogą dostarczać szybkich, praktycznych rozwiązań oraz aktualnych informacji o najnowszych trendach w danej branży; – Mogą ograniczać się do odzwierciedlenia subiektywnych poglądów autora lub też promocji niezweryfikowanych pomysłów.	W zbiorach zamkniętych lub w wolnym dostępie.

Typ źródła	Krótka charakterystyka	Zalety i ograniczenia źródła	Dostęp do źródła
Artykuły z konferencji i sympozjów	Artykuły te są referatami, które zostały zaprezentowane podczas konferencji naukowych lub sympozjów. Zawierają zwykle wstępne wyniki badań lub szkice koncepcji, które mogą być dalej rozwijane i analizowane.	+ Publikowane na bieżąco, zawierają najnowsze odkrycia i pomysły; + Mogą być cennym źródłem inspiracji i wstępnych danych do wykorzystania w badaniach ewaluacyjnych, szczególnie w przypadku tematów nowych i szybko rozwijających się; – Ze względu na wczesny etap prac przedstawione wyniki mogą być niepełne lub niewystarczająco zweryfikowane.	W wolnym dostępie.
Artykuły w czasopismach / gazetach	Artykuły publikowane w czasopismach i gazetach zazwyczaj dostarczają aktualnych informacji na temat bieżących wydarzeń, nowych trendów oraz popularnych tematów. Ich celem jest szybkie przekazanie informacji szerokiej publiczności.	+ Oferują najnowsze informacje: przydatny punkt wyjścia w badaniach ewaluacyjnych, zwłaszcza przy analizie współczesnych trendów, opinii publicznej lub szybko zmieniających się tematów; + Są pisane w zrozumiałym języku; – Nie zawsze są oparte na rzetelnych badaniach czy danych; – Mogą zawierać uproszczenia lub błędne interpretacje, przez co konieczne jest ich weryfikowanie przed wykorzystaniem w badaniach.	W zbiorach zamkniętych lub w wolnym dostępie.
Strony internetowe, social media, blogi, podcasty	Strony internetowe oraz różnego rodzaju portale społecznościowe, blogi i podcasty to obszerne zbiory informacji, które obejmuje szeroką gamę tematów. Można w nich znaleźć zarówno rzetelne i fachowe informacje (w szczególności na stronach instytutów, wyspecjalizowanych organizacji, w tym think tanków, firm consultingowych, uznanych ekspertów lub ich stowarzyszeń), jak i mniej wiarygodne treści. Jakość treści jest bardzo zróżnicowana i wymaga krytycznej oceny źródeł.	+ Duża różnorodność dostępnych informacji i tematów; + Strony prowadzone przez instytucje rządowe, organizacje międzynarodowe czy renomowane uczelnie zawierają zazwyczaj wiarygodne i aktualne informacje; – Niektóre są płytkie, mylące, stronnicze lub nawet zawierają nieprawdziwe informacje; – Brak jednolitego standardu weryfikacji treści na wielu stronach wymaga dużej ostrożności i krytycznego podejścia przy ich wykorzystaniu.	Zbiory płatne lub w wolnym dostępie.

Typ źródła	Krótka charakterystyka	Zalety i ograniczenia źródła	Dostęp do źródła
Akty prawne, dokumenty strategiczne i programowe (krajowe i unijne), oceny skutków regulacji, sprawozdania i raporty okresowe, fiszki projektowe	Oficjalne dokumenty opracowywane przez administrację publiczną na szczeblu krajowym, regionalnym lub unijnym. Ich celem jest ustanawianie ram prawnych, wyznaczanie strategicznych celów polityki czy ocena potencjalnych skutków regulacji. Są kluczowymi dokumentami dla planowania i wdrażania polityk publicznych. Innym źródłem wiedzy są fiszki projektowe przedstawiające założenia konkretnych projektów realizujących założenia polityk publicznych, stanowiące tym samym praktyczne narzędzie ich implementacji.	+ Stanowią podstawę prawną, strategiczną lub operacyjną dla podejmowania decyzji i wdrażania polityk; + Dostarczają wiarygodnych, szczegółowych informacji opracowanych na podstawie szerokich analiz; + Fiszki projektowe dostarczają szczegółowego opisu projektów, które mogą być poddawane analizie w procesie ewaluacji; – Mogą być czasochłonne w analizie ze względu na ich rozległość, szczegółowość lub specjalistyczny język; – Fiszki projektowe prezentują tylko założenia, a nie rzeczywiste efekty realizacji projektów.	W wolnym dostępie (z wyjątkiem fiszek projektowych i niektórych raportów okresowych).

Źródło: Opracowanie własne w oparciu o Thomas (2019, s. 24–27).

7.4. Dostępność źródeł dla gen AI

Szybkość odpowiedzi udzielanych przez generatywną AI (gen AI) oraz bogactwo cytowanych źródeł mogą stwarzać iluzję, że sztuczna inteligencja przeszukuje wszystkie istniejące zasoby informacyjne. Zrozumienie, na jakiej bazie informacji faktycznie budowane są odpowiedzi, jest jednak kluczowe z perspektywy rzetelności przeglądów.

Generatywna AI może bazować na trzech głównych pulach informacji:

A. Dane treningowe

Są to teksty, na których dany model był trenowany w procesie uczenia maszynowego. Obejmują one szerokie spektrum materiałów – od treści publicznie dostępnych po dane licencjonowane, jednak szczegółowy skład tych zbiorów rzadko jest w pełni ujawniany przez firmy, które zbudowały dany model.

B. Źródła dostępne przez Internet

W zależności od konfiguracji niektóre modele gen AI mogą przeszukiwać zasoby online w czasie rzeczywistym. Są to najczęściej materiały w otwartym dostępie, w tym publikacje naukowe, raporty oraz inne dokumenty specjalistyczne. Ważne jest jednak, że zwykle takie wyszukiwarki gen AI nie mają dostępu do płatnych baz danych (np. Elsevier, JSTOR, archiwum Rzeczpospolitej, raporty OECD itd.).

C. Źródła dostarczone przez użytkownika

Użytkownik może załadować własne zasoby, takie jak kolekcje publikacji naukowych, raportów, dokumentów branżowych, czy dokumentacji projektowej lub programowej. Ta opcja pozwala na znaczne skupienie analiz na konkretnej bazie źródeł i dopasowanie analiz AI do specyficznych potrzeb badawczych lub sektorowych.

Z tej charakterystyki płyną dla nas trzy ważne wnioski praktyczne.

Po pierwsze, domyślnie modele gen AI pracują na źródłach A i B. To oznacza, że są ograniczone do materiałów dostępnych publicznie i w otwartym dostępie. Wiele kluczowych źródeł, takich jak raporty OECD, artykuły z renomowanych czasopism naukowych czy rozdziały specjalistycznych podręczników, pozostaje niedostępnych dla tych narzędzi z powodu barier płatnych subskrypcji. Przykładem tego ograniczenia może być najnowszy specjalistyczny podręcznik ewaluacyjny: K.E. Newcomer & S. Mumford (red.) (2024). *Research Handbook on Program Evaluation*, Edward Elgar Publishing. Jedynie 6 z 36 rozdziałów jest w otwartym dostępie. Taka dysproporcja ukazuje skalę wartościowych materiałów niewidocznych dla większości modeli AI.

Po drugie, istnieją potencjalne problemy z przejrzystością. Użytkownicy nie znają mechanizmu wyboru źródeł przez AI. Istnieje ryzyko, że algorytmy AI pomijają istotne treści lub priorytetyzują materiały, które nie zawsze odpowiadają potrzebom badawczym. Brak przejrzystości w doborze źródeł może prowadzić do sytuacji, w której użytkownik nieświadomie opiera swoje wnioski na niepełnym lub niewłaściwie dobranym zestawie informacji.

Po trzecie, możemy znacząco wzmocnić solidność bazy informacji, na której gen AI pracuje, dostarczając mu przygotowane przez nas pakiety wyselekcjonowanych źródeł (opcja C). Ten sposób działania jest szczególnie atrakcyjny i pożądany w badaniach ewaluacyjnych, bo dotyczą one specyficznych (uwarunkowanych określonym zestawem dokumentacji), często wąskich tematów. Zasilenie gen AI eksperckimi źródłami znacząco zwiększa jego skuteczność pracy.

Poszczególne narzędzia gen AI mogą bazować na różnych pulach informacji i oferować różne funkcjonalności przydatne w celu przeglądów źródeł. W tabeli 7.3. przedstawiamy krótkie opisy platform i narzędzi gen AI, sygnalizując ich zastosowania w planowaniu przeglądu, wyszukiwaniu, analizie i syntezie tekstów.

Tabela 7.3. Zestawienie wybranych narzędzi gen AI wspierających wyszukiwanie oraz analizę źródeł tekstowych

Narzędzia gen AI	Charakterystyka
ChatGPT, Claude, Meta Llama, Gemini	Umożliwiają szybkie przetwarzanie i analizę dokumentów, generowanie streszczeń i odpowiedzi na pytania. Są też „silnikiem" – podstawą dla pozostałych platform omawianych w tej tabelce. Mogą być pomocne w planowaniu przeglądów (dostarczyć pomysły na temat typów źródeł czy procedury badawczej), zwłaszcza przy odpowiednim promptowaniu. Słabo sprawdzają się przy wyszukiwaniu źródeł. LLM nie mają bezpośredniego dostępu do aktualnych baz danych i pełnych tekstów chronionych subskrypcją lub dostęp ten jest ograniczony. Ich odpowiedzi są generowane na podstawie głównie danych, na których były trenowane, co oznacza, że mogą korzystać z publikacji sprzed pewnego okresu. Mogą „halucynować", a więc „wymyślać" nieistniejące publikacje, lub łączyć różne publikacje w jedną referencję bibliograficzną. W analizie i syntezie sprawdzają się w miarę dobrze, przy mniej specjalistycznych tematach. Mogą mieć limity dotyczące ilości przetwarzanych danych i pojawia się tu czasem problem z poprawnością odczytu i przetwarzania tekstów wgrywanych przez użytkownika.
Perplexity	To rodzaj nowoczesnej przeglądarki internetowej. Może przeszukiwać Internet i dokonywać syntezy źródeł tak znalezionych. Zaletą tej platformy jest to, że przy swoich odpowiedziach podaje dokładne źródła pochodzenia informacji. To narzędzie jest bardzo przydatne przy przeszukiwaniu różnych źródeł internetowych, w tym wszystkich treści w otwartym dostępie, a także materiałów zawartych na podstronach internetowych. Znajduje informacje i dokonuje ich syntezy – w postaci streszczeń. Odpowiedzi Perplexity, tak jak w przypadku innych gen AI, można pogłębiać, zadając dodatkowe pytania. W najnowszej wersji użytkownicy mają do dyspozycji opcję Spaces – to pozwala zasilić platformę własnymi tekstami i prowadzić analizę oraz syntezę w oparciu o własny zbiór informacji.

Narzędzia gen AI	Charakterystyka
Semantic Scholar	Semantic Scholar to platforma badawcza oparta na sztucznej inteligencji, opracowana przez organizację non-profit Allen Institute for Artificial Intelligence, służąca do nawigacji i przyspieszania wyszukiwania literatury we wszystkich dziedzinach naukowych. Jej funkcjonalność wyszukiwania jest bardzo ograniczona (tylko proste zapytania, brak możliwości wyszukiwania według zaawansowanych operatorów logicznych). Platforma jest jednak bardzo cennym zbiorem danych, z którego korzystają inne wyspecjalizowane wyszukiwarki.
Elicit, Scite, Consensus, SCISPACE	Wyszukiwarki publikacji akademickich, które udzielają syntetycznych odpowiedzi na zadane pytanie badawcze, opierają się na informacjach z publikacji otwartego dostępu. Tak więc nie tylko wyszukują źródła, ale od razu dokonują ich analizy i syntezy – w formie krótkich esejów, odpowiedzi i podsumowań artykułów. Platforma Consensus dodatkowo (przy pytaniu zadanym w formie tezy albo hipotezy) podaje odpowiedź i sygnalizuje, czy według dostępnych dla niej źródeł istnieje zgoda co do danego stwierdzenia. Wszystkie cztery platformy bazują na korpusie publikacji katalogowanym przez Semantic Scholar. Scite ma dodatkowo dostęp do wybranych periodyków naukowych i wydawców w płatnym dostępie. Główne ich ograniczenia to – co do zasady – brak dostępu do publikacji odpłatnych, niejasność co do sposobu selekcji i indeksowania wyszukiwanych tekstów. Dodatkowo gen AI może nie wychwycić wszystkich niuansów i błędnie wnioskować z treści wyszukanych tekstów.
NotebookLM, Petal AI, SCISPACE	Platformy oferujące możliwość pracy z pulą tekstów wgranych przez użytkownika. Zasadniczo oferują trzy możliwości interakcji z tekstem: 1. rozmowę z gen AI o pojedynczym źródle, 2. rozmowę z gen AI o pakiecie źródeł, 3. tworzenie przez gen AI tabel porównujących każde ze źródeł pod względem zadanych przez nas pytań. Ich zaletą jest to, że wskazują przypisy – jakie konkretne fragmenty i miejsca w dokumentach były podstawą ich odpowiedzi. NotebookLM i Petal AI pracują na dokumentach dostarczonych im przez użytkownika. SCISPACE dodatkowo może operować na zbiorze artykułów naukowych, które sama zidentyfikuje w Internecie (tylko artykuły w otwartym dostępie).

Źródło: Opracowanie własne.

7.5. Perspektywy zastosowań gen AI w przeglądach źródeł

Jak pokazano w poprzedniej części, rodzina przeglądów źródeł jest bardzo zróżnicowana. Sama praca nad wszystkimi przeglądami, niezależnie od typu, przebiega jednak według

trzech głównych etapów. Tym, co różnicuje przeglądy, jest poziom systematyczności, z jaką prowadzona jest procedura w ramach każdego z tych etapów. Trzy uniwersalne etapy to:

1. Zaplanowanie przeglądu;
2. Szukanie i selekcja źródeł;
3. Analiza i synteza źródeł.

W tej części rozdziału krótko charakteryzujemy działania podejmowane na każdym z tych etapów i skupiamy się na rozważaniach, które z tych działań możemy wesprzeć gen AI.

Zaplanowanie przeglądu

Planowanie przeglądów obejmuje trzy elementy: sformułowanie pytania badawczego, ustalenie strategii poszukiwań i selekcji źródeł oraz wybór ram koncepcyjnych.

Podstawą jest sformułowanie pytania badawczego, na które szukamy odpowiedzi. W praktyce polityk publicznych zwykle pojawiają się potocznie sformułowane potrzeby wiedzy, które trzeba przekształcić w formalne pytanie badawcze, definiując wszystkie terminy użyte w pytaniu. Trzeba podkreślić, że sprecyzowanie pytania badawczego jest absolutnie kluczowe, bo to jego kształt determinuje zakres wszystkich pozostałych działań na tym etapie.

Działaniem drugim jest sformułowanie strategii szukania i selekcji źródeł. Obejmuje ona wybór typów źródeł i baz, z których chcemy skorzystać, sekwencję poszukiwania źródeł, formuły wyszukiwania w konkretnych bazach (ang. *search query*), zakres czasowy, jaki chcemy objąć badaniem, a także sposób i kryteria selekcji zidentyfikowanych źródeł (np. wstępna selekcja abstraktów, selekcja w oparciu o pełny tekst). W skrócie, są to decyzje o tym, czego szukać, gdzie i w jaki sposób.

Trzeci element planowania przeglądu to wybór ram koncepcyjnych. To swoiste „okulary” analityczne, przez które chcemy analizować i porządkować zgromadzony materiał. Ramy mają nam docelowo pomóc w tworzeniu syntezy. Na przykład popularną ramą analityczną stosowaną do porównywania interwencji i polityk jest CIMO (*Context, Intervention, Mechanism, Outcome*) (Lemire et al., 2020). Zidentyfikowane w przeglądzie interwencje porównujemy, pisząc o kontekście, w którym zostały wdrożone, formie interwencji, mechanizmie zmiany i obserwowalnych efektach.

Generatywne AI może być potencjalnie pomocne w kilku zadaniach związanych z planowaniem przeglądów.

Precyzowanie pytania badawczego

Gen AI możemy wykorzystać przy zamienianiu potrzeb wiedzy, artykułowanych przez interesariuszy, na konkretne pytanie badawcze, które będzie prowadziło nasz przegląd źródeł tekstowych. Narzędzia gen AI (np. ChatGPT, Claude), jeśli zostaną obsadzone w roli badacza i otrzymają dobry kontekst badania oraz potrzebę wiedzy, proponują różne wersje pytań badawczych, z których możemy wybrać tę najbardziej odpowiadającą naszym potrzebom. Dodatkowo narzędzia te [Perplexity, ChatGPT, Claude] są pomocne przy wstępnym definiowaniu konkretnych terminów, które chcemy wykorzystać w pytaniu badawczym.

W naszym Projekcie 1. (patrz Aneks 1) pojawiło się wyzwanie. Ponieważ przedmiot naszych studiów jest nowy i funkcjonuje w praktyce pod różnymi nazwami, musieliśmy doprecyzować, jaki będzie główny termin poszukiwania. Używając gen AI [Perplexity], spytaliśmy o terminy pokrewne dla transformacji cyfrowej (ang. *Digital Transformation*, DT) – AI podało między innymi *4th industrial revolution, digitalization* etc. Dopytaliśmy, który z tych terminów jest najbardziej popularny. AI wskazało termin DT. To wskazanie potwierdziliśmy, porównując częstotliwość występowania terminów w Google Trends, gdzie jednoznacznie dominowało DT – *Digital Transformation*. Nasza konkluzja: DT to termin najbardziej popularny i zawierający w sobie inne, pokrewne terminy (tzw. *umbrella term*). W rezultacie ten termin zastosowaliśmy w naszym pytaniu badawczym i w dalszych wyszukiwaniach.

Pomoc przy wyborze typów źródeł

Projekt 1. wymagał sięgnięcia po bieżące (najnowsze) źródła, które identyfikują wschodzące trendy transformacji cyfrowej (jeszcze nieopisane i nieprzeanalizowane w literaturze akademickiej). Zapytaliśmy więc ChatGPT i Claude o sugestie źródeł, z których moglibyśmy czerpać informacje o nowych trendach transformacji cyfrowej. Gen AI poleciły konkretne typy i linki do źródeł: przede wszystkim trendbooki firm konsultingowych i czołowych firm technologicznych, konkretne magazyny biznesowe, technologiczne (np. MIT Sloan Management Review, Wired itd.), blogi i podcasty technologiczne.

Podobnie projekt 4. wymagał sięgnięcia po nieakademickie źródła tekstowe, ponieważ wiele z indywidualnych praktyk współprojektowania rozwiązań dotyczących problematyki marnotrawienia żywności nie jest publikowanych w literaturze naukowej. AI Perplexity polecił ciekawe, specjalistyczne bazy, jak np. Food and Agriculture Organization (FAO) Database, ReFED Insights Engine czy World Resources Institute (WRI) Food Loss and Waste Database.

W obydwu przypadkach sugestie gen AI były bardzo wartościowym punktem wyjścia do budowania strategii poszukiwań źródeł tekstowych.

Pomoc gen AI w znalezieniu ram analitycznych

Przy przeglądach interwencji publicznych przydatne są ramy analityczne, które pozwalają porządkować interwencje z różnych krajów i kontekstów w większe grupy.

Jedną z klasycznych typologii instrumentów jest ta stworzona przez M. Bemelmans-Videc, R. Rist and E. Vedung (1998). Porządkuje ona narzędzia publiczne według mechanizmu wywoływania zmian wśród adresatów polityki. Wyróżnia trzy grupy instrumentów: kije (czyli regulacje), marchewki (czyli zachęty ekonomiczne) i prawienie kazań (czyli instrumenty informacyjne i komunikacyjne).

W Projekcie 3. sprawdziliśmy, czy gen AI pytany o typologie instrumentów publicznych zaproponuje ugruntowane w literaturze typologie, w tym zacytuje typologię „kijów, marchewek i prawienia kazań" (szczegóły patrz: Aneks 2). Bazując na tym teście, uważamy, że gen AI oferuje ograniczony potencjał w identyfikowaniu ugruntowanych ram analitycznych. LLM mają tendencję do mieszania typologii lub błędnego przypisywania źródeł. RAG radzą sobie lepiej, dokładnie cytując i podsumowując literaturę, ale zwykle opierają się na wtórnych źródłach tekstowych, zamiast bezpośrednio odwoływać się do oryginalnych prac (wynika to prawdopodobnie z ograniczonego dostępu cyfrowego do starszych publikacji). Tak więc korzystanie z gen AI może być pomocne w uzyskaniu solidnych ram analitycznych, jednak propozycje AI koniecznie trzeba zderzyć z wiedzą ekspercką i własnym rozeznaniem literatury oraz testować zapytania w różnych narzędziach.

Szukanie i selekcja źródeł

Cel tego etapu to znalezienie w przestrzeni cyfrowej (Internecie, konkretnych bazach akademickich, bibliotekach i innych zbiorach cyfrowych publikacji) źródeł tekstowych, które będą przydatne do odpowiedzi na postawione pytanie badawcze.

Zwykle etap ten składa się z trzech głównych działań: 1. szukanie i gromadzenie źródeł tekstowych z Internetu lub baz, 2. wstępna selekcja (*screening*) znalezionych źródeł na podstawie ich abstraktów lub streszczeń, 3. szczegółowa selekcja źródeł opierająca się na pełnej treści danego źródła. W skrócie, na tym etapie szukamy źródeł i odrzucamy te, które w najmniejszym stopniu pasują do naszych potrzeb informacyjnych. Samą strategię poszukiwania (gdzie i jak szukamy, jakimi słowami-kluczami, jakie mamy kryteria selekcji) zaplanowaliśmy w poprzednim etapie, a teraz ją realizujemy.

Standardowe wyszukiwanie opiera się na wykorzystaniu słów kluczowych oraz operatorów logicznych (ang. *boolean operators*) w celu identyfikacji całej puli źródeł, które spełniają określone kryteria. Wprowadzamy słowa kluczowe do wyszukiwarek, a te identyfikują źródła wg słów kluczy i podają nam pełną listę źródeł pasujących do wyszukiwania, w tym informacje bibliograficzne o źródłach, ich abstrakty oraz coraz częściej nr DOI (unikalny adres internetowy, który pozwala dotrzeć do pełnego tekstu publikacji).

Zaawansowane wyszukiwarki pozwalają także budować całe formuły zapytań za pomocą operatorów – łączników (np. „AND”, który łączy różne warunki i zwraca tylko te wyniki, które spełniają oba kryteria, „OR”, który poszerza zakres wyszukiwania, oraz „NOT”, który wyklucza określone słowa lub pojęcia z wyników). Tak działają kluczowe bazy danych, takie jak Web of Science, SCOPUS, JSTOR, PubMed, które są powszechnie wykorzystywane przez badaczy i analityków na całym świecie[3].

Jedną z głównych zalet tego tradycyjnego podejścia jest jego replikowalność oraz przejrzystość procesu. Mamy więc kontrolę nad procesem selekcji źródeł tekstowych.

[3] Możliwości wyszukiwania są dużo większe niż tu sygnalizujemy. Są jeszcze co najmniej operatory bliskości – *proximity*: W/n (*Within n words*), PRE/n (*Precede by n words*) oraz znaki wieloznaczne (*wildcards*). Dodatkowo dużą różnicą baz *stricte* naukowych względem np. Google Scholar jest to, że można zaznaczyć, w jakich polach szukamy (*title/abstract/keywords/affiliation* itd.).

Możemy dokładnie opisać swoje działania i użyte bazy oraz słowa klucze, co pozwala na odtworzenie wyników przez innych badaczy.

Tradycyjne wyszukiwanie ma też ograniczenia. Często trudno opisać konkretne zagadnienie polityk publicznych wg słów kluczy – w rezultacie w wyszukiwaniu dostajemy albo tylko kilka źródeł, albo tysiące ogólnych źródeł. Ponadto ułomności baz – takie jak złe „otagowanie" publikacji wg słów kluczy, nierozpoznawalność znaków specyficznych dla języka, kwestie tłumaczeń językowych, specyficzne formatowania zapewniające zgodność WCAG (ang. *Web Content Accessibility Guidelines)* najczęściej spotykane w opracowaniach instytucji publicznych – sprawiają, że niektóre publikacje mogą być niewidoczne dla wyszukiwarek.

Gen AI wykonujące wyszukiwania semantyczne

Pierwszym możliwy sposób, w jaki możemy użyć gen AI, to zdać się całkowicie na tego typu narzędzia zarówno w wyszukaniu, jak i w selekcji źródeł. Gen AI oferują zupełnie nowy sposób wyszukiwania źródeł cyfrowych. Ma on cztery istotne różnice względem standardowego wyszukiwania, które opisaliśmy na początku tego rozdziału.

Po pierwsze, wyszukiwarki gen AI pracują na nieco innych zbiorach niż tradycyjne wyszukiwarki akademickie (szczegóły – patrz tabela 7.3.). Web of Science, Open Alex mają swoje własne zbiory publikacji obejmujące setki tysięcy periodyków. Natomiast wyszukiwarki AI pracują zwykle na korpusie źródeł zbudowanym przez Semantic Scholar (Elicit, Research Rabbit, Consensus) i źródłach otwartego dostępu (np. SCISPACE), ewentualnie umowach z wybranymi wydawcami (np. Scite). To sprawia, że jak na razie gen AI generują swoje odpowiedzi na dużo węższym zakresie źródeł. Materiał z otwartych zasobów Internetu jest też często niższej jakości i obarczony dużym szumem informacyjnym (robocze wersje publikacji, artykuły, które nie przeszły jeszcze recenzji, referaty konferencyjne, niskiej jakości periodyki), a to z kolei rodzi poważny problem, jakim jest wątpliwa wiarygodność odpowiedzi[4].

[4] Na przykład AI Consensus w przypadku pytań o to, czy szczepionki dla dzieci powodują autyzm lub czy ludzie przyczyniają się do globalnego ocieplenia, zwracał odpowiedzi utrwalające dezinformację i powielające niezweryfikowane twierdzenia (Heidt, 2023). Deweloperzy AI próbują wyeliminować te ograniczenia. W 2024 r. Consensus wprowadził rangowanie źródeł – system pokazuje, w oparciu o co zbudował odpowiedź, a dodatkowo zaznacza, ile i jakich rodzajów źródeł użył (czy dany artykuł to studium przypadku, czy systematyczny przegląd, czy jest wysoko cytowany itp.).

Druga różnica między gen AI a tradycyjnymi wyszukiwarkami to sposób wyszukiwania przez użytkownika. Zamiast używania słów kluczy zadajemy gen AI pytanie w formie pełnego zdania (gen AI często podpowiada też treść pytania). Niektóre z wyszukiwarek preferują pytania badawcze (Scite, SCISPACE) albo nawet tezy do weryfikacji (np. Consensus).

Po trzecie, wyszukiwarki gen AI różnią się – i to jest fundamentalna różnica – samym mechanizmem wyszukiwania. Tradycyjne wyszukiwarki (np. Google Scholar, Scopus, Open Alex) i inne standardowe narzędzia wyszukiwania (np. bazy OECD) wykorzystują słowa klucze do lokalizowania podobnych artykułów. Gen AI wykorzystują natomiast porównania wektorowe. Poszczególne teksty są reprezentowane przez wektory liczb, których bliskość w „przestrzeni wektorowej" odzwierciedla podobieństwo tekstów. Gen AI dobiera teksty w oparciu o poziom takich podobieństw (więcej o tym w rozdziale 1 tej książki). Potencjalnie, Gen AI może więc lepiej wychwycić sens naszego zapytania, ponieważ w wektorach jest więcej informacji na temat kontekstu niż w tradycyjnych słowach kluczach (Heidt, 2023).

Po czwarte, wyszukiwania z gen AI różnią się produktem. Tradycyjna wyszukiwarka pokaże nam listę artykułów (i ich abstraktów), które pasują do naszych słów kluczy. Taką bazę danych będziemy mogli pobrać i dalej pracować na abstraktach artykułów. Gen AI natomiast pokaże nam: a) streszczenie – odpowiedź na nasze pytanie, zbudowaną na podstawie 5–10 kluczowych w ocenie gen AI źródeł oraz b) listę tych źródeł, które wg gen AI najlepiej pasowały do naszego pytania. Dodatkowo, zwykle każde z tych pojedynczych źródeł będzie również opatrzone syntezą stworzoną przez AI (np. informacja o typie źródła, *key insights*, 1-akapitowe TL;DR)[5]. Możemy wówczas poprosić gen AI o wskazanie dalszych pasujących źródeł – wtedy pokazuje kolejne 10–20 przypadków.

Podsumowując, przy tradycyjnych wyszukiwaniach otrzymujemy surową listę źródeł tekstowych, których streszczenia musimy przeanalizować. W wyniku przeszukiwania z gen AI zbiorów internetowych lub zbiorów naukowych otrzymujemy zwykle zgrabne podsumowanie jakiegoś zagadnienia i krótką listę źródeł.

[5] TL;DR to skrót od ang. *Too long; didn't read* („za długie, nie przeczytałem"). W gwarze miejskiej używane jest w sytuacjach, kiedy coś było zbyt długie i nudne, byśmy mieli ochotę to przeczytać, albo kiedy w kilku słowach chcemy zaznaczyć streszczenie jakiegoś tekstu (naturalnie, gen AI stosują ten skrót w tej drugiej sytuacji).

To oznacza, że wyszukiwarki gen AI łączą wyszukiwanie i selekcję z analizą i syntezą, a więc etapy, które w tradycyjnej procedurze przeglądów są oddzielne. Jednak tak zintegrowany proces jest nieprzejrzysty i niereplikowalny, a konkretnie: 1. nie wiemy, w jakim zbiorze źródeł gen AI szukało i ile źródeł sprawdziło, 2. ile i na jakiej podstawie źródeł wykluczyło oraz 3. nie znamy ostatecznej wielkości subpopulacji źródeł pasującej do zapytania (wyniki gen AI ujawnia stopniowo).

Naszym zdaniem te ograniczenia praktycznie dyskwalifikują używanie semantycznych gen AI do zadania szukania źródeł tekstowych w ramach procedury przeglądu systematycznego (por. tabela 7.1.). Semantyczne wyszukiwarki mogą być jednak przydatne do pobieżnych przeglądów, wstępnego, szybkiego zorientowania się w danym w temacie lub do poszukiwań uzupełniających – dodatkowych źródeł do głównego przeglądu systematycznego. Pamiętać jednak trzeba, że ten wstępny obraz oferowany przez gen AI może być wykrzywiony, bo źródła, na których jest budowany, mogą być niereprezentatywne dla tematu, a nawet marginalne.

Gen AI wyszukujące powiązania

Platformy gen AI mogą posłużyć nam do poszukiwania powiązań między całymi grupami publikacji lub pojedynczymi publikacjami.

Na przykład Open Knowledge Maps tworzy swoje mapy na podstawie słów kluczowych, a nie centralnego artykułu i opiera się na semantycznym podobieństwie tekstu i metadanych, aby ustalić, w jaki sposób artykuły są ze sobą powiązane. Narzędzie układa sto artykułów w podobnych poddziedzinach w „bąbelki", których względne pozycje sugerują podobieństwo (Matthews, 2021).

Na polu ewaluacji i polityk publicznych takie wyszukiwania mogą być pomocne: a) do mapowania struktury danego zagadnienia polityki publicznej, b) do identyfikowania nieoczywistych powiązań między tematami albo c) do znajdowania zbiorów źródeł spoza tradycyjnej puli źródeł ewaluacyjnych, które faktycznie analizują dany problem polityki publicznej.

Z kolei Connected Papers (oparte na bazie gen AI – Semantic Scholar) identyfikuje i wizualizuje powiązania pojedynczego źródła (np. artykułu, raportu, książki) ze źródłami, na których był

on oparty, z kolejnymi źródłami, które go cytują, oraz z innymi grupami źródeł, które są z nim pośrednio – tematycznie – powiązane. Podobnie działa Research Rabbit.

Z perspektywy zastosowań w ewaluacji takie „detektywistyczne" wyszukiwania mogą być pomocne: a) w szybkim zrozumieniu, jakie badania i raporty są podstawą dla danego dokumentu – co może pomóc w identyfikacji kluczowych prac w danym obszarze polityki; b) w śledzeniu wpływu badań – zobaczeniu, jak wybrane źródło wpłynęło na dalsze publikacje, c) w analizie zależności między źródłami – jak są one ze sobą powiązane tematycznie. Mogą one pomóc w tworzeniu bardziej złożonych i wieloaspektowych ewaluacji polityk publicznych.

Pomoc gen AI w usuwaniu zduplikowanych źródeł tekstowych

Przy wyszukiwaniu źródeł z wielu baz czasochłonnym wstępnym krokiem do ich faktycznej selekcji jest usuwanie „dubli" – rekordów powtarzających się w kilku bazach. Ich identyfikacja i usunięcie nie są proste, bowiem mimo dużego podobieństwa baz w zakresie struktury rekordów, treść poszczególnych pól tego samego rekordu potrafi się różnić, np. w abstraktach i tytułach pojawiają się lub nie wielkie litery, znaki szczególne, skróty, a w opisie autorów – pełne imiona lub tylko inicjały.

Do poszukiwania i eliminacji dubli nadaje się ChatGPT. Do czata można załadować pliki z rekordami z poszczególnych baz, sformułować prośbę o zintegrowanie ich w jedną bazę i usunięcie dubli. Choć proste polecenia zintegrowania baz i usunięcia dubli nie muszą prowadzić do zadowalających wyników[6], dobre rezultaty daje polecenie: *Znajdź dla każdego rekordu w połączonej bazie, x rekordów*[7] *najbardziej podobnych pod względem treści abstraktu.* Czat stosuje do tego zadania miarę kosinusową z wektoryzacją TF-IDF. Następnie polecenie: *Usuń rekordy zduplikowane, przyjmując, że podobieństwo powyżej 0,8 oznacza duplikat*, tworzy bazę unikalnych rekordów. Zaproponowane podejście jest użyteczną alternatywą dla osób, które nie mają kompetencji pozwalających na skorzystanie z narzędzi wymagających kodowania w R czy Pythonie.

[6] Zobacz pełny opis przypadku w Aneksie 2.

[7] Ta liczba zależy od liczby baz, z których rekordy integrujemy.

Wsparcie przez gen AI filtrowania abstraktów – ranking trafności

Obiecującym narzędziem wspierającym badacza w procesie „przesiewania" rekordów na etapie analizy danych z baz indeksowych (tytuł, abstrakt, słowa kluczowe) są aplikacje do klasyfikacji tekstu oparte na algorytmach uczenia maszynowego. Istotą ich działania jest oszacowanie prawdopodobieństwa włączenia kolejnego rekordu, bazując na wcześniejszych decyzjach o włączeniu/odrzuceniu innych rekordów, dokonane przez badacza. Narzędzia te tworzą ranking wg prawdopodobieństwa i podsuwają badaczowi jako pierwsze te rekordy, które oceniają jako najbardziej trafne. Nie zastępują więc pracy człowieka, ale mogą ją ułatwić.

Przykładowo, zadanie tego typu może wykonać narzędzie ASReview, co bywa pomocne przy przeglądach quasi-systematycznych, przede wszystkim szybkich.

Wsparcie przez gen AI „przesiewania" abstraktów – ocena kryteriów włączenia

„Przesiewanie" rekordów w przeglądzie systematycznym odbywa się zwykle na podstawie analizy tytułów i abstraktów w oparciu o ustalone wcześniej kryteria włączenia. W ocenie spełnienia kryteriów wykorzystany może być ChatGPT. Po załadowaniu bazy rekordów z przykładową instrukcją: *Act as a researcher. Based on the Title (TI) and abstract (AB) examine each record and decide whether the article describe serious game used for food-related issues. Answer Yes or No in the new column added to the file*[8], otrzymujemy gotowy plik z uzupełnionymi ocenami spełnienia kryterium. Zgodność tych automatycznie generowanych ocen z ocenami ekspertów waha się w przedziale 60–70%, co można uznać za wystarczające dla przeglądu quasi-systematycznego.

Analiza i synteza źródeł

Ostatni etap pracy nad przeglądem źródeł poświęcony jest dwóm działaniom. Najpierw trzeba wykonać żmudną pracę analityczną, obejmującą indywidualną analizę każdego ze źródeł informacji. Przypomnijmy, że przy przeglądach systematycznych mogą być to setki artykułów.

[8] *Działaj jako badacz. Na podstawie tytułu (TI) i abstraktu (AB) przeanalizuj każdy rekord i zdecyduj, czy artykuł opisuje poważną grę wykorzystywaną w kwestiach związanych z żywnością. Odpowiedz „Tak" lub „Nie" w nowej kolumnie dodanej do pliku.*

Po analizie całego materiału następuje synteza, a więc próba podsumowania i wyciągnięcie całościowych wniosków z badanej populacji źródeł.

W obydwu działaniach pomagają przyjęte na wstępie ramy koncepcyjne. Określają one wymiary i listę kwestii, które chcemy wyłowić ze źródeł w toku analizy, a w syntezie sugerują strukturę narracji wniosków oraz główne tezy do weryfikacji.

Wsparcie przez gen AI analiz pełnych treści publikacji

Narzędzia gen AI mogą pomóc podczas analizy tekstów na trzy różne sposoby. Pierwszy to analiza pojedynczego tekstu. Po wgraniu np. jednego artykułu do agenta gen AI (np. Petal, NotebookLM, Perplexity – Space) możemy rozmawiać z gen AI o tym tekście i zadawać pytania odnoszące się do treści tego konkretnego tekstu. Gen AI odpowiada w oparciu o tekst źródłowy, zwykle podając odesłania do fragmentów, na których oparł swoją odpowiedź.

Sposób drugi to rozmowa o kilku, kilkunastu tekstach. Wgrywamy do narzędzia gen AI cały pakiet źródeł (np. Petal, NotebookLM do 50 źródeł), a następnie zadajemy mu pytania. W tym przypadku gen AI odpowiada w oparciu o cały pakiet tekstów źródłowych, podając odesłania do tych tekstów i fragmentów, na których oparł swoją odpowiedź.

Trzeci sposób to tabela porównawcza. Wgrywamy duży pakiet publikacji (w naszym przypadku było to ponad 130 artykułów na konkretny temat – zastosowania poważnych gier/*serious games* w polityce publicznej). Następnie formułujemy pytania porównawcze, np. a) jak dany artykuł definiuje konkretną kwestię, b) czy jest to artykuł empiryczny, c) jaka grupa docelowa była przedmiotem badania, d) jakiego kraju dotyczył artykuł etc. Następnie prosimy gen AI (np. Petal AI, SCISPACE) o stworzenie tabeli porównawczej – w kolumnach są nasze pytania (a, b, c, d), a w wierszach artykuły. Gen AI tworzy w oparciu o treść każdego z artykułów odpowiedzi do każdego z pytań. Dzięki takiej macierzy możemy porównywać artykuły pod kątem każdego z naszych pytań i na przykład zobaczyć, ile z artykułów i w jaki sposób zdefiniowało daną kwestię, które artykuły są empiryczne, jakich krajów dotyczą. Takie tabele porównawcze są szczególnie pomocne przy pracy z dużą grupą w miarę jednolitych dokumentów (np. fiszek projektowych).

Pierwsze i drugie zastosowanie gen AI do analizy treści zostało bardziej szczegółowo omówione w rozdziale książki poświęconym analizom jakościowym (rozdział 6). W tym skupimy się na trzecim sposobie analizy – testowaniu rzetelności tabel porównawczych.

Na danych z Projektu 2., które przeszły już przez etap przesiewania abstraktów, przetestowaliśmy użyteczność Petal AI w analizie porównawczej (sposób trzeci w powyższej liście) treści pełnych publikacji.

Z naszych ogólnych obserwacji wynika, że Petal gorzej radzi sobie z pytaniami o kwestie, które w artykułach opisywane są pobieżnie, niedokładnie. Ogólna ocena jest taka, że Petal nie zastępuje badacza, ale może pełnić rolę asystenta, którego trzeba sprawdzić. Weryfikacja odpowiedzi Petal zajmuje dużo mniej czasu niż samodzielne ich szukanie, co usprawnia proces analizy.

Pomoc gen AI w syntezie materiałów

W ramach Projektu 5. zdecydowano się na wykorzystanie gen AI do analizy dokumentów stanowiących podstawy prawne wdrażania funduszy unijnych. Jest to złożony i wielopoziomowy system instrumentów wsparcia, różniących się m.in. sposobem alokacji środków, źródłami finansowania czy mechanizmami zarządzania. Każdy z tych funduszy podlega odrębnym zasadom wdrażania, które szczegółowo opisane są w odpowiednich rozporządzeniach unijnych – są to wielostronicowe dokumenty prawne, które regulują kluczowe zasady funkcjonowania funduszy.

W pierwszej kolejności zidentyfikowano wszystkie rozporządzenia dotyczące zasad wdrażania funduszy. Dokumenty zostały udostępnione czatowi GPT 4.o wraz z poleceniem przygotowania syntezy informacji. Do chata wprowadzono prompt o następującej treści: *Dokonaj syntezy informacji zawartych w załączonych dokumentach w celu przygotowania porównania następujących funduszy: (1) Europejskich Funduszy Strukturalnych i Inwestycyjnych, (2) Funduszu Sprawiedliwej Transformacji, (3) Społecznego Funduszu Klimatycznego oraz (4) Instrumentu na rzecz Odbudowy i Zwiększenia Odporności w zakresie następujących kryteriów: źródło finansowania, okres wdrażania, główne cele, metoda zarządzania, beneficjenci docelowi, rodzaj udzielanego wsparcia, metoda alokacji środków.* Chat bez problemów (i błędów) przygotował zestawienie informacji wraz ze wskazaniem

numerów artykułów, w których znajdowały się poszczególne informacje, co zdecydowanie ułatwiło dalsze analizy.

7.6. Wnioski końcowe

W rozdziale omówiliśmy perspektywy wykorzystywania gen AI przy przeglądach źródeł tekstowych dostępnych w formie cyfrowej (ang. *literature reviews*).

Fundamentem naszych rozważań było stwierdzenie, że jakość przeglądów źródeł zastanych opiera się na dwóch filarach: a) puli informacji, na których budowane są wnioski (rozległość i jakość źródeł) oraz b) rzetelności przeprowadzonego procesu – jego przejrzystej, uporządkowanej i możliwej do replikacji procedurze. Dlatego pierwszą część rozdziału poświęciliśmy wyjaśnieniu szczegółów tych dwóch filarów. Przedstawiliśmy zasadnicze różnice w jakości między typami przeglądów. Naszym głównym przesłaniem jest to, że najwyższym standardem jest przegląd systematyczny, który dostarcza rzetelnych dowodów i wniosków. Przeglądy pobieżne są natomiast słabą metodą, bo ich wybiórczość i przypadkowość doboru źródeł nie pozwalają na rzetelne wnioskowanie. Pokazaliśmy także pełne spektrum rodzajów źródeł tekstowych, które mogą być wykorzystywane w ewaluacji. Krótko omówiliśmy zalety i ograniczenia każdego ze źródeł. Skupiliśmy się na wyjaśnieniu ich potencjalnej widoczności i dostępności dla platform gen AI.

Użyteczność zastosowań gen AI testowaliśmy w odniesieniu do trzech standardowych etapów prowadzenia każdego przeglądu (niezależnie od jego typu). Chodzi o: 1. zaplanowanie przeglądu, 2. szukanie i selekcję źródeł oraz 3. analizę i syntezę źródeł. Oto pięć naszych głównych obserwacji dotyczących tego, w jakich zadaniach gen AI okazało się przydatne, a w jakich nie znajduje obecnie zastosowania.

Po pierwsze, gen AI (ChatGPT, Claude) sprawdzają się – co może być zaskoczeniem – w działaniach kreatywnych i koncepcyjnych, związanych z planowaniem przeglądu źródeł: formułowaniu pytań badawczych, precyzowaniu definicji pojęć użytych w badaniu, sugerowaniu rodzajów źródeł i baz, w których można poszukiwać materiałów.

Wyjątkiem w pracy koncepcyjnej – i to nasza druga obserwacja – jest dobór ram koncepcyjnych. Adaptacja teorii, na podstawie której będzie realizowane badanie, dobór

klasycznej literatury oferującej ramy analityczne – te kwestie pozostają w gestii badaczy i gen AI tu się nie sprawdzi (prawdopodobnie z braku dostępu do literatury z zakresu polityk publicznych).

Obserwacja trzecia: semantyczny sposób wyszukiwania źródeł oferowany przez gen AI sprawdza się w sytuacji, gdy chcemy mieć pierwszy ogląd danego tematu, czyli przy wstępnym, pobieżnym wyszukiwaniu źródeł. Gen AI takie jak Perplexity mogą być wyjątkowo pomocne w poszukiwaniu źródeł internetowych (w tym materiałów umieszczanych na podstronach). Część z gen AI pomaga zidentyfikować pierwsze źródła i uzyskać wstępny ogląd tematu na podstawie publikacji naukowych. Jest to jednak ograniczone do publikacji otwartego dostępu. Ponadto trzeba sprawdzać źródła, na których gen AI buduje swoje syntezy i konkluzje (jest to o tyle łatwiejsze, że takie platformy jak Site, SCISPACE, Perplexity podają źródła i linki do nich), bo zdarza się, że są to „halucynacje".

W przypadku prowadzenia przeglądów systematycznych asystenci gen AI nie nadają się do etapu szukania i selekcji źródeł. Sama logika pracy narzędzi gen AI na obecnym etapie ich rozwoju jest odstępstwem od zasad przeglądu systematycznego[9]. Co do zasady, gen AI powiązane z RAG akademickimi nie mają też dostępu do publikacji płatnych, co uniemożliwia kompleksowe wyszukiwanie literatury. Wreszcie brak transparentności co do mechanizmu indeksowania powoduje, że użytkownicy nie mogą mieć pewności, na jakiej podstawie gen AI dobiera źródła.

Gen AI okazuje się natomiast pomocne dla wszystkich typów przeglądów (włącznie z przeglądem systematycznym) na etapie analizy i syntezy źródeł (np. Petal AI, NotebookLM). Chodzi zarówno o analizę pojedynczych tekstów (identyfikowanie i streszczanie interesujących nas wątków), jak i porównawcze analizy dużej liczby materiałów. Gen AI jest szczególnie przydatne w sytuacjach, gdy zasilimy je własnym, wcześniej wyselekcjonowanym materiałem. Trzeba jednak zaznaczyć, że choć gen AI mogą być pomocne w analizie i syntezie informacji, to nie należy od nich oczekiwać rozstrzygających wniosków ani krytycznego (rozumiejącego) podejścia do skomplikowanych i niestandardowych treści.

[9] Przykładem tego jest odsiewanie źródeł z wykorzystaniem narzędzi szacowania prawdopodobieństwa włączenia kolejnych rekordów w oparciu o wcześniejsze decyzje badacza, co sprawia, że nie możemy deklarować, że wszystkie rekordy zostały przesiane w oparciu o kryteria włączenia / wyłączenia.

Podsumowując użyteczność gen AI w przeglądach źródeł, stwierdzamy, że gen AI mogą wspomóc, ale nie zastępują badaczy. Pomoc gen AI przypomina pomoc asystenta – jego praca jest użyteczna, ale muszą to być zaplanowane przez badacza, stosunkowo proste, a przynajmniej rozbite na małe składowe zadania, poprzedzone bardzo precyzyjnymi poleceniami. Co więcej, efekty muszą być za każdym razem weryfikowane, a często poprawiane przez badacza. Pokazują to nasze przykłady formułowania zapytań do baz, usuwania zduplikowanych źródeł, analizy treści pełnych publikacji. Mimo ryzyka błędów, narzędzia gen AI sprawiają, że badacz może skupić się na ocenie i sprawdzaniu wyników, zamiast wykonywać całą pracę od podstaw, co znacząco zwiększa efektywność procesu badawczego.

Na zakończenie mamy też dwie szersze obserwacje związane z możliwymi trendami rozwoju.

Testując przez ostatni rok różne narzędzia gen AI, widzimy wyraźny trend do tworzenia zautomatyzowanych, wielozadaniowych platform. Zadanie pytania o źródła w Perplexity i uzyskanie odpowiedzi trwa kilka sekund, a łączy etapy formułowania pytania badawczego, szukania i selekcji źródeł; ‘rozmowa’ z Petal jest rozmyciem granicy między analizą i syntezą źródeł; a sformułowanie pytania w Scite czy SCISPACE i uzyskanie odpowiedzi w postaci podsumowania czterech wybranych publikacji to w zasadzie skondensowanie całego procesu przeglądu źródeł. Te możliwości gen AI są fascynujące, ale równocześnie stwarzają tylko złudzenie zaawansowania, a w praktyce odbierają użytkownikom kontrolę nad procesem badania i uniemożliwiają zweryfikowanie faktycznej wartości wiedzy wyprodukowanej w tak „spakowanych” procesach[10]. Aby uniknąć tych pułapek, zdecydowanie radzimy wykorzystywać konkretne narzędzia gen AI do konkretnych, jasno określonych zadań w ramach szerszego procesu zaprojektowanego i prowadzonego przez badaczy. To oznacza też dodatkowy koszt – wyjście poza ogólne platformy gen AI (jak ChatGPT, Claude) i zainwestowanie w bardziej wyspecjalizowane narzędzia bazujące na tych silnikach.

[10] Eksperymenty pokazują też, że takie „spakowane”, wielozadaniowe procesy prowadzone mają też niską jakość. Na przykład Gwon et al (2024) testowali wykorzystanie ChatGPT i Binga jako narzędzi wyszukiwania literatury z ogólnodostępnych baz (PubMed, Google Scholar, Cochrane Library, Clinical Trials), tzn. prosili gen AI o samodzielny dostęp do baz i identyfikowanie konkretnych źródeł. Wyniki okazały się dalece niesatysfakcjonujące – ChatGPT zaproponował 1287 źródeł, z których tylko 7 (0,5%) odpowiadało źródłom ze stanowiącego punkt odniesienia przeglądu wykonanego przez badaczy, a kolejne 18 (1,4%) mogło zostać uznane za trafne. Dużo lepiej wypadł Bing – 19 (40%) trafnych źródeł – ale to nadal za mało, by mówić o rzetelnej podstawie systematycznego przeglądu.

Druga obserwacja dotyczy możliwych scenariuszy wykorzystania gen AI w polskiej praktyce ewaluacji. Optymistyczny scenariusz jest taki, że dzięki wykorzystywaniu gen AI w najbliższych latach, a nawet miesiącach, podniesie się znacząco jakość i użyteczność przeglądów źródeł tworzonych na potrzeby polityk publicznych. Scenariusz pesymistyczny jest zaś taki, że polski rynek zaleją słabej jakości przeglądy, generowane za pomocą gen AI z ograniczonych, praktycznie przypadkowych materiałów o niskiej jakości.

To, który z tych scenariuszy się ziści, zależy od ludzi i dynamiki organizacyjnej naszego systemu, w tym od przyjmowanych przez nas standardów, autokorekty środowiska i uznania pewnych praktyk za rzetelne i użyteczne lub nierzetelne i bezwartościowe. W tym rozdziale zaproponowaliśmy typologie przeglądów i rodzaje źródeł, a także przedstawiliśmy nasze sugestie dotyczące transparentności procesu i preferowania określonych typów przeglądów. Mamy nadzieję, że ten materiał będzie pomocny dla aktorów polskiego systemu ewaluacji w uprawdopodobnieniu scenariusza optymistycznego.

Informacja o finansowaniu

Praca powstała w wyniku realizacji projektu badawczego o nr 2021/43/B/HS5/01935 finansowanego ze środków Narodowego Centrum Nauki.

Rozdział został dołączony do publikacji przez Polską Agencję Rozwoju Przedsiębiorczości, na podstawie bezpłatnej licencji niewyłącznej, nieograniczonej czasowo i terytorialnie, udzielonej PARP przez Karola Olejniczaka, Tomasza Kupca i Dominikę Wojtowicz na postawie umowy nr 3/DAS/2025/SEPR/ABK z dnia 21.03.2025.

Bibliografia

- Bemelmans-Videc, M.-L., Rist, R., Vedung, E. (red.) (1998). *Carrots, Sticks & Sermons. Policy Instruments and Their Evaluation*. Transaction Publishers (dostęp: 31.03.2025).
- Booth, A., James, M.S., Clowes, M., et al. (2021). *Systematic approaches to a successful literature review*. Sage Publications.
- Chalmers, I., Hedges, L.V., Cooper, H. (2002). A brief history of research synthesis. *Evaluation & the health professions*, 25(1), 12–37.
- Cooper, H., Hedges, L.V., Valentine, J.C. (red.) (2019). *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation.

- Glickman, M., Zhang, Y. (2024). AI and Generative AI for Research Discovery and Summarization. *Harvard Data Science Review*, *6.2*(Spring), 1–29 (dostęp: 31.03.2025).
- Grant, M.J., Booth, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. *Health information & libraries journal*, 26(2), 91–108.
- Gwon, Y.N., Kim, J.H., Chung, H.S., et al. (2024). The Use of Generative AI for Scientific Literature Searches for Systematic Reviews: ChatGPT and Microsoft Bing AI Performance Evaluation. *JMIR Med Inform*, *12*, e51187 (dostęp: 31.03.2025).
- Heidt, A. (2023). Artificial-intelligence search engines wrangle academic literature. *Nature*, *620*(7973), 456–457 (dostęp: 31.03.2025).
- Kastner, M., Antony, J., Soobiah, C., et al. (2016). Conceptual recommendations for selecting the most appropriate knowledge synthesis method to answer research questions related to complex evidence. *Journal of clinical epidemiology*, 73, 43–49.
- Matthews, D. (2021). Drowning in the literature? These smart software tools can help. Search engines that highlight key papers are keeping scientists up to date. *Nature*, *597*(7874), 141–142 (dostęp: 31.03.2025).
- Mazur, Z., Orłowska, A. (2018). Jak zaplanować i przeprowadzić systematyczny przegląd literatury. *Polskie Forum Psychologiczne*, 23(2), 235–251.
- Page, M.J., McKenzie, J.E., Bossuyt, P.M., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, 372.
- Petticrew, M., Roberts, H. (2003). Evidence, hierarchies, and typologies: horses for courses. *Journal of epidemiology & community healt*h, 57(7), 527–529.
- Sutton, A., Clowes, M., Preston, L., et al. (2019). Meeting the review family: exploring review types and associated information retrieval requirements. *Health Information & Libraries Journal*, 36(3), 202–222.
- Thomas, G. (2019). *Find your source*. Thousand Oaks, CA: Sage.
- Tricco, A.C., Soobiah, C., Antony, J., et al. (2016). A scoping review identifies multiple emerging knowledge synthesis methods, but few studies operationalize the method. *Journal of Clinical Epidemiology*, 73, 19–28.
- van Dijk, S.H.B., Brusse-Keizer, M.G.J., Bucsan, C.C., et al. (2023). Artificial intelligence in systematic reviews: promising when appropriately used. *BMJ Open*, *13*(7), e072254 (dostęp: 31.03.2025).
- Wagner, G., Lukyanenko, R., Pare, G. (2021). Artificial intelligence and the conduct of literature reviews. *Journal of Information Technology*, *37*(2), 209–226 (dostęp: 31.03.2025).
- Ye, A., Maiti, A., Schmidt, M., et al. (2024). A Hybrid Semi-Automated Workflow for Systematic and Literature Review Processes with Large Language Model Analysis. *Future Internet*, *16*(5), 167–189 (dostęp: 31.03.2025).

8. Analizy ilościowe wspierane gen AI

Kamil Filipek

8.1. Wprowadzenie: ewolucja dużych modeli językowych

Jednym z najbardziej spektakularnych przykładów postępu, jaki dokonał się w ostatnich latach w obszarze metod sztucznej inteligencji, jest rozwój tak zwanych dużych modeli językowych (ang. *Large Language Models* – LLM) (Brzeziński et al 2025). Przejście od stosunkowo prostych technik statystycznych oraz programowania regułowego[1] do kontekstowego osadzania słów (ang. *word embeddings*) przy pomocy sieci neuronowych umożliwiło wektorową reprezentację języka. Dzięki temu znacząco poprawiły się wyniki modeli językowych w obszarze: analizy wydźwięku, klasyfikacji tekstu, rozpoznawania nazw własnych (ang. *Named Entity Recognition*), tłumaczenia maszynowego.

W ostatnim czasie modele językowe coraz częściej wykorzystywane są też do analiz ilościowych w obszarze szeroko rozumianej ewaluacji (Smolić et al, 2024, Nejjar et al, 2023). Doskonale radzą sobie z generowaniem kodu, który służy do rozwiązywania prostych zadań wchodzących w zakres tzw. statystki opisowej, ale wykazują również dużą sprawność w obszarze modelowania statystycznego. Dodatkowo, generowany przez nie kod pozwala efektywnie łączyć dane pochodzące z różnych źródeł, oceniać ich jakość, proponować odpowiednie metody analityczne, a także dostarczać rzetelne interpretacje uzyskanych wyników. Ich rozwój bez wątpienia otwiera nowe możliwości, ale jednocześnie generuje problemy związane z etyką, prywatnością oraz odpowiedzialnym wykorzystaniem sztucznej inteligencji. Celem niniejszego rozdziału jest udzielenie praktycznych wskazówek: a) w jaki

[1] Programowanie regułowe (ang. *rule-based programming*) to paradygmat programowania, w którym decyzje są podejmowane na podstawie zbioru zdefiniowanych reguł. Reguły te składają się z warunków (lub przesłanek) i odpowiednich akcji (lub wniosków).

sposób dokonać wyboru właściwego modelu?, b) jak przygotować dane wejściowe?, c) jak analizować przygotowane wcześniej dane?, d) jak oceniać i interpretować uzyskane wyniki?

8.2. Wybór środowiska analitycznego

Na rynku dostępnych jest wiele modeli językowych, które można wykorzystać do mniej lub bardziej zaawansowanych ilościowych analiz ewaluacyjnych. Część modeli dostępna jest poprzez: a) gotowe interfejsy użytkownika (np. ChatGPT od OpenAI), b) interfejsy programowania aplikacji API (np. Watsonx APIs), c) platformy hostingowe, z których można pobrać artefakty modelu (np. Github) lub d) biblioteki języków programowania (np. TensorFlow Hub lub PyTorch Hub). Ze względu na rozbudowany i intuicyjny interfejs dużą popularnością cieszą się modele udostępnianie przez firmę OpenAI (ChatGPT) w ramach bezpłatnej (mniej zaawansowane modele) i płatnej subskrypcji (patrz dalej), rozwiązania firmy Anthropic (Claude) udostępnianie bezpłatnie oraz komercyjnie, czy rozwiązania firmy Google (Gemini, Gemma) udostępniane również w modelu komercyjnym i bezpłatnym. Warto jednak podkreślić, że na platformie HuggingFace.com dostępnych jest wiele bezpłatnych modeli (np. modele firm Mistral AI czy Meta), które sprawnością dorównują modelom komercyjnym.

Praca z kodem

W przypadku analiz ilościowych kluczowe znaczenie ma odpowiednia organizacja pracy z kodem programistycznym. Wybrane modele językowe oferują zaawansowane funkcje generowania i wykonywania kodu, które mogą znacząco usprawnić proces analizy danych. Część analityków wykorzystuje LLM-y (lub rozszerzenia wykorzystujące LLM-y, takie jak Copilot) do generowania kodu, który następnie kopiuje i wykonuje lokalnie lub zdalnie w kontrolowanym przez nich środowisku programistycznym (np. PyCharm, Jupyter, RStudio). Inni preferują prace z bardziej kompleksowymi interfejsami (np. ChatGPT, Claude, Gemini), które oprócz generowania kodu są w stanie zwrócić wynik w postaci tabelarycznej, graficznej lub tekstowej. Warto jednak podkreślić, że jedno i drugie podejście niesie ze sobą określone korzyści oraz zagrożenia.

Generowanie kodu. Podejście polegające na generowaniu kodu i jego wykonaniu w kontrolowanym środowisku pozwala na pełną kontrolę nad wszystkimi elementami procesu analitycznego w obszarze ewaluacji, w tym nad przygotowaniem danych,

wyborem bibliotek i konfiguracjami środowiska. Dodatkowo, co do zasady, dane pozostają na lokalnych urządzeniach, co zwiększa bezpieczeństwo i minimalizuje ryzyko wycieku wrażliwych informacji. W takim podejściu analityk ma też szerokie możliwości debugowania i optymalizacji kodu. Jest to szczególnie istotne w przypadku złożonych analiz. Podejście to wymaga jednak od analityka znajomości języka programowania, środowisk programistycznych i ich konfiguracji. Proces ten jest często bardziej czasochłonny w porównaniu z interfejsami zintegrowanymi, które oferują szybszy dostęp do wyników.

Generowanie i wykonywanie kodu. Podejście wykorzystujące kompleksowe interfejsy, takie jak ChatGPT, Claude czy Gemini, opiera się na realizacji analiz w środowisku, które nie tylko generuje kod, ale także go wykonuje i natychmiast zwraca wyniki w różnych formatach, takich jak tabele, wykresy czy teksty. Jest to rozwiązanie szybkie i wygodne, ponieważ eliminuje konieczność ręcznego przenoszenia kodu oraz uruchamiania go we wskazanym środowisku. Intuicyjność takich narzędzi sprawia, że zaawansowane analizy można prowadzić bez potrzeby zaawansowanej znajomości programowania. Co ważne, możliwość generowania wyników w wielu formatach ułatwia ich interpretację i prezentację, a zintegrowane środowisko automatyzuje wiele zadań, odciążając użytkownika. Korzystanie z takich narzędzi wiąże się jednak z pewnymi zagrożeniami, takimi jak wyciek danych w przypadku pracy na chmurowych usługach. Dodatkowo użytkownik ma ograniczoną kontrolę nad środowiskiem analitycznym, a trudność w weryfikacji generowanych procesów obliczeniowych może prowadzić do mniejszej przejrzystości analizy. Warto również zauważyć, że niektóre modele mogą nie obsługiwać zaawansowanych funkcji lub bibliotek, co ogranicza ich zastosowanie w bardziej złożonych projektach.

Warto jednak podkreślić, że zarówno w pierwszym, jak i drugim podejściu to analityk ma za zadanie zweryfikować wyniki oraz – jeśli jest to możliwe – kod, który do nich doprowadził.

Wybrane modele

Nie wszystkie dostępne na rynku LLM-y sprawdzą się dobrze w zadaniach ilościowych w obszarze szeroko rozumianej ewaluacji (Liu, 2024). Ze względu na wszechstronność modele takie jak: ChatGPT, Gemini oraz Claude będą optymalnym wyborem do szerokiego spektrum zadań analitycznych, takich jak analiza danych, budowanie modeli predykcyjnych czy generowanie raportów. W zadaniach analitycznych w obszarze ewaluacji możemy wykorzystać również inne modele, na przykład:

1. **Llama3** od firmy Meta.
   Model jest dostępny na platformie Hugging Face (różne wersje). Model, który instaluje się lokalnie, bez konieczności uruchamia skryptu programistycznego, można pobrać z tej strony. Model sprawdza się w generowaniu kodu.
2. **Gemma 2** od firmy Google.
   Model jest dostępny na platformie Google AI Studio (różne wersje modeli i różne modele). W środowisku Google AI Studio można generować kod.
3. **Mistral** od firmy Mistral AI.
   Model jest dostępny na tej stronie. Model stwarza możliwość generowania kodu.
4. **DeepSeek** od firmy DeepSeek.
   Model jest dostępny na tej stronie. Model sprawdza się w generowaniu kodu.
5. **Commad R+** od firmy Cohere.
   Model jest dostępny na tej stronie. Model stwarza możliwość generowania kodu.
6. **Perplexity** od firmy Perplexity AI.
   Model jest dostępny na tej stronie.
7. **Grok** od firmy xAI.
   Model jest dostępny na tej stronie (wymagana jest rejestracja na platformie społecznościowej X). Model sprawdza się w generowaniu kodu.

Warto podkreślić, że generowanie i wyświetlanie wykresów online bez kopiowania kodu do środowiska programistycznego (np. Jupyter Notebook) możliwe jest na platformach ChatGPT oraz Claude.

Promptowanie

Wbrew obiegowym opiniom, większości dużych modeli językowych nie można dostrajać (ang. *fine tuning*) zgodnie z klasycznym rozumieniem tego terminu, polegającym na zmianie wag w mechanizmie uwagi (ang. *attention*) (Vasvani et al., 2017). Proces tak rozumianego „fine-tuningu” jest zbyt kosztowny obliczeniowo. Istnieje jednak alternatywny sposób dostosowywania dużego modelu językowego do własnych potrzeb obliczeniowych, którym jest tzw. „kontekstualizacja”. W przeciwieństwie do „fine-tuningu” w tym podejściu nie zmienia się wag modelu. Model pozostaje w swojej oryginalnej formie, a odpowiednie wyniki uzyskuje się poprzez manipulowanie kontekstem dostarczanym modelowi. Formą kontekstualizacji dużych modeli językowych, która oprowadziła do powstania nowej profesji rynkowej, jest tzw. *prompt engineering* (White et al., 2023).

Najogólniej *prompt engineering* to formułowanie instrukcji, które prowadzą do osiągnięcia założonego celu. Sztuka tworzenia odpowiednich instrukcji obejmuje nie tylko formułowanie pytania, ale także przygotowanie informacji wejściowej w sposób, który pozwala modelowi lepiej rozpoznać kontekst, intencje użytkownika oraz dostarczyć bardziej precyzyjne, użyteczne lub szczegółowe wyniki. W przypadku analiz ilościowych warto:

- rozbić złożone zadania na etapy, aby model mógł podejść do rozwiązania sekwencyjnie;
- podać kontekst liczbowy, tzn. przygotować prompt w taki sposób, aby przedstawić dane w postaci tabelarycznej lub tekstowej (np. jako ciąg liczb z opisem);
- iteracyjnie testować rozwiązanie, co pozwoli lepiej dostosować zmienne takie jak: długość prompta, kolejność instrukcji i sposób prezentacji danych liczbowych;
- posługiwać się przykładami, które są zbliżone do oczekiwanej formy odpowiedzi lub interpretacji wyników.

8.3. Ocena i przygotowanie danych wejściowych

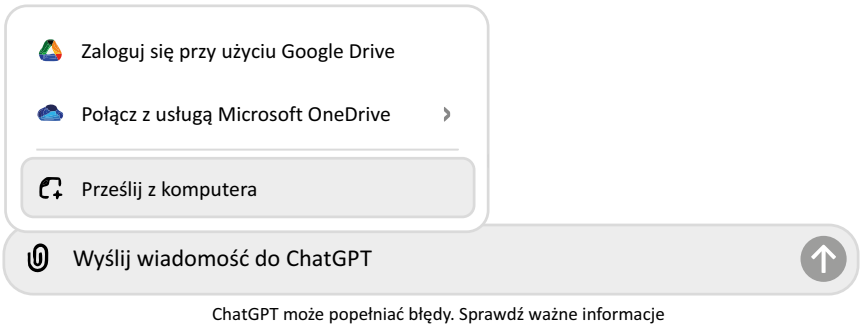
Jak już wcześniej wspomniano, niektóre LLM-y dobrze sprawdzają się w analizie danych ilościowych. Nie są to jednak narzędzia wolne od błędów, dlatego warto wypracować pewne reguły pracy z LLM (na poziomie indywidualnym, ale również zespołowym), które zminimalizują ryzyko wystąpienia niepożądanych efektów. Ważnymi krokami są: wgrywanie, ocena i korekta danych. W kolejnych podrozdziałach omówione zostaną dotąd najbardziej popularne LLM-y: a) Claude 3.5 Sonnet (bezpłatny), b) ChatGPT 4o (płatny).

Wczytywanie danych

ChatGPT w wersji płatnej (od wersji 4.0) stwarza możliwość pracy na zbiorach danych, które są dołączane do serwisu przez użytkownika. Bezpłatny ChatGPT dotąd nie dawał możliwości wgrywania własnych danych (usługa ta dostępna jest dla użytkowników wersji ChatGPT Plus). Co ważne, dane mogą być załączone w wielu formatach, np. .txt, .csv, .xls, .pdf, a nawet .png i .gif. Warto jednak pamiętać, aby dane nie były zgromadzone w zbyt dużej liczbie obiektów wejściowych. Do ChatGPT możemy jednorazowo załączyć maksymalnie 20 plików wejściowych. Załączany plik graficzny nie może przekroczyć 20 MB, plik .csv lub arkusz kalkulacyjny – 50 MB, a plik tekstowy – 2 milionów tokenów. Wartości te mają

jednak charakter zmienny. Jeśli plik jest bardzo złożony lub analiza wymaga dużej mocy obliczeniowej, zdolność modelu do szybkiej odpowiedzi może zostać bardzo ograniczona.

Rycina 8.1. Interfejs wgrywania plików do ChatGPT

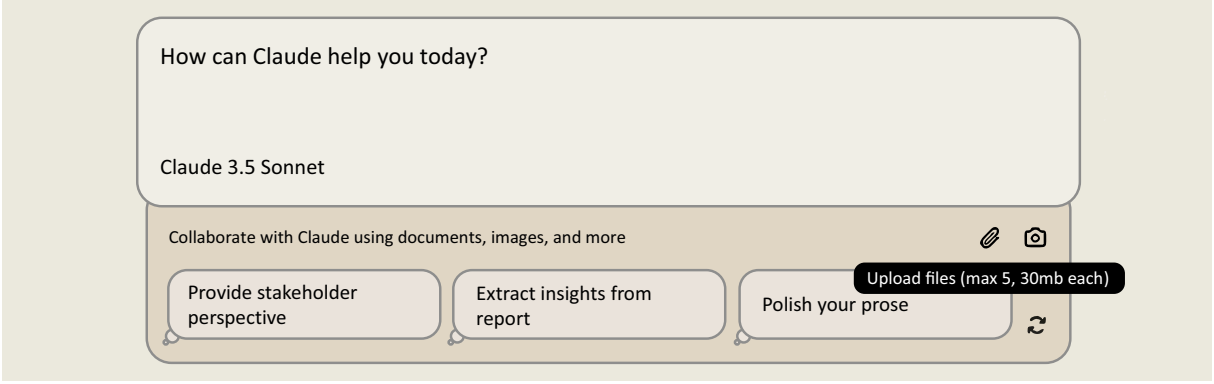


Źródło: Opracowanie własne.

Interfejs ChatGPT w wersji płatnej stwarza dodatkowo możliwość przesyłania plików z chmury: Google Drive oraz Microsoft OneDrive. Wymagana do tego jest jednorazowa zgoda na połączenie dysków chmurowych z ChatGPT.

Możliwość analizowania załączonych plików stwarza użytkownikowi bezpłatny model Claude 3.5 Sonnet. Maksymalny rozmiar załączanego pliku wynosi 30 MB (choć warto podkreślić, że limity się dynamicznie zmieniają). Można załączyć do 5 plików za jednym razem.

Rycina 8.2. Interfejs wgrywania plików Calude.ai



Źródło: Opracowanie własne.

Wstępna ocena danych wejściowych

Dane do celów ćwiczeniowych można pozyskać z wielu źródeł, przykładowo – Kaggle, GitHub, czy Otwarte Dane. W tych celach wgramy dane w uniwersalnym formacie .csv (ang. *comma separated value*) pochodzące z Banku Danych Lokalnych. Do analiz wykorzystano wskaźnik „Podmioty w działalności B+R wg sektorów gospodarczych" dla 16 województw za ostatnie 10 lat, który można wykorzystać np. w ewaluacji programów samorządu terytorialnego (źródło).

Przed wgraniem danych do modelu warto wstępnie ocenić, czy struktura pliku jest poprawna (np. czy wszystkie kolumny mają nazwy, czy dane się nie powtarzają). Przy odpowiednio sformułowanym prompcie model (w tym przypadku model Claude 3.5 Sonnet) jest w stanie zidentyfikować błędy.

Przykład prompta:

> *Chciałbym, abyś sprawdził poprawność danych w załączonym pliku pod kątem:*
> - *Kompletności: Czy w danych występują braki (np. wartości puste, NaN)?*
> - *Zakresów wartości: Czy wartości mieszczą się w oczekiwanych przedziałach (określ, jakie są dopuszczalne zakresy, jeśli znasz)?*
> - *Spójności: Czy dane są logiczne (np. brak wartości ujemnych tam, gdzie nie powinny występować)?*
> - *Typów danych: Czy typy danych w każdej kolumnie są zgodne z oczekiwaniami (np. liczby całkowite, liczby zmiennoprzecinkowe)?*
> - *Outlierów: Czy w danych występują wartości odstające, które mogą wskazywać na błędy lub wyjątkowe przypadki?*
>
> *Chciałbym otrzymać raport w punktach, wskazujący potencjalne problemy z danymi i sugestie, jak je poprawić. Jeśli to możliwe, wskaż również brakujące kroki, które powinienem podjąć, aby zwiększyć jakość danych.*

Wynikiem tak sformułowanego prompta jest szczegółowy raport oceniający dane przedstawiony na rycinie 8.3.

Rycina 8.3. Raport na tematy danych z Claude.ai

Przeanalizuję dane według podanych kryteriów i przygotuję szczegółowy raport.

1. Kompletność danych:
   - Dane są kompletne dla lat 2014–2022 dla wszystkich województw
   - Rok 2023 jest pusty dla wszystkich województw – jest to zrozumiałe, jeśli dane są w trakcie zbierania
   - Nie występują wartości NaN w pozostałych latach
2. Zakresy wartości:
   - Wszystkie wartości są dodatnie, co jest logiczne dla wskaźników na 100 tys. ludności
   - Wartości mieszczą się w przedziale od 4,3 do 31,1
   - Najniższa wartość: 4,3 (ZACHODNIOPOMORSKIE, 2014)
   - Najwyższa wartość: 31,1 (MAZOWIECKIE, 2022)
3. Spójność danych:
   - Dane są logiczne - nie ma wartości ujemnych
   - Widoczny jest generalny trend wzrostowy w większości województw
   - Zmiany rok do roku są stopniowe, bez nieuzasadnionych skoków
   - Wyjątek: PODKARPACKIE – skok z 7,4 (2014) na 14,8 (2015) – warto zweryfikować tę zmianę
4. Typy danych:
   - Wartości liczbowe są zapisane z użyciem przecinka jako separatora dziesiętnego
   - Wszystkie wartości są liczbami zmiennoprzecinkowymi z jednym miejscem po przecinku
   - Kody województw są spójne i mają format sześciocyfrowy
5. Outliery:
   - MAZOWIECKIE wykazuje konsekwentnie wyższe wartości niż pozostałe województwa (16,8–31,1)
   - ZACHODNIOPOMORSKIE ma konsekwentnie niższe wartości (4,3–10,6)
   - Wspomniany wcześniej skok w PODKARPACKIM można uznać za potencjalny outlier

Źródło: Opracowanie własne.

Co ważne, model zaproponował szczegółowe kroki, które mogą służyć poprawie danych wejściowych. W praktyce sugestie te pozwalają na ich skuteczną ocenę i modyfikację bez konieczności posiadania przez użytkownika specjalistycznej wiedzy.

Rycina 8.4. Sugestie dotyczące poprawy jakości danych z Claude.ai

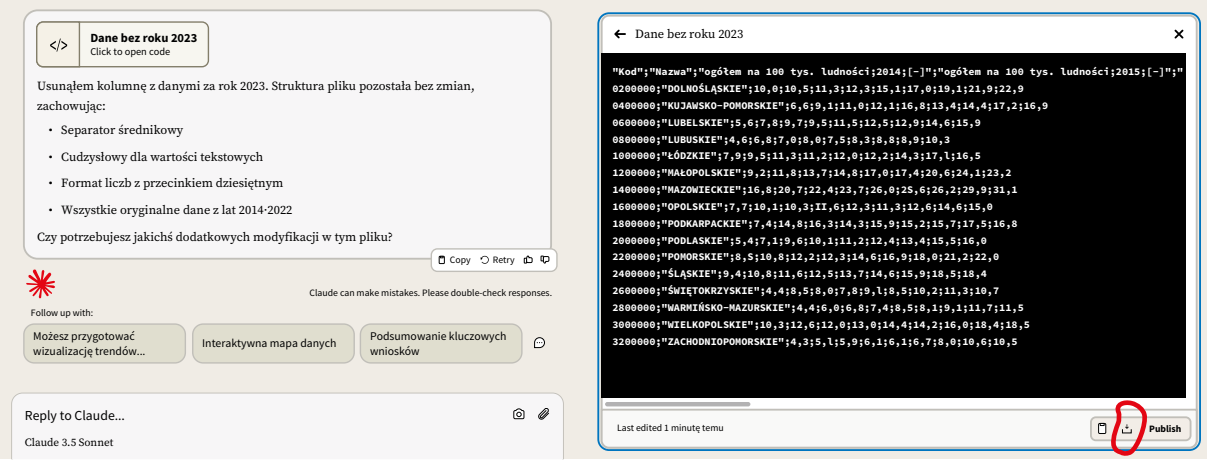
Sugestie dotyczące poprawy jakości danych:

1. Uzupełnienie brakujących danych:
   - Określić, kiedy będą dostępne dane za rok 2023
   - Ustalić harmonogram aktualizacji danych
2. Weryfikacja nietypowych zmian:
   - Sprawdzić przyczynę skoku wartości w województwie podkarpackim między 2014 a 2015
   - Zweryfikować, czy wysokie wartości dla MAZOWIECKIEGO są prawidłowe
3. Standaryzacja formatu:
   - Rozważyć użycie kropki zamiast średnika jako separatora kolumn dla lepszej kompatybilności
   - Ujednolicić format zapisu wartości liczbowych (kropka/przecinek)
4. Dodatkowe kroki dla zwiększenia jakości:
   - Dodać metadane opisujące źródło danych
   - Dodać informację o jednostce miary
   - Wprowadzić system walidacji danych przed ich publikacją
   - Dodać znaczniki jakości danych (np. dane tymczasowe, szacunkowe, zweryfikowane)
   - Wprowadzić system kontroli wersji dla kolejnych aktualizacji danych
5. Dokumentacja:
   - Stworzyć dokumentację opisującą metodologię zbierania danych
   - Dodać informację o potencjalnych ograniczeniach w interpretacji danych
   - Opisać procedury weryfikacji i kontroli jakości danych

Źródło: Opracowanie własne.

W kolejnym korku można poprosić model o przygotowanie poprawionego pliku. W tym przypadku poprosiliśmy tylko o usunięcie kolumny z danymi za 2023 rok.

Rycina 8.5. Zapis poprawionego pliku Claude.ai



Źródło: Opracowanie własne.

Zmienione dane można pobrać i przejść do kolejnych kroków analizy. Co ciekawe, model płatny ChatGPT 4o nie zidentyfikował braków danych w analizowanym pliku przy zastosowaniu identycznego prompta.

Imputacja danych

Istnieją bardziej zaawansowane metody poprawy jakości danych wejściowych, a także techniki ich uzupełniania oraz modyfikacji (tzw. metody imputacji). Odpowiednio sformułowany prompt pozwala identyfikować i uzupełniać brakujące rekordy.

Dla celów ćwiczeniowych przygotowano plik .csv z 300 rekordami, w których istnieją braki. Dane zawierają następujące kolumny: LP, Imię, Nazwisko, Zarobki, Wykształcenie, Miejsce zamieszkania, Zawód. W prompcie poproszono model o zidentyfikowanie braków danych w pliku, zaproponowanie metody imputacji i uzupełnienie bazy danych.

Przykład prompta:

> *Zidentyfikuj braki danych w załączonym pliku. Zaproponuj metodę imputacji i uzupełnij braki danych. Uzupełniony plik zapisz jako .csv.*

Tym razem wykorzystano model ChatGPT 4o. Model zidentyfikował braki danych (rycina 8.6).

Rycina 8.6. Braki danych zidentyfikowane przez ChatGPT 4o

Plik zawiera 8 kolumn i 300 wierszy, z czego kilka kolumn zawiera brakujące dane:

- Zarobki: 68 braków
- Wykształcenie: 52 braki
- Wiek: 58 braków
- Miejsce zamieszkania: 32 braki
- Zawód: 47 braków

Źródło: Opracowanie własne.

Zgodnie z instrukcją, model zaproponował metody uzupełniania danych.

Rycina 8.7. Kroki służące poprawie jakości danych wg ChatGPT 4o

Proponowane metody imputacji danych:

1. Zarobki (ciągła zmienna): Imputacja medianą, aby zminimalizować wpływ odstających wartości.
2. Wykształcenie (kategoryczna zmienna): Imputacja najczęściej występującą kategorią (modą).
3. Wiek (ciągła zmienna): Imputacja medianą.
4. Miejsce zamieszkania (kategoryczna zmienna): Imputacja modą.
5. Zawód (kategoryczna zmienna): Imputacja modą.

Źródło: Opracowanie własne.

Poprawiony model został przygotowany do zapisu.

Rycina 8.8. Zapis poprawionych danych z ChatGPT 4o

Braki danych zostały uzupełnione, a zaktualizowany plik zapisano jako CSV. Możesz go pobrać tutaj:

Braki_danych_imputowane.csv [>-]

Źródło: Opracowanie własne.

8.4. Analiza danych

Wybrane modele językowe (LLM-y) mają zdolność do generowania kodu, który pozwala wykonać zarówno proste statystyki opisowe, jak i zaawansowane modele statystyczne. Chodzi tutaj o miary podstawowe, takie jak średnia czy mediana, ale również bardziej złożone modele statystyczne, takie jak regresja wieloraka, modele klasyfikacyjne czy modele oparte na architekturze sieci neuronowych.

Statystyki opisowe

Bez wątpienia obszar, w którym efektywność modeli językowych w zakresie generowania kodu do analiz ilościowych jest relatywnie duża, to statystyka opisowa. W praktyce chodzi tutaj o miary tendencji centralnej (średnia, mediana, modalna), miary rozproszenia (odchylenie standardowe, wariancja, rozstęp, kwartyle) oraz miary kształtu rozkładu (skośność i kurtoza). Imputowany wcześniej zbiór przeanalizowany zostanie przy wykorzystaniu trzech wymienionych grup miar. Poniżej znajduje się przykładowy prompt, uruchamiający prace modelu w tym zadaniu:

> *Przygotuj statystyki opisowe dla 2 kolumn w załączonym zbiorze danych: Zarobki, Wiek. Chodzi o następujące statystyki: miary tendencji centralnej (średnia, mediana, modalna), miary rozproszenia (odchylenie standardowe, wariancja, rozstęp, kwartyle) oraz miary kształtu rozkładu (skośność i kurtoza). Dodaj je jako wiersze do wymienionych kolumn i eksportuj wynik jako oddzielny plik .csv.*

Poniżej zaprezentowano wynik analiz wykonanych przez model:

Tabela 8.1. Statystyki opisowe z ChatGPT 4o

Miara	Zarobki	Wiek
Średnia	6373,33	43,25
Mediana	5000,00	45,00
Modalna	5000,00	45,00
Odchylenie standardowe	2738,3	10,13
Wariancja	7500624,30	102,70

Miara	Zarobki	Wiek
Rozstęp	9000,00	35,00
Q1	5000,00	35,00
Q3	7500,00	50,00
Skośność	0,83	-0,19
Kurtoza	-0,46	-0,80

Źródło: Opracowanie własne.

Modele statystyczne

Zbiór, którym posługiwaliśmy się dotychczas, nie jest na tyle rozbudowany i zróżnicowany, aby tworzyć na jego podstawie ciekawe modele statystyczne. Do budowania modeli statystycznych wykorzystamy dane dotyczące rejestracji marek samochodów osobowych w okresie styczeń–wrzesień 2024. Dane dostępne są na stronie portalu Otwarte Dane. Dane zostały pobrane w formie plików .csv dla każdego miesiąca i połączone w jedną całość w celu dalszych analiz.

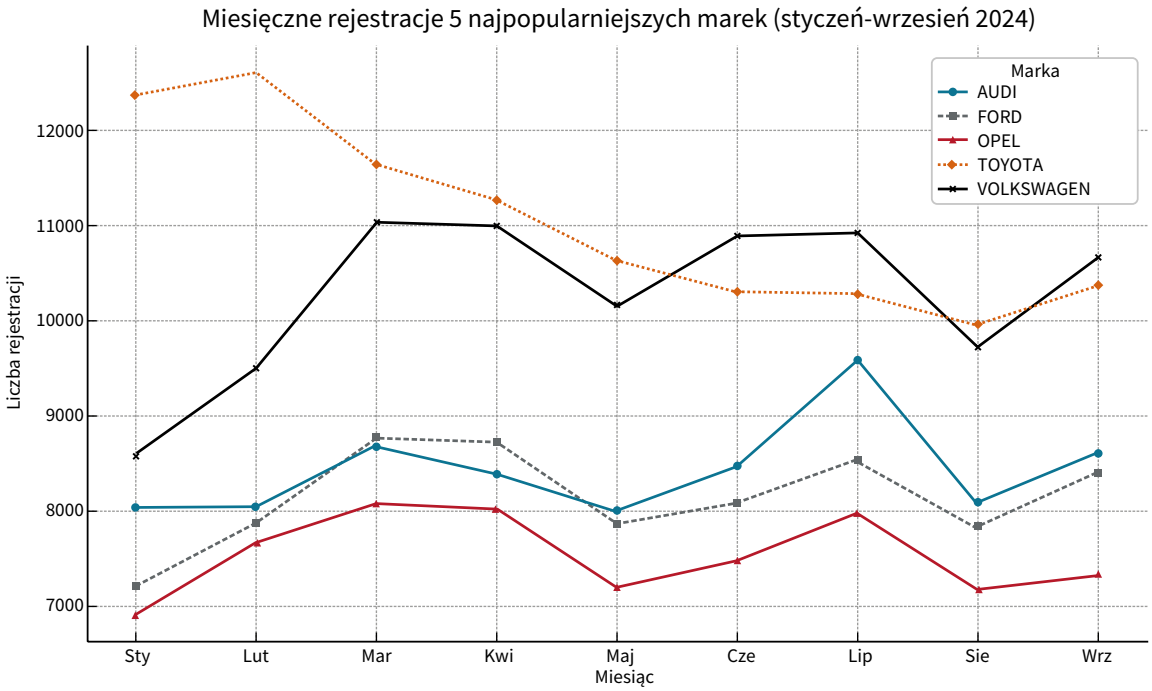
Dane połączono, wykorzystując do tego model ChatGPT 4o. Poniżej prompt:

> *Połącz dane w jeden plik. Dodaj kolumnę, która zawiera informacje o miesiącu, w którym nastąpiła rejestracja. Oznacz miesiąc cyfrą.*

Spróbujmy lepiej zrozumieć analizę, prosząc model gen AI o wykres rejestracji dla pięciu najpopularniejszych marek samochodów w Polsce:

> *Przygotuj wykres rejestracji 5 najbardziej popularnych marek na jednym wykresie dla okresu od stycznia do września 2024. Wprowadź kolory, które się odróżniają od siebie, a także odmienne kształty linii na wykresie.*

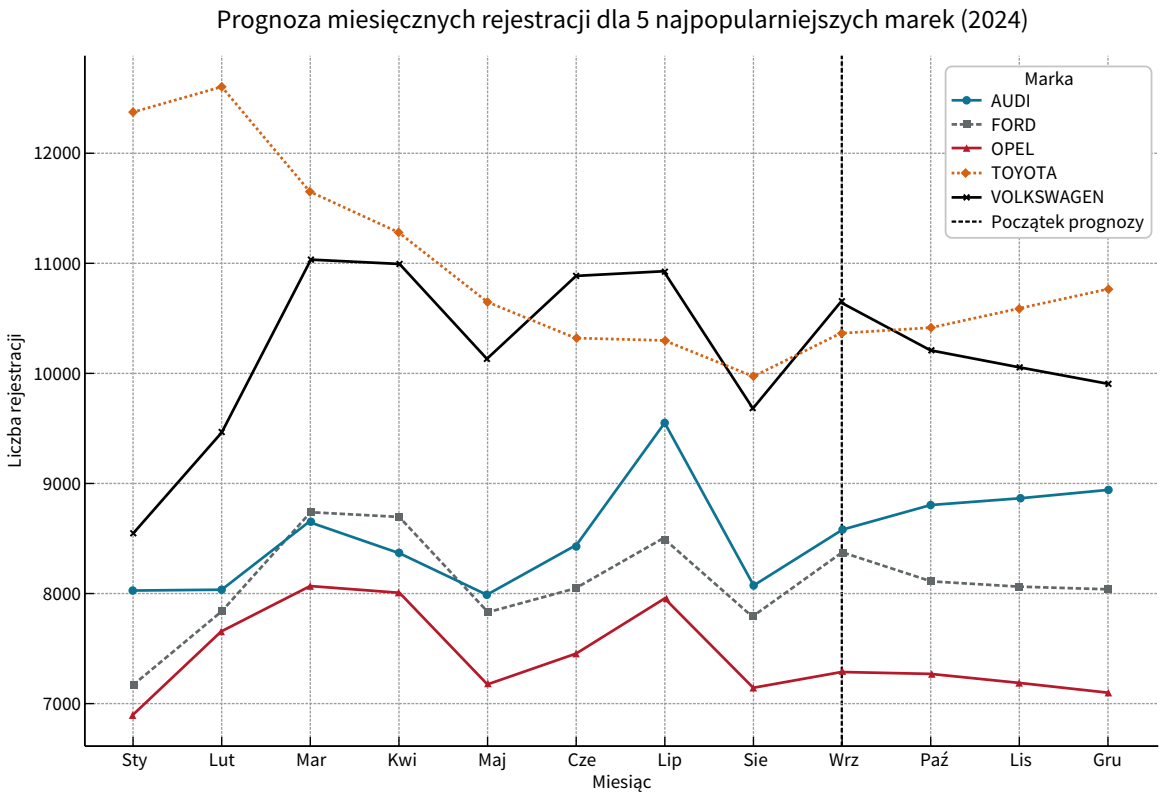
Rycina 8.9. Rejestracja pięciu najpopularniejszych marek samochodów w Polsce (styczeń–wrzesień 2024). Wykres pozyskany z ChatGPT 4o



Źródło: Opracowanie własne.

Następnie na podstawie danych z dziewięciu miesięcy przygotowana zostanie prognoza sprzedaży dla pięciu wybranych marek na kolejne trzy miesiące. Taki model statystyczny nie będzie dokładny, pomijamy w nim bowiem dane historyczne z poprzednich lat (w szczególności obrazujące tendencje wskaźnika w ostatnim kwartale roku), które mogłby wpłynąć na jakość prognozy. Sprzedaż w końcówce roku w wybranych grupach towarowych zazwyczaj bowiem nie odzwierciedla tendencji sprzedażowej z pierwszych trzech kwartałów. Warto pamiętać, że koniec roku cechuje zwykle większa sprzedaż samochodów. Wynika to zwykle z wyprzedaży rocznika oraz aktywności firm, które decydują się na zakupy przed zamknięciem roku obrotowego.

Rycina 8.10. Prognoza rejestracji pięciu najpopularniejszych marek samochodów w Polsce (październik–grudzień 2024). Wykres wykonany z ChatGPT 4o



Źródło: Opracowanie własne.

Kluczowe jest jednak zrozumienie, w jaki sposób powstała prognoza, czyli jaki model prognostyczny wykorzystał ChatGPT 4o. Formułujemy zatem prompt:

W jaki sposób dokonałeś prognozy? Jaki model prognostyczny został wykorzystany?

Model ChatGPT 4o wyjaśnił metodę, którą zastosował. Co więcej, w odpowiedzi znalazło się uzasadnienie wyboru metody.

Rycina 8.11. Wyjaśnienie metody prognostycznej przez model ChatGPT 4o

Do prognozy wykorzystałem model wygładzania wykładniczego (Exponential Smoothing), który jest jednym z klasycznych modeli szeregów czasowych. Konkretnie, zastosowałem wersję z trendem addytywnym (ang. `trend='add'`) bez sezonowości (`seasonal=None`), ponieważ dane historyczne obejmują zbyt krótki okres (tylko 9 miesięcy), aby dokładnie modelować wzorce sezonowe.

Dlaczego model wygładzania wykładniczego?

1. Szerokie zastosowanie: Model ten jest powszechnie stosowany w prognozowaniu szeregu czasowego o umiarkowanej zmienności, zwłaszcza gdy dane historyczne są ograniczone.
2. Obsługa trendu: Dzięki parametrowi `trend='add'` model uwzględnia zmiany w wartościach rejestracji w czasie (np. wzrost lub spadek w poszczególnych miesiącach).
3. Prostota i skuteczność: Exponential Smoothing jest modelem stosunkowo prostym, ale skutecznym w przypadku krótkoterminowego prognozowania.

Kluczowe kroki w prognozowaniu

1. Przygotowanie danych: Dane dla każdej marki zostały zagregowane na poziomie miesięcznym.
2. Modelowanie: Dla każdej marki utworzono oddzielny model Exponential Smoothing.
3. Dopasowanie modelu: Model został dopasowany do danych historycznych (styczeń-wrzesień).
4. Prognoza: Wykorzystano model do prognozy liczby rejestracji na październik, listopad i grudzień.

Dodatkowe uwagi

- Model nie uwzględnia czynników zewnętrznych, takich jak zmiany w preferencjach konsumentów, polityce cenowej czy wpływ zdarzeń makroekonomicznych.
- Jeśli dostępne byłyby dane z kilku lat, można by zastosować bardziej zaawansowane modele, np. ARIMA, SARIMA, lub modele uczenia maszynowego (np. LSTM, Prophet).

Źródło: Opracowanie własne.

Możliwe jest również uzyskanie informacji na temat matematycznych podstaw modelu. Warto podkreślić, że w zależności od modelu możliwe są różne metody i podejścia do jego

wyjaśnienia (np. SHAP, LIME, macierz pomyłek). Poniżej znajduje się przykładowe pytanie – prompt:

Jakie są matematyczne podstawy modelu?

Rycina 8.12. zawiera wyjaśnienie matematycznych podstaw modelu.

Rycina 8.12. Wyjaśnienie podstaw matematycznych metody prognostycznej przez model ChatGPT 4o

Matematyczne komponenty modelu

1. Poziom (level) (lt):

$$l_t = \alpha(y_t) + (1 - \alpha)(l_{t-1} + b_{t-1})$$

Poziom (l_t) odzwierciedla podstawową wartość szeregu w danym momencie. Względne znaczenie bieżącej obserwacji (y_t) i trendu w poprzednim kroku jest regulowane przez parametr $\alpha \in (0, 1)$.

2. Trend (trend) (b_t):

$$b_t = \beta(l_t - l_{t-1}) + (1 - \beta)b_{t-1}$$

Trend (b_t) opisuje zmiany poziomu w czasie. Parametr $\beta \in (0, 1)$ kontroluje wrażliwość trendu na nowe dane.

3. Prognoza (forecast) ($\hat{y}_{t+h}$):

$$\hat{y}_{t+h} = l_t + hb_t$$

Prognoza na h-ty krok w przyszłość to suma obecnego poziomu (l_t) i h-krotnego trendu (b_t).

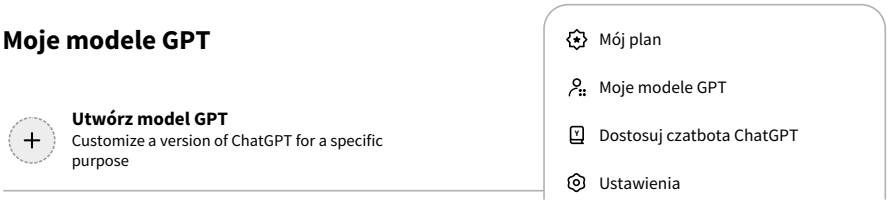
Źródło: Opracowanie własne.

Automatyzacja analiz

Analitycy często korzystający z ChatGPT w wersji płatnej mają możliwość stworzenia własnego modelu GPT (ang. *custom GPTs*). Własny model to dostosowana do specyficznych potrzeb wersja chatu, którą definiuje użytkownik. W praktyce tworzenie własnej instancji polega na umieszczeniu szczegółowych instrukcji, zgodnie z którymi powinien działać LLM. W opcjach definiujących zachowania takiej instancji można zaznaczyć funkcje „Interpreter

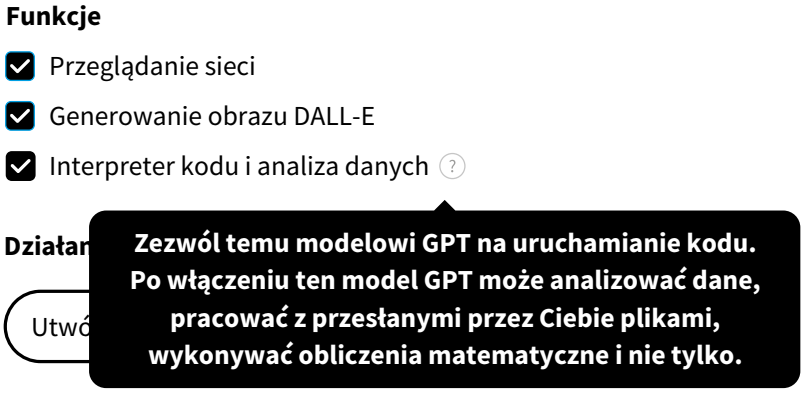
kodu i analiza danych”. W przypadku modelu GPT4o „Interpreter kodu i analiza danych” są opcją domyślną.

Rycina 8.13. Wybór własnych (*custom*) modeli GPT



Źródło: Opracowanie własne.

Rycina 8.14. Cechy własnego (*custom*) modelu GPT



Źródło: Opracowanie własne.

Pracując z interpreterem kodu, najlepsze wyniki można osiągnąć, stosując proste instrukcje, pozwalające modelowi szybko dostosować się do naszych potrzeb. Oto wybrane z nich:

- Formułuj precyzyjne pytania dotyczące analiz (np. zamiast pytać *Jak zrobić wykres?*, możesz napisać: *Jak stworzyć wykres liniowy w R, który pokazuje zmiany temperatury w czasie?*);
- Złożone zadana podziel na mniejsze kroki (np. zacznij od wczytania danych, potem przejdź do ich czyszczenia, a na końcu – do analizy.);
- Jeśli czegoś nie rozumiesz, poproś LLM o wyjaśnienie (jeśli jest to możliwe, prosząc model o analizę, posługuj się przykładami);
- Pracuj z modelem krokowo i dziel kod na bloki;
- Proś o alternatywne rozwiązania problemu w celu porównania wyników;
- *Wyjaśnij mi, co robi ten kod; jakie są błędy w poniższym kodzie*.

Rycina 8.15. Definiowanie własnego modelu w interfejsie ChatGPT

Źródło: Opracowanie własne.

8.5. Ocena wyników

Najbardziej efektywnym sposobem weryfikacji wyników jest szczegółowa analiza kodu przygotowanego przez duży model językowy (Moshin et al., 2024). Kod zwrócony przez niektóre LLM-y może być modyfikowany i dostosowywany przez użytkowników online (na platformie analitycznej) lub we wskazanym środowisku analitycznym.

Interpretacja

Modele językowe wykorzystywane do analiz ilościowych pozwalają szczegółowo śledzić ich pracę. Po wykonanych analizach warto sformułować prompt, który pozwoli prześledzić szczegółowe kroki modelu:

> *Wygeneruj kod w Pythonie, który pokazuje wszystkie kroki zastosowane w predykcji sprzedaży. Opisz każdy krok bardzo szczegółowo, tak aby osoba z niewielką znajomością Pythona była w stanie zrozumieć wykonaną analizę.*

Jak widać w przedstawionej wyżej instrukcji, wyjaśnienie kodu adresowane jest do osób, które nie dysponują wiedzą z zakresu inżynierii oprogramowania. Poniżej przeanalizowano wybrany fragment wyjaśnienia kodu, na podstawie którego powstała zaprezentowana wyżej predykcja sprzedaży pięciu najpopularniejszych marek samochodów w Polsce.

Rycina 8.16. Opis kodu przygotowany przez ChatGPT.

Skopiuj kod

```
# KROK 3: Wizualizacja prognozy
# Tworzenie wykresu dla każdej marki
plt.figure(figsize=(12, 8))

# Definiowanie unikalnych kolorów, stylów linii i markerów
line_styles = ['-', '--', '-.', ':','-']
markers = ['o', 's', '^', 'D', 'X']
colors = ['blue', 'green', 'red', 'orange', 'purple']

# Rysowanie danych dla każdej marki
for i, (brand, forecast) in enumerate (forecasts.items()):
    plt.plot(
        range(1, 13),  # Oś x: miesiące od stycznia (1) do grudnia (12)
        forecast values,  # Oś y: liczba rejestracji
        linestyle=line_styles[i % len(line_styles)], # Styl linii
        marker=markers[i % len(markers)], # Marker
        color=colors[i % len (colors)], # Kolor
        label=brand  # Etykieta marki
    )

# Dodanie pionowej linii, która oddziela dane historyczne od prognozy
plt.axvline(x=9, color='gray', linestyle='--', label='Początek prognozy')
```

Źródło: Opracowanie własne.

W celu lepszego zrozumienia kodu warto skopiować wygenerowany przez model językowy skrypt do preferowanego przez użytkownika środowiska programistycznego (np. Jupyter Notebook, PyCharm, VS Code, Google Colab). Podejście takie pozwala: a) lepiej zrozumieć działanie kodu, b) modyfikować analizę, c) testować różne scenariusze.

W przypadku pracy z danymi wrażliwymi warto rozważyć strategię, w której analizy wstępne oraz eksploracje przeprowadza się na próbkach danych w ChatGPT. Podejście takie pozwala na szybkie testowanie różnych strategii oraz stosowanie sekwencji zapytań (iteracji). Z kolei szczegółowe analizy (np. na dużych zbiorach danych) mogą być już realizowane lokalnie, w kontrolowanym środowisku, które gwarantuje bezpieczeństwo i poufność danych. Na przykład w modelu LLM można zaimplementować małą próbkę zanonimizowanych danych w celu przetestowania kodu, a po jego dopracowaniu – zaaplikować kod do pełnego zbioru lokalnie, korzystając z narzędzi takich jak Python lub R. Taki sposób pracy zapewnia zgodność z wymogami ochrony danych i pozwala lepiej optymalizować zasoby obliczeniowe.

Rycina 8.17. Analiza kodu w lokalnym środowisku Jupyter Notebook

Szczegóły skrypt z wyjaśnieniem

In [1]:
```
# Importowanie bibliotek
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.tsa.holtwinters import ExponentialSmoothing
```

KROK 1: Wczytanie danych i wstępna analiza

In [2]:
```
# Ścieżka do pliku CSV (zastąp 'dane.csv' swoją nazwą pliku)
file_path = 'MARKI_OSOBOWE_2024.csv'

# Wczytanie danych do DataFrame
data = pd.read_csv(file_path)
```

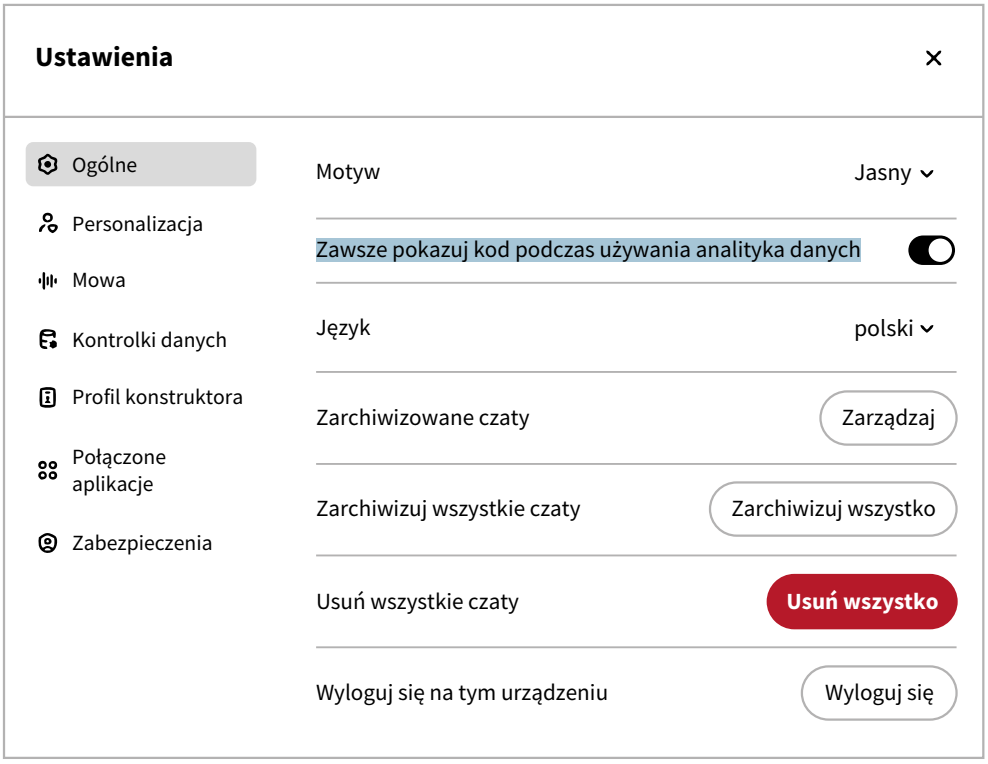
In [3]:
```
# Załaduj dane (plik powinien być wcześniej wczytany jako "data")
# Dane muszą zawierać kolumny: 'MARKA','Miesiąc Rejestracji','LICZBA'

# Filtrowanie danych dla miesięcy od stycznia do września
data_filtered = data[data['Miesiąc Rejestracji'].between(1,9)]

# Agregowanie danych według marek i miesięcy
# Pozwala obliczyć sumaryczną liczbę rejestracji dla każdej marki w każdym miesiącu
monthly_data = data_filtered.groupby(['MARKA','Miesiąc Rejestracji'])['LICZBA'].sum().unstack(fill_value=0)

# Wybór 5 najpopularniejszych marek (tych, które miały najwięcej rejestracji w okresie analizowanym)
top_5_brands = monthlLdata.sum(axis=l).sort_values(ascending=False).head(5).index
top_5_data = monthly_data.loc[top_5_brands]
```

KROK 2: Przygotowanie modelu prognozy

Źródło: Opracowanie własne.

Warto również pamiętać, korzystając z modelu ChatGPT, że dostępne jest ustawienie systemowe, które zawsze pozwala dotrzeć do kodu źródłowego. W tym celu należy kliknąć na profil użytkownika/ustawienia i włączyć opcję „Zawsze pokazuj kod podczas używania analityka danych”.

Rycina 8.18. Włączanie opcji generowania kodu w ChatGPT



Źródło: Opracowanie własne.

Wizualizacja danych

Wygodnym i efektywnym sposobem weryfikacji wyników dostarczonych przez model gen AI (LLM) jest ich wizualizacja (Vázquez, 2024). Pozwala ona na szybkie zidentyfikowanie wzorców, anomalii oraz potencjalnych błędów w dostarczonych wynikach. Wizualizacja pozwala lepiej zrozumieć złożone relacje i zależności występujące w danych, co wydaje się być szczególnie ważne w kontekście interpretowalności modeli i transparentności procesów generatywnych. W praktyce wizualizacja może obejmować np. przedstawianie rozkładu odpowiedzi modelu, analizę cech mających największy wpływ na wyniki (np. przy użyciu metod SHAP) czy analizę korelacji. Dzięki temu można łatwiej ocenić zgodność wyników (z rzeczywistością, z oczekiwaniami użytkownika itp.) oraz skutecznie identyfikować obszary wymagające poprawy.

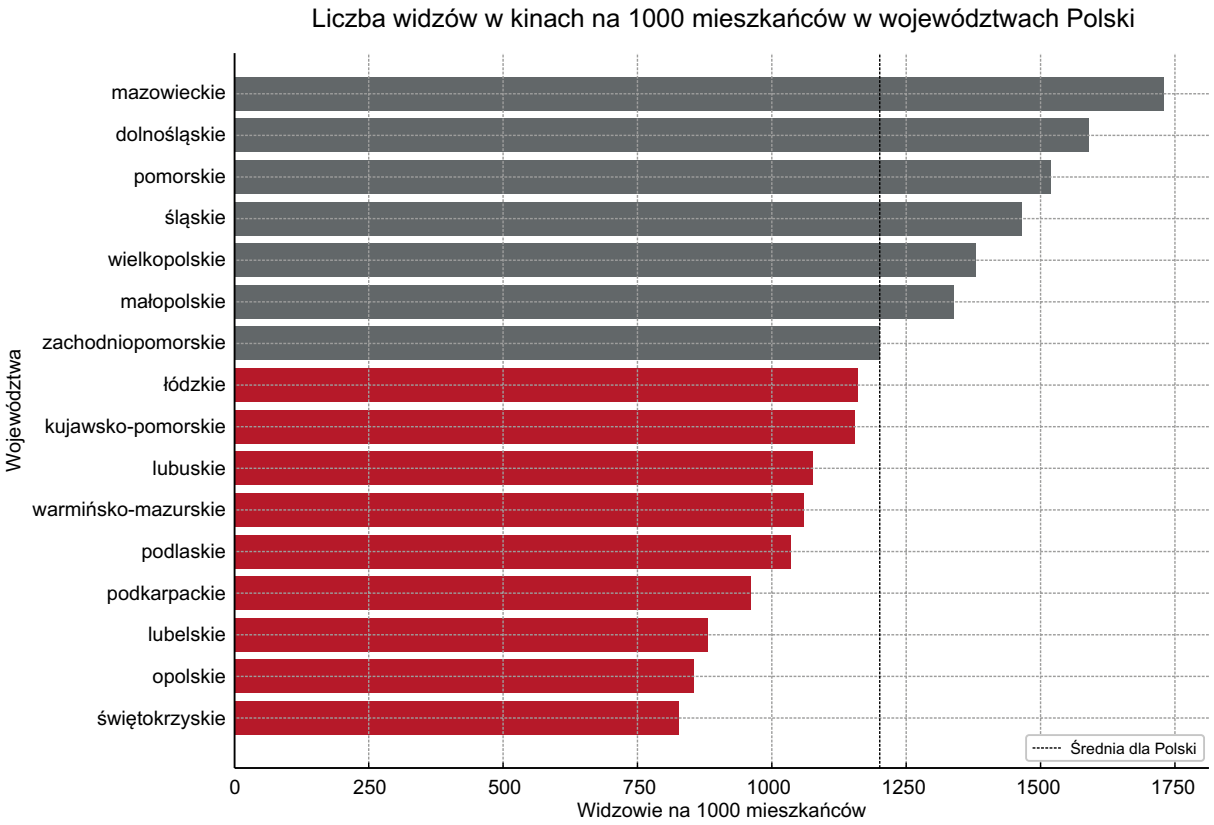
W niniejszym podrozdziale posłużono się danymi z Banku Danych Lokalnych, które ukazują liczbę widzów kin na 1000 mieszkańców w polskich województwach w 2023 roku. Do dwóch

modeli ChatGPT 4o i Claude 3.5 Sonnet wpisano taki sam prompt i załączono identyczne dane w formacie csv:

> *Przygotuj wykres słupkowy dla danych w pliku. Województwa, które uzyskały wynik wyższy od średniej dla Polski, oznacz na niebiesko. Te z niższym wynikiem oznacza na czerwono.*

Model ChatGPT 4o zwrócił poniższa wizualizację.

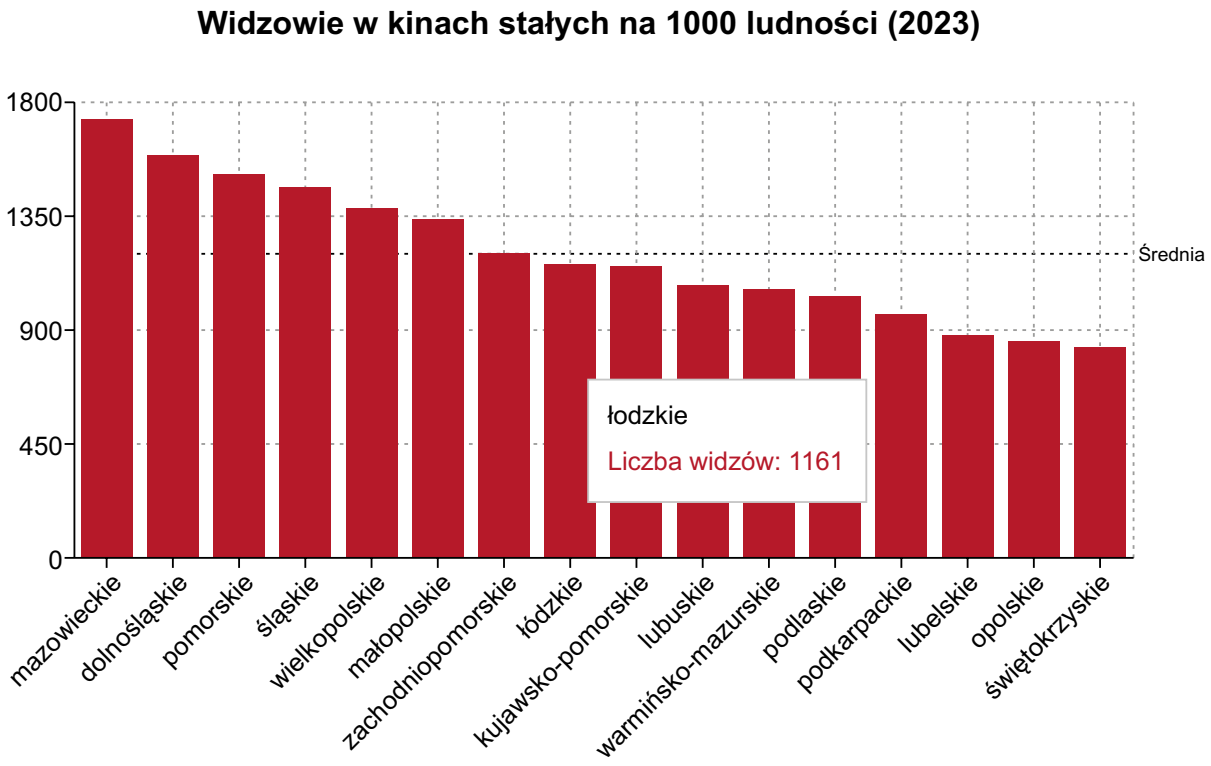
Rycina 8.19. Wizualizacja wyników w modelu ChatGPT 4o



Źródło: Opracowanie własne.

Model Claude 3.5 Sonnet nie dodał kolorów do wykresu, ale udostępnił interaktywną grafikę.

Rycina 8.20. Wizualizacja wyników w modelu Claude 3.5 Sonnet



Źródło: Opracowanie własne.

Testy

W analizie danych ilościowych ważna jest ocena poprawności i jakości uzyskanych wyników (Spiess et al., 2024). Bez odpowiednich testów i metryk trudno określić, czy model statystyczny lub analiza spełniają swoje zadanie i dostarczają wiarygodnych informacji. Testy i metryki pełnią zatem rolę „narzędzi diagnostycznych" – pomagają sprawdzić, czy uzyskane wyniki są dokładne, stabilne i użyteczne. Ich wybór zależy od rodzaju analizy oraz celu badania. Na przykład w prognozowaniu przyszłych wartości istotne są metryki mierzące precyzję przewidywań (np. średni błąd absolutny – MAE), natomiast w klasyfikacji kluczowe będzie sprawdzenie, jak często model poprawnie przypisuje etykiety (np. *accuracy, precision, recall*).

W przypadku oceny kodu wygenerowanego przez model o charakterze analitycznym i ilościowym szczególnie istotne jest sprawdzenie, czy kod poprawnie realizuje założenia analizy oraz generuje wyniki zgodne z oczekiwaniami. W tym celu można wykorzystać:

- **Testy jednostkowe.** Weryfikują poprawność działania poszczególnych funkcji obliczeniowych lub analitycznych. Przykładowo, test jednostkowy może sprawdzać, czy funkcja obliczająca średnią, medianę lub odchylenie standardowe zwraca poprawne wyniki dla znanych, kontrolowanych danych wejściowych.
- **Testy integracyjne.** Sprawdzają, czy różne moduły analityczne poprawnie współpracują w ramach całego procesu analizy. Na przykład test integracyjny mógłby sprawdzić, czy dane przechodzą poprawnie przez kolejne etapy sekwencji analitycznej, takie jak: wczytywanie danych, czyszczenie, transformacja, modelowanie i raportowanie wyników.
- **Testy wydajnościowe.** Ich celem jest upewnienie się, że algorytmy obliczeniowe działają efektywnie nawet przy dużych zbiorach danych. Na przykład test wydajnościowy może ocenić czas działania funkcji agregujących dane, takich jak obliczanie sum lub średnich, na zbiorze danych z milionami rekordów. Można także testować algorytmy bardziej zaawansowane, takie jak analizy klastrowania lub modele regresyjne, aby ocenić, czy są w stanie szybko przetworzyć dane i zwrócić wyniki.
- **Testy walidacji wyników.** Celem takich testów jest sprawdzenie, czy wyniki generowane przez model analityczny są zgodne z oczekiwaniami i poprawne z punktu widzenia założeń analizy. Na przykład, wyniki funkcji statystycznej, takiej jak test t-Studenta czy analiza wariancji (ANOVA), można porównać z wynikami uzyskanymi za pomocą innych narzędzi analitycznych, takich jak specjalistyczne oprogramowanie statystyczne (np. SPSS lub R). Innym przykładem jest weryfikacja zgodności wyników modeli predykcyjnych z rzeczywistymi danymi w testowym podzbiorze, np. przez porównanie wartości przewidywanych z rzeczywistymi za pomocą metryk takich jak średni błąd absolutny (MAE) czy współczynnik determinacji (R^2). Testy te mogą także obejmować sprawdzenie, czy wyjściowe raporty i wizualizacje są zgodne z oczekiwaniami, poprawnie interpretując dane oraz prezentując wyniki w sposób precyzyjny i czytelny dla użytkowników końcowych.

Na przykład w prognozowaniu szeregów czasowych metryki wchodzące w zakres testów walidacji wyników, takie jak średni błąd absolutny (MAE), pierwiastek z błędu średniokwadratowego (RMSE) czy procentowy błąd średni absolutny (MAPE), pozwalają

zweryfikować, na ile model trafnie odwzorowuje rzeczywiste dane. W przypadku modeli regresji, współczynnik determinacji (R^2) może służyć do oceny, jak dobrze model wyjaśnia zmienność danych, natomiast średni błąd kwadratowy (MSE) i błąd średni absolutny (MAE) wskazują na precyzję przewidywań.

Poniżej zaprezentowano prompt, który zdejmuje z badacza obowiązek znajomości tego typu testów i metryk.

> *Na podstawie przeprowadzonej analizy danych, oceń jej cel, jakość i skuteczność. Zaproponuj odpowiednie testy, które pomogą zweryfikować wyniki oraz ich wiarygodność. Weź pod uwagę charakter danych (ilościowe, jakościowe, szereg czasowy itp.), rodzaj zastosowanego modelu (np. regresja, klasyfikacja, prognozowanie, analiza klastrów) oraz główne założenia analizy. Jeśli analiza miała określony cel, np. prognozowanie, klasyfikację lub wykrywanie trendów, dobierz testy zgodne z tym celem. Wskaż także, czy istnieją prostsze modele bazowe, z którymi można porównać wyniki. Zwróć szczególną uwagę na:*
>
> - *weryfikację jakości dopasowania modelu do danych,*
> - *sprawdzenie ewentualnych założeń statystycznych (np. normalność reszt, autokorelacja),*
> - *metryki dopasowania (np. MAE, RMSE, accuracy, precision),*
> - *możliwość porównania z alternatywnymi metodami lub modelami bazowymi.*
>
> *Przedstaw wyniki w sposób zrozumiały dla użytkownika niezaznajomionego z zaawansowaną statystyką. Jeśli to możliwe, zaproponuj wizualizacje, które pomogą w interpretacji wyników i wskaż, jak można dalej ulepszyć analizę.*

Następnie prompt ten został wykorzystany do oceny prognozy sprzedaży samochodów osobowych. Dane z prognozą zostały zapisane i wgrane ponownie do modelu ChatGPT 4o. Oto interpretacja wyników prognozy zwrócona przez model:

Rycina 8.21. Ocena jakości prognozy przez model ChatGPT 4o

Interpretacja wyników:

1. Wizualna ocena:
   - Wykresy historycznych danych i prognoz wskazują na ogólną zgodność trendów, choć widoczne są różnice w wartości średniej (szczególnie dla marek AUDI i OPEL).
   - Średnie błędy (na wykresie słupkowym) sugerują, że prognozy mogą nieco niedoszacowywać lub przeszacowywać wyniki w zależności od marki.
2. Różnice w średnich:
   - Prognozowane wartości dla TOYOTA są bliższe historycznym danym niż dla innych marek, co może świadczyć o większej dokładności modelu dla tej marki.
3. Możliwości ulepszenia:
   - Dalsza analiza reszt modelu, np. sprawdzenie ich autokorelacji, pomoże lepiej ocenić jakość prognozy.
   - Rozważenie użycia bardziej zaawansowanych modeli bazujących na szeregach czasowych.

Źródło: Opracowanie własne.

Tak uzyskana interpretacja pozwala na szybkie podsumowanie wyników prognozy, wskazując zarówno zgodności, jak i potencjalne różnice między wartościami historycznymi a przewidywanymi. Dzięki temu badacz może efektywnie identyfikować wzorce i nieścisłości w prognozach, co przyspiesza proces ulepszania modelu. Dodatkowo, model podpowiada konkretne kroki analizy, takie jak ocena autokorelacji czy wykorzystanie zaawansowanych metod prognozowania. Problemy takie jak niedoszacowanie lub przeszacowanie prognoz dla konkretnych marek, mogą być trudne do zauważenia bez wsparcia modelu AI. Dzięki temu model działa jako „drugi analityk-ewaluator", wspierający interpretację wyników i sugerujący dalsze kierunki badawcze.

8.6. Zakończenie

Zaprezentowane w rozdziale analizy, prompty i różnego typu metody mają charakter autorski (subiektywny). Stanowią one jedną z wielu możliwych ścieżek wykorzystania wybranych

modeli językowych (LLM) w ilościowych analizach ewaluacyjnych. Proponowana przez autora sekwencja działań pozwala jednak kompleksowo uchwycić proces analityczny i zrozumieć najważniejsze problemy, z którymi spotykają się badacze prowadzący analizy ilościowe. Celem niniejszego opracowania było zaprezentowanie tej wiedzy w sposób względnie prosty i zrozumiały dla ewaluatora posiadającego niewielkie doświadczenie w badaniach ilościowych. Warto jednak podkreślić, że obserwujemy dynamiczny przyrost, zarówno dużych (LLM), jak i małych modeli językowych (SLM), specjalizujących się w analizach ilościowych (np. poprzez „destylację" logiki generowania kodu z dużych modeli, patrz Sun et al., 2024). Kluczowe staje się więc dla ewaluatorów nie tylko bieżące śledzenie rozwoju tej technologii, ale również jej adaptacja do konkretnych potrzeb badawczych.

Bibliografia

- Brzeziński, D., Filipek, K., Piwowar, K., et al. (2024). *Theorising the Algorithmic Condition: Challenges, Strategies, and Perspectives*. W: Algorithms, Artificial Intelligence and Beyond, 1–14. Routledge.
- Ceny LLM (dostęp: 31.03.2025).
- Liu, X., Wu, Z., Wu, X., et al. (2024). Are llms capable of data-based statistical and causal reasoning? benchmarking advanced quantitative reasoning with data. *arXiv preprint arXiv:2402.17644*.
- Modele LLM (dostęp: 31.03.2025).
- Mohsin, A., Janicke, H., Wood, A., et al. (2024). Can We Trust Large Language Models Generated Code? A Framework for In-Context Learning, Security Patterns, and Code Evaluations Across Diverse LLMs. *arXiv preprint arXiv:2406.12513*.
- Nejjar, M., Zacharias, L., Stiehle, F., et al. (2023). LLMs for science: Usage for code generation and data analysis. *Journal of Software: Evolution and Process*, e2723.
- Praca w środowisku Claude.ae (dostęp: 31.03.2025).
- Smolić, E., Pavelić, M., Boras, B., et al. (2024, May). Llm generative ai and students' exam code evaluation: Qualitative and quantitative analysis. W: *2024 47th MIPRO ICT and Electronics Convention (MIPRO),* 1261–1266). IEEE.
- Spiess, C., Gros, D., Pai, K.S., et al. (2024). Calibration and correctness of language models for code. *arXiv preprint arXiv:2402.02047.*
- Sun, Z., Lyu, C., Li, B., et al. (2024). Enhancing Code Generation Performance of Smaller Models by Distilling the Reasoning Ability of LLMs. *arXiv preprint arXiv:2403.13271*.
- Vasvani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is All You Need. *arXiv preprint arXiv:1706.03762* (dostęp: 31.03.2025).
- Vázquez, P.P. (2024, April). *Are LLMs ready for Visualization?.* W: 2024 IEEE 17th Pacific Visualization Conference (PacificVis), 343–352. IEEE.
- White, J., Fu, Q., Hays, S., et al. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

9. Analizy sieciowe wspierane gen AI

Marta Lesiak
Teresa Wyszyńska

9.1. Wprowadzenie

Celem rozdziału jest zaprezentowanie i wstępna ocena możliwości wykorzystania modeli językowych AI w analizie danych relacyjnych. Ocena ta zostanie przeprowadzona na przykładzie analizy przepływów odbiorców usług oraz usługodawców, ze szczególnym uwzględnieniem identyfikacji popularnych ścieżek szkoleniowych w Bazie Usług Rozwojowych PARP. Przedstawione studium przypadku ilustruje, w jaki sposób generatywna AI może wspierać proces pobierania danych z publicznych rejestrów za pośrednictwem API (ang. *Application Programming Interface*) oraz ich analizy, w tym analizy sieci społecznych.

Analiza sieci społecznych, rozwijana przez badaczy takich jak Barnes (1954), Knoke i Kuklinski (1982) czy Wasserman i Faust (1994), znalazła zastosowanie w wielu dziedzinach – od badania struktur społecznych, poprzez analizę organizacyjną, aż po modelowanie przepływów informacji w sieciach technologicznych. Dzięki rosnącej dostępności danych relacyjnych oraz zaawansowanych narzędzi analitycznych analiza sieciowa staje się coraz bardziej popularna, choć jej wykorzystanie wciąż wymaga specjalistycznej wiedzy. W tym kontekście sztuczna inteligencja, w szczególności modele językowe, oferuje nową jakość, umożliwiając automatyzację procesów przekształcania danych relacyjnych w struktury „analizowalne" matematycznie i graficznie. Takie podejście pozwala nie tylko na efektywniejsze badanie sieci społecznych, ale także na ich zastosowanie w różnych obszarach praktycznych – od marketingu, przez zarządzanie, po nauki przyrodnicze.

Według definicji Olechnickiego i Załęskiego (1999) sieć społeczna to „określenie systemów i układów relacji międzyosobowych występujących w zbiorowościach społecznych, w które uwikłane są różne jednostki". Z kolei wg Wassermana i Fausta (1994) sieć społeczna to

zbiór aktorów społecznych (węzłów) i relacji (powiązań) między nimi. Aktorzy mogą być jednostkami, grupami, organizacjami lub innymi zbiorowościami społecznymi. Relacje między nimi reprezentują przepływy lub transfer zasobów materialnych bądź niematerialnych, interakcje, podobieństwa, powiązania fizyczne, role społeczne lub pokrewieństwo. Sieci społeczne można przedstawić matematycznie za pomocą macierzy relacji, takich jak macierz sąsiedztwa lub macierz incydencji, które opisują zależności między poszczególnymi elementami sieci. Graficznie natomiast sieci społeczne mogą być prezentowane jako grafy, gdzie węzły odpowiadają poszczególnym jednostkom, a krawędzie między nimi symbolizują istniejące relacje. Taka dwutorowa reprezentacja – matematyczna i graficzna – pozwala na analizę struktury i dynamiki sieci, wspierając zarówno ilościowe badania zależności, jak i wizualizację wzorców zachowań w analizowanych społecznościach. Analiza danych sieciowych różni się od tradycyjnych metod opartych na danych tabelarycznych, gdyż koncentruje się na relacjach i zależnościach między elementami, a nie wyłącznie na ich indywidualnych właściwościach. Wymaga to zastosowania specjalistycznych metod oraz narzędzi, które umożliwiają modelowanie interakcji i dynamiki w sieci.

W niniejszym rozdziale przedstawiono możliwości zastosowania modeli językowych w połączeniu z bazami ilościowych danych relacyjnych. Rozdział ten podzielono na trzy główne podrozdziały, które wspólnie opisują kompleksowy proces pracy z tego typu danymi przy użyciu gen AI.

Pierwszy podrozdział dotyczy przede wszystkim przygotowania środowiska pracy z modelami językowymi i API. Przedstawiono, jak modele AI, takie jak ChatGPT, mogą wygenerować kod analityczny w języku Python, umożliwiając automatyczne pobieranie danych z publicznie dostępnych baz poprzez API. W drugim podrozdziale omówiono wykorzystanie modeli językowych w analizie danych sieciowych. Wyjaśniono, w jaki sposób gen AI wspiera organizację i przekształcanie surowych danych w uporządkowane struktury, które umożliwiają przeprowadzanie zaawansowanych analiz sieciowych. Opisano również, jak automatyzacja kodowania może przyspieszyć procesy analityczne, pozwalając na identyfikację kluczowych elementów w sieciach oraz przepływów między jednostkami. Trzeci podrozdział poświęcony jest weryfikacji wyników, co ma kluczowe znaczenie dla poprawności i wiarygodności przeprowadzonych analiz. Omówiono, jak modele gen AI mogą wskazywać na potencjalne nieścisłości lub braki w zbiorach danych oraz jak zastosowanie odpowiednich technik poprawia jakość danych wejściowych i interpretację wyników. Podrozdział podkreśla znaczenie krytycznego podejścia do danych w analizie sieciowej.

9.2. Pozyskanie danych

Mimo że coraz więcej danych jest upublicznianych w Internecie, ich pozyskanie nie zawsze jest łatwe i bezproblemowe. Na przykład wiele publicznych baz danych udostępnia swoje zasoby w sposób wymagający znajomości programowania, API czy specjalistycznych narzędzi analitycznych. Jeszcze nie tak dawno osoby, które chciały korzystać z takich danych samodzielnie, musiały zdobyć wiedzę z zakresu statystyki, analizy danych i programowania, co nierzadko wymagało ukończenia specjalistycznych kursów, szkoleń czy studiów. Były to czasochłonne i kosztowne ścieżki rozwoju. Obecnie dzięki modelom językowym generatywnej AI, takim jak ChatGPT, wiele z tych barier zostało obniżonych. Modele pomagają bowiem nie tylko generować kod analityczny w sposób zrozumiały dla osób bez zaawansowanych umiejętności programistycznych, ale także pełnić rolę interaktywnego wsparcia w nauce, poprzez pomoc w interpretacji wyników i wizualizacji danych w przystępnej formie. Gen AI może podpowiadać, jak podejść do problemu, inspirować do poszukiwania rozwiązań, a nawet symulować rozmowę z wirtualnym mentorem, co jest szczególnie cenne dla osób pracujących samodzielnie, bez możliwości konsultacji z innymi analitykami czy ekspertami. Tego typu „rozmowy” pozwalają uporządkować myśli, spojrzeć na problem z różnych perspektyw i zyskać inspirację do dalszego rozwoju. Dzięki temu proces zdobywania wiedzy i umiejętności staje się bardziej dostępny i elastyczny. Warto jednak pamiętać, że choć gen AI jest potężnym narzędziem, wyniki i porady, które uzyskamy, powinny być traktowane krytycznie – należy zawsze weryfikować uzyskane informacje i wyniki, aby uniknąć potencjalnych błędów.

Dlatego aby skutecznie korzystać z gen AI w procesie pozyskiwania i analizowania danych, konieczne jest posiadanie przynajmniej podstawowej wiedzy na temat zagadnienia, które chcemy zgłębić. Bez tej wiedzy trudno jest sformułować precyzyjne instrukcje, co jest warunkiem efektywnego wykorzystania gen AI. W naszym przypadku już na początkowym etapie kluczowe było przygotowanie jasnych wytycznych dotyczących rejestracji, uzyskiwania dostępu do API oraz przetwarzania danych. Nie wystarczy tutaj sama prośba do modelu gen AI o wygenerowanie kodu do pobrania danych przez API – konieczne jest dokładne określenie parametrów oraz kontekstu.

Proces pozyskiwania danych z wykorzystaniem AI i Pythona

W pierwszym etapie pracy z gen AI wykorzystano dane pozyskane za pomocą *Application Programming Interface*. API umożliwiło dostęp do uporządkowanych danych w sposób zautomatyzowany i ustrukturyzowany. W celu komunikacji z API wykorzystano model językowy gen AI, który wygenerował kod w Pythonie służący do pobierania danych.

Etapy pobierania danych z API za pomocą gen AI i języka Python:

1. Przygotowanie środowiska i bibliotek.
   Skonfigurowano środowisko programistyczne, instalując wymagane biblioteki, takie jak: *requests* (do komunikacji z API) i *pandas* (do przetwarzania i zapisywania danych). Model gen AI, taki jak ChatGPT, mógł wygenerować kod do instalacji tych bibliotek, jednak sam proces instalacji przeprowadzono w lokalnym środowisku Python.
2. Generowanie kodu za pomocą gen AI.
   Do wygenerowania kodu Pythona wykorzystano model gen AI, który uwzględnił kluczowe elementy, takie jak:
   - wysyłanie żądań do API z wykorzystaniem klucza API i tokenu,
   - obsługa paginacji (w przypadku danych podzielonych na strony),
   - przetwarzanie danych w formacie JSON,
   - zapis danych do pliku csv z użyciem biblioteki pandas.

   Wygenerowany kod został następnie przeniesiony do środowiska Python, gdzie przeprowadzono dalsze etapy pracy. Warto jednak podkreślić, że proces generowania kodu z wykorzystaniem gen AI często bywa iteracyjny – wygenerowany kod nie zawsze działa poprawnie za pierwszym razem. W takich przypadkach konieczne jest weryfikowanie i testowanie wygenerowanego skryptu, a także dostarczanie modelowi bardziej szczegółowych instrukcji, które pozwolą poprawić wynik. Iteracyjne podejście umożliwia sukcesywne eliminowanie błędów i dostosowanie kodu do specyficznych wymagań analizy.
3. Uzyskanie dostępu do API.
   Klucz API oraz token dostępu zostały pozyskane manualnie, ponieważ próby zautomatyzowania tego procesu za pomocą gen AI nie przyniosły efektu. W tym celu zarejestrowano się w Bazie Usług Rozwojowych, co pozwoliło na uzyskanie niezbędnych danych uwierzytelniających.

4. Pobieranie danych.
   Kod stworzony przez gen AI został uruchomiony w środowisku Python, co umożliwiło:
   - wysyłanie żądań do odpowiednich endpointów API,
   - obsługę błędów,
   - przetwarzanie danych zwróconych przez API,
   - zapis danych do strukturalnych plików (csv).

   Dzięki kodowi wygenerowanemu przez AI możliwe było znaczące uproszczenie tego procesu, jednak niezbędne było wprowadzenie ręcznych korekt, np. uwzględnienie podziału danych na strony.

W procesie pozyskiwania danych z API model gen AI odegrał kluczową rolę w automatyzacji tworzenia kodu, co uprościło zadania związane z uwierzytelnianiem, przetwarzaniem i zapisywaniem danych. Aby zapewnić poprawność działania wygenerowanego kodu, konieczne było dostarczenie precyzyjnych instrukcji oraz uzupełnienie technicznych parametrów, takich jak liczba stron danych czy specyfikacja endpointów API, aby wygenerowany kod działał poprawnie.

Wyzwania i rola AI w procesie pozyskiwania danych

Głównym wyzwaniem dotyczącym wykorzystania ChatGPT do wygenerowania kodu umożliwiającego pobranie klucza API i tokenu był etap pozyskania danych uwierzytelniających. Mimo kilkukrotnego wysyłania zapytań do gen AI nie osiągnięto oczekiwanego efektu i w ostateczności zastosowano do tego celu tradycyjne metody, takie jak: samodzielne założenie konta na stronie internetowej, wygenerowanie klucza API oraz tokena dostępu. Na tej podstawie można stwierdzić, że modele językowe AI wykazują ograniczenia w interakcji z zewnętrznymi systemami, zwłaszcza w procesach uwierzytelniania i pozyskiwania danych dostępowych, które często są specyficzne dla każdego dostawcy.

Poniżej przedstawiono kod wygenerowany przez ChatGPT, na podstawie którego pobrano dane[1]:

[1] Nie podano adresu strony internetowej, aby nie narażać jej na ryzyko zbytniego obciążenia, co może spowodować niedostępność usług. Tokenu nie podano ze względu na poufność tych danych. Kod zawiera oznaczone kolorem szarym komentarze do linijek kodu, aby można było odczytać, w jaki sposób dana linijka kodu działa i do czego służy.

Rycina 9.1. Kod wygenerowany przez ChatGPT

```
import requests
import pandas as pd
api_token = 'your_api_token_here' # Wstawienie tokenu API
 # Bazowy URL API
base_url = 'https://' #Uzupełnienie właściwym adresem API
 headers = {
 'Authorization': f'Bearer {api_token}',
 'Content-Type': 'application/json'
}
 # Lista do przechowywania wszystkich danych
all_data = []
 # Ustal liczbę elementów na stronę
items_per_page = 25
 # Pierwsze żądanie, aby uzyskać dane i liczbę stron
params = {'page': 1}
response = requests.get(base_url, headers=headers, params=params)
 if response.status_code == 200:
 initial_data = response.json()
 total_items = 3204 # Podana całkowita liczba elementów
 total_pages = (total_items // items_per_page) + (1 if total_items % items_per_page
!= 0 else 0)
 # Dodaj dane z pierwszej strony do listy
 all_data.extend(initial_data['lista'])
 # Iteruj przez resztę stron i dodawaj dane do listy
 for page in range(2, total_pages + 1):
 params = {'page': page}
 response = requests.get(base_url, headers=headers, params=params)
 if response.status_code == 200:
 page_data = response.json()
 all_data.extend(page_data['lista'])
```

```
else:
 print(f"Nie udało się pobrać danych dla strony {page}: {response.status_code} -
{response.text}")
else:
 print(f"Nie udało się pobrać danych początkowych: {response.status_code} -
{response.text}")
 # Konwertuj wszystkie dane do ramki danych pandas
df = pd.DataFrame(all_data)
 # Wyświetl pierwsze 5 wierszy ramki danych

 # Zapisywanie danych do pliku CSV
df.to_csv('dane_dostawcy_uslug.csv', index=False)
 # Grupowanie i sumowanie
grouped_df = df.groupby('status').size()
print(grouped_df)
 # Unikalne wartości w kolumnie 'status'
unique_statuses = df['status'].unique()
print(unique_statuses)
```

Źródło: Opracowanie własne przy pomocy ChatGPT.

Wyjaśnienie kodu:

- Importowanie bibliotek: *requests* do komunikacji z API oraz *pandas* do manipulacji danymi.
- Uwierzytelnienie: ustawienie nagłówków z tokenem API.
- Pobieranie danych: pobranie danych z API, z uwzględnieniem paginacji.
- Przetwarzanie danych: konwersja danych do ramki danych *pandas*, zapisanie do pliku csv.
- Analiza danych: grupowanie danych według kolumny status, wyświetlanie unikalnych wartości.

9.3. Analiza pozyskanego zbioru danych

Dane pobrane za pomocą wcześniej przedstawionego kodu zostały przeanalizowane pod kątem dostępnych informacji. Zawierały one informacje dotyczące usługodawców

i posiadanych przez nich certyfikatów. Jednak brakowało w nich kluczowych danych o uczestnikach usług rozwojowych, ponieważ te informacje nie były dostępne w API, z którego korzystano. Ponieważ celem niniejszego rozdziału było zmapowanie przepływów między odbiorcami usług a usługodawcami, dostępny zbiór należało rozszerzyć.

W związku z tym autorki pozyskały z właściwego departamentu PARP uzupełniający i zanonimizowany zbiór danych, który uzupełnił wcześniejsze informacje o usługobiorcach. Na tej podstawie rozpoczęto właściwe analizy, wykorzystując techniki analizy sieciowej.

Kwestie etyczne w pracy z AI – ochrona danych osobowych

W organizacjach obowiązują różnorodne zasady wykorzystywania danych, dotyczące m.in. zapewnienia poufności, ochrony danych osób czy minimalizacji ryzyka wynikającego z uprzedzeń w danych[2]. W niniejszym opracowaniu autorki przyjęły podejście mające na celu minimalizację tych ryzyk oraz zapewnienie zgodności z obowiązującymi regulacjami i politykami macierzystej organizacji (PARP).

Zamiast przekazywania zbioru danych rzeczywistych bezpośrednio do zewnętrznego serwisu udostępniającego model AI, wykorzystano gen AI do wygenerowania kodu w języku Python, który następnie uruchomiono w kontrolowanym środowisku lokalnym. Chociaż niektóre modele AI oferują możliwość wgrania zbioru danych i ich bezpośredniej analizy, w tym przypadku zrezygnowano z takiego rozwiązania, kierując się koniecznością zachowania pełnej kontroli nad danymi i procesem analizy, mimo że pozyskany zbiór został zanonimizowany. Jednocześnie wygenerowany kod wymagał wprowadzenia ręcznych modyfikacji, takich jak wskazanie ścieżki do katalogu z danymi, co pozwoliło na jego dostosowanie do konkretnego przypadku. Brak takich modyfikacji uniemożliwiłby prawidłowe działanie programu.

Przy czym takie podejście – jak wspomniano powyżej – zapewniło pełną kontrolę nad procesem analizy i jednocześnie pozwoliło efektywne wykorzystać możliwości gen AI do automatyzacji tworzenia kodu analitycznego. Wygenerowany kod spełnił swoje zadanie, umożliwiając odczytanie i przetwarzanie danych w sposób zgodny z wymaganiami stawianymi przed analizą.

[2] Więcej na ten temat w rozdziale pt. „BHP pracy z gen AI – bezpieczeństwo danych i poufność”.

9.4. Analizy sieciowe z użyciem modeli gen AI

Proces analizy danych przebiegał w kilku etapach:

- Generowanie kodu do opisu struktury danych: model gen AI został poproszony o wygenerowanie kodu Python, który umożliwił odczytanie kluczowych informacji o zbiorze danych i przedstawienie jego struktury w formie zrozumiałej dla dalszych analiz.
- Generowanie kodu do analizy danych: na podstawie wyników pierwszego etapu model AI wygenerował kod przeznaczony do konkretnych analiz sieciowych.

Porządkowanie danych

Etap wstępny polegał na analizie zbioru danych w celu jego uporządkowania. W tym procesie konieczne było szczegółowe opisanie każdej zmiennej, kolumna po kolumnie, w zakresie charakteru przechowywanych wartości oraz sprawdzenie typu danych (liczbowy, zmiennoprzecinkowy, łańcuchy znaków czy daty). W celu automatyzacji tego etapu wykorzystano model językowy gen AI, który został poproszony o wygenerowanie kodu w języku Python, realizującego wymienione operacje.

Instrukcja skierowana do gen AI zawierała precyzyjnie określone punkty, które powinny zostać uwzględnione w kodzie:

- uzyskanie listy nazw zmiennych, zawartych w zbiorze danych,
- ustalenie liczby zmiennych i obserwacji,
- określenie typów zmiennych (numeryczne, opisowe, zmiennoprzecinkowe, daty),
- informacja o liczbie braków danych w każdej zmiennej,
- sugestie dotyczące postępowania z brakami danych.

Ważne było, aby pytanie skierowane do gen AI było odpowiednio sformułowane i uwzględniało specyfikę analizowanego problemu. Ogólne pytanie, takie jak: *Przygotuj analizę sieciową*, będzie prowadziło do uzyskania nieadekwatnych lub zbyt ogólnych odpowiedzi. Dlatego najpierw zwrócono się do modelu z prośbą o wygenerowanie kodu, który pozwoliłby opisać zbiór.

W efekcie wygenerowano następujący kod przedstawiony na ryc. 9.2.[3]:

Rycina 9.2. Kod wygenerowany przez ChatGPT do opisu zbioru

```
Import pandas as pd
# Wczytanie danych z pliku CSV
df = pd.read_csv('ścieżka_do_pliku.csv')
# Wczytanie danych z pliku Excel
# df = pd.read_excel('sciezka_do_pliku.xlsx')
# Wyświetlenie pierwszych kilku wierszy, aby zobaczyć strukturę danych
print(df.head())
# Wyświetlenie podstawowych informacji o danych
print(df.info())
# Statystyki opisowe dla danych liczbowych
print(df.describe())
# Sprawdzenie liczby braków danych w każdej kolumnie
print(df.isnull().sum())
```

Źródło: Opracowanie własne przy pomocy ChatGPT

Kod wygenerowany przez model gen AI spełnił oczekiwania i nie różnił się zasadniczo od kodu, który mógłby zostać napisany manualnie, np. w skrypcie SPSS. Skrypt umożliwiał:

- wczytanie danych z plików w formacie csv lub xls,
- wyświetlenie pierwszych pięciu wierszy danych,
- uzyskanie podstawowych informacji o wczytanym pliku, takich jak liczba zmiennych i obserwacji,
- generowanie statystyk opisowych,
- wskazanie braków danych oraz podstawowe sugestie dotyczące ich obsługi.

Ważne jest, że zastosowanie AI do automatyzacji tego etapu pozwoliło zaoszczędzić czas, jednocześnie zapewniając odpowiednią jakość analizy. Takie podejście umożliwiło szybkie przeprowadzenie wstępnej analizy w sposób wydajny i zgodny z najlepszymi praktykami analitycznymi.

[3] Kolorem szarym są oznaczone komentarze dotyczące kodu.

Na ryc. 9.3. przedstawiono fragment wyniku uzyskanego za pomocą wygenerowanego kodu.

Rycina 9.3. Wynik uzyskany za pomocą kodu uzyskanego z ChatGPT

```
[5 rows x 36 columns]
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 458683 entries, 0 to 458682
Data columns (total 36 columns):
#   Column                      Non-Null Count   Dtype
--  -------                     --------------   -----
0   id                          458683 non-null  int64
1   id_uslugi                   458683 non-null  int64
2   numer_uslugi                458683 non-null  object
3   wojewodztwo_usluga          458683 non-null  object
4   miejscowosc_usluga          458683 non-null  object
5   powiat_usluga               458683 non-null  object
6   data_rozpoczecia_uslugi     458683 non-null  object
7   data_zakonczenia_uslugi     458683 non-null  object
```

Źródło: Opracowanie własne przy pomocy ChatGPT.

Wykorzystanie AI w przygotowaniu kodu

Tak jak już wcześniej wspomniano, wyznaczenie gen AI zadania związanego z przeprowadzeniem analiz sieciowych, bez wcześniejszej prośby o wygenerowanie kodu dotyczącego opisu zmiennych czy braków danych, spowoduje, że gen AI wygeneruje ogólne pomysły na analizy sieciowe, które nie są dostosowane do naszego zbioru danych. Dlatego kluczowe jest „wymuszenie" na modelu szczegółowego opisu danych i określenie konkretnych potrzeb.

Przykład proponowanych przez gen AI analiz, bez podania informacji o zbiorze danych przedstawiono na ryc. 9.4.

Rycina 9.4. Proponowane przez ChatGPT analizy z wyłączeniem informacji o zbiorze danych

- Analiza dyfuzji informacji – zbadanie, jak informacje o usługach rozwojowych rozprzestrzeniają się w sieci.
- Analiza preferencji i dopasowań – identyfikacja najczęściej wybieranych szkoleń i wzorców wyborów szkoleniowych wśród różnych grup demograficznych.
- Analiza dynamiczna sieci – badanie, jak sieć zmienia się w czasie, którzy dostawcy zyskują na popularności, a którzy tracą.

Źródło: Opracowanie własne przy pomocy ChatGPT

Chociaż powyższe pomysły mogą być teoretycznie interesujące, są one nieadekwatne do analizowanego zbioru danych, ponieważ nie uwzględniają specyfiki zmiennych, które w tym zbiorze są przechowywane. Brak precyzyjnych informacji o danych powoduje, że gen AI proponuje zbyt ogólne analizy, które nie odpowiadają potrzebom analitycznym. Dlatego też dostarczenie szczegółowego opisu zbioru jest kluczowe dla uzyskania wartościowych wyników analizy.

Proponowane analizy sieciowe

Po upewnieniu się, że dane są odpowiednio przygotowane i przekształcone, przystąpiono do zaprojektowania odpowiednich analiz. Zgodnie z celem rozdziału, analiza skupiła się na przepływach odbiorców zarówno między usługami, jak i między usługodawcami, aby zidentyfikować popularne ścieżki szkoleniowe.

W doborze odpowiednich analiz pomógł model AI, który na podstawie opisu zbioru danych wygenerował propozycje[4]:

1. Identyfikacja najpopularniejszych ścieżek szkoleniowych.
   - wygenerowanie statystyk dotyczących najczęściej wybieranych par podkategorii (pierwsza i kolejna podkategoria),
   - zidentyfikowanie najpopularniejszych ścieżek szkoleniowych odbiorców usług rozwojowych.

[4] Nie wszystkie wymienione analizy zostały zrealizowane na potrzeby niniejszego rozdziału, skupiono się na wybranych. Decyzja ta wynikała z ograniczonej objętości rozdziału oraz jego głównego celu – przedstawienia pracy z gen AI, a nie pełnego wykonania analiz na wybranym zbiorze czy szczegółowego opisu uzyskanych wyników.

2. Analiza przepływów dla różnych grup demograficznych:
   - wykorzystanie zmiennych takich jak: wiek, płeć, miejscowość zamieszkania czy kod pocztowy,
   - przeprowadzenie filtrowanej analizy przepływów dla określonych grup demograficznych, np. różnice w ścieżkach szkoleniowych dla osób w różnym wieku.
3. Analiza centralności przepływów:
   - użycie centralności pośredniczenia (ang. *betweenness centrality*) do sprawdzenia, które podkategorie były kluczowe w ścieżkach szkoleniowych i pełnią rolę „mostów" między innymi podkategoriami,
   - ujawnienie kluczowych usług wybieranych jako przystanki w dłuższych ścieżkach edukacyjnych.
4. Grupowanie odbiorców usług na podstawie preferowanych ścieżek szkoleniowych:
   - klasteryzacja odbiorców usług i stworzenie grup o podobnych preferencjach edukacyjnych.

Propozycje wygenerowane przez model AI okazały się zgodne z metodologią analizy sieciowej i dobrze dopasowane do specyfiki danych po dostarczeniu szczegółowego opisu zbioru. W szczególności sugestie dotyczące analizy centralności przepływów i identyfikacji najpopularniejszych ścieżek były trafne i wartościowe. Natomiast niektóre propozycje, takie jak analiza przepływów dla grup demograficznych, wymagałyby bardziej szczegółowych danych i w związku z tym nie zostały zrealizowane na potrzeby niniejszego opracowania. Można stwierdzić, że dzięki gen AI możliwe było przyspieszenie procesu planowania analiz oraz ich dostosowanie do specyfiki zbioru danych.

Przegląd wyników analiz sieciowych

Analizy przeprowadzono z wykorzystaniem gen AI, co umożliwiło przetestowanie efektywności tego narzędzia w kontekście analizy sieciowej usług rozwojowych. Głównym ich celem było zbadanie, jak gen AI wspiera proces przekształcania danych, automatyzacji generowania kodu oraz interpretacji wyników, w porównaniu z tradycyjnymi metodami analitycznymi.

W trakcie realizacji analiz gen AI zaproponował kod Pythona do przeprowadzenia szeregu operacji, takich jak: obliczanie centralności, klasteryzacja społeczności, analiza przepływów między usługami oraz przepływów między społecznościami. Dzięki temu możliwe było

zautomatyzowanie procesów wymagających zwykle ręcznego kodowania, takich jak przekształcanie danych na strukturę grafu czy identyfikacja kluczowych elementów w sieci.

Testowanie gen AI wykazało również ograniczenia. AI wymagała dostarczenia szczegółowych informacji o danych wejściowych, takich jak struktura pliku czy format danych. W przypadku bardziej zaawansowanych operacji, takich jak klasteryzacja społeczności, konieczne było wprowadzanie ręcznych poprawek w generowanym kodzie. Pokazało to nie tylko potrzebę współpracy człowieka z AI w kontekście precyzyjnych wymagań analizy, ale także wskazało, że osoby bez podstawowej znajomości wykorzystywanych metod mogą mieć trudności z ich właściwym zastosowaniem. Niemniej AI dostarcza wyjaśnień i wskazówek, które ułatwiają zrozumienie i praktyczne wykorzystanie tych metod.

Na podstawie tych doświadczeń wyciągnięto wnioski dotyczące pracy z AI, które wskazują na jej wartość jako narzędzia wspierającego automatyzację oraz eksplorację danych. Wykorzystanie gen AI w analizie sieciowej nie tylko usprawniło realizację poszczególnych etapów badania, ale także otworzyło możliwości testowania różnych scenariuszy analizy w znacznie krótszym czasie niż w przypadku metod tradycyjnych.

Przeprowadzone analizy pokazują różnorodność zastosowań generatywnej AI w pracy z danymi sieciowymi. Skupiono się na takich aspektach jak: badanie struktury sieci, w tym identyfikacja kluczowych usług i dostawców, analiza przepływów odbiorców usług między różnymi elementami sieci, a także badanie popularności usług i ich wyborów w różnych kontekstach. Dodatkowo przeanalizowano, w jaki sposób usługi grupują się w społeczności, kiedy są często wybierane razem przez odbiorców usług.

Dla zilustrowania zastosowanych metod i wyników szczegółowo omówiono wybrane przykłady analiz. Pozostałe prace, choć również istotne, przedstawiono bardziej ogólnie, co pozwoliło na podkreślenie najważniejszych wniosków i praktycznych korzyści płynących z wykorzystania generatywnej AI w tym zadaniu.

Identyfikacja i charakterystyka kluczowych dostawców usług

Na podstawie analizy sieci przepływów odbiorców usług między dostawcami zidentyfikowano kluczowe podmioty, które odgrywają centralną rolę w strukturze sieci. Analizę przeprowadzono z wykorzystaniem miar centralności, takich jak: centralność pośredniczenia,

centralność bliskości[5] oraz centralność stopnia[6], co pozwoliło określić znaczenie poszczególnych dostawców w przepływach między nimi odbiorców usług.

Sieć dostawców usług została zrekonstruowana na podstawie przepływu odbiorców usług między nimi. Relacje w tej sieci definiuje liczba odbiorców, którzy skorzystali z usług różnych dostawców. Węzły reprezentują dostawców, a krawędzie – przepływy między nimi. Waga krawędzi odzwierciedla liczbę odbiorców przemieszczających się między dostawcami.

Do analizy sieci dostawców wykorzystano dane odbiorców usług wskazujące, z których usług i których dostawców korzystali. Dane przetworzono w formie macierzy sąsiedztwa, w której wartości komórek odpowiadają liczbie wspólnych odbiorców usług między parami dostawców.

Tabela 9.1. Top dostawcy według centralności pośredniczenia

Dostawca	Centralność pośredniczenia
Dostawca A	0,0530
Dostawca B	0,0513
Dostawca C	0,0429
Dostawca D	0,0406
Dostawca E	0,0388
Dostawca F	0,0285
Dostawca G	0,0278
Dostawca H	0,0269
Dostawca I	0,0233
Dostawca J	0,0227

Źródło: Opracowanie własne.

[5] Centralność bliskości to miara, która odzwierciedla średnią odległość węzła do wszystkich innych węzłów w sieci. Węzeł o wysokiej centralności bliskości jest „blisko" innych węzłów, co oznacza, że może szybko wpływać na resztę sieci lub szybko otrzymywać informacje (Freeman, 1979). Miara była obliczana jako odwrotność sumy najkrótszych odległości od danego węzła do wszystkich innych węzłów w sieci. Zastosowano normalizację, która pozwoliła na porównywanie wartości między węzłami.

[6] Centralność stopnia pozwala na identyfikację węzłów o największej liczbie bezpośrednich połączeń. W kontekście badania pomaga zrozumieć, którzy dostawcy usług są najbardziej połączeni w sieci szkoleniowej, co może wskazywać na ich popularność, znaczenie rynkowe lub zdolność do „przyciągania" odbiorców usług. Została ona obliczona jako liczba bezpośrednich połączeń węzła podzielona przez maksymalną możliwą liczbę połączeń. Pozwala na identyfikację węzłów najbardziej połączonych bezpośrednio z innymi.

Tabela 9.2. Top dostawcy według centralności bliskości

Dostawca	Centralność bliskości
Dostawca B	0,3751
Dostawca E	0,3670
Dostawca A	0,3625
Dostawca K	0,3563
Dostawca C	0,3562
Dostawca L	0,3545
Dostawca M	0,3530
Dostawca F	0,3527
Dostawca I	0,3524
Dostawca N	0,3500

Źródło: Opracowanie własne.

Tabela 9.3. Top dostawcy według centralności stopnia

Dostawca	Centralność stopnia
Dostawca A	0,0682
Dostawca B	0,0663
Dostawca E	0,0654
Dostawca D	0,0589
Dostawca K	0,0561
Dostawca L	0,0557
Dostawca C	0,0547
Dostawca M	0,0519
Dostawca G	0,0492
Dostawca I	0,0487

Źródło: Opracowanie własne.

Prompt, z którego skorzystano: *Przygotuj, proszę, kod Pythona, za pomocą którego będzie możliwe zbudowanie sieci dostawców na podstawie danych wejściowych o przepływach odbiorców usług między dostawcami usług. Policz miary centralności sieciowej dla każdego*

dostawcy: centralność pośredniczenia centralność bliskości, centralność stopnia. Wyświetl ranking top 10 dostawców według każdej z miar centralności w formie tabeli.

Wyniki, które otrzymano na podstawie analiz, dostarczają informacji na temat kluczowych dostawców w analizowanej sieci przepływów odbiorców usług między dostawcami.

Dostawcy o wysokiej centralności pośredniczenia, jak Dostawca A, pełnią rolę pośredników w sieci, łącząc różne części struktury. Dostawcy z wysoką centralnością bliskości, tacy jak Dostawca B, są strategicznie położeni, co oznacza, że odbiorcy usług mają do nich najkrótszą drogę w sieci. Z kolei dostawcy o wysokiej centralności stopnia, jak Dostawca A, mają najwięcej bezpośrednich połączeń z innymi dostawcami, co wskazuje na ich popularność, szeroki zasięg usług oraz ich atrakcyjność.

Usługi kluczowych dostawców:

- Dostawca A – usługi związane z urodą, medycyną estetyczną i kosmetologią;
- Dostawca B – usługi z zakresu marketingu, sprzedaży i zarządzania przedsiębiorstwem;

wskazują, jakie usługi odpowiadają na najważniejsze potrzeby rozwojowe odbiorców. Analiza kluczowych dostawców wskazuje, że szczególne zainteresowanie odbiorców budzą usługi o dużym znaczeniu zawodowym:

- medycyna estetyczna i kosmetologia (2245 odbiorców),
- zarządzanie przedsiębiorstwem (956 odbiorców),
- kurs prawa jazdy kategoria C (608 odbiorców).

Oferowane przez kluczowych dostawców usługi przyciągają szerokie grono uczestników, co tłumaczy ich centralną pozycję w strukturze przepływów. Popularność tych dostawców wynika z ich zdolności do zaspokajania potrzeb w takich obszarach jak edukacja, rozwój zawodowy czy specjalistyczne kwalifikacje.

Analiza przepływów w sieci usług

Przeprowadzone analizy przepływów w sieci usług koncentrowały się na identyfikacji ruchu odbiorców usług między poszczególnymi usługami, społecznościami oraz dostawcami. Badania te pozwoliły na lepsze zrozumienie dynamiki sieci oraz identyfikację najważniejszych elementów, takich jak usługi i społeczności, które mają największy wpływ na przepływy odbiorców usług. Na potrzeby badania zaplanowano wykonanie kilku rodzajów analiz

przepływów. Pierwszym krokiem była analiza przepływów między poszczególnymi usługami, w celu zidentyfikowania najczęstszych migracji odbiorców usług w sieci. Następnie skupiono się na analizie przepływów między społecznościami – grupami usług często wybieranych razem przez odbiorców usług. Ostatnią grupę analiz stanowiły przepływy wewnętrzne oraz między dostawcami, co pozwoliło na ocenę lojalności odbiorców usług wobec dostawców oraz relacji między nimi. AI została wykorzystana do wygenerowania kodu w Pythonie, który miał wspierać realizację tych zadań.

Analiza przepływów między usługami

Z wykorzystaniem gen AI zbadano kierunkowe przepływy odbiorców miedzy usługami. Analiza ta polega na identyfikacji przejść odbiorców od jednej usługi do kolejnej. Każde przejście można traktować jako krawędź w grafie skierowanym, gdzie węzłami są usługi, a krawędzie reprezentują liczbę odbiorców przemieszczających się między nimi.

Ponownie skierowano pytanie do gen AI o wygenerowania kodu Pythona do analizy przepływów:

> *Wygeneruj, proszę, kod w Pythonie do analizy przepływów między usługami, skup się na identyfikacji powiązań między kategoriami i ogranicz analizę do przepływów między kategoriami na podstawie danych z pliku CSV zawierającego kolumny: usługa początkowa, usługa końcowa, liczba odbiorców. Kod powinien obliczać przepływy między usługami, wyłącznie bezpośrednie przejścia i wyświetlić wyniki w formie tabeli.*

Ze względu na ograniczone ramy niniejszej publikacji w tabeli przedstawiono najczęstsze przejścia między usługami.

Tabela 9.4. Dwadzieścia najczęstszych przejść między usługami

Usługa początkowa – start	Usługa docelowa – przejście	Liczba uczestników (przejście)
Zarządzanie przedsiębiorstwem	Zarządzanie zasobami ludzkimi	4400
Zarządzanie zasobami ludzkimi	Zarządzanie przedsiębiorstwem	3734
Organizacja	Zarządzanie przedsiębiorstwem	2369
Zarządzanie przedsiębiorstwem	Organizacja	2299

Usługa początkowa – start	Usługa docelowa – przejście	Liczba uczestników (przejście)
Zarządzanie przedsiębiorstwem	Marketing	2059
Zarządzanie przedsiębiorstwem	Sprzedaż	2032
Sprzedaż	Zarządzanie przedsiębiorstwem	1944
Marketing	Zarządzanie przedsiębiorstwem	1712
Organizacja	Zarządzanie zasobami ludzkimi	1392
Zarządzanie zasobami ludzkimi	Organizacja	1385
Sprzedaż	Marketing	1375
Zarządzanie zasobami ludzkimi	Sprzedaż	1327
Sprzedaż	Zarządzanie zasobami ludzkimi	1289
Marketing	Sprzedaż	1227
Sprzedaż	Organizacja	1194
Organizacja	Sprzedaż	1129
Organizacja	Psychologia i rozwój osobisty	937
Psychologia i rozwój osobisty	Organizacja	885
Psychologia i rozwój osobisty	Edukacja	872
Zarządzanie przedsiębiorstwem	Psychologia i rozwój osobisty	824

Źródło: Opracowanie własne.

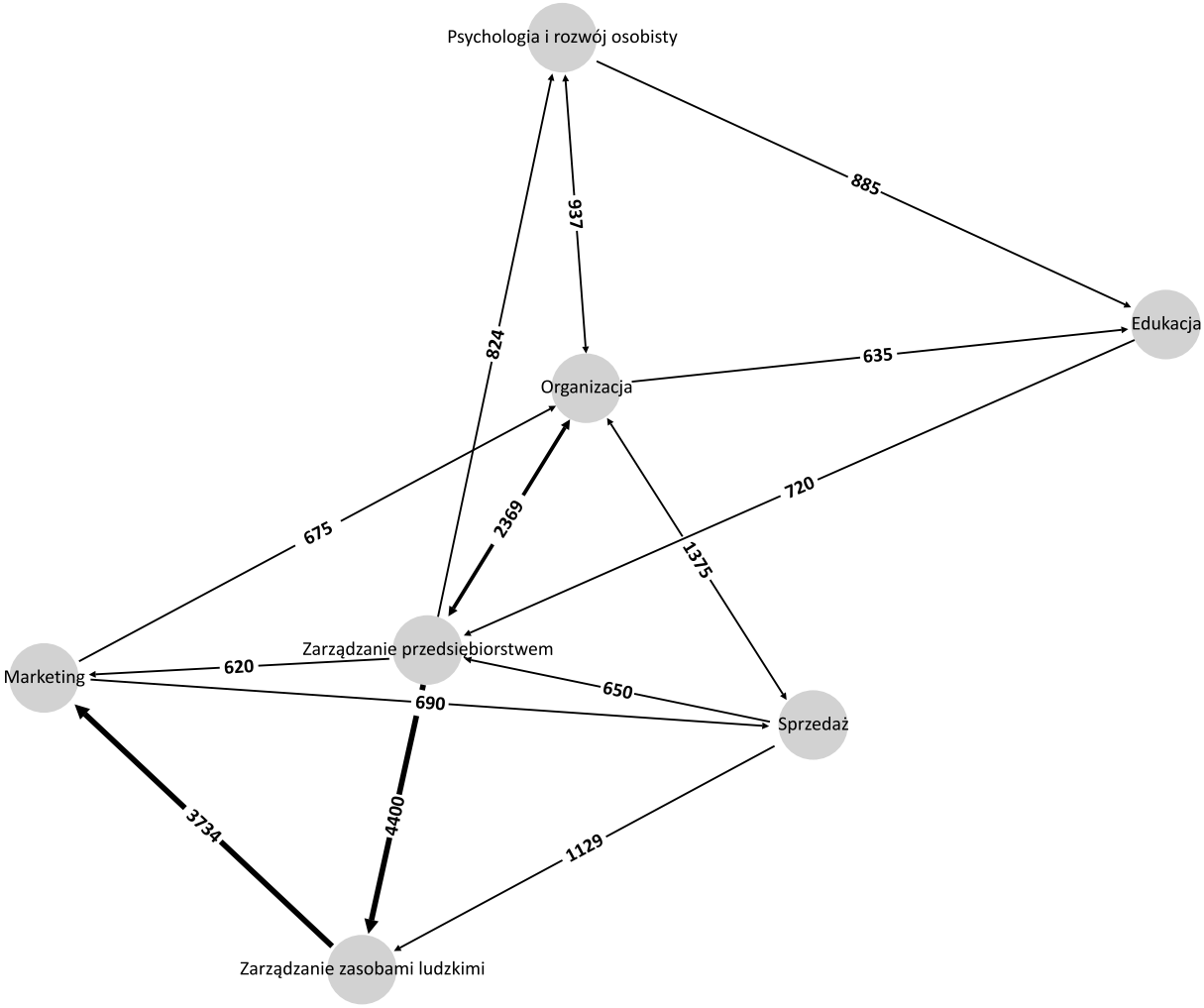
Najczęstsze przepływy zaobserwowano między zarządzaniem przedsiębiorstwem a zarządzaniem zasobami ludzkimi oraz między sprzedażą a marketingiem. W tabeli nie uwzględniono przepływów w obrębie tej samej kategorii ani mniej popularnych przejść.

Następnie zwrócono się do gen AI z prośbą o wygenerowanie kodu w Pythonie, który wizualizuje przepływy między usługami z wagami na krawędziach. Wykorzystano prompt:

> *Napisz, proszę, kod w Pythonie, który stworzy wizualizację grafu przepływów między usługami szkoleniowymi. Wygeneruj kod w Pythonie, który tworzy graf przepływów między 20 najczęstszymi usługami szkoleniowymi. Każdy węzeł reprezentuje usługę, a każda krawędź reprezentuje przepływ odbiorców. Grubość krawędzi ma być proporcjonalna do liczby odbiorców przepływających między węzłami. Węzły i krawędzie powinny być szare, a graf powinien być zapisany jako obraz PNG.*

Graf wygenerowany w ten sposób pokazuje strukturę sieci przepływów, w tym węzły (usługi) i ich powiązania.

Rycina 9.5. Graf obrazujący strukturę sieci przepływów, w tym węzły (usługi) i ich powiązania



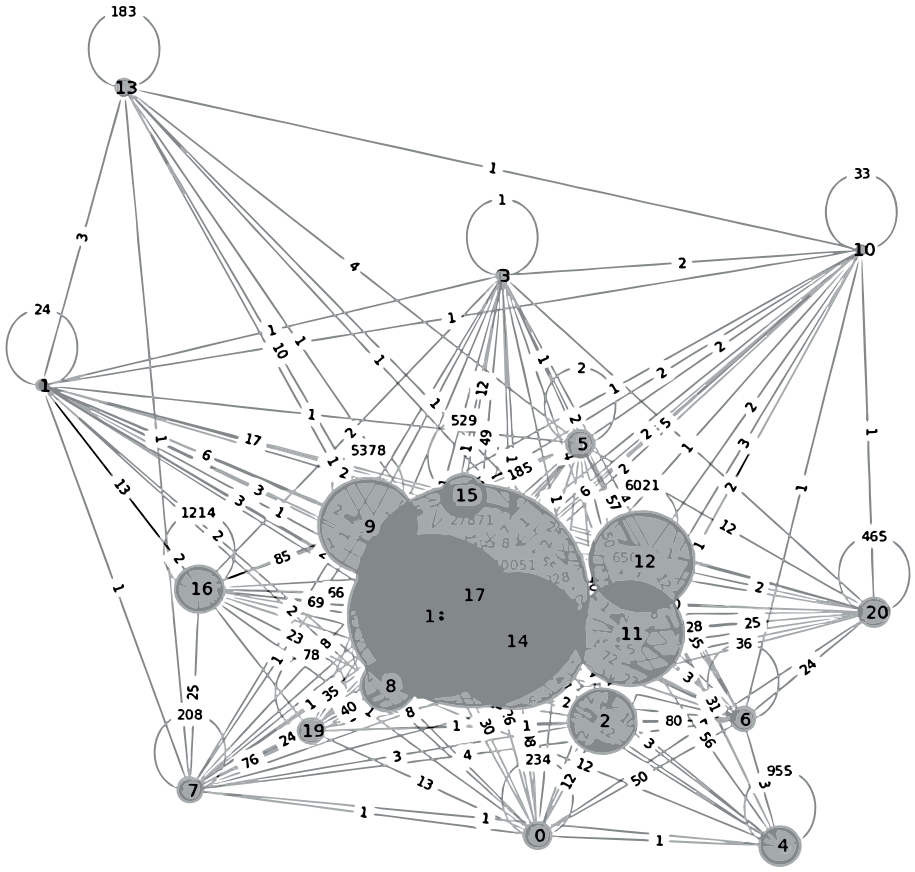
Źródło: Opracowanie własne.

Wykorzystanie gen AI pozwoliło pozyskać kod do analizy kierunkowych przepływów oraz ich wizualizacji, co ułatwiło identyfikację usług pełniących rolę centralnych punktów w sieci. Próbowano również wygenerować kod umożliwiający przedstawienie zróżnicowanej wielkości węzłów w zależności od ich centralności, jednak efekty nie były zadowalające. Wygenerowany kod nie spełniał oczekiwań pod względem czytelności i dokładności wizualizacji.

Analiza przepływów między społecznościami

Aby lepiej zrozumieć, jak odbiorcy przemieszczają się między różnymi grupami usług, przeanalizowano przepływy między społecznościami. Analiza miała na celu identyfikację relacji między społecznościami oraz wskazanie tych, które pełnią centralne role w sieci przepływów. Wyniki pokazują hierarchię społeczności oraz ich wzajemne powiązania, co dostarcza informacji o preferencjach odbiorców i charakterystyce oferowanych usług. Największe przepływy zaobserwowano między takimi społecznościami jak: Zarządzanie przedsiębiorstwem (społeczność 17), Usługi marketingowe i sprzedażowe (społeczność 5) oraz Zarządzanie zasobami ludzkimi (społeczność 18). Relacje między tymi społecznościami pokazują, że odbiorcy usług często wybierają usługi z różnych grup tematycznych, co może sugerować ich złożone potrzeby rozwojowe. „Społeczność 17" pełni szczególnie ważną rolę jako centralny punkt w sieci przepływów, co podkreśla znaczenie usług z tej grupy dla wielu odbiorców usług.

Rycina 9.6. Przepływy między społecznościami



Źródło: Opracowanie własne.

W przygotowaniu kodu, a także grafu, wsparcie zapewniła gen AI. Kod dotyczący przygotowania przepływów między społecznościami był poprawny, jednak stworzenie wykresu okazało się większym wyzwaniem. Problemy obejmowały niewyraźne przedstawienie wykresu, nachodzenie etykiet na siebie, niepoprawne kolory węzłów oraz trudności z uzyskaniem zróżnicowanych rozmiarów węzłów. Osoba biegła w Pythonie mogłyby zrezygnować z korzystania z gen AI na rzecz samodzielnego napisania kodu. Z kolei osoby bez umiejętności programistycznych, ale posiadające solidne podstawy merytoryczne w analizowanym obszarze, muszą zmierzyć się z tymi wyzwaniami, korzystając z wygenerowanego kodu i szukając rozwiązań w ramach iteracyjnej pracy z gen AI.

Analiza przepływów wewnętrznych i między dostawcami usług

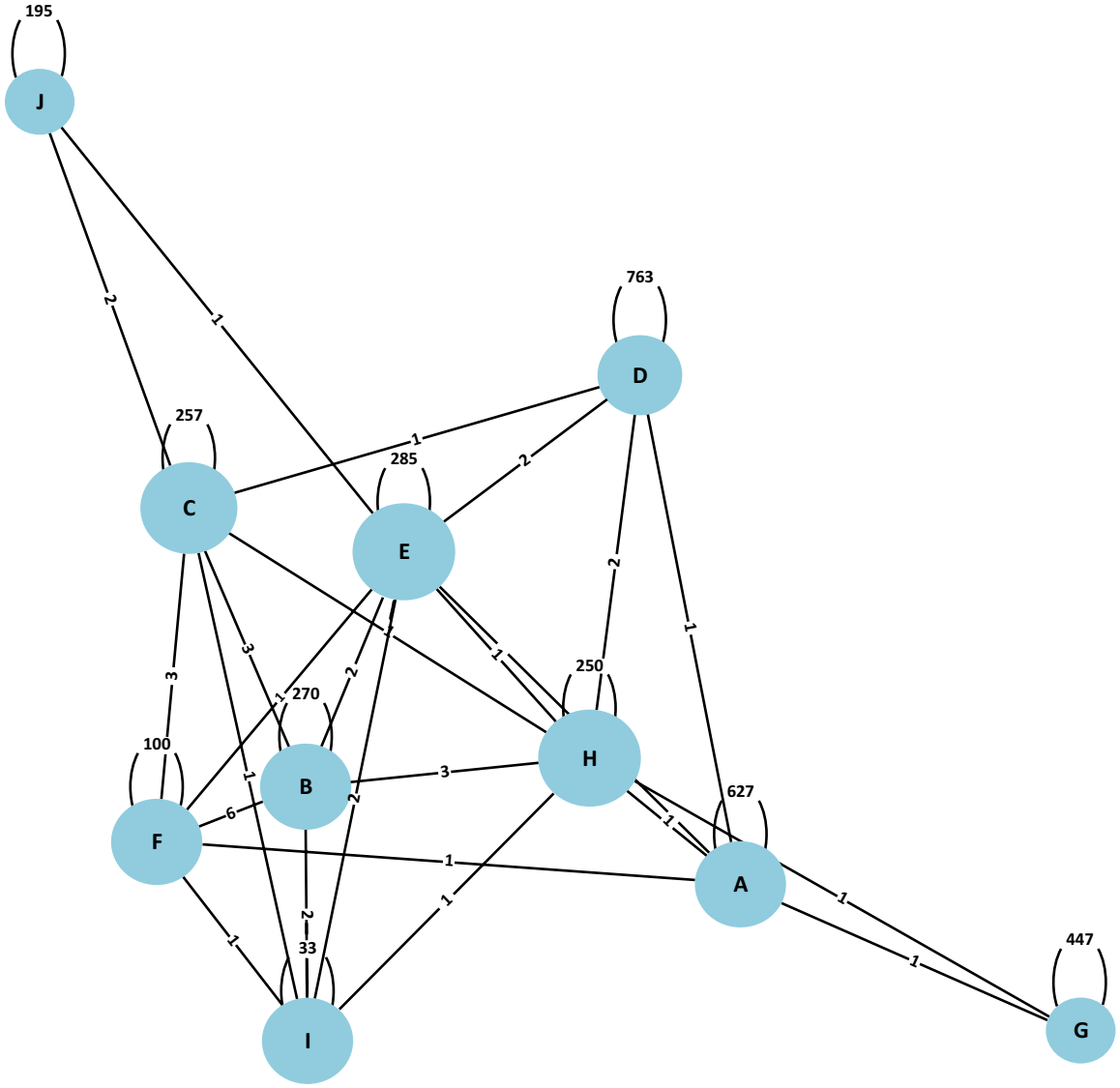
Ostatnia prezentowana w niniejszym rozdziale analiza sieciowa dotyczy przepływów wewnętrznych w obrębie jednego dostawcy oraz między różnymi dostawcami usług rozwojowych.

Generatywna AI otrzymała następujące zapytanie:

> *Przygotuj, proszę, kod Pythona, na podstawie którego będzie możliwa analiza przepływów wewnętrznych oraz między dostawcami w sieci usług. Oblicz, proszę, liczbę przepływów wewnętrznych dla każdego dostawcy, gdzie przepływ wewnętrzny to liczba odbiorców korzystających z różnych usług tego samego dostawcy. Stwórz tabelę z największymi przepływami wewnętrznymi, w której węzły reprezentują dostawców, a kolumna zawiera liczbę przepływów.*

Na podstawie otrzymanych wyników zaobserwowano, że największe przepływy wewnętrzne odnotowano u Dostawcy D (763), co świadczy o wysokiej lojalności jego klientów i szerokim zakresie usług oferowanych przez tego dostawcę. Przepływy między kluczowymi dostawcami zostały zilustrowane na wykresie, choć przy generowaniu grafu z wykorzystaniem modelu AI również wystąpiły problemy ze skalowaniem wielkości węzłów oraz grubości krawędzi (co można zauważyć na rycinie 9.7.).

Rycina 9.7. Przepływy między kluczowymi dostawcami usług



Źródło: Opracowanie własne.

Klasteryzacja społeczności

Za pomocą algorytmu Louvain'a zidentyfikowano społeczności, czyli grupy usług, które są często wybierane razem przez obiorców usług rozwojowych. Algorytm ten jest jedną z najczęściej stosowanych metod w analizie sieci społecznych (gron, klastrów) ze względu na efektywność i zdolność do wykrywania hierarchicznej struktury społeczności. Dzięki klasteryzacji było możliwe zrozumienie, jak różne usługi grupują się w sieci oraz jakie wspólne cechy mogą łączyć usługi należące do tych samych społeczności.

Prompt, który pomógł w generowaniu odpowiedniego kodu w modelu AI, to: *Stwórz, proszę, kod w Pythonie do przeprowadzenia analizy klasteryzacji społeczności w sieci usług rozwojowych na podstawie podanych ci wcześniej danych. Wykorzystaj algorytm Louvain'a. Przeprowadź klasteryzację społeczności w sieci.*

W wyniku przeprowadzonej klasteryzacji wyodrębniono 21 klastrów, z których każdy reprezentuje grupę usług rozwojowych często wybieranych razem (w pewnej sekwencji czasowej lub równolegle) przez odbiorców. Największe klastry obejmowały: klaster 0 – związany z IT i systemami komputerowymi, czyli takie usługi jak administracja IT, programowanie, obsługa komputera, projektowanie graficzne; klaster 1. – kursy językowe, takie jak: angielski, hiszpański, niemiecki, francuski; klaster 5. obejmujący usługi z zakresu marketingu, negocjacji i sprzedaży; klaster 13. zawierający usługi techniczne, taki jak kursy prawa jazdy różnych kategorii, obsługa maszyn, transport i logistyka.

Aby określić znaczenie poszczególnych usług w ramach klastrów, za pomocą kodu Pythona wygenerowanego przez gen AI, obliczono miary centralności dla każdego węzła:

- Centralność stopnia – usługi o wysokiej centralności stopnia są najbardziej połączone z innymi usługami w ramach klastra.
- Centralność pośrednictwa – usługi, które łączą różne części sieci, pełniąc rolę mostów.
- Centralność wektora własnego – usługi, które są najbardziej wpływowe w klastrze, biorąc pod uwagę znaczenie ich sąsiadów.

Dla największych klastrów wyniki przedstawiono w tabeli 9.5.:

Tabela 9.5. Miary centralności węzła

Klaster	Najważniejsze Usługi	Centralność
Klaster 0	Administracja IT, programowanie	Wysoka centralność stopnia i pośrednictwa
Klaster 1	Angielski, niemiecki	Wysoka centralność wektora własnego
Klaster 13	Transport i logistyka, kursy prawa jazdy	Wysoka centralność stopnia

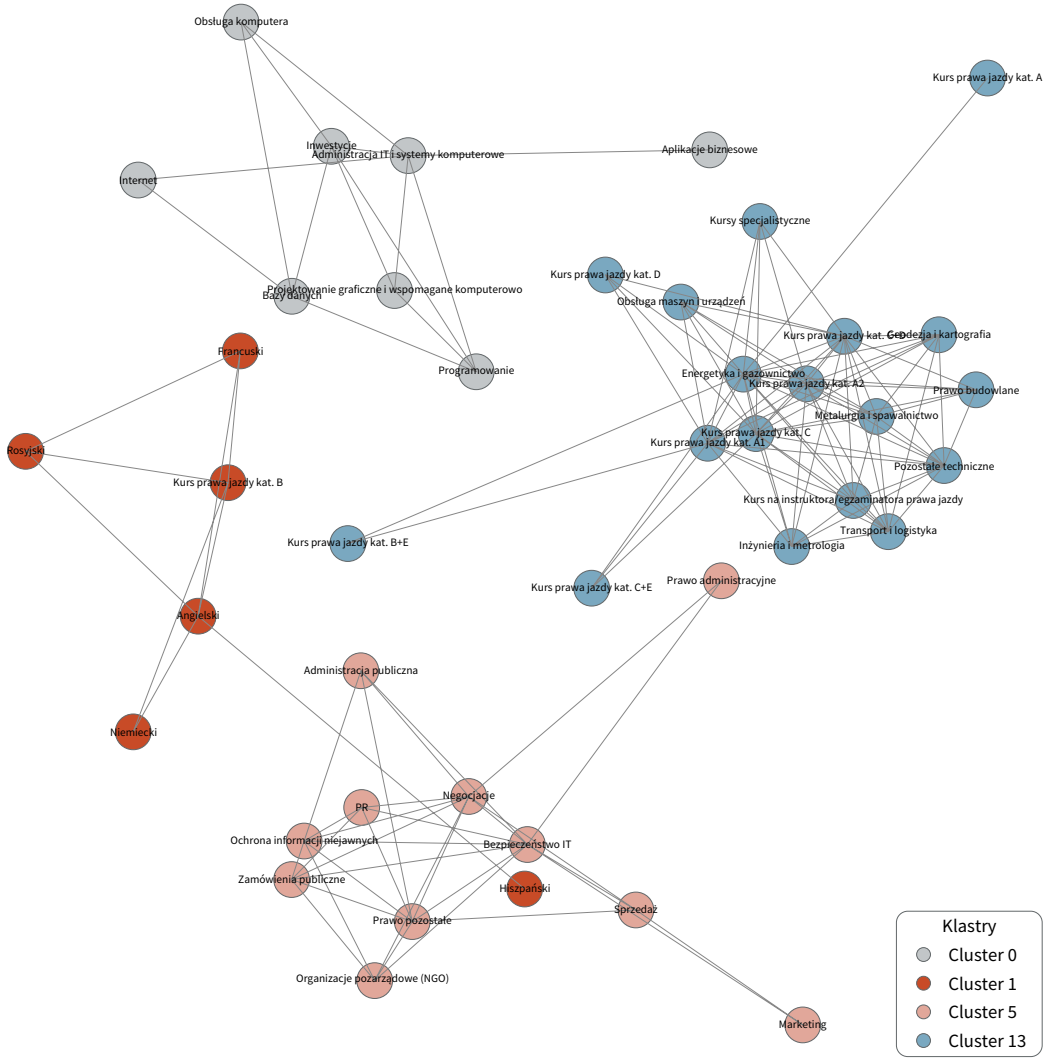
Źródło: Opracowanie własne.

Zwrócono się do gen AI o przygotowanie kodu Pythona do wizualizacji wyników.

Prompt, którego użyto: *Stwórz, proszę, kod w Pythonie, który wizualizuje sieć społeczności wybranych usług. Sieć powinna uwzględniać 4 wybrane klastry: 0, 1, 5 i 13. Każdy klaster ma być oznaczony innym kolorem. Rozmiar węzłów ma być uzależniony od centralności stopnia (degree centrality), a krawędzie mają łączyć węzły w obrębie klastrów. Dodaj etykiety węzłów z nazwami usług. Wyświetl legendę, która pokazuje kolory odpowiadające klastrom.*

Wynikiem działania kodu był poniższy graf. Jak widać, nie udało się wyskalować wielkości węzłów na podstawie centralności stopnia, gdzie większe węzły oznaczałyby bardziej centralne usługi, niemniej rezultat wykonanego zadania bez wątpienia wnosi wartość ilustracyjną dla odnotowanych zależności sieciowych.

Rycina 9.8. Wizualizacja klastrów



Źródło: Opracowanie własne.

Po uzyskaniu wizualizacji zwrócono się do gen AI o interpretację jej wyników:

Prompt: *Podaj, proszę, interpretację wyników uzyskanych na podstawie podanego przez ciebie kodu Pythona.*

W ramce poniżej przedstawiono wynik zapytania.

Rycina 9.9. Otrzymany wynik zapytania

Wnioski dotyczące społeczności: Społeczności wyodrębnione w wyniku analizy ukazują:

Powiązania między usługami: Usługi wybierane razem często należą do tego samego sektora lub odpowiadają na pokrewne potrzeby odbiorców (np. społeczność IT i programowania).

Hierarchię popularności: Większe społeczności obejmują szeroką gamę usług, co może świadczyć o ich uniwersalnym charakterze i wysokim popycie na związane z nimi kompetencje.

Możliwość personalizacji ofert: Wyniki wskazują na potencjał dostosowywania ofert szkoleniowych do potrzeb odbiorców wybierających konkretne grupy usług.

Źróło: Opracowanie własne.

Następnie zadano bardziej szczegółowe pytanie o interpretację:

> *Na podstawie przedstawionej wizualizacji grafu oraz wyników analizy miar centralności dla klastrów, podaj, proszę, analizę struktury wykrytych społeczności (klastrów), interpretację znaczenia kolorów w grafie, wyjaśnienie, dlaczego usługi zostały zgrupowane w takie klastry, charakterystykę połączeń wewnątrz klastrów versus między klastrami.*

Wówczas otrzymano następujący wynik:

Rycina 9.10. Otrzymany wynik zapytania

Analiza struktury klastrów:

Graf pokazuje wyraźny podział na 4 główne społeczności (oznaczone różnymi kolorami). Każdy klaster reprezentuje powiązaną tematycznie grupę usług. Widoczna jest relatywnie silna spójność wewnątrz klastrów i rzadsze połączenia między nimi.

Interpretacja miar centralności:

Klaster 0 (IT): administracja IT i programowanie jako węzły o wysokiej centralności stopnia sugerują, że te usługi są „hubami" w swojej dziedzinie. Wysoka centralność pośrednictwa wskazuje, że te usługi często służą jako punkt przejściowy do innych specjalizacji IT. Może to oznaczać, że odbiorcy usług często zaczynają od tych podstawowych kursów przed przejściem do bardziej specjalistycznych.

Klaster 1 (Języki): wysoka centralność wektora własnego dla kursów angielskiego i niemieckiego wskazuje na ich strategiczne znaczenie. Usługi te są połączone z innymi ważnymi węzłami w sieci, co sugeruje, że znajomość języków jest często łączona z innymi kompetencjami.

Klaster 13 (Transport): wysoka centralność stopnia dla transportu i logistyki oraz kursów prawa jazdy pokazuje silne wzajemne powiązania. Wskazuje na kompleksowe podejście odbiorców usług do edukacji w tym sektorze.

Rekomendacje: Warto rozwijać ścieżki edukacyjne łączące kursy IT z innymi specjalizacjami. Warto rozwijać pakiety łączące naukę języków z kursami branżowymi. Istnieje przestrzeń dla tworzenia spersonalizowanych ścieżek rozwoju, bazujących na wzorcach przepływów.

Kluczowe obserwacje: Widoczna jest naturalna segmentacja odbiorców usług według zainteresowań. Istnieją wyraźne „ścieżki kariery" w ramach każdego klastra. Niektóre usługi pełnią rolę pomostową między różnymi obszarami kompetencji.

Źróło: Opracowanie własne.

Można więc zauważyć, że im bardziej skonkretyzowane jest zapytanie, tym bardziej dopracowana, precyzyjna interpretacja jest generowana przez model.

Podsumowując, w procesie klasteryzacji gen AI wygenerowała kod do przeprowadzenia analizy, automatycznie przypisując usługi do społeczności na podstawie wyników algorytmu.

Gen AI dostarczyła też propozycje interpretacji wyników, które zostały zweryfikowane i dopracowane. Mimo że wnioski w pierwszej serii zapytań są względnie ogólne, to jednak poprawne. Dzięki wsparciu AI proces klasteryzacji był bardziej efektywny, a wyniki mogły zostać szybciej przekształcone w użyteczną formę. Kod wygenerowany przez AI oceniono jako prawidłowy, umożliwiający uzyskanie odpowiedzi na zadane pytanie, dotyczące klasteryzacji społeczności. Najwięcej problemów było z przedstawieniem wyników na grafie i dostosowaniem go do danych (w tym takich problemów jak brak etykiet, nieprawidłowe skalowanie węzłów czy brak połączeń).

Wnioski

Przeprowadzone analizy pozwoliły lepiej zrozumieć strukturę sieci usług rozwojowych, przepływy odbiorców usług oraz kluczowe elementy wpływające na ich wybory. Zidentyfikowano najpopularniejsze ścieżki szkoleniowe, w tym te związane z zarządzaniem przedsiębiorstwem oraz zarządzaniem zasobami ludzkimi, które odgrywają centralną rolę w sieci. Usługi sprzedażowe i marketingowe tworzą kolejne istotne grupy, wskazując na zapotrzebowanie na kompetencje wspierające działalność biznesową.

Klasteryzacja społeczności wykazała, że usługi można podzielić na wyraźnie zdefiniowane grupy, w których dominują specyficzne tematy, takie jak zarządzanie, finanse czy edukacja. Analiza przepływów między społecznościami ujawniła, że odbiorcy usług często migrują między społecznościami związanymi z zarządzaniem przedsiębiorstwem, zasobami ludzkimi oraz marketingiem. Wyniki te wskazują na potrzebę kompleksowych ścieżek rozwoju zawodowego, które łączą różne obszary tematyczne.

Przeprowadzone analizy uwidoczniły również ograniczenia danych wejściowych. Problemy takie jak brak etykiet dla niektórych dostawców usług czy trudności w analizie społeczności jednowęzłowych wskazują na potrzebę lepszej kontroli jakości danych w przyszłych tego typu analizach.

Doświadczenia z analizy mogą być także przydatne w kontekście innych branż czy typów danych, szczególnie tam, gdzie istnieje potrzeba identyfikacji przepływów między podmiotami lub ścieżek wyborów odbiorców. Zastosowana metodologia sieciowa i praca z gen AI mogą zostać rozszerzone na inne obszary, takie jak analiza rynków usług, procesów czy logistyki.

9.5. Podsumowanie

Celem niniejszego opracowania było przetestowanie możliwości generatywnej AI w realizacji zadań analitycznych, takich jak zmapowanie przepływów odbiorców usług i identyfikacja popularnych ścieżek szkoleniowych. Analizy miały ocenić, w jakim stopniu AI może wspierać proces przetwarzania danych sieciowych oraz jakie wyzwania wynikają z jej zastosowania. Rozdział ukazuje zarówno potencjał, jak i ograniczenia modeli językowych AI, takich jak ChatGPT, w analizie danych sieciowych.

Aby skutecznie wykorzystać generatywną AI, kluczowe jest posiadanie podstawowej wiedzy merytorycznej w przedmiocie i w prowadzeniu tego typu analiz. Praca z tym narzędziem opiera się na precyzyjnym formułowaniu instrukcji oraz krytycznym podejściu do generowanych wyników. Bez odpowiedniego przygotowania analityka istnieje ryzyko ograniczenia potencjału narzędzia oraz obniżenia jakości uzyskiwanych wyników. Praca z gen AI w analizie sieciowej pokazała, że modele językowe mogą znacząco przyspieszyć proces analizy danych, jednak wymagają aktywnego zaangażowania analityka. Technologia AI okazała się szczególnie pomocna w generowaniu kodu Pythona do wczytywania, przetwarzania i analizy danych, co pozwoliło zautomatyzować wiele procesów i skoncentrować się na interpretacji wyników.

Jednocześnie napotkano pewne ograniczenia – wygenerowany kod często wymagał dostosowania do specyfiki analizy, np. eliminacji społeczności jednowęzłowych, poprawy błędów związanych z typami danych czy dopracowania wizualizacji. W szczególności brakowało uwzględniania przez gen AI kluczowych parametrów, takich jak skalowanie węzłów czy odpowiednie rozmieszczenie etykiet. W efekcie w początkowych iteracjach generowane przez model grafy były mało czytelne i wymagały znaczącej ręcznej interwencji analityka. Wskazuje to na potrzebę dalszego rozwoju funkcji AI w zakresie tworzenia wizualizacji.

Badanie wykazało, że efektywność pracy z gen AI zależy od precyzyjnego formułowania zapytań (promptów) oraz znajomości podstaw programowania. Im lepiej określono cel i kontekst analizy, tym trafniejsze były generowane wyniki. Dla mniej doświadczonych analityków kluczowe jest rozwijanie umiejętności tworzenia zapytań oraz krytyczna ocena rezultatów, aby uniknąć błędnych interpretacji.

Wnioski z pracy z gen AI wskazują na jej potencjał w analizie danych sieciowych, szczególnie dla analityków posiadających doświadczenie w tej dziedzinie. Osoby mniej zaawansowane mogą napotkać trudności w ocenie poprawności kodu i wyników, co podkreśla znaczenie umiejętności analitycznych w pracy z gen AI. Mimo to rozwijanie kompetencji w zakresie analizy sieciowej oraz świadome korzystanie z gen AI może zwiększyć efektywność i jakość analiz. Model AI stanowi wsparcie w realizacji złożonych zadań analitycznych, ale jego pełne wykorzystanie wymaga odpowiedniego przygotowania i zaangażowania analityka.

Bibliografia

- Barnes, J.A. (1954). Class and Committees in a Norwegian Island Parish. *Human Relations*, 7, 39–58.
- Blondel, V.D., Guillaume, J.L., Lambiotte, R., et al. (2008). *Fast unfolding of communities in large networks* (dostęp: 31.03.2025).
- Borgatti, S.P., Everett, M.G., Johnson, J.C. (2013). *Analyzing Social Networks*, SAGE, 2018.

10. Strategie komunikacji i prezentacji wyników

Andrzej Gołoś
Paulina Grabowska

10.1. Wprowadzenie

Skuteczna komunikacja wyników badań ewaluacyjnych jest kluczowym elementem procesu ewaluacyjnego, ponieważ to od niej zależy, czy wyniki zostaną zrozumiane i wykorzystane przez różnorodne grupy odbiorców. Dobór odpowiednich strategii komunikacyjnych pozwala nie tylko na lepsze przedstawienie danych, ale również na ich wykorzystanie przy podejmowaniu decyzji, kształtowaniu opinii publicznej czy planowaniu dalszych działań.

Generatywna sztuczna inteligencja (gen AI) otwiera nowe możliwości w tym obszarze. Dzięki swoim zaawansowanym funkcjom może wspierać proces komunikacji na wielu etapach – od personalizacji treści raportów, przez ich wizualizację, aż po dostosowanie form i kanałów komunikacji do specyfiki odbiorców. To sprawia, że gen AI może znacząco podnieść jakość i efektywność całego procesu.

W niniejszym rozdziale postawione zostało pytanie: **w jaki sposób gen AI może wspomóc proces skutecznej komunikacji wyników ewaluacji do różnych grup odbiorców, z uwzględnieniem specyfiki ich potrzeb**? Aby na nie odpowiedzieć, przeprowadzono eksperyment, w którym krok po kroku zbadano, jak gen AI wspiera proces tworzenia person (profili) odbiorców oraz dostosowywania treści i wizualizacji wyników raportu ewaluacyjnego do ich potrzeb.

Nacisk został położony na ten etap projektu ewaluacyjnego, w którym raport został już opracowany i wymaga dostosowania do potrzeb różnych odbiorców. Analizowano, jak gen AI może wspierać użytkowników w prezentacji wyników, uwzględniając specyficzne wymagania

konkretnych grup odbiorców. Rozdział dotyczy komunikacji wyników, a nie analizy danych źródłowych, która jest tematem osobnego rozdziału książki.

W zadaniu polegającym na dostosowaniu przekazu i wizualizacji treści raportu do potrzeb odbiorcy wykorzystana została koncepcja **persony**. Persona to narzędzie koncepcyjne, którego celem jest zrozumienie potrzeb i oczekiwań użytkowników. Koncepcja wywodzi się ze świata projektowania produktów cyfrowych, gdzie została spopularyzowana przez Alana Coopera w książce *The Inmates are Running the Asylum*. Persona jest skróconym opisem profilu danego odbiorcy i systematyzuje kluczowe informacje na jego temat. Tym samym pozwala odpowiedzieć na pytanie: „Czy rozwiązanie, które proponujemy, odpowiada na potrzeby tego użytkownika i wspiera go w realizacji jego celów?”

W kontekście badań ewaluacyjnych persony umożliwiają dopasowanie komunikacji do specyfiki odbiorców wyników. Najistotniejsze elementy opisu persony obejmują:

- Tożsamość użytkownika – kim jest dana osoba, jej podstawowe role społeczno-zawodowe i cechy istotne w kontekście korzystania z raportu.
- Kontekst działania – sytuacje, w których użytkownik korzysta z wyników (np. presja czasu, dostępność narzędzi).
- Cele i potrzeby – motywacje użytkownika, które mogą skłonić go do zapoznania się z wynikami badań.
- Preferowane kanały i formy komunikacji (w tym np. poziom szczegółowości treści, dobór języka, form wizualizacyjnych i sposobu przedstawienia danych).

W eksperymencie zastosowano iteracyjny proces tworzenia person i dostosowywania raportów do ich potrzeb. Procedura obejmowała następujące kroki:

Krok 1: Tworzenie person

Rozpoczęto od stworzenia szkiców protoperson za pomocą ChatGPT (Open AI). Wprowadzone kategorie opisu obejmowały następujące zagadnienia: kim jest persona, kontekst działania, cele i potrzeby oraz preferowane formy i kanały komunikacji. ChatGPT wygenerował listę siedmiu propozycji protoperson, które poddaliśmy ocenie.

Krok 2: Zebranie danych z wywiadów

Równolegle przeprowadzono wywiady z pracownikami administracji publicznej, którzy na co dzień przetwarzają raporty ewaluacyjne. Rozmowy te dostarczyły szczegółowych informacji dotyczących potrzeb finalnych odbiorców, sytuacji, w których powstają materiały, oraz preferencji w zakresie form komunikacji. Umożliwiło to ocenę sensowności propozycji stworzonych przez ChatGPT.

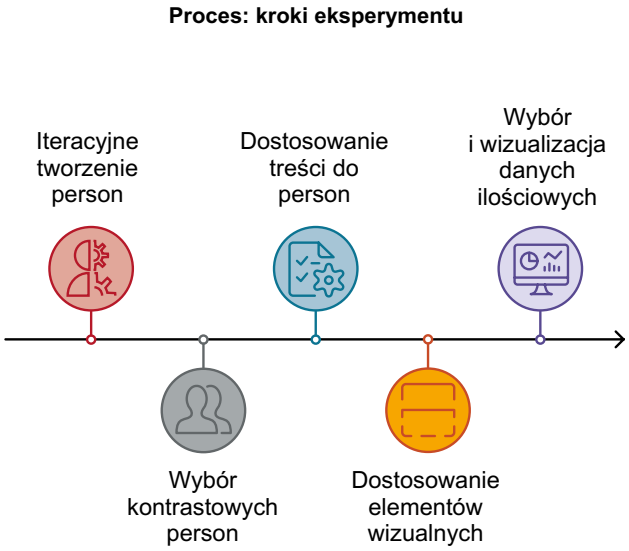
Krok 3: Wybór dwóch kontrastowych person

Do dalszej analizy wybrano dwie persony – decydenta politycznego i dziennikarza ekonomicznego – z uwagi na to, że reprezentują wyraźnie różne grupy celów i oczekiwań. Decydent polityczny potrzebuje zwięzłych, strategicznych informacji, które wspierają podejmowanie decyzji i komunikację sukcesów polityki publicznej. Dziennikarz ekonomiczny oczekuje materiałów medialnych, które mogą zostać szybko przekształcone w atrakcyjne publikacje.

Krok 4: Dostosowanie raportów do person

Każda persona została uwzględniona w procesie personalizacji treści raportu, jego prezentacji oraz wizualizacji danych. Dla celów eksperymentu wybrano raport pt. „Wpływ funduszy europejskich perspektywy finansowej 2007–2013 na rozwój społeczno-gospodarczy Polski Wschodniej”, opracowany na zlecenie Ministerstwa Infrastruktury i Rozwoju.

Rycina 10.1. Schemat przedstawiający procedurę tworzenia person i dostosowywania raportów do ich potrzeb



Źródło: Napkin AI, narzędzie, które przekształca nawet złożony tekst w proste diagramy.

Schemat ten stanowi podstawę dla dalszych eksperymentów z wykorzystaniem gen AI, które zostaną omówione w kolejnych częściach rozdziału, czyli:

- Podrozdział 2: Opis przebiegu procesu, w tym iteracyjne tworzenie person oraz dostosowywanie treści, elementów wizualnych i danych ilościowych raportu z wykorzystaniem gen AI.
- Podrozdział 3: Podsumowanie wyników i szczegółowe wnioski oraz rekomendacje, które mogą stanowić praktyczne wskazówki dla innych użytkowników.

Dzięki takiej strukturze nie tylko przedstawiono tok badania, ale również zaproponowano metodykę pracy, która może być użyteczna w przyszłych zastosowaniach gen AI w komunikacji wyników ewaluacji. Te odkrycia mogą przydać się się zarówno badaczom i ewaluatorom, którzy chcą efektywniej docierać do różnych grup odbiorców, jak i instytucjom publicznym oraz organizacjom poszukującym innowacyjnych sposobów na zwiększenie przejrzystości i skuteczności komunikacji opartej na danych.

10.2. Wykorzystanie generatywnej AI w dostosowywaniu treści raportów ewaluacyjnych – krok po kroku

Przebieg całego procesu wykorzystania różnych programów i platform bazujących lub wykorzystujących gen AI został w kompleksowy sposób przedstawiony w graficznym podsumowaniu, które w szczególności pokazuje, jak zastosowano sztuczną inteligencję do tworzenia person, dostosowywania treści raportów oraz opracowywania metod prezentacji wyników i wizualizacji danych[1].

Iteracyjny proces tworzenia person

Sprawdzono, jak dobrze poradzi sobie ChatGPT (w wersji 4o) z zadaniem stworzenia protopersony odbiorcy ewaluacji. Nie planowano przekazywać zbyt wielu informacji „na wejściu", aby sprawdzić inicjatywę modelu w zadawaniu pytań i ocenić trafność jego „intuicji". Zarysowano jedynie najważniejsze ramy zadania i zamierzano uzależnić ciąg dalszy od ewentualnych dodatkowych pytań modelu oraz pierwszych rezultatów prac.

W kategoriach opisu persony jako najważniejsze wskazane zostały na początek następujące wymiary:

- kim jest persona,
- kontekst, w którym funkcjonuje,
- cele persony, z powodu których może interesować się wynikami badań ewaluacyjnych,
- preferowane formy i kanały komunikacji.

Wybrane kategorie w sposób wzajemnie powiązany pozwalają określić:

- jakie potrzeby informacyjne ma dana osoba oraz jak przyswaja informacje,
- jak te informacje mogą wpłynąć na jej decyzje lub działania, co z kolei pomaga w selekcji odpowiednich informacji i wyników do wyeksponowania,
- jak persony mogą efektywnie odbierać wyniki.

[1] Wizualizacja procesu wykorzystania generatywnej AI w dostosowywaniu treści raportów ewaluacyjnych (Canva, n.d.) jest dostępna online.

Dookreślono jeszcze, że interesują nas dodatkowe dwa typy person. Pierwszy powinien odnosić się do osoby „wewnętrznej", czyli pracującej w administracji publicznej. Z kolei drugi typ to osoba „zewnętrzna", czyli spoza świata administracji publicznej. Ponadto wskazano, że oczekiwana jest wyczerpująca i kompletna listy. Następnie określono iteracyjny sposób wykonania zadania: po wstępnym promptcie miały nastąpić dodatkowe pytania ze strony modelu, a po udzieleniu odpowiedzi – kreacja listy person w ustalonym standardzie.

Model wygenerował dodatkowe pytania, które dotyczyły:

- roli sztucznej inteligencji w komunikacji wyników badań, co w kolejnej instrukcji pozwoliło doprecyzować, że nie chodzi o wsparcie w analizie, a jedynie w komunikacji wyników, polegającej na przetworzeniu gotowego raportu;
- rodzaju ewaluowanych programów, co jest istotne, bo dookreśla dziedzinę i tematykę raportów, a to z kolei ułatwia przewidzenie, jakie kategorie interesariuszy mogą interesować się taką tematyką;
- przykładów kanałów komunikacji – to ostatnie okazało się szczególnie cenne w kontekście odróżnienia tej kategorii od form komunikacji – pojęcia te bywają mylone, a formy mieszają się z kanałami, co utrudnia planowanie dalszych działań.

Dodatkowo wyjaśniono, jak należy rozumieć „raport ewaluacyjny" i co to znaczy, że może on być dalej wykorzystywany i przetwarzany.

Po takim „dopromptowaniu" ChatGPT zwrócił wstępną listę siedmiu protoperson odbiorców ewaluacji, którą uznaliśmy za całkiem przekonującą.

Wśród odbiorców „wewnętrznych" model wyodrębnił trzy, różnicując je pod względem zakresu odpowiedzialności (**specjalista ds. ewaluacji, menedżer programu oraz decydent polityczny**), z czego logicznie wynikał poziom szczegółowości lub ogólności materiału oczekiwanego przez tego typu osobę. Persony wydawały się reprezentować rzeczywiste funkcje i prawdziwych interesariuszy obecnych w procesie ewaluacyjnym, ale bez dodatkowych, zewnętrznych informacji trudno było to zweryfikować. Na liście zabrakło być może pracownika komórki komunikacji i promocji w instytucji organizującej ewaluacje – ostatecznie uznano jednak, że jest to odbiorca pośredni, który sam przetwarza materiał, dostosowując go do rzeczywistych odbiorców końcowych.

Wśród odbiorców zewnętrznych wszystkie propozycje: **eksperta-naukowca, przedstawiciela organizacji branżowej, przedsiębiorcy i dziennikarza**, wydały się adekwatne. Łatwo było sobie wyobrazić sytuacje, w których każda z takich osób mogłaby potrzebować informacji pochodzących z ewaluacji. Do rozważenia pozostawałaby jeszcze nieco paradoksalna persona tzw. „szerokiej publiczności" (paradoks polega na tym, że takiej kategorii trudno jest nadać przekonujące wspólne cechy, z uwagi na jej ogromne wewnętrzne zróżnicowanie). Faktem jest, że „ogół społeczeństwa" bywa wymieniany jako jedna z grup docelowych działań informacyjno-promocyjnych wokół programów finansowanych z funduszy UE. Prawdopodobnie drogą eliminacji dałoby się wskazać, jakiego rodzaju i jak podanych materiałów taka publiczność z pewnością nie oczekuje oraz jakimi kanałami warto się z nią komunikować. Trudniej byłoby wskazać elementy kontekstu czy celów, z powodu których mogłaby chcieć zapoznawać się z odkryciami ewaluacji.

Opis protoperson pod ustalonymi względami wydawał się trafny, zwięzły, a mimo to posiadający dużą wartość informacyjną. Repertuar rekomendowanych kanałów i form komunikacji był bogaty. Przede wszystkim jednak propozycje były wewnętrznie spójne: można było łatwo ustalić związek pomiędzy różnymi elementami opisu persony – z jej celów i kontekstu wynikały oczekiwania co do form i kanałów komunikacji. Niekiedy stało się to kosztem pewnej stereotypizacji modelowych sylwetek (decydent pracuje pod presją czasu, przedsiębiorca chce się przekonać, jak jego firma wypadła na tle innych), co jest zresztą jednym z głównych zarzutów podnoszonych wobec podejścia wykorzystującego koncepcję person.

Podstawowym problemem jest oczywiście fakt, że protopersona nie została jeszcze w żaden sposób zweryfikowana i nie mamy możliwości ustalenia, jak dobrze opisuje rzeczywistość. Wyraźnie podkreślamy, że persona powinna być stale udoskonalana w świetle wiedzy o realnym użytkowniku, zwłaszcza o jego reakcjach na pierwsze wersje naszego „produktu". Bez takiego stałego uczenia się pozostajemy w kręgu własnych przypuszczeń, wspieranych jedynie ogólnymi danymi treningowymi modelu, nieuwzględniającego niuansów lokalnych i specyfiki poszczególnych organizacji.

Ogólnie jednak można uznać, że pierwsza iteracja zakończyła się sukcesem. Osoba, która nie ma żadnej wiedzy na temat możliwych odbiorców ewaluacji, z pomocą gen AI jest w stanie pewnie postawić pierwszy krok i stworzyć praktyczny materiał do dalszego rozwoju. Ingerencja człowieka (*human-in-the-loop*) była minimalna, a dodatkowo częściowo była

ukierunkowana przez pytania zadane przez sam model, co znakomicie ułatwiło wykonanie pierwszej iteracji. Rezultaty ćwiczenia skłaniają do zadania istotnych pytań, których często się nie stawia – kto będzie odbiorcą, dlaczego może potrzebować materiału i jak efektywnie można mu go przekazać – co samo w sobie jest wartościowe i pożyteczne.

Równocześnie z procesem tworzenia protoperson przeprowadzono trzy wywiady z pracownikami administracji publicznej przygotowującymi (samodzielnie lub przy pomocy zespołu, którym kierują) materiały na podstawie raportów z badań ewaluacyjnych. Celem wywiadów było lepsze zrozumienie potrzeb docelowych odbiorców, sytuacji, w których pojawia się potrzeba przygotowania materiału i oczekiwań odbiorców co do formy materiału. Rezultatem tych rozmów miała być weryfikacja listy protoperson oraz finalny wybór tych, którym poświęcono uwagę w dalszej części eksperymentu.

I tak w rozmowie z pracownikiem Departamentu Analiz i Strategii Polskiej Agencji Rozwoju Przedsiębiorczości (PARP) wstępnie zmapowaliśmy listę interesariuszy korzystających z ewaluacji i sytuacje, w których ma to miejsce. W rozmowie z przedstawicielem Krajowej Jednostki Ewaluacji Ministerstwa Funduszy I Polityki Regionalnej (MFiPR) rozmawiano o tym, w jaki sposób i przy jakich okazjach z wyników ewaluacji korzystają decydenci (np. ministrowie i wiceministrowie). Z kolei w wywiadzie z przedstawicielami Departamentu Komunikacji i Marketingu PARP ustalono najczęstsze zadania i wyzwania związane z korzystaniem z ewaluacji i popularyzacją płynących z niej wniosków.

Wywiady te doprowadziły do decyzji, by do dalszych prac wybrać dwie sylwetki – **polityka i dziennikarza**. Przemawiały za tym następujące argumenty:

- Na potrzeby dalszego eksperymentowania, polegającego na próbie opracowania wyników przykładowego raportu z uwzględnieniem potrzeb modelowych odbiorców, wystarczająca była krótka lista dobrze wyodrębnionych person. Interesowało nas sprawdzenie możliwości uzyskania wariantów materiału, a nie przetestowanie wszystkich możliwych kombinacji tej wariantowości.
- Istotne było wybranie jednej persony wewnętrznej i jednej zewnętrznej. Polityk potrzebuje skondensowanej, strategicznej informacji dostosowanej do bieżących decyzji, a dziennikarz musi komunikować wyniki w przystępnej formie dla szerokiej publiczności. Dzięki tej różnicy można sprawdzić, jak gen AI radzi sobie z tworzeniem materiałów dla odbiorców o odmiennych potrzebach.

- W trakcie wywiadów uzyskano stosunkowo najwięcej szczegółowych informacji o tych właśnie sylwetkach, a także przykładów konkretnych materiałów. Pojawiła się wobec tego szansa sprawdzenia, jak materiał przygotowany przez gen AI przystaje do faktycznych potrzeb.

W tym samym czasie w trakcie przeglądu literatury zidentyfikowano opisy odbiorców wiedzy, opracowane w ramach projektu „Brokerzy Wiedzy" przez EGO s.c. i Pracownię Gier Szkoleniowych w 2015 r. Opis dotyczył trzech sylwetek odbiorców wiedzy w obrębie administracji publicznej: menedżera, polityka i dyrektora. Uznano, że zarówno struktura tego opisu, wzbogacająca zaproponowany w opracowaniu układ o kategorię „tematów, które mogą interesować personę", jak i jego szczegółowość, mogą być dobrymi punktami odniesienia w dalszych pracach. Najważniejszym argumentem było to, że sylwetki te opracowano na podstawie pogłębionych badań o znacznie większej skali, niż możliwa do podjęcia w ramach prac nad niniejszym artykułem. Spełniały one więc to, co postuluje się w tym rozdziale: były zweryfikowane na podstawie informacji pochodzących od użytkowników.

W kolejnej konwersacji z ChatGPT potwierdzono wybór dwóch person, zażądano uzupełnienia ich opisu o tematy interesujące dla persony oraz zachowanie szczegółowości i stylu opisu, jak w udostępnionym gen AI materiale wzorcowym („Brokerzy Wiedzy").

W rezultacie wszystkich opisanych działań powstał opis dwóch person. Ich zestawienie zawiera poniższa tabela – łatwo przy jej pomocy wychwycić kluczowe różnice pomiędzy obydwoma profilami.

Tabela 10.1. Porównanie dwóch person wybranych do dalszych prac

Kategoria	Decydent polityczny	Dziennikarz ekonomiczny
Kim jest persona	Polityk, odpowiedzialny za kształtowanie polityki publicznej, szczególnie w kontekście funduszy UE	Dziennikarz specjalizujący się w ekonomii i rozwoju regionalnym
Kontekst	Pracuje pod presją czasu, potrzebuje szybkiego dostępu do kluczowych informacji	Tworzy materiały medialne, potrzebuje danych z potencjałem do wzbudzenia zainteresowania publicznego
Cele i potrzeby	Informacje wspierające podejmowanie decyzji i komunikację sukcesów polityki publicznej	Dane medialne, które można szybko przekształcić w atrakcyjne publikacje

Kategoria	Decydent polityczny	Dziennikarz ekonomiczny
Preferowane formy	Zwięzłe raporty strategiczne, kluczowe wskaźniki, syntetyczne podsumowania	Gotowe wizualizacje, streszczenia, dane medialne w przystępnej formie
Preferowane kanały	Briefingi, raporty przygotowywane na potrzeby negocjacji i spotkań	Artykuły prasowe, media cyfrowe, infografiki

Eksperyment pokazał, że ChatGPT sprawnie generuje początkowe wersje person, zwłaszcza kiedy otrzymuje dobrze zdefiniowane kategorie opisu. Dzięki dodatkowym pytaniom model potrafił skierować proces w odpowiednim kierunku, dostosowując swoje wyjściowe założenia do naszych wskazówek.

Główną zaletą gen AI było szybkie tworzenie zarysów person na podstawie minimalnych założeń wyjściowych. Protopersony wydawały się trafne, choć czasem stereotypowe, co jest jednym z wyzwań pracy z AI. Modele gen AI dobrze radzą sobie z tworzeniem profili prostszych grup odbiorców, jednak w przypadku bardziej złożonych interesariuszy, takich jak przedstawiciele różnych funkcji w obrębie administracji publicznej, AI może potrzebować więcej kontekstu i zdecydowanie wymaga weryfikacji przez człowieka przy pomocy badań z użytkownikami. Materiał z takich badań gen AI potrafi wykorzystać do poprawy i rozbudowy pierwszych propozycji.

W dalszej części eksperymentu postanowiono sprawdzić, jak będą wyglądały materiały przygotowane przez gen AI z uwzględnieniem specyfiki obu person.

Dostosowanie treści raportów

W tym podrozdziale opisano, w jaki sposób dostosować treści wybranego raportu „Wpływ funduszy europejskich perspektywy finansowej 2007–2013 na rozwój społeczno-gospodarczy Polski Wschodniej. Raport końcowy” do potrzeb person decydenta i dziennikarza. Dla każdej z dwóch person wykorzystano różne metody pracy z gen AI, aby pokazać szersze spektrum dostępnych rozwiązań.

Dla persony decydenta politycznego założono, że są potrzebne precyzyjne i rzetelne informacje wspierające podejmowanie strategicznych decyzji. W tym celu wykorzystaliśmy ChatGPT do szczegółowej analizy raportu, włącznie z interpretacją danych przedstawionych

na wykresach. Na podstawie tej analizy przygotowano kluczowe elementy do stworzenia prezentacji, która nie tylko w przejrzysty sposób przedstawia najważniejsze wnioski z raportu, ale także zawiera dopasowane wizualizacje w formie wykresów danych oraz ilustracje wygenerowane z promptów.

Dla persony dziennikarza ekonomicznego opracowano dedykowanego chatbota opartego na GPT, zaprojektowanego do personifikacji dziennikarza i dostosowywania treści raportu do jego specyficznych potrzeb. W efekcie powstała zwięzła informacja prasowa oraz scenariusz krótkiej animacji, która podsumowuje tę informację w sposób atrakcyjny i przystępny.

Oba przedstawione rozwiązania w interesujący sposób wspierają dostosowanie języka, formy oraz treści do specyficznych oczekiwań wybranych grup odbiorców badań, a także zapewniają skuteczne i bardziej angażujące przekazywanie wyników.

Aby we współpracy z ChatGPT przygotować treści raportu do prezentacji dla decydenta politycznego, konieczne było zrealizowanie pięciu kluczowych etapów. Każdy z nich został szczegółowo i precyzyjnie opisany w formie prompta, co umożliwiło skuteczne dopasowanie treści do potrzeb tej grupy odbiorców.

Podsumowanie kluczowych etapów:

1. **Zaznajomienie ChatGPT z opracowaną już wcześniej personą decydenta politycznego:** AI przeanalizowało personę decydenta i wymieniło jego główne potrzeby, takie jak szybki dostęp do kluczowych wniosków i rekomendacji, które mogą mieć wpływ na podejmowane decyzje.
2. **Zaznajomienie ChatGPT z treścią raportu ewaluacyjnego:** AI przeprowadziło analizę danych i wniosków zawartych w raporcie, aby wybrać te informacje, które są najbardziej istotne dla decydenta politycznego.
3. **Symulacja persony decydenta politycznego przez ChatGPT:** AI poproszone, by wcielić się w personę, wygenerowało w punktach zwięzłe podsumowanie raportu, dostosowane do oczekiwań decydenta politycznego.
4. **Dopasowanie wizualizacji danych:** Aby każdy punkt podsumowania raportu był pełniejszy o kluczowe wizualizacje danych, ChatGPT podjął próbę dopasowania wykresów zawartych w raporcie.
5. **Dopasowanie wizualizacji w formie ilustracji:** W celu uzupełnienia prezentacji o walory estetyczne dla każdego punktu podsumowania ChatGPT wygenerował opis prompta,

dzięki któremu mogły powstać dopełniające prezentację ilustracje. (Ten etap został szerzej opisany w podrozdziale pt. „Dostosowanie sposobu prezentacji wyników”).

Ewaluacja procesu wykazała, że generatywna AI może szybko dostosować treści raportów do potrzeb decydenta politycznego, ale kluczowe dla efektywności są dwie kwestie: odpowiednia strategia pracy z AI oraz stałe monitorowanie jej wyników w celu zapewnienia rzetelności. Aby skutecznie zaplanować współpracę, należy bardzo jasno określić cel – w naszym przypadku była to prezentacja z kluczowymi wnioskami, uzupełniona wykresami i ilustracjami, atrakcyjna i czytelna dla decydenta[2].

W przypadku dostosowywania treści raportu do potrzeb persony dziennikarza ekonomicznego zastosowano inne rozwiązanie Chat GPT, a mianowicie GPTs (*Generative Pre-trained Transformers*). GPTs otrzymał szczegółowe wytyczne oparte na wcześniejszej analizie potrzeb dziennikarza, aby mógł wcielić się w jego rolę i dynamicznie dostosowywać treści raportu ewaluacyjnego.

Oczekiwania dziennikarza ekonomicznego obejmowały m.in.: zwięzłą informację prasową o raporcie i krótką animację podsumowującą, którą można wykorzystać na stronie internetowej jako uzupełnienie artykułu.

Wprowadzenie GPTs opartego na szczegółowo opracowanej personie dziennikarza znacznie bardziej usprawniło proces dostosowywania treści raportów ewaluacyjnych do specyficznych potrzeb persony aniżeli wcześniej opisany sposób wykorzystujący jedynie konwersację z ChatGPT. Kluczowe korzyści tego podejścia to:

- **Zwięzłość i klarowność odpowiedzi** – GPTs raz dostosowany do preferencji persony dziennikarza ekonomicznego generuje krótkie i precyzyjne odpowiedzi, łatwe do wykorzystania w informacjach prasowych i scenariuszach animacji. Model koncentruje się na kluczowych danych liczbowych i trendach, co ułatwia przygotowanie zwięzłych, treściwych materiałów prasowych, spełniających oczekiwania tej grupy odbiorców.
- **Dostosowanie do wymagań medialnych** – GPTs opracowany dla dziennikarza ekonomicznego uwzględnia potrzebę prezentowania informacji w formie przystępnej dla szerokiej publiczności. Generowane treści są uproszczone i zrozumiałe, co ułatwia

[2] ChatGPT, OpenAI, komunikacja osobista, wrzesień 2024; pełny przebieg promptowania.

zarówno ich interpretację, jak i publikację w mediach, zwiększając ich dostępność i atrakcyjność dla odbiorców.

- **Szybkość uzyskania odpowiedzi** – Wykorzystanie modelu GPTs umożliwia błyskawiczne generowanie odpowiedzi, co jest kluczowe zarówno dla tworzenia zarysu, jak i pełnych publikacji wyników. W porównaniu ze standardową konwersacją z ChatGPT, jak w przypadku decydenta politycznego, gdzie model musi być każdorazowo instruowany, aby wcielić się w rolę określonej persony, GPTs specjalnie zaprojektowany pod kątem konkretnej persony działa znacznie szybciej. Taka specjalizacja usprawnia przetwarzanie danych i znacząco skraca czas pracy nad projektem.
- **Sugestie tematów i wizualizacji** – Model GPTs generuje propozycje promptów wspierających wizualizację w programach takich jak Midjourney. Dzięki temu dostarcza nie tylko dane, ale także pomysły na animacje i infografiki, które mogą wzbogacić artykuły medialne oraz uatrakcyjnić finalne publikacje, wspierając kompleksowe podejście do prezentacji treści.
- **Wielokrotne wykorzystanie GPTs** – GPTs raz stworzony może być używany wielokrotnie do opracowywania różnych raportów i treści. W przeciwieństwie do standardowej rozmowy z ChatGPT nie wymaga każdorazowego przypominania, w jaką personę ma się wcielić i jakie są jej wymagania. Dzięki temu proces jest bardziej efektywny i spójny.

Tworzenie GPTs dla dziennikarza ekonomicznego, jak i dla każdej innej wybranej persony, jest procesem kilkuetapowym. Kluczowe i zalecane do dalszych eksperymentów kroki obejmowały:

1. **Tworzenie persony** – Pierwszym krokiem było zdefiniowanie szczegółowej persony dziennikarza ekonomicznego, bazowano przy tym na opracowanych wcześniej preferencjach persony dotyczących zwięzłości, klarowności oraz potrzeby uzyskiwania informacji, które są łatwe do publikacji. Kluczowe elementy persony obejmowały: specjalizację w analizie gospodarczej, potrzebę szybkiej dostawy danych, preferencje dotyczące wizualizacji (animacje, infografiki) oraz zrozumiały styl komunikacji.
2. **Doprecyzowanie odpowiedzi GPTs** – Kluczowym etapem było określenie, jak model powinien generować odpowiedzi, aby sprostać oczekiwaniom dziennikarzy. Odpowiedzi musiały być krótkie, klarowne i oparte na konkretnych danych liczbowych, jednocześnie należało unikać nadmiaru szczegółowych, technicznych informacji. Ważnym elementem było także uwzględnienie sugestii dotyczących przekształcania danych w atrakcyjne publikacje medialne, takie jak animacje i wizualizacje, które mogłyby wzbogacić przekaz.

3. **Dostosowanie tonu wypowiedzi** – Ton odpowiedzi GPTs został dopasowany do wymagań dziennikarza ekonomicznego: formalny, ale przystępny. Odpowiedzi musiały być profesjonalne i zrozumiałe, co jest kluczowe w pracy z raportami gospodarczymi i politycznymi.
4. **Dodanie funkcji kreatywnych** – Wprowadzono zdolność modelu do sugerowania scenariuszy animacji, które wizualnie wyjaśniają złożone dane. GPTs generował pomysły na narracje, które można przekształcić w krótkie animacje, co znacząco zwiększyło atrakcyjność publikacji.

Stworzony model GPTs wspiera tworzenie różnorodnych treści, takich jak informacje prasowe, publikacje, artykuły, wytyczne do infografik oraz scenariusze animacji i filmów. Może być więc wszechstronnym narzędziem dla działu odpowiedzialnego za promocję w instytucji.

Na podstawie doświadczeń związanych z tworzeniem GPTs dla dziennikarza ekonomicznego opracowano następujące zalecenia:

- **Zwięzłość i precyzja** – model GPTs musi generować krótkie, precyzyjne odpowiedzi, które są łatwe do przekształcenia w publikacje.
- **Fokus na dane liczbowe i wizualizacje** – ważne jest, aby GPTs koncentrował się na dostarczaniu twardych danych oraz sugerował możliwe formy wizualizacji, które można wykorzystać w artykułach.
- **Dostosowanie do specyfiki pracy medialnej** – model musi być elastyczny i potrafić dostosować swoje odpowiedzi do różnych typów raportów, w zależności od ich charakteru.
- **Monitorowanie i weryfikacja odpowiedzi** – regularna kontrola jakości generowanych odpowiedzi jest kluczowa dla upewnienia się, że GPTs spełnia wymagania i jest zgodny z danymi zawartymi w raportach.

Narzędzie, jakim jest GPTs, okazało się skuteczne w szybkim dostosowywaniu treści raportów do potrzeb medialnych, generowało odpowiedzi zrozumiałe dla dziennikarzy i gotowe do publikacji. Jednak należy pamiętać, że każda odpowiedź wymaga weryfikacji przez człowieka, a regularne monitorowanie jakości jest niezbędne, aby uniknąć błędnych interpretacji danych[3].

[3] Stworzony na cele tej publikacji GPTs dziennikarza ekonomicznego jest dostępny pod tym adresem. GPTs dziennikarza ekonomicznego, OpenAI, komunikacja osobista, wrzesień 2024; pełny przebieg promptowania.

Dostosowanie sposobu prezentacji wyników

Generatywne modele AI mogą z powodzeniem wspierać proces tworzenia multimedialnych prezentacji wyników raportów, obejmujący generowanie slajdów, grafik oraz animacji wraz z ich scenariuszami. Na rynku istnieje wiele takich narzędzi, których możliwości rozwijają się dynamicznie. W tym przypadku kluczowe było skupienie się na metodologii i traktowanie wybranych narzędzi jako elementu eksperymentu.

Podobnie jak w poprzednim podrozdziale, w którym dla każdej persony dostosowano treści raportów, opracowano różne metody prezentacji wyników zgodne z preferencjami określonymi na etapie tworzenia person. Dla decydenta politycznego przygotowano prezentację, a dla dziennikarza ekonomicznego – informację prasową w formie tekstowej oraz jej podsumowanie w postaci animacji.

W celu wykonania graficznej prezentacji dla decydenta politycznego wykorzystano ChatGPT do stworzenia scenariusza prezentacji, z podziałem na slajdy zawierające kluczowe wyniki raportu, krótkie opisy, wykresy danych i ilustracje. Następnie scenariusz przetestowano w Canvie, wykorzystując jej narzędzia AI.

Tworzenie scenariusza przy pomocy ChatGPT obejmowało takie zadania jak:

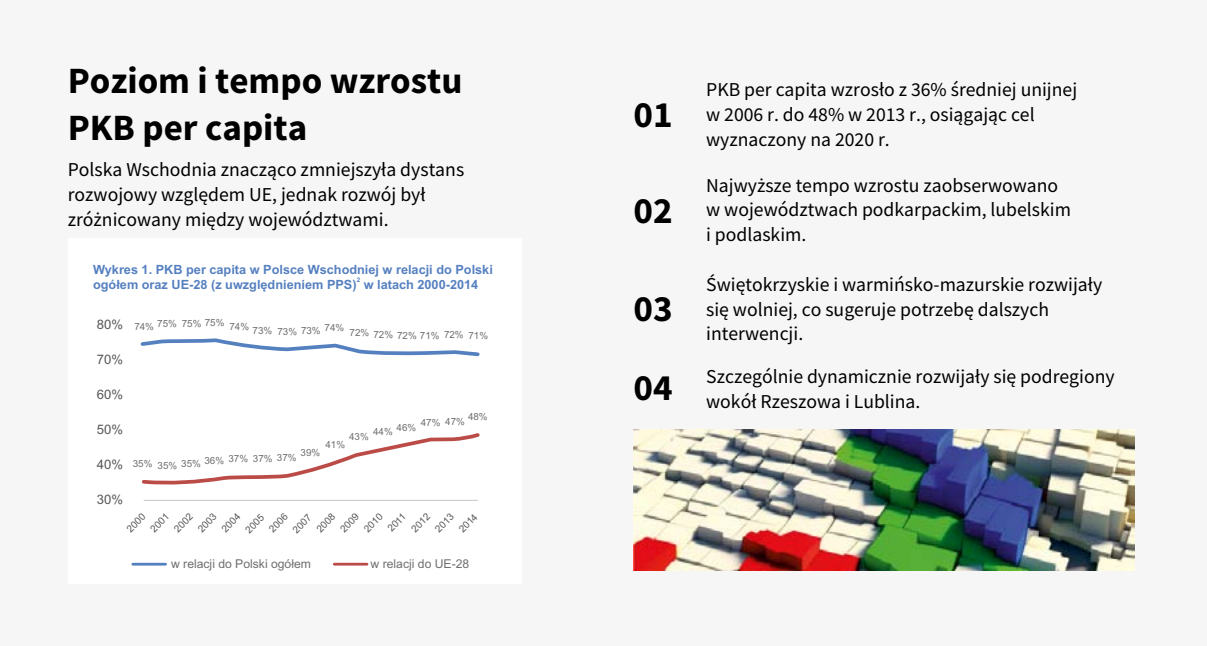
- **Podział na slajdy z przyporządkowanymi do nich informacjami i kluczowymi danymi** – co pozwoliło na szybki wybór szablonu prezentacji w programie Canva.
- **Dopasowanie wizualizacji danych z raportu do poszczególnych slajdów** – co pozwoliło stworzyć spójną informacyjnie prezentację.
- **Opis promptów do wykorzystania w generatorach AI obrazów** – co pozwoliło na wygenerowanie grafiki dla każdego slajdu przy użyciu narzędzi gen AI dostępnych w Canva.

Z kolei wykorzystanie scenariusza w programie graficznym wspieranym AI, na przykładzie Canva, obejmowało:

- **Wygenerowanie różnych szablonów prezentacji** – za pomocą krótkiego prompta w programie Canva stworzono kilka wariantów, aby ocenić, który najlepiej sprawdzi się do prezentacji wyników.

- **Dopasowanie szablonu do scenariusza** – dostosowano liczbę slajdów oraz rozmieszczenie informacji zgodnie z założeniami scenariusza. Na slajdach umieszczono takie elementy jak: tytuł, krótki opis, wykres danych oraz punkty przedstawiające kluczowe informacje.
- **Dodanie ilustracji na każdym slajdzie** – ilustrację wygenerowano w narzędziu AI dostępnym w Canva na podstawie promptów zawartych w scenariuszu.
- **Dopracowanie prezentacji pod kątem wizualnym**[4] – w tym weryfikacja czytelności prezentacji.

Rycina 10.2. Jeden ze slajdów prezentacji dla decydenta politycznego wykonany w programie Canva, przy pomocy dostępnych w nim narzędzi AI. Scenariusz całej prezentacji oraz prompty do wykonania ilustracji wykonano w Chat GPT.



Źródło (dostęp: 29.11.2024).

Z kolei w przypadku zadania dostosowania treści raportu do potrzeb persony dziennikarza ekonomicznego wykorzystano dedykowany chatbot GPTs do wykonania dwóch zadań polegających na napisaniu informacji prasowej o raporcie oraz scenariusza do animacji (wraz z promptami do Midjourney – narzędzia gen AI umożliwiającego tworzenie grafik). GPTs wykonał analizę raportu, następnie przedstawił informacje kluczowe z punktu widzenia

[4] Z efektem wykorzystania scenariusza stworzonego przy pomocy ChatGPT i narzędzi AI dostępnych w Canva można zapoznać się pod tym adresem.

persony dziennikarza ekonomicznego. W następnym kroku przy użyciu GPTs przygotowano opis informacji prasowej, który może zostać wykorzystany na stronie instytucji.

Ostatnim krokiem było użycie GPTs do stworzenia scenariusza do animacji wybranego przez nas zagadnienia: „Ewolucja PKB oraz inwestycje w infrastrukturę transportową, z podkreśleniem różnic między województwami i pozytywnego wpływu funduszy UE". Zadanie polegało na opracowaniu czterech krótkich zdań opisujących temat oraz czterech promptów, które wykorzystano do generowania grafik Midjourney. Następnie, aby nadać temu elementowi wizualizacji danych formę animacji, użyto narzędzia AI dostępnego w programie Canva[5]. Jeśli wygenerowane grafiki nie spełniają oczekiwań pod względem jakości wizualnej, warto skorzystać z narzędzi opartych na AI do ich ulepszania, takich jak Krea czy Magnific AI.

Rycina 10.3. Klatki z animacji podsumowującej informację prasową przygotowaną dla dziennikarza ekonomicznego. Wizualizacja została stworzona za pomocą narzędzia Midjourney, a animacja opracowano w Canva przy wykorzystaniu dostępnych w tym programie narzędzi AI. Całość, łącznie z tekstami, została przygotowana na podstawie scenariusza opracowanego z pomocą GPTs personifikującego dziennikarza ekonomicznego

Źródło (dostęp: 29.11.2024).

[5] Uzyskana w ten sposób animacja w Canva jest dostępna pod tym adresem.

Generatywne modele AI, takie jak ChatGPT, Midjourney czy funkcje AI dostępne w Canva, mają ogromny potencjał w procesie budowania argumentacji i scenariuszy wizualizacji wyników dla różnych grup odbiorców. AI znacząco przyspiesza proces tworzenia treści oraz wizualizacji, jednocześnie umożliwiając precyzyjne dostosowanie przekazu do specyficznych potrzeb odbiorców takich jak decydenci polityczni i dziennikarze ekonomiczni.

Jednak mimo korzyści wynikających z automatyzacji, narzędzia gen AI wymagają stałego monitorowania przez człowieka, aby zapewnić rzetelność i kompletność danych. Modele gen AI mogą znacznie zwiększyć efektywność pracy, ale kluczowe pozostaje zaangażowanie profesjonalistów, którzy potrafią efektywnie korzystać z tych narzędzi, jednocześnie kontrolując ich wyniki. Ważnym aspektem jest również dopasowanie estetyki wizualizacji do potrzeb danych person.

Dostosowanie wizualizacji danych

Ostatnią część rozdziału poświęcono zastosowaniom gen AI w obszarze wizualizacji danych. Pytanie, na które zostanie udzielona odpowiedź, brzmi: „**Czy generatywna sztuczna inteligencja jest w stanie ułatwić wizualizację danych pochodzących z raportów ewaluacyjnych?**”. W rozdziale zostanie opisane, czego udało się dowiedzieć, próbując przy pomocy gen AI:

- zidentyfikować w raporcie dane liczbowe, kluczowe z punktu widzenia person,
- wybrać odpowiedni sposób wizualizacji takich danych,
- stworzyć rekomendowane sposoby wizualizacji.

Przykładowy raport ewaluacyjny, dotyczący wpływu funduszy europejskich perspektywy finansowej 2007–2013 na rozwój społeczno-gospodarczy Polski Wschodniej, został „przepuszczony” przez ChatGPT. W serii promptów poproszono o:

- dobór najważniejszych danych ilościowych i wskaźników z uwzględnieniem potrzeb i zainteresowań obu person: decydenta politycznego i dziennikarza ekonomicznego,
- uzasadnienie wyboru w odniesieniu do konkretnych cech persony,
- rekomendację adekwatnych sposobów wizualizacji danych z uwzględnieniem wymagań obu person,
- ekstrakcję rekomendowanych danych z raportu,
- stworzenie wizualizacji na podstawie danych.

W ten sposób pokazano, że wizualizację danych rozumie się szeroko – nie tylko jako techniczną czynność tworzenia określonych graficznych układów danych, ale jako szerszy proces komunikacji. Obejmuje on także decyzję, co i dlaczego zaprezentować potencjalnemu odbiorcy.

W każdym z tych kroków oceniono, jaki potencjał kryje się w wykorzystaniu gen AI oraz z jakimi ryzykami należy się liczyć.

Dobór danych i wskaźników oraz jego uzasadnienie

Postawiony przed zadaniem wyboru kluczowych danych i wskaźników, które mogą być interesujące dla obu person, ChatGPT zaproponował w przypadku decydenta politycznego:

- PKB *per capita* Polski Wschodniej w stosunku do średniej unijnej rok po okresie objętym analizą (2014).
- Udział funduszy unijnych we wzroście PKB *per capita* Polski Wschodniej w analizowanym okresie.
- Łączna wartość środków europejskich skierowanych do Polski Wschodniej w latach 2007–2013.

ChatGPT uzasadnił dobór tych wskaźników, wskazując, że każdy z nich ilustruje interwencję na poziomie makro-, analizuje całościowy wpływ funduszy na gospodarkę, dostarcza argumentów dla procesów decyzyjnych (np. związanych z negocjacjami skali wsparcia w kolejnym okresie programowania) oraz „paliwo" dla komunikacji sukcesów polityki publicznej.

Na tej podstawie można sformułować kilka obserwacji i wniosków:

- Dobór wskaźników jest trafny i uwzględnia specyfikę persony. W syntetycznym ujęciu raportu te wskaźniki prawdopodobnie znalazłyby się na liście kluczowych dla decydenta politycznego. Można by łatwo rozszerzyć listę tych wskaźników, np. poprzez dobranie trzech kluczowych danych ilustrujących każdy z wniosków z rozdziałów raportu lub zagadnień tematycznych.
- Wszystkie zaproponowane wskaźniki oraz ich wartości są prawidłowe i faktycznie zawarte w raporcie – w jego części tekstowej.

- Nasza modelowa persona jest bardzo ogólna i silnie zestereotypizowana, nie uwzględnia kontekstu i celów konkretnego zadania. Kolejne prompty mogłyby poprawić to dopasowanie.
- Efekt stereotypizacji widać w koncentracji na pozytywnym obrazie (zgodnie z zasadą doboru „paliwa" dla komunikacji sukcesów). Raport zawiera jednak także opis problemów i wyzwań, zwłaszcza w kontekście rosnącego dystansu rozwojowego między Polską Wschodnią a resztą kraju. Te problematyczne wątki zostały pominięte w pierwszej iteracji, ale ChatGPT zidentyfikował je w kolejnych.

W przypadku persony dziennikarza ekonomicznego ChatGPT zaproponował następujące dane:

- Spadek PKB *per capita* Polski Wschodniej w porównaniu ze średnią krajową.
- Dwukrotny wzrost odsetka powiatów o najniższym poziomie rozwoju w badanym okresie.
- Ranking województw Polski Wschodniej według wsparcia *per capita* z funduszy europejskich.

ChatGPT wyjaśnił, że wskaźniki zostały dobrane ze względu na ich potencjał medialny (wzbudzenie zainteresowania odbiorcy), wzbogacenie narracji przez zastosowanie kontrastów i porównań oraz prezentację sukcesów wybranych regionów.

Również w tym przypadku dobór wskaźników wydaje się właściwy. Przy trafnym opisie persony ChatGPT mógłby zaproponować użyteczny zestaw danych do materiału dla dziennikarza. Warto jednak zwrócić uwagę na efekt stereotypizacji, skutkujący wyborem przede wszystkim danych kontrowersyjnych lub negatywnych, nawet kosztem adekwatnej reprezentacji ogólnego wydźwięku raportu.

Ponadto, w zadaniu wyboru danych dla dwóch person jednocześnie ChatGPT miał tendencję do rozłącznego doboru wskaźników (inne dla każdej persony) i uzasadniał słuszność tej zasady w dalszej części konwersacji. W rzeczywistości większość wskaźników mogłaby znaleźć się w skrótowym materiale dla każdej z person. Sugeruje to, że opracowanie materiału z wykorzystaniem podejścia opartego na personach powinno dotyczyć jednorazowo tylko jednej persony.

Podsumowując ten pierwszy etap ćwiczenia, można stwierdzić użyteczność najpopularniejszego obecnie narzędzia gen AI w zakresie personalizacji komunikatów,

doboru adekwatnych danych ilościowych. Proponowane wskaźniki reprezentują kluczowe tezy materiału źródłowego i są dopasowane do modelowych sylwetek odbiorców. Jednakże selekcja wymaga krytycznej refleksji, bez której istnieje ryzyko wzmocnienia efektu stereotypizacji person. Ponadto, praktycznym zaleceniem jest przygotowywanie doboru danych dla jednej persony na raz, a nie równocześnie dla kilku.

Dobór sposobów wizualizacji danych i wskaźników

Właściwy dobór sposobu wizualizacji danych to złożone zagadnienie, szeroko omawiane w literaturze specjalistycznej. Wraz ze wzrostem ilości dostępnych danych i ich powszechnego wykorzystania znaczenie tej dziedziny staje się coraz większe. Najważniejsze kryteria, które powinna spełniać dobra wizualizacja danych, to:

- Przejrzystość – wizualizacja powinna ułatwiać zrozumienie danych i przekazywać informacje w sposób klarowny, bez zbędnego komplikowania, które mogłoby zdezorientować odbiorcę.
- Adekwatność do danych – sposób wizualizacji musi być dopasowany do rodzaju danych, np. wykresy liniowe najlepiej sprawdzają się do prezentacji trendów w czasie, natomiast wykresy kołowe czy słupkowe lepiej ilustrują proporcje.
- Celowość – wizualizacja powinna odpowiadać na konkretne pytania lub wspierać odbiorcę w podejmowaniu decyzji, skupiając się na kluczowych elementach, zamiast przytłaczać nadmiarem informacji.
- Estetyka i czytelność – dobrze zaprojektowana wizualizacja powinna być atrakcyjna wizualnie, ale jednocześnie czytelna. Zbyt bogata grafika może utrudniać odbiór danych, a zbyt prosty design może nie przyciągnąć uwagi odbiorcy.
- Dostępność – wizualizacja powinna być dostosowana do potrzeb różnych grup odbiorców, w tym osób z niepełnosprawnościami. Należy uwzględnić m.in. dostępność kolorystyczną czy teksty alternatywne.

Użytkownik pośredni raportu ewaluacyjnego, przygotowujący na jego podstawie kolejne materiały, niekoniecznie musi być specjalistą w dziedzinie wizualizacji danych. Dlatego podpowiedzi gen AI, dotyczące doboru odpowiednich wizualizacji, mogą okazać się cenne. Przyjrzyjmy się, jak ChatGPT poradził sobie z tym zadaniem.

Dla decydenta politycznego za najbardziej odpowiednie uznano wykresy słupkowe i linie trendów. Pierwsze – ze względu na ich czytelność i możliwość szybkiego zapoznania się

z kluczowymi tendencjami, drugie – z uwagi na istotną dla tej persony obserwację trendów w czasie.

Dla persony dziennikarza ekonomicznego ChatGPT zaproponował mapy cieplne, uznając je za adekwatne do pokazania kontrastów i lokalnych różnic, oraz wykresy punktowe do przestrzennej ilustracji dystansów między regionami.

Co do zasady, te rekomendacje wydają się trafne i dopasowane – zarówno do typu konkretnych danych, jak i potrzeb person. Jednak zadanie zostało wykonane w sposób maksymalizujący różnice między personami, co skutkowało rozłącznymi rekomendacjami typów wizualizacji. Nie musi to być optymalne rozwiązanie. Zasada zrozumiałości i łatwości szybkiego zapoznania się z danymi jest istotna także dla dziennikarza i jego odbiorców. Dlatego mapy cieplne, a zwłaszcza wykresy punktowe, choć atrakcyjne wizualnie, w pewnych warunkach mogą okazać się nieodpowiednie. Z kolei wykresy prezentujące linie trendu i wykresy słupkowe z powodzeniem mogłyby zostać wykorzystane również w przypadku wskaźników proponowanych dla dziennikarza. Po raz kolejny wskazuje to na potrzebę przygotowywania materiału dla jednej persony na raz.

Ekstrakcja danych z raportu

W przyjętym scenariuszu użytkownik pośredni (np. pracownik jednostki ewaluacyjnej w instytucji) korzysta z gotowego raportu, opracowanego przez ewaluatora. Jego zadaniem jest dostosowanie materiałów bazujących na tym raporcie do konkretnego kontekstu i potrzeb odbiorców końcowych. Nie zakładano, że użytkownik ten jest analitykiem, mającym dostęp do danych źródłowych (np. surowych wyników badania ankietowego czy tabeli z rezultatami modelowania ekonometrycznego).

Dlatego podjęto próbę ustalenia, jak ChatGPT radzi sobie z odczytywaniem danych przedstawionych w raporcie (w tekście, na wykresach i w tabelach) do celów ich wtórnego wykorzystania. Prompt zawierał polecenie stworzenia tabeli na podstawie zidentyfikowanych danych, odnoszących się do rekomendowanych wskaźników. W przypadku trudności z taką ekstrakcją zadaniem ChatGPT miało być podanie tytułu rozdziału, w którym znajduje się wykres lub tabela zawierająca dane, by ułatwić użytkownikowi samodzielne spisanie danych.

W tym zadaniu napotkaliśmy istotne ograniczenia ChatGPT:

- Narzędzie nie było w stanie przypisać znajdujących się w raporcie danych do rekomendowanych wskaźników. Mimo że niektóre rekomendowane wskaźniki były zilustrowane w raporcie w formie prawidłowo zatytułowanego wykresu, ChatGPT nie potrafił ich zidentyfikować.
- Brak możliwości ekstrakcji danych z wykresu (obrazu) w momencie prowadzenia eksperymentu, był znanym ograniczeniem ChatGPT (brak OCR). Warto zauważyć, że zdecydowaną większość danych w analizowanym raporcie zaprezentowano właśnie na wykresach nieposiadających tekstu alternatywnego ani tekstowej tabeli z danymi, a w niektórych przypadkach – nawet nieposiadających etykiet wartości, zaprezentowanych przy poszczególnych punktach serii danych. Jest to, niestety, dość powszechna praktyka w procesie przygotowywania raportów z badań ewaluacyjnych, której skutkiem jest ograniczenie możliwości realizacji zadań wspieranych przez gen AI, takich jak te w naszym eksperymencie. Warto zauważyć, że niektóre inne narzędzia gen AI, takie jak np. Claude (Anthropic), z pewnością poradziłyby sobie lepiej z zadaniem ekstrakcji danych dzięki zaawansowanym możliwościom OCR.
- Odczytanie danych z tabel, nawet z uwzględnieniem dużego stopnia ich złożoności (tabela z trzema wymiarami), nie powodowało trudności. Jednak łatwo sobie wyobrazić, że w określonych warunkach mogłoby być problematyczne, na przykład wtedy, gdy wiersze i kolumny tabeli nie byłyby dostatecznie jasno opisane, tabela zawierałaby scalone pola, zagnieżdżone struktury, niestandardowe znaki czy symbole (np. dodatkowe graficzne oznaczenie trendu) lub brak by było odpowiednich separatorów między elementami tabeli.
- ChatGPT mylił się w identyfikacji rozdziałów, w których miały znajdować się dane rekomendowane do wizualizacji. Podawał tytuły nieistniejących rozdziałów i generował fałszywe informacje na temat ich zawartości. Doprecyzowanie prompta (ograniczenie listy tytułów rozdziału tylko do tych wymienionych w spisie treści) w niektórych przypadkach pomogło w prawidłowej identyfikacji rozdziału, ale w innych nadal skutkowało błędnymi wskazaniami – rekomendowanych danych nie można było znaleźć we wskazywanym rozdziale. Ponownie niektóre inne narzędzia wykorzystujące generatywną sztuczną inteligencję (np. NotebookLLM, Claude 3 Sonnet czy Gemini 2.0, a także systemy AI integrujące generatywne modele językowe z bazami wiedzy w podejściu typu RAG) prawdopodobnie lepiej poradziłyby sobie z prawidłową identyfikacją rozdziałów.

Prowadzi to do następujących wniosków:

- W tej części zadania (ekstrakcja danych) ChatGPT nie jest obecnie użytecznym narzędziem i nie może być wykorzystany przez osoby zajmujące się wtórnym wykorzystaniem raportów ewaluacyjnych.
- Aby w przyszłości ułatwić wsparcie ChatGPT przy takich zadaniach, należałoby wprowadzić nowy standard przygotowywania raportów ewaluacyjnych przez ich autorów. Mógłby on obejmować takie elementy jak: posługiwanie się wykresami wyłącznie wraz z tabelą danych, dodawanie metadanych do wykresów ze szczegółowymi danymi oraz określenie wymagań związanych z formatem tabel (np. wyłącznie dwa wymiary z jednoznacznie opisanymi kategoriami, brak scalania pól itp.). Warto zauważyć, że niektóre z tych standardów są już obecnie stosowane przez instytucje w ramach wdrażania zasad dostępności cyfrowej (*Web Content Accessibility Guidelines*, WCAG), m.in. raportów ewaluacyjnych.
- Nawet po takich zabiegach niezbędne byłoby bardzo dokładne sprawdzenie poprawności ekstrakcji danych przez człowieka.

Do tej pory ta część naszego eksperymentu przebiegała w sposób ciągły i dawała obiecujące rezultaty. W obszarze ekstrakcji danych ta ciągłość została przerwana. Aby zrealizować ostatni etap – stworzyć wizualizację – niezbędne było przygotowanie danych o strukturze odpowiadającej wybranym miernikom. Wykonano to również przy pomocy ChatGPT i komend w języku naturalnym.

Tworzenie wizualizacji

ChatGPT Plus ma wbudowane możliwości przygotowania wykresów na podstawie danych. Dane mogą być podane zarówno w oknie konwersacji (zwłaszcza jeśli mają prostszą strukturę i nie są zbyt rozbudowane), jak i w pliku tekstowym, w formie załącznika. Na tej podstawie, przy użyciu komend w języku naturalnym, możemy przygotować prostsze wykresy i nadać im pożądaną stylistykę[6].

[6] Przykład takiego użycia.

W ten sposób ChatGPT poradzi sobie m.in.:

- z przygotowaniem wykresów słupkowych czy kołowych – ale nie poradzi sobie z zastosowaniem bardziej złożonych wykresów w rodzaju map (potrzebuje dostępu do danych dotyczących geokodowania) czy przepływów (np. wykres Sankey'a),
- z edycją kolorów wykresów w pełnym zakresie,
- ze zmianą domyślnych ustawień wykresu – np. dodanie etykiet wartości, obramowanie serii danych, usunięcie lub zmiany tytułu wykresu; nie poradzi sobie np. z zadaniem dostosowania kroju czcionki.

ChatGPT oferuje zatem równoważny zakres funkcjonalności, taki jak rozbudowane narzędzia analityczne z modułami wizualizacji danych wykorzystującymi AI (np. ThoughtSpot, Pyramyd czy Tableau Pulse). Równoważność polega na tworzeniu wizualizacji danych przy pomocy komend w języku naturalnym oraz zdolności do autonomicznej rekomendacji typu wykresu do rodzaju danych.

Bardziej złożone wizualizacje nie są możliwe do uzyskania przy pomocy wbudowanego modułu analizy i wizualizacji danych. Jednak popularne narzędzia oparte o LLM potrafią wspierać użytkownika w tworzeniu kodu, np. Python, R czy Javascript, i wykorzystaniu rozbudowanych bibliotek tych programów, wyspecjalizowanych w wizualizacji danych. W ten sposób osoby nieposiadające umiejętności programistycznych są w stanie bez większych problemów osiągnąć zamierzony cel, wyłącznie komunikując swoje potrzeby w języku naturalnym.

W toku prac spróbowano przetestować taką możliwość. Wyzwanie polegało na stworzeniu postulowanej przez ChatGPT wizualizacji z wykorzystaniem „mapy cieplnej" (mapy, która natężeniem kolorów przekazuje informacje o natężeniu zjawiska lub cechy). W dodatku powierzchnią takiej mapy cieplnej miało być terytorium geograficzne, a wizualizowanymi kategoriami – jednostki administracyjne. Dla uproszczenia zdecydowano, że będą to województwa Polski, a nie powiaty czy gminy województw wschodniej Polski, jak w przykładzie wykorzystanym w eksperymencie, ale nie zmienia to istoty zadania.

W stosunkowo krótkim czasie przy pomocy ChatGPT i komend w języku naturalnym wypracowano kod, który umożliwił:

- stworzenie pliku wsadowego ze sztucznymi danymi, przypisującymi województwom Polski losowe (testowe) dane z określonego zakresu,

- uzyskanie zakładanej wizualizacji, w kilku iteracjach poprzez dobranie właściwych kolorów kodujących natężenie prezentowanej cechy,
- otrzymanie gotowej prezentacji w Power Point, zawierającej wypracowaną grafikę, tytuł oraz podstawowy opis zawartości (listę trzech województw z najwyższą i najniższą wartością wskaźnika).

Rezultat tej pracy został przedstawiony na ryc. 10.4.

Rycina 10.4. Wykres typu mapa ciepła, wygenerowany w ChatGPT na podstawie danych testowych dla województw

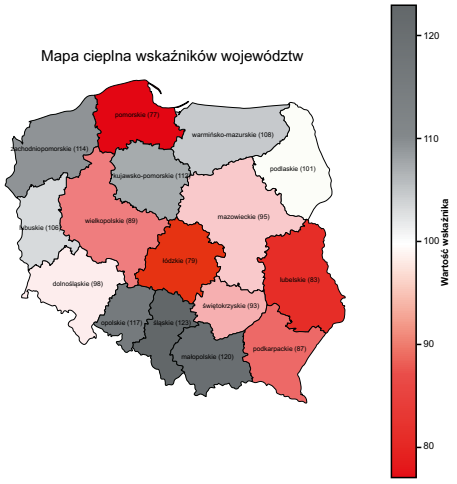
MAPA CIEPLNA WSKAŹNIKÓW WOJEWÓDZTW

Trzy województwa z najwyższą wartością wskaźnika to:

- śląskie - 123.0
- małopolskie - 120.0
- opolskie - 117.0

Trzy województwa z najniższą wartością wskaźnika to:

- pomorskie - 77.0
- łódzkie - 79.0
- lubelskie - 83.0



Źródło: Opracowanie własne.

10.3. Podsumowanie

Eksperymenty przedstawione w ostatniej części rozdziału dostarczyły informacji na temat możliwości i ograniczeń gen AI w kontekście wizualizacji danych. Okazało się, że aktualnie sztuczna inteligencja ma znaczny potencjał w zakresie:

- Wsparcia w wyborze danych – ChatGPT jest w stanie zidentyfikować kluczowe wskaźniki z raportu i przypisać je do konkretnych potrzeb person, co pomaga użytkownikom pośrednim raportu w dalszym personalizacji komunikatów.
- Wsparcia w doborze wizualizacji – ChatGPT może dostarczyć rekomendacje dotyczące odpowiedniej wizualizacji dla danego rodzaju danych. Jest przy tym w stanie uwzględnić potrzeby i ograniczenia odbiorców komunikacji.
- Wykonania określonych wizualizacji – dzięki wbudowanym możliwościom ChatGPT Plus, można przygotować proste wizualizacje i dostosować je do podanych wymagań, a następnie wydatnie przyspieszyć proces produkcji dużej liczby wystandaryzowanych wykresów, adekwatnych do rodzaju danych. ChatGPT jest też w stanie przygotować kod programistyczny, który użytkownikowi-laikowi pozwala uzyskać niemal nieograniczone możliwości wizualizacji danych przy pomocy specjalistycznych bibliotek programu Python, R czy Javascript.

Nasz eksperyment wykazał jednak kilka ważnych ograniczeń:

- Problemy z ekstrakcją danych z wykresów i obrazów – ChatGPT ma trudności z odczytaniem danych z wykresów i innych obiektów graficznych. Biorąc pod uwagę dominujący jeszcze obecnie standard przygotowywania raportów ewaluacyjnych, nie może on służyć jako narzędzie do uzyskiwania danych z raportów do celu ich dalszej obróbki.
- Ograniczenia w identyfikacji danych – ChatGPT często nie był w stanie precyzyjnie zlokalizować rekomendowanych danych w treści raportu, co skutkowało błędnymi wskazaniami rozdziałów i wskaźników.
- Brak obsługi zaawansowanych wizualizacji – choć ChatGPT potrafi generować proste wykresy słupkowe i kołowe, nie jest w stanie obsłużyć bardziej skomplikowanych wizualizacji bez integracji z dodatkowymi narzędziami.

Aby zwiększyć użyteczność gen AI w procesie wizualizacji danych pochodzących z raportów ewaluacyjnych, warto rozważyć:

1. Standaryzację raportów – autorzy raportów powinni uwzględniać dobre praktyki, takie jak: dodawanie metadanych do wykresów, wprowadzanie tabel z danymi uzupełniającymi wykresy oraz rezygnacja ze skomplikowanych struktur tabel, które mogą stanowić problem w procesie automatycznej ekstrakcji danych. Pomocne w tym zakresie może okazać się upowszechnianie i dalsze rozwijanie standardów dostępności cyfrowej (WCAG) raportów.

2. Używanie prostych wizualizacji – o ile to możliwe, dane powinny być prezentowane w formie prostych wykresów słupkowych lub liniowych, aby maksymalnie ułatwić ich interpretację przez AI oraz końcowych dobiorców raportów.

10.4. Zakończenie

Eksperyment pozwolił przeanalizować skuteczność generatywnej AI na kilku kluczowych etapach: tworzenia person, dostosowywania treści, tworzenia form graficznych reprezentujących treść oraz wizualizacji danych ilościowych.

Nasze główne wnioski z eksperymentu są następujące:

- Persona jest wartym popularyzacji narzędziem ułatwiającym przygotowanie raportu ewaluacyjnego lub jego wariantów (prezentacji, broszur, notatek prasowych czy dla zarządu) dla potrzeb odbiorców końcowych. Generatywna AI sprawdza się jako narzędzie wspierające wczesne etapy tworzenia person (protoperson). Warto jednak pamiętać o konieczności weryfikacji wygenerowanych treści oraz iteracyjnym podejściu do ich rozwijania, które pozwala uniknąć uproszczeń (stereotypów) i lepiej dopasować persony do rzeczywistych potrzeb odbiorców.
- Generatywna AI może znacznie przyspieszyć proces dostosowywania treści raportów, szczególnie w przypadkach, gdy liczy się szybkość i zwięzłość przekazu. Niemniej jednak jej użycie wymaga jasnego zdefiniowania celów komunikacyjnych i ręcznego nadzoru nad wynikowymi treściami, aby zapewnić ich zgodność z założeniami raportu.
- Generatywna AI może być używana jako narzędzie wspierające w procesie tworzenia form graficznych, szczególnie na etapie koncepcyjnym. Jej zastosowanie przyspiesza pracę, ale wymaga integracji z bardziej zaawansowanymi narzędziami wizualizacyjnymi w celu osiągnięcia profesjonalnego rezultatu.
- Generatywna AI wspiera proces doboru danych oraz ich wizualizacji, szczególnie w kontekście ich prostszych form. Jednak jej skuteczność zależy od odpowiedniego przygotowania danych (nie każde narzędzie gen AI potrafi odczytać dane z obiektów graficznych) i manualnej weryfikacji rekomendacji, zwłaszcza w przypadku bardziej skomplikowanych projektów.

Eksperyment potwierdził, że generatywna AI stanowi wszechstronne wsparcie w personalizacji i wizualizacji raportów ewaluacyjnych. Jest szczególnie przydatna

w zadaniach wymagających szybkości i elastyczności, jednak jej użycie na chwilę obecną wymaga zaangażowania człowieka na każdym etapie procesu, aby zapewnić zgodność i skuteczność wygenerowanych treści.

Bibliografia

- Cooper, A. (1999). *The Inmates Are Running the Asylum: Why High-Tech Products Drive Us Crazy and How to Restore the Sanity*. Sams Publishing, Indianapolis.
- EGO & PGS (2015). *Brokerzy wiedzy – gra planszowa*. Warszawa (dostęp: 20.08.2024).
- Evergreen, S.D.H. (2018). *Presenting Data Effectively: Communicating Your Findings for Maximum Impact*. SAGE Publications, Thousand Oak.
- IMAPP sp. z o.o., PC++ (2016). Wpływ funduszy europejskich perspektywy finansowej 2007–2013 na rozwój społeczno-gospodarczy Polski Wschodniej. Raport końcowy. Ministerstwo Inwestycji i Rozwoju, Warszawa (dostęp: 20.08.2024).
- Nussbaumer Knaflic, C. (2015). *Storytelling with data. A data visualization guide for business professionals*. Wiley & Sons, Inc., Hoboken, New Jersey.
- Olejniczak, K. (2017). The Game of Knowledge Brokering: A New Method for Increasing Evaluation Use. *American Journal of Evaluation*, 38(4), 554–576 (dostęp: 20.08.2024).

CZĘŚĆ III

Gen AI w administracji publicznej – implikacje systemowe

11. Gen AI w organizacjach publicznych – regulacje i *codes of conducts*

Igor Lyubashenko

11.1. Wprowadzenie

Dlaczego temat generatywnej AI jest ważny?

Wyobraź sobie zespół analityczny w urzędzie, który musi przetworzyć setki stron dokumentów, przeanalizować złożone dane i przygotować syntetyczny raport w ciągu kilku dni. To codzienność polskiej administracji publicznej, gdzie niewielkie zespoły ewaluacyjne mierzą się z rosnącą liczbą zadań analitycznych przy ograniczonych zasobach kadrowych. W tym kontekście pojawia się generatywna sztuczna inteligencja (gen AI) – narzędzie, które może fundamentalnie zmienić sposób pracy analityków sektora publicznego.

Gen AI oferuje możliwość automatyzacji czasochłonnych, rutynowych zadań, takich jak wstępne przetwarzanie danych, generowanie pierwszych wersji raportów czy synteza informacji z wielu źródeł. Pozwala to analitykom skupić się na zadaniach wymagających głębszej ekspertyzy merytorycznej, krytycznego myślenia i kontekstowego rozumienia specyfiki sektora publicznego.

Jednak wykorzystanie gen AI w administracji publicznej niesie ze sobą szczególną odpowiedzialność. Instytucje publiczne, działające na podstawie zasad przejrzystości i zaufania publicznego, muszą szczególnie ostrożnie podchodzić do wdrażania nowych technologii. Kluczowe staje się wypracowanie jasnych zasad i procedur korzystania z gen AI, które z jednej strony pozwolą czerpać korzyści z tej technologii, a z drugiej – zapewnią zgodność z wymogami prawnymi i etycznymi służby publicznej.

Ta daleko idąca zmiana jest porównywalna z wcześniejszymi przełomowymi momentami – wprowadzeniem komputerów czy systemów zarządzania dokumentacją elektroniczną. Każda z tych zmian wymagała nie tylko adaptacji technologicznej, ale przede wszystkim przemyślanego podejścia do zarządzania zmianą organizacyjną i kompetencjami pracowników.

Jaka jest bieżąca luka w praktyce zarządzania AI?

Temat zarządzania sztuczną inteligencją nie jest nowy, ale w ostatnich latach zyskał na znaczeniu. Analiza bazy artykułów naukowych Scopus, obejmująca publikacje od 1995 roku, pokazuje pojedyncze artykuły związane z hasłem *AI governance* (zarządzanie sztuczną inteligencją) w latach 1995–2017, po czym następuje gwałtowny wzrost liczby takich publikacji: w 2018 roku opublikowano już 50 artykułów, a w 2023 roku – aż 705.

Dotychczasowe badania koncentrowały się głównie na trzech obszarach:

1. Kwestie etyczne – jak zapewnić, żeby AI działało zgodnie z zasadami takimi jak: transparentność, sprawiedliwość czy poszanowanie godności ludzkiej (Veale et al., 2023)?
2. Ramy prawne – jak stworzyć regulacje, które będą skuteczne, ale jednocześnie elastyczne wobec szybko rozwijającej się technologii (Wirtz et al., 2022)?
3. Bezpieczeństwo i prywatność – jak chronić dane i zapobiegać potencjalnym nadużyciom AI (Chen et al., 2023)?

Te zagadnienia dotyczą głównie tego, co możemy nazwać **wymiarem regulacyjnym** – czyli tworzenia ogólnych ram i zasad dla rozwoju i wdrażania systemów AI.

Pojawienie się pod koniec 2022 roku powszechnie dostępnych narzędzi gen AI dodało nowy wymiar do tej dyskusji. Wcześniej kwestie etycznego wykorzystania AI dotyczyły głównie ekspertów i twórców systemów. Teraz każdy z nas może korzystać z potężnych narzędzi AI, co sprawia, że problemy etyczne stają się bardziej powszechne i „demokratyczne". W tym rozdziale skupimy się na **wymiarze organizacyjnym** – czyli na tym, jak różne instytucje i organizacje próbują wewnętrznie uregulować korzystanie z ogólnodostępnych narzędzi gen AI przez swoich pracowników.

Należy podkreślić, że rozróżnienie na wymiar regulacyjny i organizacyjny, choć pomocne, nie jest idealne. W praktyce te dwa wymiary często się przenikają. Na przykład głębokie zintegrowanie narzędzi gen AI w procesy organizacyjne może *de facto* prowadzić do

stworzenia systemu opartego na AI, co z kolei podlega szerszym regulacjom. Wyobraźmy sobie urząd miasta, który wykorzystuje gen AI do automatycznego przetwarzania wniosków mieszkańców. Z pozoru jest to kwestia organizacyjna – jak efektywnie wykorzystać nowe narzędzie. Jednak w momencie, gdy system zaczyna autonomicznie podejmować decyzje wpływające na życie obywateli, wkraczamy w obszar regulacyjny dotyczący praw jednostek i odpowiedzialności administracji publicznej.

Zatem w praktyce te dwa wymiary często się przenikają, tworząc złożony ekosystem zarządzania AI. Kluczowe jest zrozumienie, że skuteczne wdrażanie wewnętrznych zasad korzystania z gen AI w organizacjach musi być osadzone w szerszym kontekście istniejących regulacji prawnych.

Jaki jest cel tego przeglądu?

Celem tego rozdziału jest zaprezentowanie szerokiego obrazu pokazującego, jak instytucje publiczne podchodzą do tworzenia wewnętrznych regulacji dotyczących korzystania z generatywnej AI. Zostaną przedstawione przykłady wytycznych i zasad opracowanych przez różne instytucje i organizacje publiczne, które mają na celu bezpieczną integrację gen AI w swoich procesach organizacyjnych. Główne pytanie badawcze brzmi: **Jakie są przykłady wewnętrznych wytycznych dotyczących etycznego i odpowiedzialnego korzystania z gen AI przez administrację publiczną?**

W kolejnych podrozdziałach zostaną przedstawione wyniki analizy zidentyfikowanych dokumentów oraz końcowe refleksje na temat stanu regulacji gen AI w administracji publicznej, ze szczególnym uwzględnieniem praktyki ewaluacyjnej. W załączniku do niniejszego podrozdziału opisano metodykę wyszukiwania odpowiednich dokumentów i ich analizy.

11.2. Zbadana populacja dokumentów

Tabela 11.1. pokazuje liczbę dokumentów, które zostały poddane analizie. Warto podkreślić, że zestawienie to nie odzwierciedla rzeczywistej liczby wszystkich istniejących dokumentów tego typu. Procedura wyszukiwania dokumentów opierała się wyłącznie na publicznie dostępnych materiałach. Co za tym idzie – brak danej organizacji lub instytucji w zestawieniu

nie oznacza, że nie opracowała ona podobnych wytycznych. Istnieje możliwość, że takie wytyczne są dostępne wyłącznie wewnętrznie. Warto także podkreślić, że na liście zidentyfikowanych przypadków dominują przykłady z USA, zarówno na poziomie stanowym, jak i miejskim.

Tabela 11.1. Liczba zidentyfikowanych dokumentów

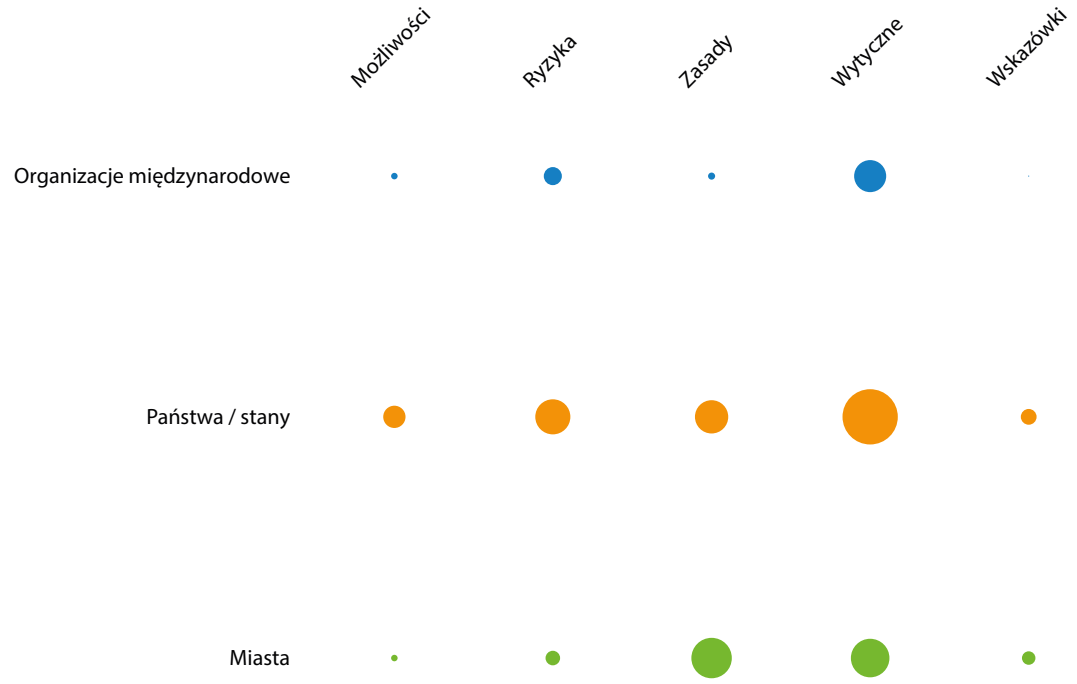
Typ instytucji	Liczba zidentyfikowanych dokumentów
Organizacje międzynarodowe	3
Państwa/stany	13
Miasta	10
Międzynarodowe NGO	0

Źródło: Opracowanie własne.

Rycina 11.1. ilustruje rezultaty pierwszego etapu analizy, skupiającego się na identyfikacji kluczowych elementów w badanych dokumentach (zob. szczegóły metodologii w „Aneks 5. Metodyka wyszukiwania i analizy dokumentów”). Wielkość poszczególnych „bąbli” na wykresie odzwierciedla częstotliwość występowania danych elementów w dokumentach opublikowanych przez analizowane typy instytucji.

Celowo zrezygnowano z prezentacji dokładnych wartości liczbowych, aby uniknąć błędnej interpretacji wyników analizy jakościowej. Należy bowiem pamiętać, że mniejsza liczba zidentyfikowanych fragmentów dla danej kategorii instytucji nie musi oznaczać mniejszego zaangażowania w kwestie wewnętrznej regulacji korzystania z gen AI. Przykładowo, w przypadku organizacji międzynarodowych analizie poddano tylko trzy dokumenty, co naturalnie przekłada się na mniejszą liczbę zidentyfikowanych fragmentów w porównaniu z liczniejszymi dokumentami na poziomie miast.

Rycina 11.1. Struktura analizowanych dokumentów



Źródło: Opracowanie własne.

Warto podkreślić, że nie ma doskonałej rozłączności między zasadami przewodnimi a wytycznymi operacyjnymi. Jak zostanie pokazane w kolejnej części rozdziału, te same kwestie są czasem formułowane jako zasada, a innym razem jako wytyczna. Wynika to z różnic w podejściu autorów dokumentów do przedstawiania kluczowych rekomendacji.

Preambuły zawarte w analizowanych dokumentach z reguły podkreślają ogromny potencjał gen AI w zakresie zwiększania efektywności i innowacyjności w pracy organizacji, a równocześnie – konieczność zarządzania ryzykiem związanym z bezpieczeństwem, prywatnością oraz zgodnością z przepisami prawnymi. Wytyczne często odwołują się do istniejących strategii i ram regulacyjnych, takich jak Narodowa Strategia AI w przypadku Wielkiej Brytanii czy Dyrektywa o Zautomatyzowanym Podejmowaniu Decyzji w przypadku Kanady. Wiele dokumentów zwraca uwagę na konieczność odpowiedzialnego używania narzędzi gen AI, aby zapewnić, że ich wykorzystanie nie narusza prawa ani nie wpływa negatywnie na dobrostan użytkowników. Preambuły często wskazują na dynamiczny

rozwój technologii gen AI i potrzebę elastycznego podejścia do jej regulacji, zaznaczając, że wytyczne będą podlegały ciągłej ewaluacji i aktualizacjom w odpowiedzi na nowe wyzwania i możliwości związane z gen AI.

Tabela 11.2. przedstawia **szanse i zagrożenia** związane z gen AI, które zostały wymienione w analizowanych dokumentach. Warto zauważyć, że więcej uwagi skupia się na zagrożeniach niż na szansach. Należy zaznaczyć, że przedstawiona kategoryzacja ma pewne ograniczenia metodologiczne – niektóre zjawiska mogą być interpretowane zarówno jako szansa, jak i zagrożenie, a część kategorii nie jest całkowicie rozłączna. Te niejednoznaczności odzwierciedlają złożoną naturę technologii gen AI i trudności w jednoznacznej klasyfikacji jej wpływu na organizacje. Analiza zebranego materiału wskazuje na wyraźną dysproporcję – w badanych dokumentach znacznie więcej miejsca poświęcono potencjalnym zagrożeniom niż szansom. Ta asymetria w opisie ryzyka i korzyści może odzwierciedlać ostrożne podejście instytucji publicznych do nowych technologii.

Tabela 11.2. Szanse i zagrożenia związane z gen AI

Szanse	Zagrożenia
Zwiększenie produktywności • Wsparcie przy złożonych zadaniach • Przyspieszenie pracy • Zmniejszenie obciążenia pracą • Ułatwienie kontaktu z obywatelami • Zwiększenie kreatywności • Polepszenie jakości pracy • Poprawienie doświadczeń klientów/ obywateli	• Ryzyko naruszenia obowiązującego prawa (w tym autorskiego) • Wzmocnienie uprzedzeń i dyskryminacja • Możliwość wycieku informacji wrażliwych • Nadmierna wiara w prawdziwość generowanej treści • Niewystarczająca transparentność i wyjaśnialność • Utrata przez człowieka odpowiedzialności za procesy • Ryzyka wizerunkowe • Podatność na cyberataki • Nieekologiczność AI • Zróżnicowana wydajność zależna od języka i tematyki • Podejmowanie niewłaściwych decyzji • Tworzenie nieprawdziwej informacji • Niejasne podstawy trenowania • Niestabilność systemów AI

Źródło: Opracowanie własne.

Wszystkie dokumenty **definiują gen AI** w sposób dość jednolity jako technologię, która wykorzystuje zaawansowane algorytmy i modele uczenia maszynowego do tworzenia nowych treści, takich jak tekst, obrazy, dźwięki, wideo i kod, na podstawie wprowadzonych przez użytkownika danych, zwanych promptami. Modele gen AI uczą się wzorców i struktur z ogromnych zbiorów danych, często pozyskiwanych z Internetu, i generują treści, które

przypominają te stworzone przez człowieka. Jest to definicja zgodna z przytoczoną we wstępie niniejszego opracowania.

Na tym etapie analizy zdefiniowano także dwie dodatkowe cechy charakterystyczne dla stanów i miast amerykańskich w kontekście zarządzania generatywną sztuczną inteligencją. Po pierwsze, wiele z nich powołało **wyspecjalizowane zespoły lub jednostki ds. AI**, których zadaniem jest tworzenie rekomendacji, nadzorowanie i wdrażanie technologii AI (nie tylko generatywnej) w sposób odpowiedzialny i zgodny z zasadami etycznymi. Po drugie, w przypadku jednego ze stanów (Maine) wprowadzono **czasowe moratorium na użycie gen AI przez pracowników** w celu dokładniejszego zbadania ich wpływu oraz opracowanie odpowiednich regulacji przed szerokim wdrożeniem.

11.3. Zidentyfikowane zasady, wytyczne i wskazówki

W tym podrozdziale omówiono wynik drugiego etapu analizy, polegającego na szczegółowym zbadaniu zasad, wytycznych i wskazówek zawartych w analizowanych dokumentach. Zasady przewodnie to fundamentalne wartości i „principia", które powinny przyświecać korzystaniu z gen AI. Są to ogólne wytyczne o charakterze normatywnym, określające „ducha" odpowiedzialnego korzystania z tej technologii. Wytyczne operacyjne to konkretne instrukcje i procedury dotyczące praktycznego wdrażania i codziennego korzystania z gen AI w organizacji. Określają one „literę prawa" – co dokładnie należy, a czego nie wolno robić. Praktyczne wskazówki to przykłady właściwego używania gen AI, często wzbogacone o scenariusze użycia i studia przypadków.

Zasady przewodnie

Analiza zgromadzonych dokumentów pozwoliła na wyodrębnienie katalogu zasad przewodnich, które mają stanowić swego rodzaju fundament dla pracowników korzystających z generatywnej sztucznej inteligencji. Zasady te przedstawiono – wraz z krótkim komentarzem – w tabeli 11.3.

Tabela 11.3. Zasady korzystania z gen AI zawarte w analizowanych dokumentach

Zasada przewodnia	Komentarz
Przejrzystość, odpowiedzialność i zaufanie	Zapewnienie transparentności w wykorzystaniu gen AI, przyjęcie odpowiedzialności za generowane treści, ustanowienie mechanizmów nadzoru i ujawnianie użycia gen AI – to zasady kluczowe dla budowania zaufania i umożliwienia zbiorowego uczenia się.
Cyberbezpieczeństwo i ochrona danych	Zabezpieczenie danych wrażliwych, wdrożenie odpowiednich kontroli dostępu i ochrona przed nieautoryzowanym użyciem. Zgodność z przepisami o ochronie prywatności i wdrożenie silnych środków bezpieczeństwa.
Równość, sprawiedliwość i inkluzywność	Zapobieganie dyskryminacji i nierównym skutkom stosowania AI. Świadomość potencjalnych historycznych uprzedzeń w systemach AI i działania na rzecz sprawiedliwości społecznej.
Zgodność z misją instytucji	Wykorzystanie AI w sposób wspierający podstawową misję służby publicznej. Jasne definiowanie problemów do rozwiązania i angażowanie różnorodnych interesariuszy.
Zgodność z obowiązującym prawem i standardami	Zapewnienie legalności, etyczności i odpowiedzialności w użyciu AI. Współpraca ze specjalistami oraz uwzględnianie kwestii prawnych, etycznych i środowiskowych.
Rozwijanie kompetencji i umiejętności pracowników	Zapewnienie niezbędnych umiejętności technicznych, udział w szkoleniach i promocja odpowiedzialnych praktyk związanych z wykorzystaniem AI wśród pracowników.
Dążenie do odpowiedzialnej innowacyjności	Proaktywne badanie i ocena wykorzystania AI przy jednoczesnym zarządzaniu ryzykiem i priorytetowym traktowaniu prywatności, bezpieczeństwa i zrównoważonego rozwoju.
Podejście skoncentrowane na człowieku	Priorytetowe traktowanie dobrostanu, bezpieczeństwa i godności ludzi. Rozwój systemów wykorzystujących AI z myślą o wzmacnianiu ludzkich możliwości i przyczynianiu się do poprawy społeczeństwa.
Wzmocnienie pozycji pracownika	Wykorzystanie AI do wspierania pracowników w świadczeniu lepszych usług i rozwijaniu umiejętności. AI ma służyć rozwijaniu kompetencji zawodowych i poprawie jakości świadczonych usług, ale ostateczne decyzje pozostają w gestii pracownika bazującego na swoim doświadczeniu i osądzie zawodowym.
Zachowanie kontroli i odpowiedzialności przez człowieka	Zapewnienie ludzkiej kontroli na wszystkich etapach cyklu życia systemu AI, w tym przeglądu wyników, testowania i monitorowania. Jasne określenie ról i odpowiedzialności operatorów – ludzi.
Wyjaśnialność	Zapewnienie, że pracownicy rozumieją ogólne możliwości i ograniczenia używanych narzędzi gen AI oraz potrafią wytłumaczyć, w jaki sposób wykorzystali je w swojej pracy.
Świadome korzystanie z AI	Zrozumienie możliwości i ograniczeń AI oraz stosowanie technik zapewniających trafność i dokładność wyników. Kompleksowe podejście do wdrażania, utrzymania i monitorowania systemów AI.
Dostosowanie narzędzi AI do właściwych zadań	Staranny dobór narzędzi AI do konkretnych potrzeb, z uwzględnieniem wymagań użytkowników, wpływu na środowisko i zgodności z podstawowymi wartościami organizacji.

Zasada przewodnia	Komentarz
Zobowiązanie do rzetelności	Priorytetowe traktowanie dokładności informacji generowanych przez AI. Wdrożenie procesów weryfikacji i audytu wyników AI oraz zachęcanie użytkowników do krytycznej oceny i sprawdzania faktów.
Współpraca między człowiekiem a AI	Promowanie synergii między ludźmi a technologią AI, w celu wykorzystania potencjału obu stron.
Monitorowanie postępu i dostosowanie się	Ciągłe monitorowanie i adaptacja użycia AI w miarę postępu technologii, pozostawanie na bieżąco z rozwojem.
Ekonomia i efektywność	Wykorzystanie AI w sposób ekonomiczny i efektywny, aby maksymalizować korzyści przy minimalnych kosztach.
Współpraca z sektorem prywatnym	Współpraca z firmami prywatnymi w zakresie rozwoju i wdrażania AI, aby korzystać z najlepszych praktyk i innowacyjnych rozwiązań.
Adaptacja do zmieniającej się rzeczywistości	Elastyczne dostosowywanie się do zmian w technologii i regulacjach, aby być przygotowanym na przyszłe wyzwania.

Źródło: Opracowanie własne.

Wytyczne operacyjne

W przypadku wytycznych operacyjnych konieczne było podzielenie ich na trzy kategorie:

1. Wskazówki dotyczące właściwego postępowania z gen AI.
2. Wskazówki dotyczące niewłaściwego postępowania z gen AI.
3. Przykłady dozwolonego użycia gen AI.

Taki podział umożliwia przejrzyste przedstawienie zaleceń mających na celu zapewnienie odpowiedzialnego i zgodnego z zasadami korzystania z gen AI. W poniższych tabelach znajdują się szczegółowe wytyczne dla każdej z tych kategorii.

Tabela 11.4. Wskazówki dotyczące właściwego postępowania z gen AI

Wskazówka	Komentarz
Weryfikuj wytwory gen AI	Należy dokładnie sprawdzać wszystkie treści generowane przez AI przed ich użyciem. Mogą one zawierać nieścisłości, uprzedzenia lub nieodpowiednie informacje. Użytkownik ponosi odpowiedzialność za jakość, dokładność i stosowność wygenerowanych treści. Konieczna jest krytyczna ocena odpowiedzi pod kątem potencjalnych problemów. Pliki generowane przez AI powinno się traktować ostrożnie, gdyż mogą zawierać złośliwy kod.

Wskazówka	Komentarz
Ujawniaj wytwory gen AI	Korzystanie z generatywnej AI musi być ujawniane w sposób przejrzysty. Należy wskazać konkretny użyty model gen AI, podane instrukcje oraz zakres ludzkiej weryfikacji i edycji. Wymagane są obowiązkowe zastrzeżenia przy wykorzystaniu gen AI w procesach decyzyjnych lub do generowania treści bezpośrednio dostępnych dla opinii publicznej. Instytucje muszą prowadzić rejestr danych wejściowych, instrukcji i wyników gen AI w celu zapewnienia odpowiedzialności i przejrzystości.
Postępuj zgodnie z obowiązującym prawem i procedurami	Należy przestrzegać odpowiednich przepisów, polityk i procesów zatwierdzania przy korzystaniu z generatywnej AI – w tym dotyczących ochrony danych, bezpieczeństwa IT i własności intelektualnej. W razie potrzeby należy konsultować się z ekspertami prawnymi i IT.
Ucz się, eksperymentuj, wymieniaj się doświadczeniami	Zaleca się edukację w zakresie szans i zagrożeń związanych z generatywną AI poprzez dostępne szkolenia. Należy korzystać z wiedzy technicznej lub ją zdobywać, aby efektywnie używać generatywnej AI. Należy zapoznawać się z nowymi narzędziami, oceniać ich przydatność i dzielić się wiedzą.
Chroń dane i informacje wrażliwe	Przy korzystaniu z narzędzi generatywnej AI należy chronić wrażliwe dane i informacje. Należy unikać wprowadzania jakichkolwiek danych osobowych, tajnych lub poufnych, ponieważ mogą one zostać wykorzystane w niewłaściwy sposób. Należy przestrzegać polityk prywatności i bezpieczeństwa oraz korzystać z generatywnej AI wyłącznie w bezpiecznych, kontrolowanych środowiskach.
Postępuj zgodnie z obowiązującymi normami etycznymi	Należy zapewnić etyczne i odpowiedzialne wykorzystanie generatywnej AI w instytucjach. Należy unikać danych wrażliwych, rozumieć wykorzystanie danych, uzyskiwać zgody prawne i dotyczące bezpieczeństwa, testować pod kątem uprzedzeń i podatności na zagrożenia, dokumentować reakcje na incydenty i wyraźnie oznaczać treści generowane przez AI. Priorytetem powinna być różnorodność, integracja i przeciwdziałanie dyskryminacji.
Dokumentuj użycie AI oraz incydenty	Pracownicy muszą dokumentować korzystanie z generatywnej AI, w tym decyzje o rozwoju lub wdrożeniu takich narzędzi, kroki podjęte w celu zapewnienia odpowiednich wyników oraz przypadki, w których gen AI jest używana do automatycznego podejmowania decyzji. Szczególnej dokumentacji wymagają incydenty, czyli niepożądane zdarzenia związane z użyciem gen AI. Wymagania dotyczące dokumentacji różnią się w zależności od kontekstu.
Zarządzaj ryzykiem	Należy zapewnić zachowanie praw dostępu przy korzystaniu z modeli generatywnej AI. Należy oceniać ryzyka, koszty i ograniczenia konkretnych modeli, biorąc pod uwagę wpływ na interesariuszy i zgodność z wartościami organizacji. Dane wejściowe powinny być starannie weryfikowane, ponieważ modele są podatne na niewłaściwe użycie. Instytucje powinny najpierw eksplorować zastosowania o niskim ryzyku, dostosowując środki łagodzące ryzyko.
Oddzielaj sprawy zawodowe od prywatnych	Przy korzystaniu z systemów generatywnej AI należy oddzielać sprawy zawodowe od prywatnych. Należy używać służbowego adresu e-mail i unikalnego hasła dla wszelkiego związanego z pracą użycia tych narzędzi, aby zapewnić właściwe prowadzenie rejestrów i zapobiec nieautoryzowanemu dostępowi do kont służbowych. Należy unikać używania prywatnych kont lub haseł do spraw służbowych.

Wskazówka	Komentarz
Wspomagaj pracę ludzi, nie zastępuj jej	Narzędzia generatywnej AI powinny być oceniane przez instytucje rządowe pod kątem tego, jak mogą pomagać pracownikom, a nie ich zastępować. Każde wdrożenie takich narzędzi musi przynieść pozytywny rezultat dla obywateli, taki jak poprawa usług, zmniejszenie biurokracji i oszczędności. Przed wdrożeniem należy przeprowadzić dokładną ocenę, w tym analizę wpływu regulacyjnego, aby upewnić się, że gen AI jest optymalnym rozwiązaniem.
Podążaj do wyjaśnialności	Należy zachować przejrzystość, odpowiedzialność i wyjaśnialność przy korzystaniu z narzędzi generatywnej AI. Pracownik powinien być w stanie wyjaśnić i uzasadnić swoje porady i decyzje. Należy unikać używania niewyjaśnialnej lub „czarnoskrzynkowej" AI do procesów decyzyjnych lub zatwierdzających.
Nie automatyzuj decyzji	Nie należy w pełni automatyzować podejmowania decyzji przy użyciu generatywnej AI. Zamiast tego należy używać tych narzędzi jako pomocy, a nie zamienników. Należy budować własne zrozumienie możliwości i ograniczeń gen AI. Należy krytycznie oceniać wyniki systemu, używać neutralnych instrukcji i zawsze przeglądać treści generowane przez AI, nawet jeśli wydają się wiarygodne.
Korzystaj z zatwierdzonych narzędzi	Należy przestrzegać warunków dotyczących nowych narzędzi IT w organizacji. Narzędzia gen AI, które nie mogą być dostępne z sieci organizacji, wymagają wcześniejszego wniosku o system IT i oceny ryzyka. Należy używać zasobów gen AI swojej agencji do pomocy w obowiązkach służbowych, a nie do użytku osobistego. Należy się spodziewać, że korzystanie z technologii gen AI może być rejestrowane i monitorowane.
Używaj swojego osądu etycznego i intelektualnego	Należy aktywnie badać odpowiedzi narzędzi generatywnej AI pod kątem występowania różnego rodzaju uprzedzeń i zapewnić, że podejmowane są działania korygujące w odniesieniu do instrukcji, wykorzystania różnych źródeł lub alternatyw dla narzędzia gen AI. Przez cały czas należy używać swojego osądu etycznego i intelektualnego, aby chronić podstawowe wartości.
Podążaj do różnorodności i inkluzywności w zespołach zajmujących się AI	Należy zachęcać do różnorodności w zespołach zajmujących się rozwojem AI, aby zapewnić uwzględnienie szerokiego zakresu pomysłów i perspektyw.
Zapewniaj odpowiednią jakość danych do uczenia modeli	Przy budowaniu i trenowaniu nowych modeli należy zapewnić, że dane używane do trenowania modelu są odpowiedniej jakości, reprezentatywne i wolne od uprzedzeń. Należy regularnie monitorować wydajność modeli i dostosowywać je lub ponownie trenować na podstawie tych wyników.
Staraj się korzystać z narzędzi przyjaznych środowisku	Należy sprawdzić, czy dostawca gen AI wyznaczył cele redukcji emisji gazów cieplarnianych. Należy przeprowadzić ocenę wpływu na środowisko jako część propozycji rozwoju lub zakupu narzędzi generatywnej AI. Należy zachęcać deweloperów do przejrzystości w kwestii zrównoważonego rozwoju, preferując tych, którzy jasno komunikują wpływ środowiskowy swoich systemów AI.

Wskazówka	Komentarz
Twórz instytucje wspomagające	Należy upewnić się, że istnieją struktury wspierające wdrożenie generatywnej AI. Struktury te nie muszą być w pełni dojrzałe przed pierwszym projektem, ale powinny zapewniać wystarczającą kontrolę, aby korzystanie z generatywnej AI było bezpieczne i odpowiedzialne. Należy rozważyć ustanowienie strategii AI i odpowiednich zasad korzystania z gen AI.
Konsultuj się z przełożonymi	Należy zgłaszać wszelkie przypadki, w których nie jest możliwe pełne zastosowanie tych wytycznych, do odpowiedniego wyższego urzędnika odpowiedzialnego w instytucji.

Źródło: Opracowanie własne.

W ramach wytycznych dotyczących właściwego użycia warto zwrócić uwagę na obecność szczegółowych rekomendacji technicznych. Są one skierowane głównie do działów IT i dotyczą bezpiecznych parametrów modeli. Ze względu na ich specjalistyczny charakter, nie zostały one omówione szczegółowo w niniejszym opracowaniu.

Tabela 11.5. Wskazówki dotyczące niewłaściwego postępowania z gen AI

Wskazówka	Komentarz
Nie wprowadzaj wrażliwych danych	Nie należy wprowadzać do narzędzi gen AI żadnych informacji poufnych, wrażliwych, klasyfikowanych lub danych osobowych. Dane udostępnione gen AI mogą być wykorzystane w wynikach dla innych użytkowników lub niewłaściwie użyte. Należy wprowadzać tylko informacje, które są już publicznie dostępne lub nie są zastrzeżone.
Nie pozwalaj gen AI działać autonomicznie	Należy zachować ostrożność przy używaniu surowych treści generowanych przez AI do podejmowania decyzji lub nienadzorowanego udzielania odpowiedzi w imieniu organizacji na zapytania publiczne. Zawsze należy włączyć ludzką weryfikację wygenerowanych treści, sprawdzając poprawność faktów i możliwe naruszenia własności intelektualnej.
Nie publikuj wytworów AI bez sprawdzenia i ujawnienia źródła	Nie należy używać kodu oprogramowania wygenerowanego przez AI bez dokładnego sprawdzenia. Ma to na celu zapewnienie, że nie wprowadza się niebezpiecznego lub nawet złośliwego kodu do systemów IT organizacji. Nie należy publikować wyników gen AI (tekstu, obrazu lub kodu) bez pełnej, świadomej recenzji i ujawnienia.
Nie używaj AI w sytuacjach o wysokim ryzyku	Gen AI nie powinna być używana w obszarach wysokiego ryzyka, które mogłyby spowodować szkodę dla zdrowia, bezpieczeństwa, praw podstawowych lub środowiska. Nie należy polegać na ogólnodostępnych modelach gen AI w procesach o charakterze krytycznym i wrażliwych pod względem terminowości.
Nie korzystaj z wytworów AI dosłownie	Pracownicy nigdy nie powinni bezpośrednio powielać wyników modelu gen AI w dokumentach publicznych i pismach urzędowych.

Wskazówka	Komentarz
Nie otwieraj linków wygenerowanych przez AI	Nie należy rozpowszechniać ani otwierać linków dostarczonych lub wygenerowanych przez publiczne narzędzia lub boty gen AI. Te linki mogą prowadzić do stron phishingowych[1] lub do pobrania złośliwego oprogramowania.
Nie korzystaj z AI do podejmowania szybkich decyzji	Gen AI działa stosunkowo wolno w porównaniu z innymi systemami komputerowymi i nie powinna być używana w przypadkach, gdy wymagana jest szybka odpowiedź na pytanie.
Nie traktuj AI jako źródła prawdy	Gen AI nie jest zoptymalizowana pod kątem dokładności i nie należy na niej polegać jako na jedynym źródle prawdy bez dodatkowych środków zapewniających dokładność.
Nie korzystaj z AI, gdy priorytetem jest wyjaśnialność	Podobnie jak w przypadku innych rozwiązań opartych na sieciach neuronowych, wewnętrzne działanie rozwiązania gen AI może być trudne lub niemożliwe do wyjaśnienia, co oznacza, że nie powinno być używane tam, gdzie konieczne jest wyjaśnienie każdego kroku w decyzji.
Nie korzystaj z nieodpowiednio wytrenowanej AI	Wydajność gen AI zależy od dużych ilości danych treningowych. Systemy, które zostały przeszkolone na ograniczonych ilościach danych, na przykład w specjalistycznych obszarach używających terminologii prawniczej lub medycznej, mogą dawać nieprawidłowe lub niedokładne wyniki.
Nie korzystaj z AI do tworzenia reguł korzystania z AI	Nie jest odpowiednie używanie publicznie dostępnych narzędzi do tworzenia dokumentów dotyczących zmian w istniejących regułach/zasadach korzystania z gen AI.
Nie używaj obrazów wygenerowanych przez AI	Nie należy używać obrazów wygenerowanych przez gen AI w materiałach organizacji. Ten obszar przedstawia zwiększone ryzyko związane z własnością intelektualną.
Nie twórz treści, które mogą być uznane za *deepfake*	Nie należy generować obrazów, dźwięków lub wideo, które mogłyby być mylone z obrazami lub nagraniami realnych osób, na przykład tworzyć fałszywe zdjęcia lub nagrania konkretnego urzędnika albo członka społeczeństwa (*deepfake*) – nawet jeśli treść będzie odpowiednio oznaczona; nie należy generować fałszywego obrazu lub nagrania, które rzekomo przedstawia obywatela lub urzędnika publicznego, nawet jeśli nie jest to konkretna osoba; nie należy generowanie fałszywych „respondentów” lub wymyślonych profili do ankiet lub innych badań.
Nie używaj narzędzi, które mogą „podsłuchiwać” na żywo	Nie należy korzystać z narzędzi, które mogą słuchać i transkrybować rozmowę „na bieżąco”.
W razie wątpliwości nie korzystaj z AI	Jeśli masz jakiekolwiek wątpliwości, nie używaj generatywnej AI.

Źródło: Opracowanie własne.

[1] *Phishing* to oszukańcza technika cyberprzestępcza polegająca na podszywaniu się pod wiarygodne podmioty (np. instytucje, firmy, osoby) w celu wyłudzenia poufnych informacji (np. danych logowania, informacji bankowych) lub nakłonienia ofiary do wykonania określonych działań (np. kliknięcia w złośliwy link czy pobrania szkodliwego oprogramowania).

Warto zauważyć, że w niektórych przypadkach wskazówki dotyczące właściwego i niewłaściwego postępowania pokrywają się, dotyczą tego samego problemu, ale są sformułowane albo pozytywnie (jako zalecenie), albo negatywnie (jako zakaz). Na przykład wskazówki dotyczące ujawniania wytworów AI i weryfikacji ich dokładności odpowiadają zaleceniom dotyczącym właściwego postępowania, podczas gdy zakazy publikowania niezweryfikowanych treści AI bez ujawnienia źródła odnoszą się do niewłaściwego postępowania. Pokazuje to pewną elastyczność w sposobie formułowania podobnych wytycznych.

Tabela 11.6. Dwuwymiarowe podejście do formułowania wskazówek dotyczących właściwego i niewłaściwego korzystania z generatywnej AI

Temat	Pozytywne sformułowanie	Negatywne sformułowanie
Weryfikacja wytworów gen AI	Weryfikuj wytwory gen AI	Nie publikuj wytworów gen AI bez sprawdzenia
Ujawnianie użycia gen AI	Ujawniaj wytwory gen AI	Nie publikuj wytworów gen AI bez ujawnienia źródła
Ochrona danych wrażliwych	Chroń dane i informacje wrażliwe	Nie wprowadzaj wrażliwych danych
Autonomia gen AI	Wspomagaj pracę ludzi, nie zastępuj jej	Nie pozwalaj gen AI działać autonomicznie
Wyjaśnialność AI	Podążaj do wyjaśnialności	Nie korzystaj z gen AI, gdy priorytetem jest wyjaśnialność
Jakość danych i modeli	Zapewniaj odpowiednią jakość danych do uczenia modeli	Nie korzystaj z nieodpowiednio wytrenowanej gen AI
Tworzenie treści	–	Nie twórz treści, które mogą być uznane za *deepfake*
Prywatność	–	Nie używaj narzędzi, które mogą „podsłuchiwać” na żywo
Zaufanie do gen AI	–	Nie traktuj gen AI jako źródła prawdy

Źródło: Opracowanie własne.

Macierz porównawcza (tabela 11.6.) ilustruje to zjawisko, zestawiając pozytywne i negatywne sformułowania dla każdego tematu. Analiza tej macierzy pozwala dostrzec, że dla większości tematów istnieją zarówno pozytywne, jak i negatywne sformułowania. Szczególnie widoczne jest to w przypadku takich tematów jak „Weryfikacja wytworów gen AI” czy „Ujawnianie użycia gen AI”, gdzie pozytywne zalecenia mają swoje odpowiedniki w formie zakazów. Interesujące jest również to, że niektóre tematy, takie jak „Tworzenie treści” czy

„Prywatność", posiadają tylko negatywnie sformułowane wytyczne, co może sugerować istnienie obszarów szczególnej ostrożności w korzystaniu z gen AI.

Analiza wymienionych przykładów dozwolonego użycia gen AI w instytucjach (tabela 11.7.) ukazuje dość szeroki zakres potencjalnych zastosowań tej technologii. Obejmują one zarówno zadania związane z przetwarzaniem i analizą informacji, jak i tworzeniem różnorodnych treści. Gen AI może znacząco usprawnić procesy komunikacyjne, analityczne i decyzyjne, oferując narzędzia do szybkiego podsumowywania długich dokumentów, generowania odpowiedzi na pytania obywateli czy tworzenia wstępnych wersji dokumentów. Technologia ta znajduje również zastosowanie w bardziej specjalistycznych dziedzinach, takich jak programowanie, analiza danych czy tłumaczenia. Warto jednak podkreślić, że we wszystkich wskazanych przypadkach podkreśla się wspomagającą rolę gen AI. Ostateczna weryfikacja, edycja i decyzje powinny pozostawać w gestii pracowników.

Tabela 11.7. Przykłady dozwolonego użycia

Przykład dozwolonego użycia	Komentarz
Generowanie podsumowań dłuższych tekstów	Umożliwia szybkie streszczenie długich dokumentów, oszczędzając czas i wysiłek pracowników.
Ułatwienie komunikacji z obywatelami	Gen AI może wspomagać interakcje z obywatelami poprzez automatyczne generowanie odpowiedzi na pytania.
Generowanie wersji roboczych kodu komputerowego	Gen AI może tworzyć fragmenty kodu, wspomagając programistów w ich pracy i przyspieszając rozwój oprogramowania.
Generowanie wersji roboczych dokumentów/korespondencji	Tworzenie wstępnych wersji dokumentów, pism urzędowych lub wiadomości e-mail, które mogą być później edytowane przez pracowników.
Generowanie obrazów, dźwięku i wideo	Gen AI może tworzyć materiały wizualne i dźwiękowe na potrzeby komunikacji, marketingu lub edukacji.
Analiza danych (ilościowych i jakościowych)	Gen AI może wspomagać analizę dużych zbiorów danych, dostarczając wartościowych informacji i wniosków.
Badania gabinetowe	Umożliwia szybkie zbieranie i analizowanie informacji na różne tematy, wspierając proces badawczy.
Tłumaczenia	Gen AI może automatycznie tłumaczyć teksty na różne języki, ułatwiając komunikację.
Wspomaganie decyzji	Gen AI dostarcza analiz i rekomendacji, które mogą ułatwiać proces podejmowania decyzji przez pracowników.

Źródło: Opracowanie własne.

Do kategorii wytycznych operacyjnych można także odnieść **procedury związane z gen AI**, które znalazły się w wytycznych z Wirginii, mianowicie:

- Procedura przeglądu dostawców technologii gen AI;
- Procedura akceptacji rozwiązań opartych na gen AI przed ich wdrożeniem;
- Zasady nabywania kompetencji przez pracowników w zakresie gen AI;
- Zasady zakupu usług i rozwiązań opartych na gen AI.

Praktyczne wskazówki

Katalog praktycznych wskazówek, które znalazły się w analizowanych dokumentach, został przedstawiony w tabeli 11.8.

Tabela 11.8. Praktyczne wskazówki dotyczące korzystania z gen AI

Wskazówka	Komentarz
Przykłady „hakowania" gen AI	Opis metod „hakowania" gen AI[2] – uświadomienie istnienia tego typu zagrożeń oraz pomoc w ich identyfikacji i zapobieganiu.
Przykładowe zadania	Opis scenariuszy i zadań, które mogą być realizowane przy użyciu gen AI, pokazujący praktyczne zastosowania technologii.
Instrukcja rezygnacji ze zbierania danych	Opis szczegółowych kroków, które należy wykonać, by zrezygnować z opcji zbierania danych przez narzędzia gen AI (śledzenia), aby chronić prywatność.
Instrukcja identyfikacji *deepfake*	Krótki przewodnik po metodach identyfikacji *deepfake*, co jest kluczowe dla ochrony przed dezinformacją.
Dodatek zawierający „podręcznik"	Jeden z dokumentów zawiera rozbudowany dodatkowy materiał edukacyjny, który pomaga pracownikom w lepszym zrozumieniu i wykorzystaniu gen AI.
Platforma wymiany informacji	Jeden z dokumentów zachęca pracowników do aktywnego udziału w wymianie doświadczeń na dedykowanej platformie.

Źródło: Opracowanie własne.

[2] Hakowanie generatywnej AI (prompt injection/prompt hacking) to techniki manipulacji instrukcjami (promptami) wysyłanymi do systemu AI, które mają na celu obejście jego wbudowanych zabezpieczeń i ograniczeń, uzyskanie dostępu do wrażliwych danych lub skłonienie systemu do generowania niedozwolonych lub szkodliwych treści – np. poprzez sprzeczne polecenia, ukryte instrukcje czy wykorzystanie specyficznych wzorców językowych.

11.4. Uwagi końcowe

Zawarty w niniejszym opracowaniu przegląd wytycznych dotyczących korzystania z generatywnej sztucznej inteligencji przez pracowników organizacji odzwierciedla aktualny stan wiedzy i praktyk w **I połowie 2024 roku**. Podsumowując niniejszy rozdział, można wyciągnąć kilka kluczowych wniosków.

Przede wszystkim zidentyfikowany katalog zasad, wytycznych i rekomendacji wyraźnie ukazuje złożoność wyzwań związanych z integracją tej technologii w procesy organizacyjne. Dokumenty pochodzące z różnych źródeł, w tym z instytucji międzynarodowych, rządów i miast, mogą stanowić cenny punkt odniesienia dla instytucji sektora publicznego w Polsce i innych organizacji planujących opracowanie własnych regulacji wewnętrznych.

Analiza wykazała, że w dokumentach dominuje podejście skupione na identyfikowaniu i zarządzaniu ryzykiem związanym z użyciem gen AI, kosztem promowania jej potencjalnych korzyści. Przeważają zalecenia o charakterze negatywnym, jasno określające, czego nie należy robić, aby uniknąć potencjalnych nadużyć i zagrożeń. Takie podejście, choć zrozumiałe i pomocne w ustanawianiu jednoznacznych granic, nie jest jedynym możliwym rozwiązaniem.

Warto podkreślić, że niemal wszystkie analizowane dokumenty zwracają uwagę na dynamiczny rozwój technologii gen AI. W związku z tym podkreśla się konieczność ciągłej adaptacji i aktualizacji zasad, aby zachować ich skuteczność i adekwatność w zmieniającej się rzeczywistości technologicznej.

Przeprowadzony przegląd dostarcza cennych obserwacji dla jednostek ewaluacyjnych. Przede wszystkim pokazuje, że skuteczne wdrożenie gen AI wymaga całościowego podejścia organizacyjnego. Jednostki nie powinny regulować wykorzystania tej technologii w oderwaniu od szerszego kontekstu instytucjonalnego. Konieczna jest współpraca z działami IT, jednostkami prawnymi i kierownictwem w tworzeniu spójnych ram wykorzystania gen AI.

Szczególnie istotna obserwacja dotyczy podejścia do zarządzania ryzykiem. W analizowanych dokumentach dominuje strategia świadomego zarządzania ryzykiem, nie zaś całkowitego zakazu używania gen AI. Dla jednostek ewaluacyjnych oznacza to potrzebę precyzyjnego określenia obszarów bezpiecznego wykorzystania tej technologii (jak zadania administracyjne czy wstępne analizy) oraz tych wymagających szczególnej ostrożności.

Kolejnym kluczowym wnioskiem jest fundamentalne znaczenie przejrzystości w wykorzystaniu gen AI. Instytucje publiczne przykładają do tej kwestii szczególną wagę, co dla jednostek ewaluacyjnych przekłada się na potrzebę wypracowania przejrzystych standardów dokumentowania i raportowania użycia gen AI w procesach ewaluacyjnych. Jest to szczególnie istotne w kontekście zachowania wiarygodności badań ewaluacyjnych i zaufania do ich wyników.

Analizowane dokumenty podkreślają również kluczową rolę rozwoju kompetencji pracowników w zakresie odpowiedzialnego wykorzystania gen AI. Jednostki ewaluacyjne mogą i powinny aktywnie włączać się w proces budowania tych kompetencji w swoich organizacjach, czerpiąc z doświadczeń innych instytucji publicznych w zakresie programów szkoleniowych i dzielenia się wiedzą.

Podsumowując, jednostki ewaluacyjne nie muszą samodzielnie wypracowywać wszystkich rozwiązań związanych z wykorzystaniem gen AI. Mogą i powinny czerpać z już istniejących w sektorze publicznym doświadczeń, adaptując je do specyfiki działalności ewaluacyjnej. Kluczowe jest jednak zachowanie równowagi między innowacyjnością a rygorem metodologicznym, między efektywnością a odpowiedzialnością, oraz między automatyzacją a ludzkim osądem w procesach ewaluacyjnych.

Bibliografia

- Australian Government, Digital Transformation Agency. (2023, November 22). *Interim guidance on government use of public generative AI tools* (dostęp: 31.03.2025).
- Baltimore City. (2024). *Mayoral Executive Order Establishing Principles and Policy Governing Use of Generative Artificial Intelligence*.
- Chen, Y.-C., Ahn, M.J., Wang, Y.-F. (2023). Artificial Intelligence and Public Values: Value Impacts and Governance in the Public Sector. *Sustainability, 15*(6), 4796 (dostęp: 31.03.2025).
- City of Boston. (2023). *City of Boston Interim Guidelines for Using Generative AI* (dostęp: 31.03.2025).
- City of Helsinki. (2022, November 22). *The City of Helsinki established principles for the ethical use of data and artificial intelligence.* City of Helsinki (dostęp: 31.03.2025).
- City of San Francisco. (2023, December 11). *San Francisco Generative AI Guidelines*. San Francisco (dostęp: 31.03.2025).
- City of San Jose. (2023). *Generative AI Guidelines* (dostęp: 31.03.2025).
- City of Tempe. (2023). *Ethical Artificial Intelligence (AI) Policy* (dostęp: 31.03.2025).
- Commonwealth of Massachusetts. (2024). *Executive Order* (dostęp: 31.03.2025).
- Commonwealth of Pennsylvania. (2023). *Executive Order* (dostęp: 31.03.2025).

- Commonwealth of Virginia. (2024). *Policy Standards for the Utilization of Artificial Intelligence by the Commonwealth of Virginia* (dostęp: 31.03.2025).
- Dencik, J., Goehring, B., Marshall, A. (2023). Managing the emerging role of generative AI in next-generation business. *Strategy & Leadership*, *51*(6), 30–36 (dostęp: 31.03.2025).
- European Commission. (2023). *Guidelines for Staff on the Use of Online Available Generative artificial Intelligence Tools* (dostęp: 31.03.2025).
- *Global Observatory of Urban Artificial Intelligence (GOUAI)*. (n.d.). Global Observatory of Urban Artificial Intelligence (GOUAI) (dostęp: 20.06.2024).
- Government of Canada. (2024). *Guide on the use of generative artificial intelligence* (dostęp: 31.03.2025).
- Kansas Office of Information Technology Services. (2024). *Generative Artificial Intelligence Policy* (dostęp: 31.03.2025).
- New York City Office of Technology & Innovation. (2024). *Preliminary Use Guidance: Generative Artificial Intelligence* (dostęp: 31.03.2025).
- Organisation for Economic Co-operation and Development. (2023, June). *Secretary-General's Guidelines for Staff on the Use of Generative Artificial Intelligence*.
- Renugadevi, R., Shobana, J., Arthi, K., et al. (2024). Real-Time Applications of Artificial Intelligence Technology in Daily Operations. W: D. Satishkumar, M. Sivaraja (red.), *Advances in Computational Intelligence and Robotics* (s. 243–257). IGI Global (dostęp: 31.03.2025).
- Soken-Huberty, E. (2023, January 16). *The 15 Biggest NGOs in the World*. Human Rights Careers (dostęp: 31.03.2025). Stadt Wien. (2024). *Compass for the use of generative artificial intelligence (AI) in a work-related context* (dostęp: 31.03.2025).
- State of California. (2024). *State of California GenAI Guidelines for Public Sector Procurement, Uses and Training* (dostęp: 31.03.2025).
- State of Maine. (2023). *Cybersecurity Directive* (dostęp: 31.03.2025).
- State of New Jersey. (n.d.). *State of New Jersey Interim Guidance on Responsible Use of Generative AI* (dostęp: 31.03.2025).
- State of Oklahoma. (2023). *Executive Order* (dostęp: 31.03.2025).
- State of Oregon. (2023). *Esectutive Order* (dostęp: 31.03.2025).
- State of Wisconsin. (n.d.). *Executive Order* (dostęp: 31.03.2025).
- Suleyman, M. (2023). *The coming wave* (First edition). Crown.
- UK Government. (2024). *Generative AI Framework for HMG* (dostęp: 31.03.2025).
- UN Chief Executives Board for oordination. (2023, July 20). *Digital & Technology Network (DTN) Meeting Report*.
- U.S. Office of Personnel Management. (n.d.). *Responsible Use of Generative Artificial Intelligence for the Federal Workforce* (dostęp: 31.03.2025).
- Veale, M., Matus, K., Gorwa, R. (2023). AI and Global Governance: Modalities, Rationales, Tensions. *Annual Review of Law and Social Science*, *19*(1), 255–275 (dostęp: 31.03.2025).
- Wirtz, B.W., Weyerer, J.C., Kehl, I. (2022). Governance of artificial intelligence: A risk and guideline-based integrative framework. *Government Information Quarterly*, *39*(4), 101685 (dostęp: 31.03.2025).

12. Perspektywy adaptacji organizacyjnych i systemowych wykorzystania AI w ewaluacji

Jacek Pokorski
Grzegorz Rzeźnik

12.1. Wstęp

W rozdziale podejmujemy refleksję nad możliwymi uwarunkowaniami zmian zachowania osób (pracowników i kierownictwa jednostek ewaluacyjnych) oraz ich adaptacji do wyzwania, jakim jest wykorzystanie technologii generatywnej sztucznej inteligencji (gen AI) w administracji publicznej. W pierwszej części naszego rozdziału omówimy czynniki, które blokują adaptację na poziomie indywidualnym. Są nimi różnego typu dysfunkcje i błędy poznawcze (ang. *biases*), które mogą występować zarówno na poziomie decydentów systemu, jak i pracowników jednostek.

Stawiamy tezę, że te ograniczenia indywidualne osób mogą przekładać się na utratę szans, jakie obecnie zapewnia gen AI, na mniejszą sprawność jednostek ewaluacyjnych i w rezultacie – na niższą efektywność systemu ewaluacji jako całości. Jesteśmy przekonani, że bariery na poziomie jednostkowym można do pewnego stopnia redukować odpowiednimi rozwiązaniami organizacyjnymi i systemowymi. Zakładamy, że modyfikacja zachowania uczestników systemu powinna jednocześnie stymulować pozytywne zmiany w tym systemie. Z kolei włączenie do procesów i praktyki jednostek ewaluacyjnych odpowiednich czynników behawioralnych – zachęt lub ograniczeń systemowych – może wywoływać oczekiwane zachowania osób (np. sięganie po narzędzia gen AI, w tym w ramach ewaluacji wewnętrznych, ograniczanie deklaratywności na rzecz wtórnej analityki danych i źródeł zastanych itp.).

W rozdziale przedstawiamy autorską diagnozę indywidualnych postaw i zachowań w kontekście wprowadzania innowacji organizacyjnych, koncentrując się na barierach

i możliwościach związanych z adaptacją technologii gen AI w sektorze publicznym. Wykorzystujemy metody analizy znane z psychologii społecznej, szczególnie te dotyczące indywidualnych decyzji, w kontekście systemowych działań promujących nowe zachowania. Inspirację czerpiemy również z literatury dotyczącej organizacji i zarządzania, w tym z teorii zarządzania zmianą, szczególnie w kontekście transformacji cyfrowej. Taka kombinacja perspektyw pozwala nam stworzyć ramę analityczną, która w pierwszej kolejności identyfikuje potencjalne czynniki blokujące pracowników jednostek ewaluacyjnych, a następnie proponuje systemowe i organizacyjne rozwiązania stymulujące wdrażanie technologii gen AI.

W naszych rozważaniach wychodzimy od założeń klasycznych modeli zarządzania zmianą, takich jak model Lewina (Clarke, 1997; Burnes, 2020), Kottera (Kotter, 1996), oraz ADKAR (Hiatt, 2006), jednak na nich nie poprzestajemy. Modele te stanowiły solidne fundamenty teoretyczne dla inicjowania transformacji organizacyjnych, w tym cyfryzacji. Jednak współczesne badania (Burnes, 2020; Todnem By, 2021) wskazują, że w dynamicznym i złożonym środowisku organizacyjnym sektor publiczny wymaga bardziej elastycznych i adaptacyjnych podejść, które wykraczają poza jednorazowe zmiany, a oferują ciągłe możliwości dostosowania do postępu technologicznego.

Model Lewina opisuje proces zmiany jako trójfazowy cykl: odmrożenie (przygotowanie do zmiany), zmiana (wdrożenie nowego rozwiązania) i zamrożenie (utrwalenie nowego stanu jako standardu). Model ten, choć może wydawać się zbyt uproszczony w kontekście współczesnych wyzwań, pozostaje użyteczną ramą dla wprowadzania technologii AI w sektorze publicznym, szczególnie w kontekście etapowego przygotowania, wdrażania i utrwalania nowych standardów w systemie ewaluacji. Z kolei Kotter w zaktualizowanej wersji swojego podejścia (Kotter, 2014) podkreśla kluczową rolę zwinności organizacyjnej (ang. *agility*) i zdolności szybkiego reagowania na zmiany. Ten ośmiostopniowy proces, począwszy od budowania poczucia pilności zmiany (np. poprzez wskazanie na przestarzałość dotychczasowych metod ewaluacji), aż po zakorzenienie nowych praktyk w kulturze organizacyjnej, stanowi wartościowe narzędzie do wprowadzania AI.

Z perspektywy indywidualnej model ADKAR oferuje unikalne podejście i wskazuje, że zmiana organizacyjna jest sumą transformacji jednostkowych. Skuteczność takiej zmiany zależy od poziomu kompetencji, świadomości, zaangażowania i motywacji pracowników. Badania

(Hornstein, 2015) potwierdzają skuteczność tego modelu w projektach transformacji cyfrowej na dużą skalę.

Na zakończenie rozdziału formułujemy wnioski dotyczące możliwości stymulowania transformacji publicznych jednostek ewaluacyjnych w stronę wykorzystania gen AI. Kluczowym elementem naszych rozważań jest model COM-B, który uwzględnia trzy podstawowe determinanty zmiany zachowań: zdolność (ang. *capacity*), możliwość (*opportunity*) i motywacja (*motivation*). Dzięki zastosowaniu tego modelu możemy zidentyfikować konkretne bariery i czynniki wspierające wdrożenie gen AI w organizacjach publicznych, w tym w jednostkach ewaluacyjnych. Model ten wspomaga analizę nie tylko na poziomie indywidualnym, ale także systemowym, wskazując na potrzebę zrównoważenia rozwoju kompetencji pracowników, zapewnienia odpowiednich zasobów oraz tworzenia zachęt motywacyjnych.

Przy opracowaniu rozdziału wykorzystano również narzędzia gen AI w celach porządkowania i syntezy danych oraz integracji podejścia teoretycznego z praktycznymi wskazówkami autorów. Umożliwiło to wypracowanie odpowiedniej strategii działań, wspierających wykorzystanie gen AI w jednostkach odpowiedzialnych za ewaluację polityk publicznych.

12.2. Czynniki blokujące transformację ewaluacji w kierunku wykorzystania gen AI

Charakterystyka zjawiska

Wykorzystanie technologii gen AI stanowi nowe źródło szans i ryzyk dla funkcjonowania administracji publicznej. Procesy cyfryzacji i automatyzacji bezpośrednio wpływają na jakość zarządzania publicznego. Obserwujemy to w takich obszarach funkcjonowania organizacji publicznych jak na przykład: realizacja zamówień publicznych, współpraca administracji rządowej i samorządowej z obywatelami czy projektowanie i ewaluacja programów.

Aktualnie możliwości wykorzystania różnego typu narzędzi analitycznych wspieranych gen AI do generowania wiedzy na potrzeby administracji, w tym do ewaluacji wykorzystania publicznych funduszy i programów, są w praktyce nieograniczone. Z kolei dotychczasowy

(aktualny) model wykonywania zadań ewaluacyjnych w większości opiera się na narzędziach badań społecznych, w tym na danych pochodzących głównie z deklaracji (od respondentów – uczestników programów, interesariuszy) i ze źródeł pierwotnych (pozyskiwanych w ramach różnego typu badań terenowych i im pokrewnych, wykorzystujących wywiady, fokusy, panele czy ankietowanie na próbach). Analityka zastanych źródeł informacji stanowi zazwyczaj tylko uzupełnienie lub wprowadzenie w kontekst zasadniczej części badania ewaluacyjnego. Tego typu badania społeczne służące ewaluacji programów najczęściej zamawiane są w postępowaniach o udzielanie zamówienia publicznego (w przetargach) i wykonywane przez agencje badawcze, konsultantów-ewaluatorów, rzadziej przez ośrodki naukowo-badawcze czy *think tanki* i ich ekspertów. Biorąc pod uwagę ogromne możliwości narzędzi gen AI oraz danych zastanych, które narzędzia te mogą integrować i przetwarzać, zasadność utrzymywania tego tradycyjnego modelu realizacji ewaluacji – stosunkowo drogiego (opartego o usługi obce, o względnie dużej skumulowanej wartości) i nierzadko obciążonego (dominacją deklaratywnych źródeł danych czy niedokładnością wynikającą z badania na próbach o wątpliwej precyzji pomiaru i wiarygodności, przy uwzględnieniu wymogów metody reprezentatywnej) – warto poddać pod rozwagę (pod kątem kryteriów racjonalności i gospodarności).

Należy przy tym zauważyć, że występuje szereg czynników hamujących transformację ewaluacji w kierunku wykorzystania gen AI (w szczególności odciągających pracowników jednostek ewaluacyjnych od sięgania w swojej pracy po tego typu innowacyjne narzędzia). Są one podobne do tych, które bywają identyfikowane jako bariery transformacji cyfrowej w organizacjach (Fundacja Digital Poland, 2020). Ponadto, nie bez znaczenia jest fakt, że w tym przypadku owa transformacja dotyka sektora publicznego, co wpływa na jej dodatkową specyfikę, odróżniającą od digitalizacji biznesu (w tym w zakresie wykorzystania AI) (Pokorski, 2024).

Odnosząc się do dorobku badań z zakresu psychologii społecznej, należy zauważyć, że co do zasady, próba modyfikacji dotychczasowych zachowań osób (np. kierownictwa i pracowników jednostek ewaluacyjnych) nierzadko napotyka na szereg ograniczeń, które paraliżują (lub co najmniej spowalniają) zmianę na poziomie jednostkowym, jak również zmiany systemowe. Kluczowe w tym zakresie pułapki (*biases*) dotyczą przede wszystkim procesu podejmowania decyzji jednostkowych i niegotowości (mentalnej) do adaptacji, a nie dostępnych zasobów instytucjonalnych. Te ostatnie, realnie rzecz biorąc, tracą na znaczeniu. Transformacja może

bowiem prowadzić do ich redukcji lub do wyższej produktywności ich wykorzystania (czasu, budżetu, infrastruktury, pracowników).

Z kolei czynniki, które mogłyby pomóc stawić czoła tym wyzwaniom i przełamać bariery, aby „AI zagościło pod strzechami ewaluacji" (dostarczając nowych rozwiązań i umożliwiając sięganie po przełomowe możliwości analizy wielu źródeł danych), znaleźć można w szeroko pojętej teorii zarządzania zmianą. W szczególności przyglądając się nieudanym projektom i dobrym praktykom transformacji cyfrowej (zwłaszcza w sektorze publicznym), można doszukać się analogii dla naszej problematyki i wypracować rekomendacje dla transformacji systemu ewaluacji w stronę adaptacji technologii AI[1].

Proces decyzyjny

Dostępność narzędzi gen AI to potencjalnie nowe, obiecujące możliwości pracy i wielka szansa dla badaczy i ewaluatorów pracujących w instytucjach publicznych (zarówno z perspektywy jakości wytwarzanych produktów merytorycznych, jak i produktywności pracy). Z drugiej strony, wkraczające „nowe" może radykalnie przemodelować system generowania wiedzy w administracji programów (czy szerzej – w sektorze publicznym) i dotychczasowy sposób funkcjonowania jednostek ewaluacyjnych, działów strategii lub równoważnych struktur analityczno-badawczych. Tradycyjne podejście jednostek opiera się przede wszystkim na działaniu planowym, względnie przewidywalnym i bezpiecznym (wieloletnie plany badawcze, projekty o znaczącym budżecie dotyczące programów lub badań ich szeroko pojętego otoczenia). Rokrocznie powiela ono wypracowane procesy (procedury) jak również względnie sprawdzone sposoby zarządzania dostawami wiedzy na rzecz systemu (wdrażania, programowania, planowania strategicznego), w szczególności zarządzanie zamówieniem publicznym i kontraktem z wykonawcą badań (ewaluatorem), przy przewidywalnych zasobach (czasu, budżetu, pracowników jednostki). Mimo że podejście tradycyjne (zachowawcze – polegające na „robieniu tego, co dotychczas, jakby nigdy nic", jakby niezauważające gen AI) z biegiem czasu będzie nie do utrzymania, wydaje się, że perspektywa spodziewanego (potencjalnego) dyskomfortu związanego ze zmianą obecnie nie zachęca jednostek do

[1] W rozdziale wykorzystano fragmenty następujących prac autora (Jacek Pokorski): *Podejmowanie decyzji a pułapki kognitywne. Studium przypadku z sektora publicznego* (2023) oraz *Transformacja cyfrowa w systemie ewaluacji wykorzystania funduszy europejskich w Polsce* (*Digital Transformation in the Evaluation System of European Funds Utilization in Poland*) (2024), przygotowanych w ramach studiów MBA Uniwersytetu SWPS, Instytutu Podstaw Informatyki PAN i Woodbury School of Business at Utah Valley University.

innowacji, jak również może hamować adaptacje organizacyjne i szeroko pojęty postęp systemowy.

Na decyzję o wyborze modelu realizacji badań ewaluacyjnych w Polsce, służących m.in. ocenie wykorzystania funduszy europejskich w ramach krajowych i regionalnych programów rozwoju, a mianowicie: czy będzie to model „konserwatywny" (tradycyjny, oparty przede wszystkim na generowaniu danych pierwotnych w ramach zamawianych przez sektor publiczny badań społecznych) czy „postępowy" (innowacyjny i rozwijający wykorzystanie gen AI i zaawansowaną analitykę danych) albo w jakim stopniu system ewaluacji będzie hybrydowo stosował elementy obu tych modeli – oddziałuje szereg obciążeń. Są to typowe pułapki kognitywne, znane z literatury dotyczącej organizacji i zarządzania, w które wpadają zarówno osoby decyzyjne w systemie, jak i poszczególni pracownicy jednostek ewaluacyjnych. W celu zidentyfikowania czynników blokujących transformację ewaluacji w kierunku wykorzystania gen AI posłużono się systematyką pułapek w procesie podejmowania decyzji (błędów poznawczych, ang. *cognitive biases*; obciążeń dla decyzji wynikających z błędnego rozpoznania, ang. *decision-making biases*), jaką przedstawili amerykańscy badacze Kinicki i Williams (Kinicki, Williams, 2020)[2]. Nasza analiza wskazuje, że w rezultacie wpadania w owe pułapki, proces decyzyjny prowadzić może do kurczowego trzymania się modelu „konserwatywnego" (lub do jego dominacji) przez większość aktorów systemu ewaluacji. Następuje to jednocześnie wbrew obserwowanym trendom i możliwościom technicznym, jakie oferuje dokonująca się na naszych oczach rewolucja cyfrowa, w tym dostępność i systematyczny rozwój narzędzi AI oraz dostęp do „pokładów" nowych danych.

W erze rewolucji gen AI struktury systemu ewaluacji stanęły (lub zdaniem autorów, powinny były stanąć) przed następującymi pytaniami:

1. **Czy korzystać z nowych narzędzi gen AI, w tym oferowanych przez nie dostępów do nowych źródeł danych i sposobów analizy** (tj. czy podjąć wysiłek pozyskania narzędzi – w tym zakupu licencji, wykorzystania, rozwijania oraz ich zabezpieczenia – czyli wysiłek prawny, dotyczący zapewniania infrastruktury IT, kompetencyjny w zakresie obsługi

[2] Analiza przeprowadzona została w odwołaniu do sześciu powszechnie występujących pułapek w podejmowaniu decyzji, które zdefiniowali autorzy 9. edycji publikacji *Management: A practical Introduction* (s. 217–2019). Należy podkreślić, że w 10. edycji tej pozycji wydawniczej (autorstwa Kinickiego i Soigneta, 2022) zdefiniowano dodatkowe cztery pułapki (*The Overconfidence Bias, The Hindsight Bias, The Framing Bias, The Categorical Thinking Bias*), jednakże nie są one przedmiotem niniejszej analizy.

nowych narzędzi analitycznych, zarządczy dotyczący odpowiedzialności za dane i działania podległych osób)?

2. **Czy podjąć skoordynowane działania prowadzące do upowszechniania narzędzi gen AI w procesach ewaluacji prowadzonych na poziomie sieci jednostek?**
3. **Czy zmienić plany ewaluacji i model działania systemu na „postępowy”** (obiecujący merytorycznie i ekonomicznie, ale eksperymentalny), **czy dalej prowadzić badania w dotychczasowym modelu „tradycyjnym”** (oswojonym, względnie przewidywalnym, ale kosztowym i coraz mniej pewnym/wiarygodnym)?

Zmiana byłaby odejściem od dotychczasowych heurystyk, o których mówią Kinicki i Williams (Kinicki, Williams, 2020)[3], tj. w pewnym sensie wyjściem poza „strefę komfortu”, co wymagałoby przebudowania struktur, zbudowania kompetencji, ponownego niełatwego zdobycia zaufania (poparcia) decydentów i uzyskania adekwatnych zasobów przez jednostki systemu. Jednocześnie przełamanie schematu (modelu tradycyjnego) zamawiania ewaluacji w dotychczasowy sposób wpychałoby te jednostki w sytuację niepewności, mimo potencjalnie obiecujących korzyści w długim okresie. Jednostki ewaluacyjne nie mają bowiem gwarancji, że zarówno nowe podejście, jak i sam proces transformacji będą efektywne oraz bezproblemowe i zastąpią czy zmodyfikują poprzednie narzędzia a także zredukują nakłady pracy[4].

Wnioski z obserwacji i dyskusji z uczestnikami systemu wskazują, że dotychczas raczej dominują obawy i opór po stronie jednostek ewaluacyjnych wobec sięgania po nowe rozwiązania, niż „strategiczna orientacja proinnowacyjna”. Można powiedzieć, że przeważają głosy typu: „to, co jest robione teraz, jest *good enough*”; „nie warto przemodelowywać struktur, procesów i zasobów – mamy wystarczająco pracy z tym, co robimy teraz, a to byłoby dodatkowe «umęczenie», przy czym nie wiadomo, czy będziemy w stanie z tego

[3] Por. tamże (Zasada kciuka: Ludzie robią coś, ponieważ tak się przyjęło, bo tak robili dotychczas i nie przysporzyło im to kłopotów).

[4] Przywołana potencjalna obawa jednostek jest podobna do sytuacji obserwowanej w służbach statystyki publicznej. Mają one wiedzę i umiejętności analityczne do produkcji statystycznej, opartej na danych pochodzących z rejestrów administracyjnych i *big data*, ale jednocześnie są zobowiązane dostarczać statystyki wg wieloletnich i międzynarodowych standardów Programów Badań Statystycznych Statystyki Publicznej, które w głównej mierze bazują na deklaracjach (wywołane dane pochodzą z badań reprezentacyjnych i sprawozdawczości różnych jednostek). Redukowanie badań opartych na angażowaniu respondenta na rzecz badań „zza biurka”, prowadzonych na danych z systemów, które rejestrują aktywność populacji respondentów, nie jest łatwe do wdrożenia i przemodelowania wieloletniej praktyki GUS, przy określonym poziomie zasobów.

skorzystać”[5]. Proces decyzyjny zdominowało zatem myślenie tendencyjne. Nie są brane pod uwagę wszystkie istotne czynniki – także te, które mogą docelowo doprowadzić do destrukcji systemu. W praktyce bowiem odejście od tego kierunku transformacji ewaluacji włączającej potencjał gen AI prawdopodobnie doprowadzić może do zastąpienia wiedzy generowanej przez jednostki ewaluacyjne wiedzą z innych źródeł (rozwiązania typu *public intelligence* (Chłoń-Domińczak, Pokorski et al., 2022) analogiczne do narzędzi znanych ze sfery biznesu, np. monitoringu w czasie rzeczywistym opartym na połączonych systemach danych, powiązanym z algorytmami AI, dokonującymi wartościowania trendów, wyławiania interesujących jednostek, formułowania komunikatów ostrzegawczych itp.). Wówczas można się spodziewać, że transformacja „przyjdzie z góry” i będzie zapewne bardziej bolesna dla sieci jednostek, w szczególności pracowników sceptycznych wobec innowacji w kierunku ww. rozwiązań włączających gen AI. Zatem decyzja o utrzymywaniu modelu „tradycyjnego”, w krótkim okresie ma charakter pozornie bezpieczny, ale długofalowo – destrukcyjny i kosztowny.

Bariery poznawcze

W wyżej zarysowanym procesie podejmowania decyzji zachodzić mogą realne bariery poznawcze (Pokorski J., 2023). Na podstawie przywołanej systematyki Kinickiego i Williamsa, są nimi:

1. Ograniczanie perspektywy poznawczej przedstawicieli jednostek ewaluacyjnych do tego, co jest dostępne, znane i rozumiane „tu i teraz” (1. ***The Availability Bias***). To z kolei szerzej łączy się z teoretycznym brakiem dostępności, jeśli chodzi o wchodzenie w nowe obszary i inwestowanie w pewnego rodzaju „eksperymenty” (tj. inwestowanie czasu; rozwój nowych kompetencji – informatycznych, analitycznych, prawnych, zarządczych; zasoby infrastruktury IT), a dodatkowo z „mentalną niegotowością” na modyfikację sprawdzonych ścieżek postępowania i poddawania weryfikacji dotychczasowego dorobku „produkcji ewaluacyjnej” (badań, raportów, rekomendacji). Ostatnia obawa kadry kierowniczej jednostek może wynikać z niepewności o swoją pozycję, przyszłość, wiarygodność, gdyby próba z nowym modelem się nie powiodła. Zgodnie z zasadą (nierzadko obowiązującą w korporacjach): „jesteś tak dobry, jak Twój ostatni projekt” (*performance*), nikt nie będzie patrzył na cały dorobek jednostek ewaluacyjnych lub proces (drogę, jaką przebyła dana jednostka). Nowy model pracy oznaczałby wystawienie się do weryfikacji „na własne życzenie” (w tym przed zmieniającymi się członkami zarządu, dyrektorami, podsekretarzami

[5] Opinie sparafrazowano, mają one charakter poglądowy.

stanu itp.). Powyższe aktualnie okazuje się nazbyt ryzykowne – pozornie lepiej jest pozostać w bezpiecznym i przewidywalnym, choć nie w pełni efektywnym modelu.

2. Poszukiwanie potwierdzenia przez liderów systemu, że wchodzenie w nową działkę i pełnienie w niej roli koordynacyjnej (np. wypracowywanie systemowych rozwiązań na rzecz korzystania z narzędzi gen AI w ewaluacji, promowanie badań opartych na nowych zasobach danych) jest „zabójczo niebezpieczne". Można domniemywać, że liderzy systemu, żywiąc podobne obawy jak te powyższe, mogą innym także sugerować (czy też upowszechniać) trudności i wyzwania, czym wpadają w kognitywną pułapkę potwierdzenia (2. ***The Confirmation Bias***). Przejawia się ona w myśleniu typu: „nowe nieznane, niepewne…, nie chcemy tego koordynować, szukajmy argumentów, aby podjąć decyzję wspierającą nasze obawy". W rezultacie przebieg dyskutowanych kwestii strategicznych może zmierzać w kierunku dyskredytowania nowych narzędzi i źródeł danych. Obawy prowadzić mogą do piętrzenia problemów i namnażania zasobów, które potencjalna zmiana musiałaby angażować („mamy mało osób w zespole, temat trudny i rozwojowy, generujący problemy prawne, IT, kompetencyjne, koordynacyjne, kontrolne").
3. Problem reprezentatywności (3. ***The Representativeness Bias***). W procesie decyzyjnym i dyskusjach strategicznych nie zawsze uczestniczą przedstawiciele jednostek o orientacji proinnowacyjnej, w tym liderzy sektorowi i regionalni. W rezultacie niedopuszczenia do głosu wszystkich lub odpowiedniej reprezentacji interesariuszy różnych szczebli decyzja może być „ucierana" na podstawie niereprezentatywnej próbki (skrzywiony dobór informatorów / potencjalnych użytkowników nowego modelu).
4. Poniesione nakłady (4. ***The Sunk-Cost Bias***). W ciągu ostatnich 20 lat jednostki systemu rozwinęły swój potencjał merytoryczny w zakresie prowadzenia badań ewaluacyjnych (lata szkoleń, udział w konferencjach, wymiana dobrych praktyk na temat standardów metodologicznych, w tym w zakresie wnioskowania statystycznego, przyczynowego, wypracowywania rekomendacji/decyzji opartej na wiarygodnych danych oraz dowodach itp.). Bazowano w tym zakresie na dorobku nauk społecznych, m.in. ugruntowanych sposobach pozyskiwania danych pierwotnych metodami ilościowymi i jakościowymi, w tym opinii różnego typu informatorów (wywiady kwestionariuszowe/ankietowanie; wywiady pogłębione, fokusy, panele itp.). Zatem kolejna pułapka w procesie decyzyjnym krążyć może wokół myśli typu: „nie będziemy teraz uczyć się nowych kompetencji i porzucać dotychczasowe – nie zbudujemy ich szybko i tanio". Ponadto, mając na uwadze, że zastane dane administracyjne są szczególnie chronione (np. zbiór systemów danych Narodowego Funduszu Zdrowia), co do zasady, powinny być one analizowane

wewnętrznie, a nie przekazywane zewnętrznym ekspertom, agencjom badawczym, *think tankom* czy konsultantom (wyciek danych mógłby pogrążyć kierownika jednostki oraz narazić instytucję na straty wizerunkowe i finansowe). To z kolei oznacza konieczność budowania wewnętrznych, ewaluacyjnych zespołów AI & *data hubs* z inżynierami danych (ang. *data engeneers*, *data scientists*) i analitykami AI (*prompt engeneers*). To okazuje się bardzo kosztowne. Trudne może być także do wyobrażenia dla przedstawicieli jednostek, że możliwe jest względnie szybkie przestawienie *mind-set* „producentów i konsumentów wiedzy" w sektorze publicznym, poprzez zmianę dotychczasowych raportów tekstowych z rekomendacjami wytwarzanymi miesiącami, na nowe formy raportów wytwarzane *ad-hoc* przy wspomaganiu AI. W pewnym sensie wybór nowego modelu pracy, mógłby sugerować potencjalne niedoskonałości poprzedniego, przyznanie się do błędu w tym modelu, w który potencjał zainwestowano duże środki. Poza tym efektywność kosztowa nowego modelu jest niepewna, zwłaszcza gdy struktury są niedostosowane, osoby niechętne a zasoby niedostępne. Jedyna „słuszna myśl" – „róbmy zatem tak, jak dotychczas: zamawiajmy kompleksowe badania na zewnątrz[6], płaćmy za ankietowanie (efekt Concorda).

5. Wyżej opisane pułapki myślowe to po części także „efekt zakotwiczenia" (5. ***The Anchoring & Adjustment Bias***), zwany inaczej „owczym pędem" („dotąd tak robiliśmy, wszystkie jednostki tak robią" – mimo że jakość ankietyzacji z roku na rok spada, a koszty rosną…). Zamawianie badań jako określonej pozycji kosztów i zadań jednostki sektora finansów publicznych względnie dobrze się planuje i nadzoruje (por. roczne plany zamówień, monitorowanie wykonania rzeczowo-finansowego), w przeciwieństwie do raportów badawczo-analitycznych, które przy wspomaganiu AI częściej mogłyby być wytwarzane wewnętrznie. Stąd, mimo że tradycyjny model „produkcji wiedzy" jest umiarkowanie efektywny, nadal się w niego inwestuje i utwierdza potrzebę jego utrzymania (także pod pozorem obiektywizmu, związanego z zewnętrznym osądem wartości programu czy wyższej jakości zamawianej ekspertyzy). Przejście na „model postępowy" w ewaluacji, generowałoby istotne problemy zarządcze. Trudno jest radykalnie i szybko zreformować personel jednostek (częściowe przekwalifikowanie, nowo pozyskane osoby o innym profilu kompetencyjnym), gdy mamy do czynienia:
   a) z „nieradykalnymi" możliwościami finansowania potrzebnych zasobów;
   b) z wątpliwą gotowością do „pójścia w nowe nieznane" i w „trudniejsze" rozwiązania (np. w większym zakresie rozliczanie za efekt/dzieło/utwór/wynik, niż za procedowanie/projektowanie/zamawianie/nadzorowanie);

[6] W nowym modelu *de facto* nie chodzi o zaprzestanie zamawiania, tylko o jego optymalizację zakresową i zasobową, przy uwzględnieniu możliwości, jakie zapewnia technologia gen AI.

c) z obawami co do szans na uzyskania odgórnego (ze strony kierownictwa instytucji) umocowania w zakresie wykonania nowej „inteligentnej” (*public intelligence*) misji jednostki.

6. ***The Excalation of commitment Bias***. Jednostki nie są skłonne, aby się przyznać, że nie do końca optymalnie produkują wiedzę (tj. zamawiają drogie badania na zewnątrz, podczas gdy system publiczny jest „zalany” danymi i dysponuje narzędziami AI, zapewniającymi inteligentne przetwarzanie *in-house*). Osoby, które tworzyły zręby systemu blisko 20 lat temu przy diametralnie innych możliwościach badawczo-analitycznych, realnie rzecz biorąc, nie chcą go porzucić, stwierdzając przy tym, że dotychczasowe podejście się wyczerpało. Realizacja ewaluacji w takim modelu wydawała się skuteczną i racjonalną decyzją (zasięganie wiedzy eksperckiej z zewnątrz i dostarczanie oceny obiektywnej), jednakże uwarunkowania się zmieniły i utrzymanie „tradycyjnego” podejścia w długim okresie nie wydaje się możliwe. Jednocześnie porzucenie czegoś, co jest względnie komfortowe (co zapewnia względną pewność pozycji, dorobku, praktyk), okazać się może bardziej bolesne niż perspektywa korzyści w przyszłości (*prospect theory*[7]).

Powyższa analiza pokazuje, jak często występować mogą pułapki poznawcze w procesie podejmowania decyzji w przypadku transformacji ewaluacji w kierunku wykorzystania gen AI. Sygnalizuje ona także, jak trudne może być reformowanie i adaptowanie struktur *public* do nowej rzeczywistości *smart*.

12.3. Czynniki katalityczne dla transformacji ewaluacji w kierunku wykorzystania gen AI

Sposoby przełamywania barier oraz wywołania zmiany

Efektywna transformacja systemu ewaluacji w kierunku wykorzystania gen AI wymaga działań na poziomie organizacyjnym, które zarówno wspierają zmiany, jak i wymuszają przełamywanie barier indywidualnych. Poniżej przedstawiono kluczowe bariery oraz

[7] Por. „teoria perspektywy” Daniela Kahnemana, dotyczących procesów decyzyjnych podejmowanych w warunkach niepewności (Nagroda Nobla w dziedzinie ekonomii w 2002 r. za „zastosowanie rozwiązań psychologii w badaniach ekonomicznych, ze szczególnym uwzględnieniem teorii perspektywy”).

czynniki, które mogą pozwolić na skuteczne wdrożenie rozwiązań gen AI w zespołach odpowiedzialnych za realizację zadań ewaluacyjnych. Proponowane podejście opiera się na strategiach ułatwiających adaptację oraz mechanizmach behawioralnych utrudniających utrzymanie dotychczasowych, nieefektywnych praktyk. Każde z nich ma wpływ zarówno na poszczególne osoby aktywne w środowisku ewaluatorów, jak i na cały ekosystem organizacji zaangażowanych w ocenę polityk i programów publicznych.

Analiza barier w adaptacji narzędzi gen AI w ewaluacji wskazuje na dominację szeregu ograniczeń poznawczych i organizacyjnych. Jednym z najczęściej występujących problemów jest ograniczenie perspektywy decydentów do informacji łatwo dostępnych i znanych z dotychczasowych praktyk. W rezultacie pracownicy jednostek ewaluacyjnych nie postrzegają potencjału AI jako realnej alternatywy dla obecnych metod badawczych. Przełamanie tej bariery wymaga demonstracji praktycznych korzyści płynących z wdrożenia AI. Organizacja krótkich form edukacyjnych i szkoleń opartych na rzeczywistych przypadkach zastosowania gen AI może przybliżyć decydentom narzędzia, które automatyzują analizę danych zastanych, przyspieszają generowanie raportów i umożliwiają predykcję wyników polityk publicznych. Kluczowe będzie także rozwijanie intuicyjnych systemów wizualizacji danych (w czym również narzędzia gen AI mogą pomóc), takich jak interaktywne raporty, które uczynią wyniki analiz bardziej zrozumiałymi i dostępnymi dla szerokiego grona odbiorców.

Drugim istotnym ograniczeniem jest tendencja do potwierdzania własnych przekonań i obaw. W kontekście systemowej transformacji ewaluacji liderzy systemu mogą z góry odrzucać AI jako zbyt ryzykowne i skupiać się na trudnościach wdrożeniowych. Aby przełamywać ten schemat, przydatne będą niezależne ewaluacje eksperymentalnych wdrożeń AI. Zewnętrzni eksperci lub zespoły ewaluacyjne mogą obiektywnie ocenić korzyści oraz ryzyka nowego podejścia, wskazując jego potencjał w kontekście oszczędności czasu, zasobów i zwiększonej efektywności. Ponadto, organizowanie spotkań z liderami innych jednostek administracyjnych, które skutecznie wdrożyły narzędzia AI w swojej pracy, pozwoli na wymianę dobrych praktyk i zmniejszenie obaw przed nieznanym.

Kolejna bariera opisana w rozdziale wynika z przywiązania systemu do dotychczasowych modeli pracy i inwestycji w rozwój tradycyjnych kompetencji badawczych. Wieloletnie doświadczenia związane z zamawianiem badań ewaluacyjnych oraz oparcie się na metodach badań pierwotnych stają się „zakotwiczeniem", które utrudnia wdrożenie innowacji.

Pracownicy jednostek obawiają się, że nowy model pracy może podważyć dotychczasowy dorobek, jednocześnie wymagany jest czas i środki na rozwój nowych umiejętności. Aby ograniczyć ten efekt, konieczne jest podkreślenie, że funkcjonalności narzędzi gen AI nie negują wcześniejszych osiągnięć, ale mogą stanowić ich naturalne rozwinięcie i wzmocnienie. Stopniowe wdrażanie gen AI – na przykład poprzez hybrydowy model ewaluacji, łączący tradycyjne badania z nowoczesną analityką danych – umożliwi ewolucję systemu bez radykalnych zmian strukturalnych. Przeprowadzanie małoskalowych projektów pilotażowych pozwoli także na ograniczenie ryzyka niepowodzenia.

W przypadku efektu zakotwiczenia widoczne jest kurczowe trzymanie się sprawdzonych procedur i przekonanie, że „dotąd tak robiliśmy i było dobrze". Zamawianie badań zewnętrznych jest dobrze zorganizowane, przewidywalne i łatwe do monitorowania, co dodatkowo wzmacnia opór wobec zmiany. Przełamanie tego schematu wymaga zmiany kultury organizacyjnej oraz rozwijania proinnowacyjnego myślenia. Konieczne jest zatem promowanie przykładów udanych transformacji w jednostkach ewaluacyjnych, zarówno pochodzących z administracji krajowej, jak i zagranicznej. Tworzenie przestrzeni do eksperymentowania, takich jak „sandboxy regulacyjne" (Fal, 2022), w których jednostki mogą testować nowe rozwiązania AI, umożliwi stopniowe wyjście poza „strefę komfortu".

Dodatkowo pułapka „eskalacji zaangażowania" powoduje, że pracownicy jednostek i decydenci niechętnie przyznają, iż utrzymywanie tradycyjnych metod ewaluacyjnych jest kosztowne i nieskuteczne w obliczu współczesnych możliwości technologicznych. Zamiast tego kurczowo trzymają się względnie przestarzałych praktyk, traktując je jako mniej ryzykowne. Aby ograniczyć ten efekt, kluczowe jest dostarczenie argumentów ekonomicznych oraz strategicznych, przemawiających za wdrożeniem narzędzie gen AI. Analiza kosztów i korzyści powinna uwzględniać długoterminowe oszczędności wynikające z automatyzacji procesów oraz zwiększonej jakości wyników ewaluacji.

Podsumowując, skuteczne przełamanie opisanych barier wymaga zintegrowanego podejścia, które łączy interwencje organizacyjne, rozwój kompetencji pracowników oraz narzędzia motywacyjne. Stopniowe wdrażanie narzędzi gen AI, budowanie świadomości ich potencjału oraz promowanie kultury otwartości na innowacje stanowią kluczowe elementy transformacji. Rozwiązania te powinny być wdrażane na podstawie uznanych koncepcji zarządzania zmianą, takich jak modele Lewina, Kottera oraz ADKAR, z uwzględnieniem indywidualnych i systemowych uwarunkowań sektora publicznego.

Rekomendacje dla systemu ewaluacji polityk publicznych

Na podstawie przeprowadzonej analizy autorzy zaproponowali rekomendacje dla systemu ewaluacji polityk publicznych.

1. **Budowanie zespołów i rozwijanie współpracy w obszarze gen AI**
   Organizacje publiczne mogą zyskać ogromne korzyści, wyznaczając liderów AI, którzy będą inspirować i koordynować działania związane z wdrażaniem tej technologii w ewaluacji. Sieci współpracy pomiędzy jednostkami, obejmujące takie inicjatywy jak wspólne projekty, rotacja stanowisk, programy stażowe czy wymiana doświadczeń, wspierają kreatywność i innowacyjność. Takie inicjatywy w systemie ewaluacji pozwoliłyby na lepsze zrozumienie możliwości gen AI oraz dostosowanie rozwiązań do specyficznych potrzeb każdej jednostki ewaluacyjnej.
2. **Przygotowanie strategii AI dla systemu ewaluacji**
   Strategia wdrożenia AI w systemie ewaluacji może pomóc uporządkować działania, określając kluczowe cele i mierzalne wskaźniki efektywności (ang. *key performance indicators*, KPI). Włączenie gen AI do procesów raportowania oraz analiz w czasie rzeczywistym znacząco zwiększyłoby ich przejrzystość i przydatność w podejmowaniu decyzji. Strategia powinna również uwzględniać potrzeby uczestników systemu, w tym rozwój kompetencji pracowników oraz sposoby na optymalizację pracy z wykorzystaniem nowych narzędzi analitycznych.
3. **Tworzenie przestrzeni dla innowacji – „Evaluation LAB"**
   Stworzenie ośrodka typu „Evaluation LAB" jako centrum innowacji może być krokiem milowym w rozwoju systemu ewaluacji. To miejsce umożliwić mogłoby projektowanie i testowanie nowych rozwiązań AI w bezpiecznym środowisku, zanim zostaną one wdrożone na większą skalę. Dzięki temu uczestnicy systemu będą mogli eksperymentować z technologią, dostosowując ją do swoich potrzeb, a jednocześnie minimalizować ryzyko. Wdrożenie „Evaluation LAB" może również pomóc w analizie skuteczności wdrażanych rozwiązań oraz w identyfikacji nowych obszarów zastosowań dla gen AI.
4. **Promowanie liderów innowacji i wspieranie kreatywnych pomysłów**
   Zachęcanie do kreatywnego wykorzystania gen AI w ewaluacji poprzez nagradzanie innowacyjnych pomysłów, buduje kulturę otwartą na zmiany. Konkursy na nowe zastosowania gen AI oraz wyróżnienia za efektywne wykorzystanie technologii mogą zmotywować pracowników do angażowania się w proces transformacji. Rozwiązania

opracowane i przetestowane (np. w „Evaluation LAB”) mogą następnie inspirować inne jednostki do adaptacji podobnych narzędzi, dostosowanych do lokalnych potrzeb.

5. **Uproszczenie i automatyzacja procedur ewaluacyjnych z wykorzystaniem gen AI**
   AI ma potencjał do uproszczenia oraz przyspieszenia wielu procedur i zadań ewaluacyjnych, takich jak analiza danych czy przygotowywanie raportów. Automatyzacja tych procesów pozwoli pracownikom skupić się na bardziej strategicznych zadaniach, oszczędzi czas i zredukuje koszty. Zastosowanie AI może również przyczynić się do większej dokładności i spójności w wynikach, co przełoży się na wyższą jakość ewaluacji.
6. **Zorientowanie na potrzeby beneficjentów z wykorzystaniem AI**
   Zastosowanie gen AI w analizie danych administracyjnych pozwala na ograniczenie konieczności angażowania uczestników programów w dostarczanie dodatkowych informacji. Dzięki integracji danych oraz automatycznemu monitorowaniu, procesy ewaluacyjne mogą być prowadzone sprawniej i mniej obciążająco dla uczestników. To podejście minimalizuje koszty transakcyjne, jednocześnie zwiększając precyzję i trafność analiz.
7. **Budowanie pozytywnej atmosfery wokół AI w ewaluacji**
   Przyjazna komunikacja na temat korzyści wynikających z zastosowania AI oraz docenianie inicjatyw pracowników związanych z tą technologią mogą zbudować zaangażowanie i pozytywne nastawienie uczestników systemu. Ponadto promowanie sukcesów i podkreślanie praktycznych korzyści płynących z wdrożenia AI w ewaluacji przyciągnie talenty oraz zachęci do aktywnego udziału w jej transformacji.
8. **Systematyczne monitorowanie postępów i ewaluacja zmian z wykorzystaniem AI**
   Regularne raportowanie wyników wdrożeń AI, takich jak oszczędności czasowe, lepsza jakość analiz czy efektywność kosztowa, wzmacnia zaufanie do procesu transformacji. AI może również pomóc w monitorowaniu wskaźników efektywności i identyfikacji obszarów wymagających usprawnień, co pozwoli uczestnikom systemu na bieżąco dostosowywać działania do zmieniających się potrzeb.

Zaproponowane rekomendacje mogą wesprzeć wdrażanie narzędzi gen AI w ewaluacji polityk publicznych, jednocześnie odpowiadając na potrzeby zarówno organizacji, jak i uczestników ewaluowanych programów w sposób efektywny i zorientowany na przyszłość.

12.4. Zakończenie

Przeprowadzona analiza wskazuje, że skuteczna adaptacja generatywnej sztucznej inteligencji w systemie ewaluacji polityk publicznych jest zarówno szansą na poprawę efektywności, jak i wyzwaniem związanym z przełamywaniem barier poznawczych, organizacyjnych i behawioralnych. Wskazane ograniczenia, takie jak przywiązanie do tradycyjnych metod, opór wobec zmiany czy obawy związane z nowymi technologiami, można przezwyciężyć dzięki zintegrowanym działaniom systemowym.

W rozdziale zidentyfikowano kluczowe obszary transformacji, które opierają się na automatyzacji procesów, tworzeniu zachęt organizacyjnych, edukacji pracowników oraz wprowadzaniu innowacyjnych modeli zarządzania zmianą. Wskazano również takie mechanizmy jak impulsy behawioralne (ang. *behavioral nudges*) czy projekty pilotażowe jako narzędzia wspierające adaptację AI. Inspiracje zaczerpnięte z międzynarodowych doświadczeń zostały uzupełnione o konkretne rekomendacje dostosowane do polskiego systemu ewaluacji. Podkreślono znaczenie jasnych strategii wdrożeniowych, spójnych reguł i stałego monitorowania postępów.

Wnioski płynące z analizy dowodzą, że sukces wdrożenia AI w systemie ewaluacji zależy od podejścia łączącego technologię, rozwój kompetencji oraz zmianę kultury organizacyjnej. Kluczowe będzie stopniowe wdrażanie narzędzi gen AI oraz tworzenie przestrzeni do eksperymentowania i testowania nowych rozwiązań. Wypracowane strategie mogą przyczynić się do zwiększenia precyzji analiz, oszczędności czasu i innych zasobów, a także do poprawy jakości rekomendacji dla polityk publicznych.

Podsumowując, wdrożenie gen AI w ewaluacji nie jest jedynie kwestią technologiczną, lecz przede wszystkim organizacyjną i kulturową. Stworzenie środowiska sprzyjającego innowacjom, opartego na zaufaniu, edukacji i motywacji pracowników, pozwoli na skuteczną transformację systemu ewaluacji i przygotuje go na wyzwania przyszłości.

Bibliografia

- Brynjolfsson, E., McAfee, A. (2014). *The Second Machine Age*.
- Burnes, B. (2020). The Origins of Lewin's Three-Step Model of Change. *Journal of Applied Behavioral Science*, 56(1), 32–59.

- Cameron, E., Green, M. (2019). *Making Sense of Change Management: A Complete Guide to the Models, Tools and Techniques of Organizational Change*. Kogan Page Publishers.
- Chłoń-Domińczak, A., Pokorski, J., et al. (2022). *Public Intelligence. Wykorzystanie danych administracyjnych do monitorowania i ewaluacji polityk publicznych*. PARP, Warszawa.
- Clarke, L. (1997). *Zarządzanie zmianą*. Warszawa: Wydawnictwo Gebethner i S-ka.
- Deloitte United States (2023). Generative AI to Transform Government Procurement (dostęp: 24.09.2024).
- Fal, K. (2022). Instytucja piaskownic regulacyjnych jako instrument wspierający innowacyjność gospodarek. W: *Business Law Journal*, t. LXXV z nr 10/2022 (892), Polskie Wydawnictwo Ekonomiczne (DOI 10.33226/0137-5490.2022.10.5).
- Fundacja Digital Poland (2020). *13 faktów o transformacji cyfrowej* (dostęp: 27.11.2024).
- Guerreiro, J., Rebelo, S., Teles P. (2023). Regulating Artificial Intelligence. *Working Paper* 31921, NBER (dostęp: 27.11.2024).
- Hiatt, J.M. (2006). *ADKAR: A Model for Change in Business, Government and our Community*.
- Hornstein, H.A. (2015). The integration of project management and organizational change management is now a necessity. *International Journal of Project Management*, 33(2), 291–298.
- IBM Center for The Business of Government (2023). *Government Procurement and Acquisition: Opportunities and Challenges Presented by Artificial Intelligence and Machine Learning*.
- Kinicki, A., Williams, B. (2020). *Management: A Practical Introduction* (wyd. 9). McGraw Hill Education, New York.
- Kotter, J.P. (1996). *Leading Change*. Harvard Business School Press.
- Kotter, J.P. (2014). *Accelerate: Building Strategic Agility for a Faster-Moving World*. Harvard Business Review Press.
- KPMG (2024). Cyberbezpieczeństwo w ESG (dostęp: 24.09.2024).
- Lampart, M. (2023). *Rozwiązania generatywnej sztucznej inteligencji – zagrożenia i aspekty prawne*. PARP Warszawa.
- McKinsey & Company (2024). Revolutionizing Procurement: Leveraging Data and AI for Strategic Advantage (dostęp: 24.09.2024).
- MIT Sloan Management Review Polska (2024). Dlaczego transformacje cyfrowe się nie udają? (dostęp: 24.09.2024).
- NBER (2018). Artificial Intelligence, Globalization, and Strategies for Economic Development. *Working Paper* 24653 (dostęp: 24.09.2024).
- NBER (2021). Artificial Intelligence, Globalization, and Strategies for Economic Development (dostęp: 24.09.2024).
- OECD (2019). *Digital Government Review*.
- OECD (2020). *AI and Public Procurement*.
- OECD (2021). *AI Monitoring in Public Administration*.
- OECD (2021). *The Digital Transformation of Public Services*.
- OECD (2021). *AI Principles and Policies in Public Administration*.
- Platforma Przemysłu Przyszłości. Przeczytaj co powinien robić menadżer odpowiedzialny za cyfrową zmianę (dostęp: 24.09.2024).
- Pokorski, J. (2023). *Podejmowanie decyzji a pułapki kognitywne. Studium przypadku z sektora publicznego*. Praca w ramach studiów MBA Uniwersytetu SWPS, IPI PAN, Woodbury UVU.
- Pokorski, J. (2024). *Transformacja cyfrowa w systemie ewaluacji wykorzystania funduszy europejskich w Polsce*. Praca dyplomowa, MBA Uniwersytet SWPS, IPI PAN, Woodbury UVU.

- Pokorski, J. (2024). *Co jest potrzebne, aby przeprowadzić ewaluację i w jaki sposób możemy ją realizować?* W: Bienias S., Kupiec T., Strzęboszewski P. (red.), Poradnik ewaluacji dla administracji publicznej, Ministerstwo Funduszy i Polityki Regionalnej, Warszawa.
- Rogers, E.M. (2003). *Diffusion of Innovations*.
- Singapore Government (2020). *AI in Public Administration Program*.
- Singapore Government (2020). *AI-Driven Public Performance Evaluation* (AIDPPE).
- Thaler, R.H., Sunstein, C.R. (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness*.
- The World Economic Forum (2019). *Guidelines for AI Procurement*.
- UK Government (2020). *Digital Leaders Awards*.
- UK Government (2020). *Policy Insight Platform*.

13. Zakończenie

Karol Olejniczak
Dominik Batorski
Jacek Pokorski

13.1. Odkrywanie potencjału gen AI

Generatywna sztuczna inteligencja to technologia o rosnącym i dynamicznie rozwijającym się potencjale, która może znacząco usprawnić procesy analityczno-badawcze i ewaluacyjne, oferując coraz bardziej zaawansowane sposoby przetwarzania danych, syntetyzowania wiedzy i wspierania decydentów.

Jak pokazały kolejne rozdziały tej książki, zastosowanie modeli językowych w jednostkach ewaluacyjnych i innych komórkach analityczno-badawczych otwiera perspektywy automatyzacji wielu czasochłonnych procesów, od analizy dokumentów po generowanie raportów i rekomendacji, a ich potencjał wciąż dynamicznie rośnie. Kluczowe pytanie nie brzmi już: czy AI znajdzie zastosowanie w ewaluacji programów i polityk rozwoju?, ale: jak zrobić to mądrze, aby technologia ta rzeczywiście wspierała podejmowanie lepszych decyzji w zarządzaniu publicznym?

Efektywne wykorzystanie narzędzi gen AI wymaga więc zarówno eksperymentowania z ich praktycznymi zastosowaniami, jak i świadomej refleksji nad ich ograniczeniami. Dopiero próby konkretnych zastosowań, w konkretnych zadaniach analityczno-badawczych pozwalają odkryć wyzwania, nieoczekiwane szanse optymalizacji, ale również zdefiniować konkretne problemy, dylematy i szersze implikacje płynące z zastosowania tej technologii. Ujmując to obrazowo, jesteśmy na etapie pionierskim – nie ma jeszcze jednej wytyczonej ścieżki działań czy mapy zastosowań. Eksplorujemy zarówno możliwości, jak i pułapki krok po kroku (*case by case*).

Patrząc przez pryzmat działań w procesie ewaluacji, w książce udało nam się rozpoznać możliwe zastosowania gen AI przy tworzeniu zakresów zamówień na ewaluację (rozdział 5.) w całym spektrum sposobów pozyskiwania wiedzy z przeglądów źródeł, analiz

jakościowych, ilościowych, analiz sieciowych (rozdziały 6.–9.) oraz w określaniu strategii komunikacji i prezentacji wyników (rozdział 10.). Najwięcej uwagi i miejsca poświęciliśmy pozyskiwaniu wiedzy – ponieważ jest to najbardziej czasochłonny etap badań ewaluacyjnych, a zastosowania narzędzi gen AI wydają się tu najbardziej oczywiste.

Dynamika rozwoju technologii gen AI pozwala jednak na zastosowania, które, gdy zaczynaliśmy pracę nad tą książką, były jedynie w sferze dyskusji. Przykładami mogą być możliwości wykorzystywania syntetycznych interesariuszy przy antycypowaniu potrzeb wiedzy dla konkretnych polityk publicznych, tworzenie wirtualnych person i recenzentów rekomendacji czy wspieranie, w czasie rzeczywistym, dyskusji z odbiorcami badań informacją zwrotną generowaną z puli dowodów lub na podstawie komentarzy *ad hoc*. Te przykłady i jeszcze wiele innych zastosowań to nowe, fascynujące przestrzenie dalszej eksploracji, do której zachęcamy naszych Czytelników.

13.2. Architektura współpracy z gen AI

Pierwsze eksperymenty praktyczne, opisane w tym tomie, pokazują, że zastosowanie gen AI wymaga podjęcia wielu, często drobnych i technicznych decyzji, które ostatecznie jednak decydują o jakości efektu pracy modelu. Roboczo tę konfigurację wyborów nazwaliśmy **„architekturą współpracy z gen AI”**. Sam proces określania takiej architektury jest wysoce iteratywny (musimy wypróbować różne opcje), to swoiste majsterkowanie z architekturą zadań, w które angażujemy AI. Wymaga to zarówno ciekawości, jak i cierpliwości. Bazując na opisanych w tej książce przypadkach i próbach, uważamy, że na architekturę produktywnej współpracy z gen AI, składają się cztery grupy zagadnień.

Pierwszą grupę stanowi **rola i zadania, jakie wyznaczymy dla gen AI**. To, jak zaprojektujemy proces pracy (ang. *workflow*) i jaki zakres zadań wyznaczymy, determinuje późniejsze efekty. Można pracować na pojedynczych prostych zadaniach, ale też możemy ułożyć całą sekwencję procesu wspieranego różnymi agentami gen AI. Dobrą ilustrację tych wyborów znajdą Czytelnicy w rozdziale 5., pokazującym całą sekwencję pracy przy projektowaniu badania i przygotowywaniu Opisów Przedmiotu Zamówienia. Podobnie w rozdziale 10. zademonstrowano zarówno zadania kreatywne (tworzenie person), jak i działania komunikacyjne połączone w jeden proces projektowy. Z kolei rozdziały 6., 7., 8. i 9. pokazały

inne podejście – konkretne zadania analityczne, które można realizować punktowo i z wykorzystaniem różnych asystentów gen AI.

Drugą grupą są **technologie narzędzi gen AI**. Chodzi zarówno o wybór dostępnych na rynku modeli LLM czy całych platform gen AI, jak również o kwestie ustawień domyślnych tych platform czy technologicznych rozwiązań za nimi schowanych (np. ustawienia tokenizacji w różnych modelach, kwestie douczania modeli, ang. *fine-tuning*). Rozdział 1. stanowi zarys spektrum technologii (modele, ich możliwości, w tym transparentność i elastyczność niektórych ustawień „fabrycznych"). W rozdziale 3. pokazano z kolei, jak ustawienia technologii determinują ważne aspekty bezpieczeństwa i poufności danych przy zastosowaniu gen AI do działań badawczych. W rozdziale 4. dobrze wyjaśnione zostały konsekwencje ustawień technicznych dla jakości analizy tekstów podłączanych z bazami informacji. Pewne ograniczenia logiki semantycznej pracy modeli omówiono w rozdziale 7., gdzie opisano ograniczenia gen AI w zastosowaniu do niektórych rodzajów przeglądów.

Trzecia grupa kwestii to **interakcje użytkownika z gen AI**. Tu mamy na myśli różne strategie promptowania gen AI, sekwencje współpracy, a także przyjęte przez użytkownika techniki krytycznej oceny odpowiedzi udzielnych przez model. Kluczowych informacji na ten temat dostarczył rozdział 2., w którym omówiono fundamenty skutecznej interakcji z gen AI. Z kolei rozdziały 6., 7., 8. i 9. prezentowały praktyczne przykłady interakcji z gen AI w różnych kontekstach analitycznych – od analiz jakościowych po ilościowe i sieciowe, pokazano tu, jak dostosować strategie współpracy do konkretnych zadań badawczych. Natomiast w rozdziale 10. wyjaśniono, jak zróżnicowanych technik promptowania wymaga efektywne przekazywanie informacji z badań różnym grupom odbiorców.

Wreszcie czwarte zagadnienie kształtujące skuteczną współpracę z gen AI to **zasoby informacyjne, na których bazuje gen AI**. Pod tym hasłem kryją się decyzje dotyczące tego, czy pracujemy z LLM, czy łączymy LLM z zasobami Internetu, czy też chcemy wzmocnić pracę modelu naszymi własnymi zbiorami danych i specjalnie dobraną pulą informacji. W rozdziale 4. szczegółowo omówiono, jak można łączyć modele gen AI z wewnętrznymi bazami dokumentów, co pozwala na znaczące rozszerzenie możliwości analitycznych. Rozdziały 7. i 8., dotyczące przeglądów źródeł i analiz ilościowych, pokazały jak gen AI może być wykorzystana do pracy z różnorodnymi zasobami danych, zarówno wewnętrznymi, jak i zewnętrznymi. W rozdziale 9. poświęconym analizom sieciowym zademonstrowano, jak gen AI może być

zastosowana do przetwarzania złożonych struktur danych, co wymaga odpowiedniego doboru i przygotowania zasobów informacyjnych.

Podsumowując, architektura współpracy z generatywną AI w badaniach ewaluacyjnych stanowi złożony ekosystem wzajemnie powiązanych elementów. Jak pokazują doświadczenia opisane w poszczególnych rozdziałach tej książki, efektywne wykorzystanie gen AI wymaga świadomego projektowania czy raczej poszukiwania konfiguracji, która zadziała przy konkretnym zadaniu. Naszym zdaniem, kluczem do sukcesu jest **zintegrowane podejście**, w którym wszystkie cztery omówione aspekty (wybór zadań, technologii, styl interakcji i zasoby informacji) są traktowane jako elementy spójnego projektu. Wydaje nam się, że w coraz większym stopniu badacze będą architektami takich procesów, umiejętnie łączącymi własną ludzką ekspertyzę z możliwościami generatywnej AI (Ethan Mollick określa to jako ***co-intelligence***).

13.3. Szersze implikacje płynące z zastosowań gen AI

W tym tomie skoncentrowaliśmy się na codziennej praktyce. Jesteśmy jednak świadomi tego, że nasza codzienna praca jest osadzona w szerszym kontekście organizacyjnym administracji publicznej, a samo pojawienie się gen AI w sektorze publicznym ma głębokie implikacje dla zespołów i organizacji realizujących zadania analityczne i ewaluacje, a nawet całego systemu realizacji polityk publicznych.

Te szersze kwestie organizacyjne staraliśmy się zasygnalizować w rozdziałach 11. i 12. zamykających książkę. Oba te rozdziały zgodnie wskazują, że skuteczne wdrożenie gen AI w politykach publicznych i praktyce ewaluacji wymaga kompleksowego (systemowego) podejścia, łączącego aspekty technologiczne, organizacyjne i kulturowe. Konieczne jest nie tylko opracowanie odpowiednich regulacji i wytycznych, ale również przełamanie barier poznawczych, rozwijanie kompetencji pracowników oraz tworzenie przestrzeni do eksperymentowania i testowania nowych rozwiązań. Szczególną uwagę chcemy zwrócić na cztery wątki problemowe.

Pierwszym wyzwaniem, wskazanym w obu rozdziałach, jest **złożoność regulacyjno-organizacyjna**. W rozdziale 11. szczegółowo przeanalizowano różnorodne podejścia do tworzenia wewnętrznych regulacji dotyczących wykorzystania gen AI w administracji publicznej, podkreślając bieżącą dominację podejścia skoncentrowanego na zarządzaniu ryzykiem. Dokumenty analizowane w tym rozdziale wyraźnie wskazują na większą koncentrację na potencjalnych zagrożeniach niż na szansach związanych z gen AI. Z kolei w rozdziale 12. zidentyfikowano bariery poznawcze i organizacyjne, które utrudniają adaptację nowych technologii w jednostkach ewaluacyjnych, w tym przywiązanie do tradycyjnych metod pracy i opór przed zmianą, oraz zaproponowano, jak można je przełamywać.

Drugim istotnym wyzwaniem jest **kwestia kompetencji i kultury organizacyjnej**. W rozdziale 12. podkreślono, że jednostki ewaluacyjne często kurczowo trzymają się dotychczasowych, oswojonych praktyk, w obawie przed wyjściem poza „strefę komfortu". Autorzy wskazują na szereg błędów poznawczych (ang. *biases*), które hamują transformację. Rozdział 11. uzupełnił tę perspektywę – wskazano tu kluczową rolę rozwoju kompetencji pracowników w zakresie odpowiedzialnego wykorzystania gen AI, co jest podkreślane w analizowanych dokumentach regulacyjnych. Aby w pełni wykorzystać możliwości gen AI, konieczna będzie nie tylko adaptacja technologii, lecz także zmiana podejścia do metodologii ewaluacyjnej, w której coraz większą rolę będą odgrywać kompetencje cyfrowe, umiejętność krytycznego myślenia oraz zdolność do pracy z dynamicznie rozwijającymi się systemami AI.

Trzecim obszarem wyzwań jest **równowaga między innowacyjnością a odpowiedzialnością**. W rozdziale 11. pokazano, że organizacje publiczne próbują znaleźć złoty środek między wykorzystaniem potencjału gen AI a zarządzaniem ryzykiem, związanym z prywatnością danych, bezpieczeństwem i zgodnością z przepisami. Rozdział 12. uzupełnił tę perspektywę, zaproponowano tu stopniowe wdrażanie narzędzi gen AI poprzez projekty pilotażowe i hybrydowe modele ewaluacji, które pozwalają na ewolucję systemu bez radykalnych zmian strukturalnych.

Czwartym wyzwaniem jest kwestia **przejrzystości i zaufania**. Instytucje publiczne przykładają szczególną wagę do przejrzystości w wykorzystaniu gen AI. „Wytłumaczalność" jest jednak dużym wyzwaniem z racji na zamknięty – co do zasady – charakter tych technologii i swoistą czarną skrzynkę procesów sztucznej inteligencji. Na poziomie codziennej praktyki jednostek ewaluacyjnych można te postulaty przełożyć na potrzebę wypracowania standardów

dokumentowania i raportowania użycia gen AI w procesach ewaluacyjnych. Budowanie pozytywnej atmosfery wokół AI oraz systematyczne monitorowanie i komunikowanie korzyści płynących z jej zastosowania są kluczowe dla budowania zaufania do procesu transformacji.

13.4. Refleksja przy eksperymentowaniu – pytania otwarte

Rdzeniem ewaluacji jest ocena. Nasze ewaluacyjne środowisko jest biegłe w kluczowych elementach naszego rzemiosła, czyli: definiowaniu kryteriów oceny, określaniu punktu odniesienia (standardu), pomiarze, a wreszcie osądzie wartości na podstawie wyników badania i dowodów (Scriven, 1991). Kuszący jest więc pomysł sformułowania kryteriów oceny zastosowania gen AI w praktyce ewaluacji.

Takie próby są już podejmowane w naszym środowisku. W *New Directions For Evaluation* B. Montrosse-Moorhead (2023) proponuje dwie grupy kryteriów: oceniające konceptualizację i implementację AI w ewaluacji (spójność projektu i wdrożenia, wydajność procesu, kwestie równościowe w procesie) oraz oceniające efekty użycia AI w ewaluacji (skuteczność, zaufanie, poprawność metodologiczna i wiarygodność, zrozumiałość, równość społeczna uzyskanych informacji i dowodów).

Naszym zdaniem, zbyt wcześnie jest jeszcze na formułowanie kryteriów oceny, ponieważ AI jest nowym fenomenem, zmienia się bardzo dynamicznie, a jego potencjalne zastosowania i oczekiwania wobec tej technologii dopiero staramy się rozpoznać i określić. Zamiast tego zachęcamy raczej do eksploracji możliwości i krytycznej refleksji.

Przy pojedynczych eksperymentach z codziennym zastosowaniem gen AI proponujemy trzy zdroworozsądkowe pytanie weryfikujące: 1. Czy dane zastosowanie gen AI jest bezpieczne? 2. Czy jest etyczne? 3. Czy jest produktywne?

Patrząc natomiast szerzej – systemowo – sygnalizujemy następujące kwestie do dyskusji w środowisku badaczy społecznych i praktyków ewaluacji. Formułując te kwestie, inspirowaliśmy się nie tylko wąską praktyką ewaluacji, ale również szerokimi dyskusjami

dotyczącymi zasad zastosowania AI w polityce publicznej i zarządzaniu publicznym (European Commission, 2019; OECD, 2019; Mikalef, 2024; Bullock et al., 2024 Sesion II).

Wydajność i skuteczność pracy

Czy zastosowanie gen AI pozwoli w najbliższych latach usprawnić naszą pracę operacyjną i wzmocni naszą misję ewaluacyjną – pomoc decydentom w podejmowaniu lepszych decyzji, dzięki wykorzystaniu wiedzy opartej na badaniach? Jakie implikacje ma używanie gen AI dla przejrzystości i rzetelności oferowanej przez nas wiedzy? Z jednej strony gen AI może radykalnie przyspieszyć produkcję wiedzy (np. Jacob (2024) pisze wręcz o automatyzacji, nie tylko ewaluacji, ale całego cyklu polityk publicznych). Z drugiej strony, poza samym zalewem informacji pojawiają się głębsze wyzwania dla jakości i trafności naszej pracy. Istnieje ryzyko metodologicznego zawężenia, gdzie łatwość przetwarzania danych tekstowych przez AI może skłaniać do marginalizowania trudniejszych do uchwycenia, lecz często kluczowych danych jakościowych czy obserwacyjnych. Ponadto nadmierne poleganie na zgeneralizowanych modelach AI może prowadzić do homogenizacji wniosków ewaluacyjnych, które choć szybko wygenerowane, mogą tracić wrażliwość na unikalny kontekst badanych interwencji. Dlatego, oprócz wypracowania procedur zapewniających rzetelność i przejrzystość analiz generowanych przez AI, kluczowa staje się ciągła krytyczna refleksja nad doborem metod i interpretacją wyników w specyfice danego badania. Niezbędna jest większa ostrożność w interpretacji wyników generowanych przez AI oraz wypracowanie procedur i standardów zapewniających rzetelność i przejrzystość analiz.

Bezpieczeństwo i prywatność danych

Jakie implikacje dla bezpieczeństwa i prywatności danych może mieć zastosowanie gen AI? Te kwestie wstępnie eksplorowaliśmy w aspekcie wykorzystania lokalnych modeli lub szyfrowania danych analizowanych poza infrastrukturą użytkownika, ale pytanie to ma szczególne znacznie w przypadku dużych rozwiązań i modeli gen AI działających w chmurze, z których korzystanie wymaga przekazywania własnych danych (w tym treści i dokumentów) do dostawcy danego modelu.

Odpowiedzialność i kwestie etyczne

Jaki zakres odpowiedzialności bierze na siebie użytkownik gen AI, stosując modele do codziennych zadań badawczo-analitycznych oraz wspierając się nimi przy formułowaniu rekomendacji dla polityk i programów? Jakie są implikacje związane z kwestiami prawa autorskiego i własności produktów współtworzonych z AI? Czy jesteśmy świadomi spraw dotyczących sprawiedliwości społecznej i niedyskryminacji? Wiemy, że modele wytworzone w procesie uczenia maszynowego mogą bowiem zawierać szereg skrzywień (*biases*), wynikających choćby z niedoskonałości danych użytych w procesie trenowania modelu. Szczególnie problematyczne są tu skrzywienia prowadzące do wykluczenia, dyskryminacji, czy powielania negatywnych stereotypów i uprzedzeń. Co się stanie w sytuacji, gdy te skrzywienia – które mogą być przez AI nie tylko powielane, ale wręcz subtelnie wzmacniane – zaczną przekładać się na ocenę interwencji publicznych i kształt przyszłych programów? Dodatkowym wyzwaniem jest tu ryzyko „iluzji obiektywizmu": wyniki generowane przez AI, często prezentowane w precyzyjny sposób, mogą wydawać się bardziej wiarygodne i neutralne niż są w rzeczywistości, maskując potencjalne błędy lub właśnie owe skrzywienia. Skłania to do bezkrytycznego przyjmowania wniosków i rodzi pilną potrzebę dyskusji nad nowymi standardami oceny jakości i trafności analiz ewaluacyjnych wspomaganych przez AI.

Dynamika produkcji i wykorzystania wiedzy

W jakim kierunku rozwinie się system ewaluacji? Jak zmieni się współpraca między zlecającymi a wykonawcami ewaluacji? Co będzie zlecane na zewnątrz, a jakie zadania realizować będą ewaluatorzy wewnętrzni? Rozwój i popularyzacja generatywnej AI nieuchronnie prowadzą do redefinicji roli ewaluatorów w ekosystemie polityk publicznych. Dotychczasowe zadania, wymagające „ręcznej" analizy dużych zbiorów danych różnego typu, mogą zostać częściowo zautomatyzowane, co oznacza, że specjaliści będą mogli skupić się na innych czynnościach – zarówno na tych, na które wcześniej nie stawiali dużego akcentu w procesie ewaluacji, jak i na zupełnie nowych zadaniach merytoryczno-technologicznych. Jednocześnie jednak wprowadzenie modeli językowych do procesu ewaluacji wiąże się z ryzykiem nadprodukcji treści i inflacji informacji, co może utrudniać selekcję i weryfikację kluczowych wniosków. Ten rodzący się paradygmat analityki wspieranej gen AI wymaga wypracowania hybrydowych metodologii badawczych (zgodnie z koncepcją *co-intelligence*), łączących analityczną precyzję algorytmów z krytycznym osądem i kontekstową wiedzą ekspertów. Istotnym wyzwaniem staje się również zmiana kompetencji ewaluatorów, którzy

oprócz tradycyjnych umiejętności metodologicznych powinni rozwijać zdolność efektywnego recenzowania i weryfikacji wiedzy generowanej przez AI. Ponadto pojawia się konieczność stworzenia nowych standardów metaewaluacyjnych, które umożliwią ocenę wiarygodności i użyteczności wniosków formułowanych przy znaczącym udziale narzędzi generatywnych.

Role i relacje uczestników procesu ewaluacji

Jak potencjalne zastosowanie gen AI wpłynie na role i relacje uczestników procesu ewaluacji? Chodzi tutaj nam o kwestie dyskursu demokratycznego i reprezentatywności. Z jednej strony wykorzystanie gen AI jako dominującego narzędzia procesu może zaburzyć partycypacyjny i demokratyczny charakter ewaluacji, wzmacniając autonomizację algorytmów AI i jednocześnie redukując rolę człowieka w procesie (jako respondenta, interesariusza, ewaluatora, decydenta). Nadmierne poleganie na AI w analizie i raportowaniu może ograniczyć kluczowy dla wartości ewaluacji dialog między badaczami a interesariuszami, niezbędny do wspólnego nadawania sensu wynikom i budowania ich zrozumienia oraz akceptacji. To może prowadzić do osłabienia uspołeczniającej funkcji ewaluacji, ze wszystkimi tego konsekwencjami (Korporowicz, 2022, 2023).

Z drugiej strony te technologie mogą podnieść dyskusje na wyższy poziom i uczynić je głębszymi – opartymi w mniejszym stopniu na emocjach i incydentalnych historiach, a bardziej na całościowych dowodach i wartościach, które tym dowodom przypisują różni interesariusze. Główną barierą w procesach deliberatywnych jest czas, który interesariusze muszą zainwestować, aby zgłębić temat i zapoznać się z dowodami. Agenty gen AI mogą więc działać jak asystenci streszczający wnioski z badań prostym językiem. Mogą również odgrywać rolę ekspertów (biegłych) odpowiadających na pytania (także na podstawie wskazanej biblioteki źródeł) pojawiające się podczas dyskusji, praktycznie w czasie rzeczywistym. Mogą na bieżąco tworzyć opcje scenariuszy, uwzględniając elementy podnoszone przez interesariuszy w dyskusji. Algorytmy AI mogą też pomóc w analizie dużych ilości danych pochodzących z szerokich konsultacji społecznych (w tym dotyczących raportów z badań ewaluacyjnych), wychwytując niezauważone wcześniej wzorce czy głosy mniejszości (lub nawet wskazując nieobecne głosy), które w tradycyjnych metodach mogą umykać z powodu typowych ludzkich skrzywień poznawczych. Gen AI może nawet służyć jako mediator w polaryzujących dyskusjach, identyfikując i sugerując punkty wspólne oraz ułatwiając osiąganie konsensusu – modele „rozumiejące” już dziś nie ograniczają się

do typowo reaktywnych odpowiedzi, a rozwijają zaawansowaną wrażliwość na stosowanie odpowiedniego języka w złożonych kontekstach i wielowymiarowego antycypowania zdarzeń.

Konkludując wątek ról i relacji uczestników ewaluacji, gen AI może być czynnikiem alienującym i osłabiającym społeczne funkcje ewaluacji albo katalizatorem pogłębionej partycypacji w procesach wypracowywania zrozumienia problemów i definiowania rozwiązań. Naszym zdaniem, kierunek zastosowań zależy od aktywności, determinacji i pomysłowości środowiska ewaluacyjnego.

Współczesne jednostki ewaluacyjne i szerzej – środowisko badaczek i badaczy polityk publicznych stoją przed wyzwaniem integracji nowoczesnych technologii w sposób, który wzmocni misję publiczną i podniesie jakość dostarczanych analiz. Nawigacja w tym nowym krajobrazie technologicznym wymaga nie tylko entuzjastycznej eksploracji możliwości, ale przede wszystkim głębokiej etycznej i metodologicznej rozwagi. Odpowiedzi na te wyzwania będziemy musieli szukać wspólnie, jako środowisko badaczy i praktyków ewaluacji, dzieląc się doświadczeniami i stopniowo wypracowując dobre praktyki adaptacji AI w naszej dziedzinie. Mamy nadzieję, że przedstawione w tej książce przykłady, metody i refleksje staną się inspiracją i praktycznym przewodnikiem dla badaczy, ewaluatorów i decydentów, którzy podejmują wyzwanie odkrywania potencjału generatywnej AI dla swojej pracy.

Bibliografia

- Bubeck, S., Chandrasekaran, V., Eldan, R., et al. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv* preprint arXiv:2303.12712.
- Bullock, J.B., Chen, Y.-C., Himmelreich, J., et al. (red.). (2024). *The Oxford Handbook of AI Governance*. Oxford University Press.
- Christiano, P.F., Leike, J., Brown, T., et al. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.
- Dobbins, M., Robeson, P., Ciliska, D., et al. (2009). A description of a knowledge broker role implemented as part of a randomized controlled trial evaluating three knowledge translation strategies. *Implementation Science*, *4*(23).
- European Commission. (2019). *High-Level Expert Group on Artificial Intelligence: Ethics Guidelines for Trustworthy AI*. European Commission.
- Frost, H., Geddes, R., Haw, S., et al. (2012). Experiences of knowledge brokering for evidence-informed public health policy and practice: three years of the Scottish Collaboration for Public Health Research and Policy. *Evidence & Policy*, *8*(3), 347–359.
- Jacob, S. (2024). Artificial Intelligence and the Future of Evaluation: from Augmented to Automated Evaluation. *Digital Government: Research and Practice*, online first, 1–12 (dostęp: 31.03.2025).

- Kojima, T., Gu, S.S., Reid, M.Y., et al. (2022). Large language models are zero-shot reasoners. *arXiv* preprint arXiv:2205.11916.
- Korporowicz, L. (2023). Sztucznej mądrości nie będzie. *Perspektywy Kultury* nr 42 (3/2023), s. 117-130, Akademia Ignatianum w Krakowie (dostęp: 31.03.2025).
- Korporowicz, L. (2022). Społeczna aksjologia ewaluacji. Uniwersytet Jagielloński w Krakowie.
- Mikalef, P. (2024). *Responsible AI governance: from ideation to implementation*. W: Charalabidis, Y., Medaglia, R., van Noordt, C. (red.), Research Handbook on Artificial Intelligence and Decision Making in Organizations. Edward Elgar Publishing, s. 126–142 (dostęp: 31.03.2025).
- Montrosse-Moorhead, B. (2023). Evaluation criteria for artificial intelligence. *New Directions for Evaluation*, *178-179*, 123–134 (dostęp: 31.03.2025).
- OECD. (2019). *The OECD AI Principles* (dostęp: 31.03.2025).
- Olejniczak, K., Kupiec, T., Raimondo, E. (2014). *Brokerzy wiedzy. Nowe spojrzenie na rolę jednostek ewaluacyjnych*. W: Haber, A., Olejniczak, K. (red.), *(R)ewaluacja 2. Wiedza w działaniu*. Polska Agencja Rozwoju Przedsiębiorczości, s. 67–112.
- Olejniczak, K., Raimondo, E., Kupiec, T. (2016). Evaluation units as knowledge brokers: Testing and calibrating an innovative framework. *Evaluation*, *22*(2), 168–189 (dostęp: 31.03.2025).
- Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. *arXiv* preprint arXiv:2203.02155.
- Pokorski, J. (2024). *Co jest potrzebne, aby przeprowadzić ewaluację i w jaki sposób możemy ją realizować?* W: Bienias S., Kupiec T., Strzęboszewski P. (red.), Poradnik ewaluacji dla administracji publicznej, Ministerstwo Funduszy i Polityki Regionalnej, Warszawa (dostęp: 31.03.2025).
- Scriven, M. (1991). *Evaluation thesaurus* (4th ed ed.). Sage.
- Vaswani, A., Shazeer, N.M., Parmar, N., et al. (2017). Attention is All you Need. *Neural Information Processing Systems*.
- Wei, J., Wang, X., Schuurmans, D., et al. (2022). Chain of thought prompting elicits reasoning in large language models. *arXiv* preprint arXiv:2201.11903.
- Ziegler, D.M., Stiennon, N., Wu, J., et al. (2019). Fine-tuning language models from human preferences. *arXiv* preprint arXiv:1909.08593.

14. Autorzy

Karol Olejniczak

Autor jest profesorem nauk o polityce publicznej na Uniwersytecie SWPS i współzałożycielem firmy badawczej Evaluation for Government Organizations (EGO). Jego prace naukowe koncentrują się na projektowaniu rozwiązań przyjaznych obywatelom oraz ewaluacji skuteczności polityk publicznych. W swojej działalności badawczej i edukacyjnej wykorzystuje gry, eksperymenty i projektowanie behawioralne. Ma dwudziestoletnie doświadczenie w prowadzeniu badań i szkoleń dla polskich, unijnych, amerykańskich i kanadyjskich instytucji publicznych. Współpracował z Komisją Europejską, Bankiem Światowym, Europejskim Bankiem Odbudowy i Rozwoju oraz wieloma polskimi ministerstwami i instytucjami samorządowymi. Przez osiem lat kierował Akademią Ewaluacji, programem studiów podyplomowych dla administracji publicznej. Obecnie kieruje Centrum Projektowania i Ewaluacji Polityk Publicznych (C4PDE) Uniwersytetu SWPS. Więcej informacji.

Dominik Batorski

Autor jest socjologiem i ekspertem w dziedzinie *data science*, który łączy pracę naukową z działalnością doradczą i biznesową. Pracuje w Interdyscyplinarnym Centrum Modelowania Matematycznego i Komputerowego (ICM) na Uniwersytecie Warszawskim. Jego badania koncentrują się na zmianach społecznych i gospodarczych związanych z upowszechnieniem technologii informacyjno-komunikacyjnych oraz sposobach korzystania z technologii i zachowań użytkowników. Jest współautorem badań takich jak „Diagnoza społeczna” oraz „Cyfrowa gospodarka”. Ma wieloletnie doświadczenie w analizie dużych danych zbieranych przez serwisy internetowe i firmy telekomunikacyjne. Jest również współtwórcą narzędzi analitycznych do monitoringu i optymalizacji marketingu w mediach społecznościowych w oparciu o dane. Ekspert jest laureatem Nagrody im. Marka Cara za badania nad rozwojem społeczeństwa informacyjnego w Polsce. Więcej informacji.

Jacek Pokorski

Autor od 2004 r. jest związany z Polską Agencją Rozwoju Przedsiębiorczości – główny specjalista Jednostki Ewaluacyjnej, od 2016 r. kierownik Sekcji Monitoringu i Ewaluacji, a od 2025 r. zastępca dyrektora Departamentu Analiz i Strategii. W latach 2010–2014 wiceprezes

Polskiego Towarzystwa Ewaluacyjnego, następnie do 2024 przewodniczący organu nadzoru PTE. Absolwent UKSW (obecnie członek Rady Biznesu Wydziału Społeczno-Ekonomicznego), SGH oraz SWPS, IPI PAN & Woodbury School of Business/Utah Valley University (dyplom MBA). Specjalizuje się w badaniach przedsiębiorstw i otoczenia biznesu oraz ewaluacjach programów rozwoju gospodarczego (m.in. wsparcia konkurencyjności i innowacyjności gospodarki, podnoszenia kwalifikacji kadr, rozwoju infrastruktury społeczno-gospodarczej). Autor i współredaktor publikacji, ewaluator, trener i konsultant w zakresie badań ewaluacyjnych. Więcej informacji.

Kamil Filipek

Autor jest dyrektorem Centrum Sztucznej Inteligencji i Modelowania Komputerowego UMCS oraz pracownikiem naukowym w Katedrze Socjologii na Wydziale Filozofii i Socjologii UMCS. Specjalizuje się w badaniach nad nowymi technologiami, cyfrową transformacją społeczeństwa oraz analizą mediów społecznościowych. Jest autorem wielu publikacji naukowych oraz uczestnikiem licznych projektów badawczych, skupiających się na wpływie technologii na współczesne społeczeństwo. Więcej informacji.

Andrzej Gołoś

Autor jest ekspertem w badaniach konsumenckich i konsultantem, specjalizującym się w wykorzystaniu danych w procesach decyzyjnych oraz realizacji złożonych projektów badawczych. Jest również service designerem i facylitatorem warsztatów design thinking. Współpracował z wiodącymi markami z sektora e-commerce i retail (m.in. Allegro, Eurocash), farmaceutycznego (m.in. Pfizer, Bayer) oraz publicznego (m.in. PARP, UOKiK). Więcej informacji.

Paulina Grabowska

Autorka bada i projektuje innowacje z zakresu zrównoważonego i regeneratywnego rozwoju z wykorzystaniem rozwiązań AI. Jest również wykładowczynią biodesignu na Wydziale Projektowania SWPS w Warszawie. W swojej pracy projektowej i naukowej wykorzystuje szereg narzędzi służących wizualizowaniu jej koncepcji i tworzeniu prezentacji, które angażują i inspirują odbiorców. Więcej informacji.

Andrzej Jędrzejowski

Autor jest badaczem-analitykiem z misją popularyzacji wykorzystania dużych zasobów danych rejestrowych i administracyjnych w badaniach prowadzonych przez sektor publiczny.

Rozwija narzędzia i procesy służące zwiększeniu efektywności zagospodarowania danych wtórnych (w tym m.in. *web scraping*, *text mining*). Związany z sektorem publicznym (PARP – Departament Analiz i Strategii oraz Biuro Informatyki) i prywatnym (infuture.institute). Więcej informacji.

Paweł Kędzia

Autor prowadzi własną działalność badawczo-rozwojową, specjalizując się w przetwarzaniu języka naturalnego (NLP). W swojej pracy doktorskiej badał metody pogłębionej analizy semantycznej w zadaniu wykrywania relacji semantycznych między fragmentami tekstu w języku polskim. Jego doświadczenie obejmuje także projekty komercyjne, takie jak narzędzia ewaluacyjne służące do metaanaliz i ekstrakcji informacji z analiz i raportów. Ma duże doświadczenie w realizacji projektów badawczych i ewaluacyjnych.

Tomasz Kupiec

Autor jest adiunktem w Centrum Europejskich Studiów Regionalnych i Lokalnych (EUROREG) na Uniwersytecie Warszawskim oraz wiceprezesem Polskiego Towarzystwa Ewaluacyjnego. Jego specjalizacja obejmuje wykorzystanie ewaluacji w procesie stanowienia i realizacji polityk publicznych, systemy monitoringu i ewaluacji, determinanty efektywności, uczenie się organizacji publicznych, wykorzystanie wiedzy w procesie podejmowania decyzji oraz zarządzanie strategicznym rozwojem lokalnym i regionalnym. Jest autorem i współautorem wielu raportów i publikacji z zakresu ewaluacji polityki spójności, determinantów wykorzystania ewaluacji, funkcjonowania systemów ewaluacji i monitoringu oraz zarządzania strategicznego rozwojem. Więcej informacji.

Bartosz Ledzion

Autor jest projektantem usług i badaczem, specjalizuje się w analizie oraz rozwoju narzędzi sztucznej inteligencji w badaniach ewaluacyjnych i marketingowych. Jest współzałożycielem firmy EGO S.C. oraz ekspertem w dziedzinie badań społecznych, service design i ekonomii behawioralnej. Ukończył kierunek User Experience Design w 2018 roku i uczestniczył w szkoleniach Cambridge Business Analytics. Jako certyfikowany moderator Design Thinking przeprowadził wiele projektów UX i service design dla sektorów publicznego i prywatnego. Ekspert ma bogate doświadczenie w branży FMCG, prowadził badania konsumenckie i UX dla takich firm jak Nestle Waters, Nestle Purina i Eurocash. Więcej informacji.

Marta Lesiak

Autorka jest pracownikiem Jednostki Ewaluacyjnej Polskiej Agencji Rozwoju Przedsiębiorczości. Ukończyła UMCS na kierunku socjologia oraz studia podyplomowe na Uniwersytecie Warszawskim (Metody Statystyczne w Biznesie, na Wydziale Nauk Ekonomicznych). Współautorka i autorka raportów z badań społecznych i ewaluacyjnych oraz analiz danych do tych raportów. Obecnie kształci się w kierunku *Data Scientist*. Założycielka strony internetowej popularyzującej statystykę.

Igor Lyubashenko

Autor jest adiunktem na Uniwersytecie SWPS, specjalizuje się m.in. w wykorzystaniu narzędzi AI w badaniach jakościowych. Uzyskał doktorat z politologii i jest autorem książki oraz licznych artykułów naukowych. Ma doświadczenie w międzynarodowych projektach badawczych, np. Horyzont 2020. Więcej informacji.

Grzegorz Rzeźnik

Autor jest doktorem nauk o polityce i socjologiem, z doświadczeniem w zarządzaniu projektami dotyczącymi wykorzystania nowych technologii w działalności biznesowej i badawczej. Jego działalność obejmuje obszary związane z rozwojem innowacji we współpracy z firmami, instytucjami publicznymi oraz jednostkami naukowymi. Ma doświadczenie w przygotowaniu raportów dotyczących wykorzystania nowych technologii, w tym analiz rynkowych technologii sztucznej inteligencji w gospodarce. Więcej informacji.

Jacek Szut

Autor jest absolwentem Uniwersytetu Adama Mickiewicza w Poznaniu. Badacz ilościowy i jakościowy, entuzjasta (choć nie bezkrytyczny) AI. Z ewaluacją związany od roku 2006. Obecnie zatrudniony w PARP, gdzie zajmuje się ewaluacją programów wsparcia przedsiębiorczości, od 2025 r. kierownik Sekcji Ewaluacji Programów Rozwojowych. W ostatnich latach współpracował z UKSW przy ewaluacji projektu Środowiskowe Centrum Zdrowia Psychicznego. Wcześniej pracował w sektorze prywatnym – w Pentor RI, TNS Pentor oraz Pro Publicum. Autor i współautor wielu raportów, w tym m.in. ewaluacyjnych. Od ponad roku testuje i wykorzystuje w badaniach (projektowanie i realizacja) rozwiązania bazujące na sztucznej inteligencji.

Dominika Wojtowicz

Autorka jest profesorką w Katedrze Ekonomii na Akademii Leona Koźmińskiego. Ma stopień doktora habilitowanego nauk ekonomicznych. W trakcie swojej kariery odbyła stypendia naukowe na Wydziale Nauk Politycznych Universita' di Siena we Włoszech oraz na Wydziale Zarządzania Universidad Politecnica de Cartagena w Hiszpanii. Jej zainteresowania koncentrują się na szeroko pojętej skuteczności interwencji publicznych ukierunkowanych na wzmacnianie rozwoju gospodarczego i społecznego. Jest ekspertką w zakresie badań ewaluacyjnych projektów i programów współfinansowanych ze środków unijnych. Koordynowała projekty badawcze dotyczące nowoczesnych metod i narzędzi oceny wybranych obszarów interwencji publicznych oraz wykorzystania praktyk ewaluacji w ocenie skutków regulacji. Więcej informacji.

Teresa Wyszyńska

Autorka jest absolwentką studiów matematycznych z wieloletnim doświadczeniem w dziedzinie ewaluacji polityk publicznych. Pracuje w Polskiej Agencji Rozwoju Przedsiębiorczości, gdzie od 2016 r. odpowiada za planowanie i projektowanie zakresu oraz metodologii badań ewaluacyjnych. W swojej pracy koncentruje się na analizie skuteczności wdrażanych programów oraz ocenie ich wpływu na rozwój społeczno-gospodarczy. Autorka zajmuje się także analizami danych, które są wykorzystywane w procesach decyzyjnych.

15. Aneksy

Aneks 1. Informacje o projektach, w których testowaliśmy użycie gen AI[1]

Tabela 15.1. Projekty, w których testowaliśmy użycie gen AI

Projekty
Projekt 1. Wsparcie transformacji cyfrowej przedsiębiorstw. Wnioski na przyszłość – badanie ewaluacyjne Jednym z celów projektu było zidentyfikowanie, jakie trendy, procesy i technologie rozwijają się i będą rozwijać się w obszarze szeroko pojętej transformacji cyfrowej w perspektywie do 2030 r. oraz jakie korzyści i wyzwania może to rodzić dla polskich przedsiębiorców. Projekt realizowała PARP we współpracy z konsorcjum spółek EGO, LBiE oraz Fundeko.
Projekt 2. Zastosowania poważnych gier (ang. *serious games*) w politykach żywnościowych Celem był systematyczny przegląd praktyk zastosowań poważnych gier w projektowaniu, wdrażaniu i ewaluacji polityk publicznych dotyczących kwestii żywności i żywienia (ang. *food policies*). Badanie zrealizowano w ramach grantu badawczego finansowanego z NCN.
Projekt 3. Typologia instrumentów publicznych – „kije, marchewki, kazania” Celem badania było sprawdzenie popularności klasycznej typologii instrumentów polityk publicznych w globalnej literaturze naukowej z zakresu polityk publicznych i ewaluacji. Badanie zrealizowano w ramach grantu badawczego finansowanego z NCN.
Projekt 4. Współtworzenie mechanizmów zmiany – studium ograniczania strat żywności Celem badania był systematyczny przegląd rozwiązań publicznych, dotyczących ograniczania strat i marnotrawstwa żywności (ang. *food loss and waste*), które zostały stworzone w procesie współprojektowania (ang. *co-design of public solutions and services*). Badanie zrealizowano w ramach grantu badawczego finansowanego z NCN.
Projekt 5. Ekspertyza dla Parlamentu Europejskiego Jednym z celów ekspertyzy było przedstawienie różnorodności funduszy, programów i inicjatyw UE, które bezpośrednio lub pośrednio przyczyniają się do osiągnięcia celów polityki spójności, a w szczególności identyfikacja i analiza ich redundancji oraz wskazanie możliwości ich konsolidacji. Zadanie było realizowane w ramach badania *Streamlining EU Cohesion funds: addressing administrative burdens and redundancy* (IPIB/REGI/IC/2024-010) na zlecenie Dyrekcji Generalnej ds. Polityk Wewnętrznych Unii oraz Dyrekcji ds. Polityki Strukturalnej i Spójności Parlamentu Europejskiego.

Źródło: Opracowanie własne.

[1] Niniejszy aneks odnosi się do rozdziału pt. „Przeglądy źródeł wspierane gen AI”.

Aneks 2. Zaplanowanie przeglądu[2]

Znajdowanie ram analitycznych

W projekcie 3. sprawdziliśmy, czy gen AI pytany o typologie instrumentów publicznych zaproponuje ugruntowane w literaturze typologie, w tym zacytuje typologię „kijów, marchewek i prawienia kazań" opublikowaną w książce pt. *Policy Instruments and Their Evaluation* pod red. M.-L. Bemelmans-Videc, R. Rist & E. Vedung (Transaction Publishers, 1998).

Zadaliśmy więc następujące pytanie (nieco odmienne dla LLM i dla RAG-ów):

> *Prompt for universal gen AI (LLM): Act as a researcher. What are the typologies of policy instruments? Provide literature sources for your answer.*
>
> *Prompt for research-specific gen AI (RAGs): What are the typologies of policy instruments?*

Opis wyników przedstawiamy w poniżej tabeli.

Tabela 15.2. Rzetelność gen AI w odwoływaniu się do ugruntowanych typologii – przykład typologii instrumentów polityk publicznych

Gen AI	Czy podało źródłową książkę	Komentarz
ChatGPT 4.o (LLM)	Tak, ALE…	Podało sześć typologii instrumentów polityki z literatury: 1. Model NATO Hooda (1983); 2. Klasyfikacja Vedunga (regulacje, środki ekonomiczne, informacja); 3. Bemelmans-Videc et al. (marchewki, kije, kazania); 4. Schneider i Ingram (1990); 5. Linder i Peters (1989); 6. Narzędzia Salamona (2002). Powołało się na główną książkę, ale jej treść pomieszało, tworząc z niej dwie różne typologie (2. i 3.).

[2] Niniejszy aneks odnosi się do rozdziału pt. „Przeglądy źródeł wspierane gen AI".

Gen AI	Czy podało źródłową książkę	Komentarz
Claude Sonnet 3.5 (LLM)	Tak, ALE…	Zaproponowało pięć typologii instrumentów polityki. 1. Typologia NATO Hooda; 2. Carrots, Sticks, and Sermons Vedunga; 3. Narzędzia publiczne wg Salamona; 4. Instrumenty merytoryczne i proceduralne Howletta (2000); 5. Instrumenty regulacyjne, ekonomiczne i informacyjne wg Bemelmans-Videc et al. Powołało się na główną książkę, jednak z cytowanego źródła wymyśliło dwie różne typologie (2. i 5.).
Meta AI: meta.ai (LLM)	Tak, ALE…	Pojawia się odniesienie do rozdziału z typologią, ale: 1. Rozdział jest błędnie przypisany do innej książki z tego okresu (Peters & van Nispen); 2. Przedstawiony jest tylko jeden aspekt typologii – instrumenty ekonomiczne. Jest to najbardziej niedokładne odniesienie do wyszukiwania gen AI. Jednak, co ciekawe, jest to jedyny przypadek, w którym gen AI przywołało najnowszy typ narzędzi – szturchnięcia (*nudges* – Thaler & Sunstein, 2008).
Elicit (RAG)	Tak	Gen AI zaczyna od typologii z książki (właściwe odniesienie do całej książki i poprawny opis typów), a następnie podsumowuje rozszerzenia lub modyfikacje typologii przez różnych autorów.
Perplexity Pro (RAG)	Tak	Przywołuje typologię jako jedną z siedmiu głównych: 1. Typologia Doerna i Phidda (1992); 2. Zaktualizowany schemat NATO Hooda i Margettsa (2007); 3. Typologia Lindera i Petersa; 4. Typologia Vedunga; 5. Typologia IPBES; 6. Instrumenty polityki środowiskowej; 7. Klasyfikacja Pala (2014).
Scite (RAG)	Pośrednio	Omawia instrumenty polityki jako „typologię Vedunga”, ale nie podaje bezpośredniego odniesienia do książki jako źródła. Agent korzysta z książki podczas kwerendy roboczej, ale w ostatecznym tekście odnosi się do publikacji, która podsumowuje oryginalną książkę (Gelius et al, 2022), czyli „cytuje z drugiej ręki”.
SCISPACE (RAG)	Tak, ALE…	Zaczyna się od wskazania typologii jako „jednej z najważniejszych” i buduje odpowiedzi wokół jej typów. Odwołuje się wyraźnie do rozdziału Vedunga, ale główne cytaty obejmują badaczy podsumowujących oryginalną książkę (czyli cytowanie „z drugiej ręki”).

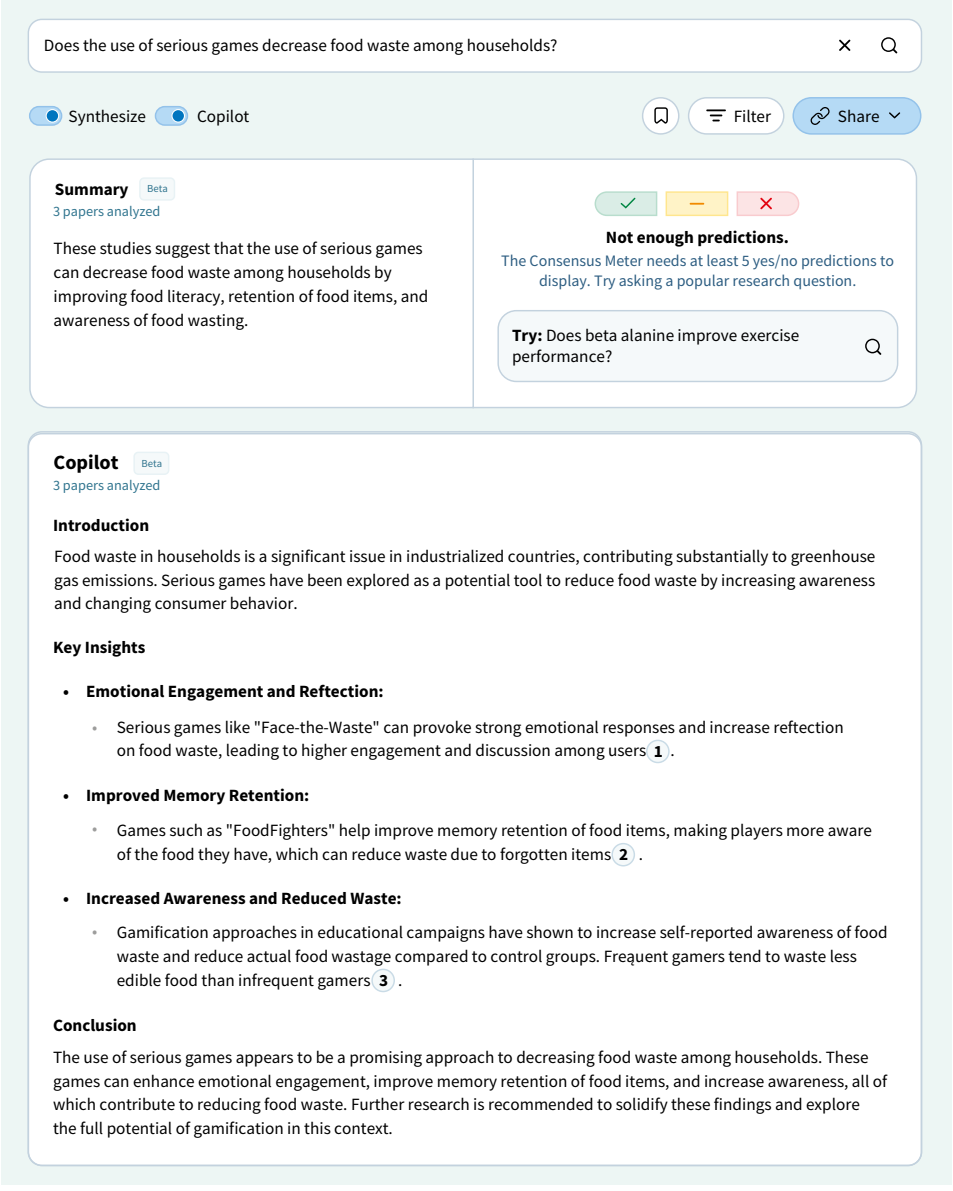
Źródło: Opracowanie własne, test wykonany 3.07.2024.

Aneks 3. Szukanie i selekcja źródeł[3]

Przykłady wyszukiwania semantycznego z gen AI

Rycina 15.1. Wyniki wyszukiwania literatury przez gen AI – Consensus

Zadane pytanie: *Does the use of serious games decrease the food waste among households?*

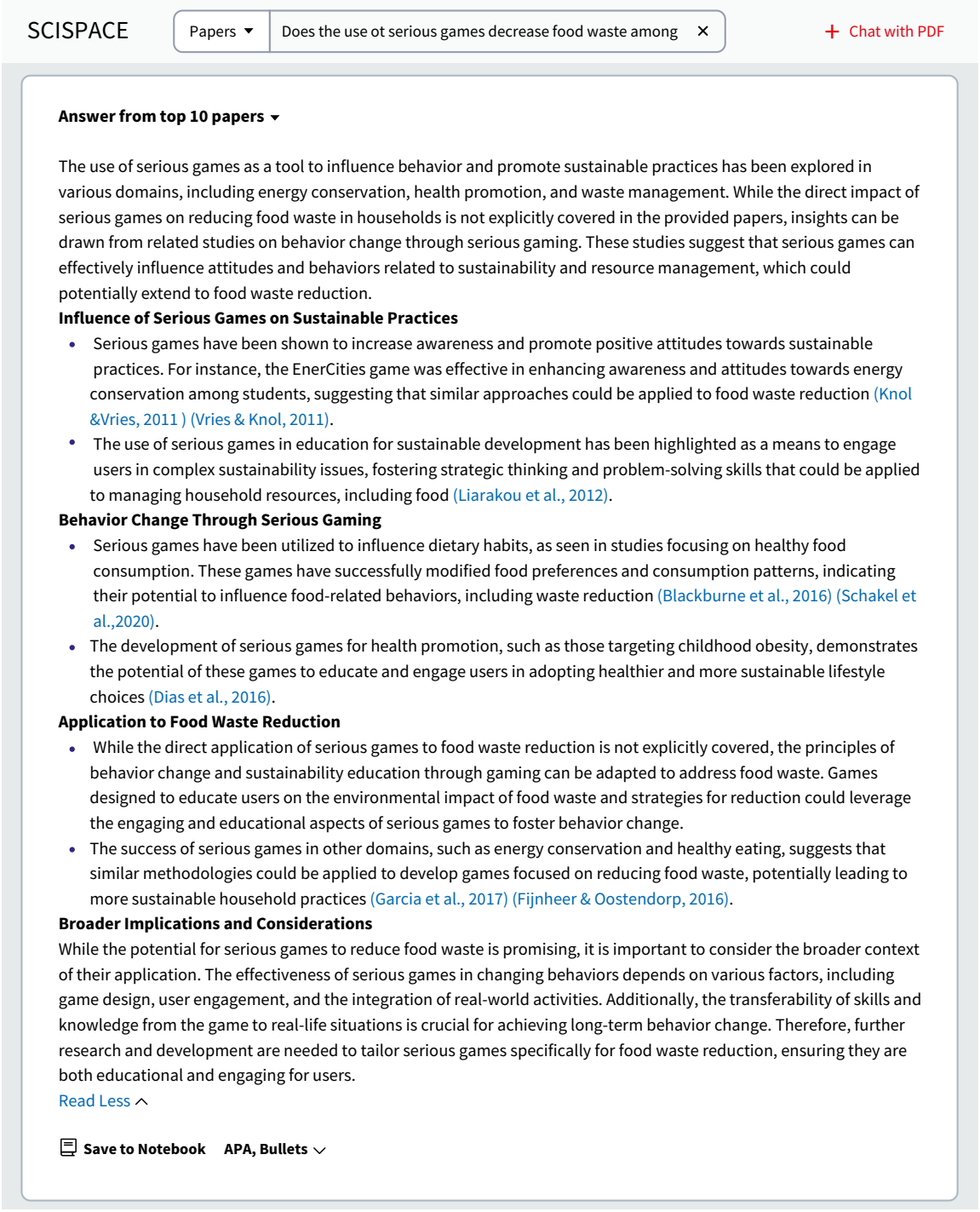


Źródło: Consensus.

[3] Niniejszy aneks odnosi się do rozdziału pt. „Przeglądy źródeł wspierane gen AI”.

Rycina 15.2. Wynik wyszukiwań literatury przez gen AI – SCISPACE

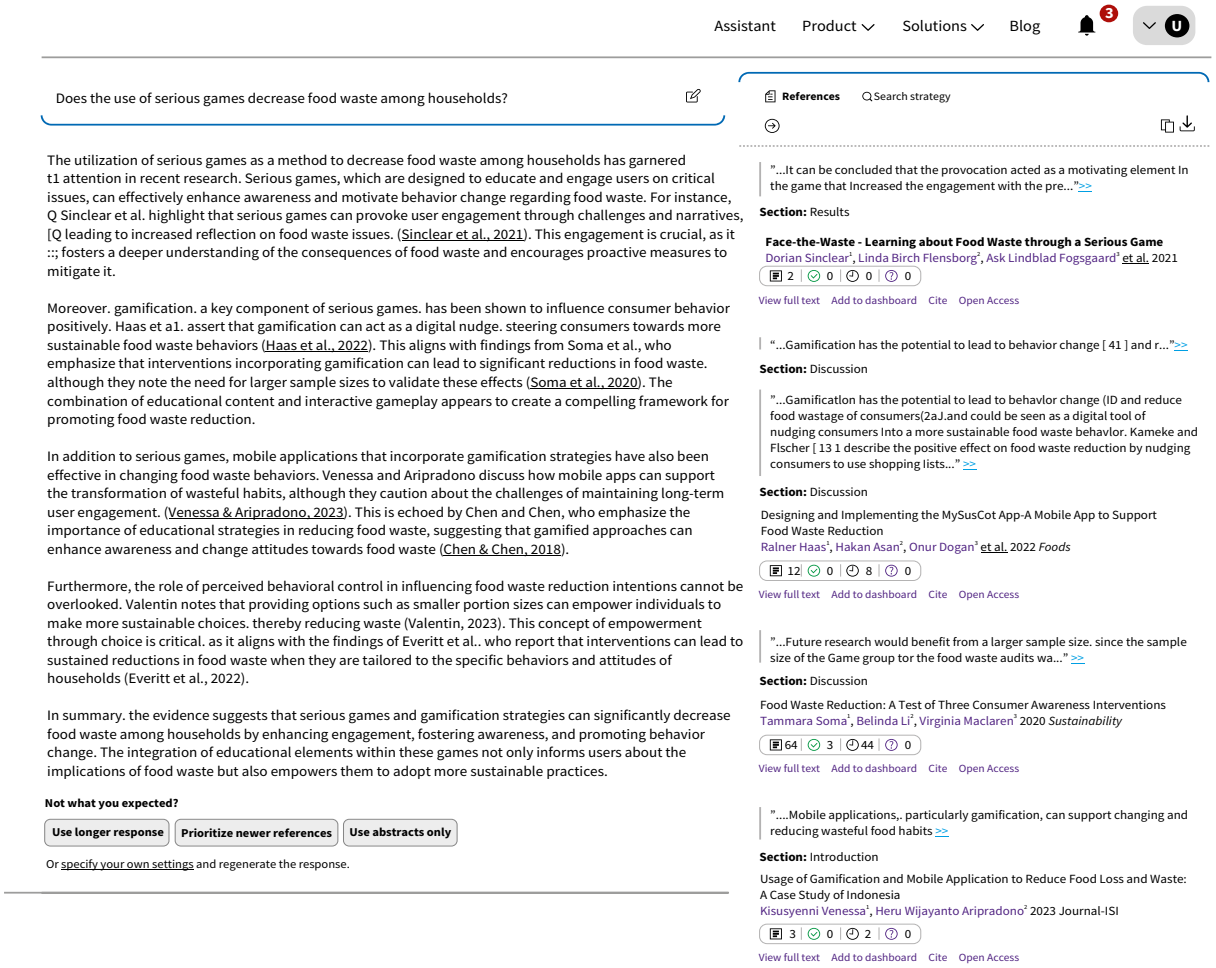
Zadane pytanie: *Does the use of serious games decrease the food waste among households?*



Źródło: SCISPACE.

Rycina 15.3. Wynik wyszukiwań literatury przez gen AI – Scite

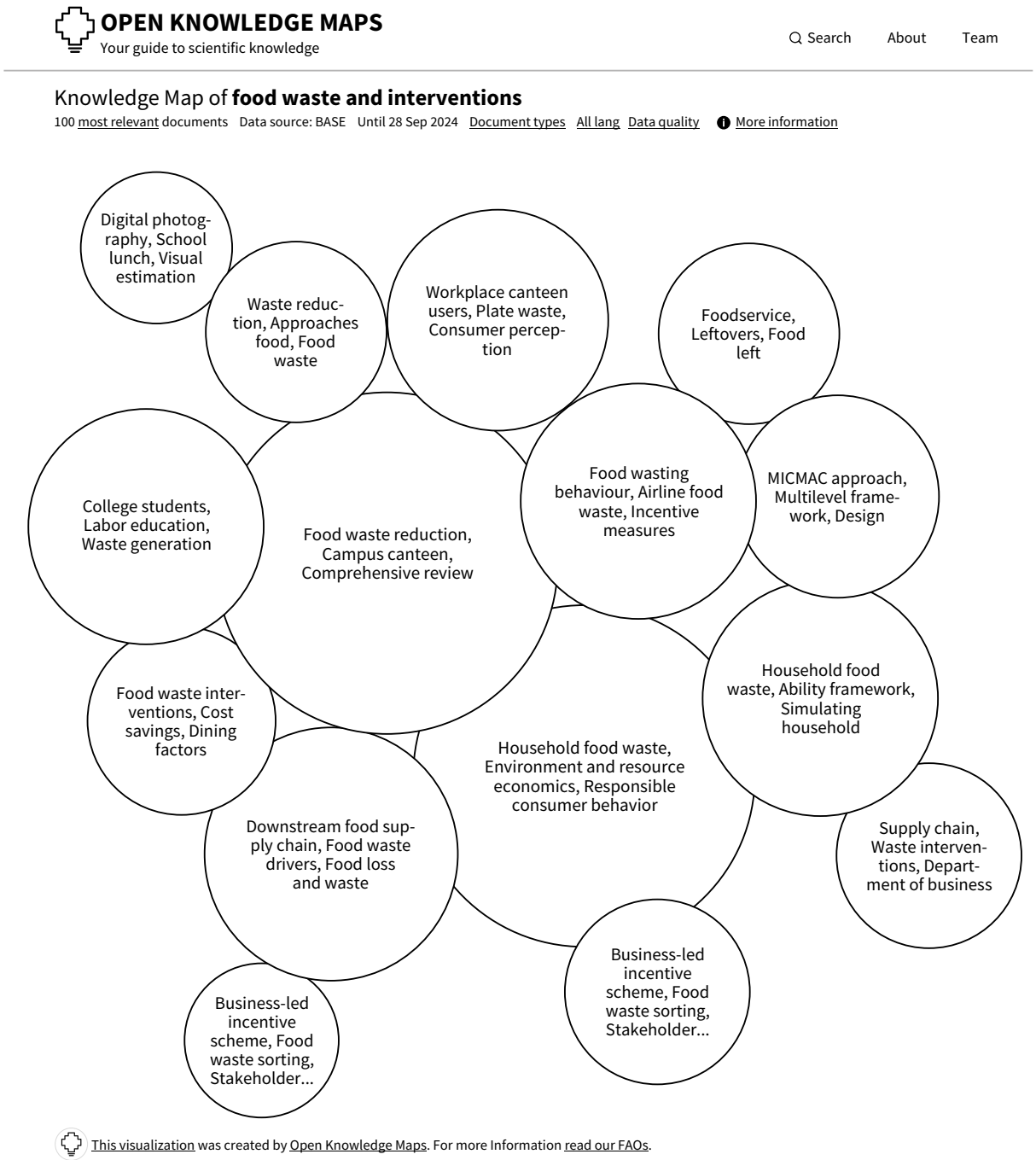
Zadane pytanie: *Does the use of serious games decrease the food waste among households?*



Źródło: Scite.

Gen AI wyszukujące powiązania

Rycina 15.4. Przykład mapy powiązań publikacji otwartego dostępu dla wyszukiwania „Food Waste AND intervention”



Źródło: Open knowledge maps.

Usuwanie zduplikowanych źródeł

Weryfikacja przydatności ChatGPT do integracji rekordów z różnych źródeł i usuwania „dubli" dokonana została poprzez porównanie osiąganych wyników z tymi uzyskanymi przy wykorzystaniu narzędzi opartych o R. W naszym projekcie źródła zostały wyszukane z trzech baz: Scopus (239 rekordów), Web of Science (165) i Lens (189) z wykorzystaniem tej samej frazy wyszukania[4]. Bazy te oczywiście mają dużą część wspólną i wyszukanie w każdej kolejnej bazie dodaje coraz mniej unikalnych rekordów, a coraz więcej dubli.

Do integracji wykorzystaliśmy pakiet bibliometrix w oprogramowaniu R, który pozwolił zidentyfikować 207 dubli. Metoda ta okazała się niedoskonała i kolejne 36 dubli znaleźliśmy, analizując podobieństwo abstraktów przy użyciu pakietu stringdist w R. W tym drugim przypadku wszystkie pary abstraktów wskazane jako podobne[5] były potwierdzane dodatkowo przez badacza. Ostateczna zintegrowana baza liczyła zatem 350 unikalnych rekordów.

Pierwsza próba wykorzystania ChatGPT 4.o polegała na załadowaniu trzech plików z rekordami z poszczególnych baz i ograniczoną liczbą kolumn o ujednoliconych nagłówkach (*Authors, Title, Source, Abstract, Year*) oraz prośbą o zintegrowanie baz w jedną i usunięcie dubli[6]. W odpowiedzi na najbardziej ogólny prompt, który nie sugeruje, na podstawie których kolumn powinno odbyć się szukanie dubli, czat stworzył bazę n = 474, opierając się jedynie na tytułach publikacji. Poproszony o wykonanie zadania na podstawie kolumny abstrakt stworzył bazę n = 548, a abstrakt, tytuł i rok n = 590.

Dalsza konwersacja wyjaśniła, że czat identyfikuje duble na zasadzie dokładnego dopasowania, czyli nie rozwiązuje problemu różnych opisów tego samego rekordu w różnych bazach. Prośba o zastosowanie przybliżonego dopasowania kończyła się błędem i komunikatem, że procedura trwa zbyt długo. Poproszony o zastosowanie technik, które analizują semantykę tekstu, czat zaproponował wykorzystanie modeli przetwarzania języka naturalnego, np. BERT, do których on nie ma dostępu.

Dopiero polecenie: *Znajdź dla każdego rekordu w połączonej bazie x rekord*ów *najbardziej podobnych pod względem treści abstraktu,* a następnie *usu*ń *rekordy zduplikowane,*

[4] Tej samej w sensie znaczenia, ale z zachowaniem odpowiedniej dla każdej bazy składni i operatorów.

[5] Według odległości Levenshteina.

[6] Testowaliśmy pliki w formatach csv i xlsx bez istotnych różnic w wynikach.

przyjmując że podobieństwo powyżej 0,8 oznacza duplikat, stworzyło bazę potencjalnie unikalnych rekordów liczącą n = 357 o zawartości zbieżnej z bazą uzyskana przy wykorzystaniu narzędzi opartych o R. Zaproponowane podejście jest użyteczną alternatywą dla osób niemających kompetencji, pozwalającą na skorzystanie z narzędzi wymagających kodowania w R czy Pythonie.

Przesiewanie abstraktów i streszczeń

Zdolność ChatGPT 4.o do „przesiewania" rekordów w oparciu o kryteria włączenia/wyłączenia zweryfikowaliśmy poprzez porównanie ocen wygenerowanych przez czat i badaczy na próbie 333 abstraktów z artykułów. Wyniki porównania pokazuje poniższa tabela.

Tabela 15.3. Zgodność ocen badacza i ChatGPT

	Znaleziony artykuł o grze poważnej dotyczący jedzenia	Znaleziona gra „była grana"
Zgodne oceny	230/333 (69%)	179/298 (60%)
Fałszywie pozytywne	89/103 (86%)	52/119 (44%)
Fałszywie negatywne	14/103 (14%)	67/119 (56%)

Źródło: Opracowanie własne.

Udział ocen zgodnych nie przekroczył 70% dla żadnego z pytań. Udziały te w zasadzie nie zmieniały się przy modyfikacji prompta. Ciekawą obserwacją jest, że w zależności od pytania / kryterium wśród ocen czata rozbieżnych z ocenami badacza znacząco wahają się udziały ocen fałszywie pozytywnych i negatywnych[7]. Nie udało nam się znaleźć wyjaśnienia tych różnic, które pozwoliłoby poprawić zgodność ocen.

[7] Przez ocenę fałszywie pozytywną rozumiemy sytuację, gdy czat uznaje kryterium za spełnione, a badacz nie. Ocena fałszywie negatywna, to sytuacja odwrotna.

Aneks 4. Analiza i synteza źródeł[8]

Analiza pełnych treści publikacji

Na danych z Projektu 2., które przeszły już przez etap przesiewania abstraktów, przetestowaliśmy użyteczność Petal w analizie treści pełnych publikacji.

Wykorzystując funkcje AI Table, stworzyliśmy tabelę, której wiersze stanowią kolejne publikacje, a kolumny odpowiadają pytaniom, które chcemy zadać do tekstu. Petal generuje odpowiedzi w każdej komórce takiej tabeli. Testem objęliśmy 30 publikacji, do których zadaliśmy dziesięć pytań. Następnie badacze samodzielnie czytali publikacje i oceniali trafność odpowiedzi Petal w skali: 0 = niedokładna, 1 = słaba, 2 = dobra, 3 = świetna.

80% odpowiedzi Petal zostało ocenionych jako co najmniej dobre, a 60% jako świetne. Jednocześnie zauważono zróżnicowanie odpowiedzi w zależności od pytania. Najtrafniejsze pojawiały się w przypadku pytań: *How does the game look and work*?, *Is the game analog or digital?* oraz *What is the name of the game analyzed in this article?.* Najgorzej wypadły odpowiedzi na pytania: *Who are the players of the game described in this article?, How are the effects of the game measured?* oraz *What are the observed effects of the game*?

W przypadku dwóch ostatnich pytań analizę postanowiono przeprowadzić dla całej populacji wybranych artykułów (N = 134). Badacze, którzy mieli wgląd do wszystkich artykułów, identyfikowali sposoby mierzenia rezultatów gier opisanych w poszczególnych publikacjach i poddawali je krytycznej ocenie. O to samo poproszono Petal. Ostatecznie wyniki obu metod – analizy przeprowadzonej przez gen AI oraz przez człowieka – zostały porównane.

Petal wskazał, że efekty gier zostały zmierzone we wszystkich analizowanych artykułach. Jednak analiza badaczy wykazała, że w 20 spośród 134 artykułów pomiary takie w ogóle nie miały miejsca. Gen AI kwalifikowało „planowane" pomiary jako faktycznie przeprowadzone, lub wymyślało metody pomiaru, których faktycznie nie opisano w artykułach. Zaznaczmy, że komenda dla Petal brzmiała: *Were the effects of the game measured? If yes, how?*, co dawało możliwość wskazania, że efekty nie były zmierzone. Dodatkowo, w 16 z 134 artykułów Petal wskazał na przeprowadzenie pomiarów efektów, podczas gdy analiza badaczy wykazała,

[8] Niniejszy aneks odnosi się do rozdziału pt. „Przeglądy źródeł wspierane gen AI".

że były to jedynie „zamarkowane” ewaluacje, które pod względem metodologicznym nie powinny być traktowane jako pomiary. Wnioski z tego porównania są następujące: Gen AI nie sprawdziło się w sytuacjach, w których pomiary efektów w ogóle nie istniały. Algorytm zdawał się za wszelką cenę przyjmować założenie, że pomiary zostały dokonane, co prowadziło do fałszywych wyników. Ponadto gen AI zawiodło w wykrywaniu niuansów metodologicznych, gdzie wymagana była krytyczna ocena wartości zastosowanych metod ewaluacyjnych. W sytuacjach gdy brakowało rzetelnej metodologii, gen AI nie potrafiło zakwalifikować tego jako faktycznego braku pomiaru. Pozostała część wyników była spójna z analizą przeprowadzoną przez człowieka.

Aneks 5. Metodyka wyszukiwania i analizy dokumentów[9]

Wybór typu instytucji tworzących wewnętrzne uregulowania praktyk korzystania z gen AI

W trakcie przeglądu literatury akademickiej okazało się, że kwestia unormowania korzystania z generatywnej sztucznej inteligencji (gen AI) w wymiarze organizacyjnym nie doczekała się jeszcze poważnych analiz naukowych. Zważywszy na nowość zagadnienia oraz typowy cykl przygotowania i publikacji artykułów naukowych, trwający nawet kilkanaście miesięcy, nie jest to zaskakujące. Wobec tego, identyfikację dokumentów oparto na wyszukiwaniu internetowym. Takie podejście pozwoliło zidentyfikować wyłącznie te dokumenty, które są dostępne publicznie.

W celu zebrania danych, zastosowano podejście funkcjonalne, skupiając się na instytucjach i organizacjach, które pełnią kluczowe role w kształtowaniu polityk publicznych i zarządzaniu publicznym na różnych szczeblach:

1. Organizacje międzynarodowe – jako podmioty ustanawiające standardy i wytyczne w skali międzynarodowej;
2. Rządy – odpowiedzialne za tworzenie ram wykonawczych (wytycznych, rozporządzeń, zarządzeń) i polityk na poziomie krajowym;
3. Miasta, instytucje samorządowe – reprezentujące lokalny szczebel administracji, najbliższy codziennemu życiu obywateli;
4. Duże organizacje pozarządowe – często pełniące istotne funkcje publiczne i mające wpływ na kształtowanie opinii oraz praktyk w zakresie wykorzystania nowych technologii.

Procedura poszukiwania dokumentów

W kolejnych sekcjach przedstawiono szczegółową procedurę wyszukiwania dokumentów dla każdej z wymienionych wcześniej kategorii. Pierwszego wyszukiwania dokonano w maju 2024 r., dodatkowa runda została wykonana w lipcu 2024 r. Proces wyszukiwania przeprowadzono zarówno w języku polskim, jak i angielskim. Jednakże zapytania w języku polskim nie przyniosły żadnych rezultatów. W związku z tym wszystkie przedstawione poniżej zapytania sformułowano w języku angielskim. Według najlepszej wiedzy autora, w momencie powstawania tego rozdziału trwają prace nad odpowiednimi rekomendacjami dla

[9] Niniejszy aneks odnosi się do rozdziału pt. „Gen AI w organizacjach publicznych – regulacje i *codes of conducts*”.

pracowników administracji publicznej w Ministerstwie Cyfryzacji, jak również nad wytycznymi dotyczącymi odpowiedzialnego wykorzystania gen AI w urzędzie m.st. Warszawy. Jednakże, ze względu na brak publicznego dostępu do tych dokumentów, nie zostały one uwzględnione w niniejszej analizie. Sytuacja ta podkreśla dynamiczny charakter regulacji w obszarze gen AI oraz potrzebę ciągłej aktualizacji badań w tym zakresie.

Organizacje międzynarodowe

W tym przeglądzie skoncentrowano się na kilku kluczowych organizacjach międzynarodowych, w których uczestniczy Polska, w szczególności: Organizacja Narodów Zjednoczonych (ONZ), Bank Światowy, Organizacja Współpracy Gospodarczej i Rozwoju (OECD) oraz Unia Europejska.

Wyszukiwanie przeprowadzono za pomocą następujących narzędzi:

- Perplexity.AI. Zapytanie: „Does [ORGANIZATION NAME] have guidelines for staff on using generative AI?”.
- Wyszukiwarki Google i Bing. Zapytanie: „[ORGANIZATION NAME] AND 'Generative AI' AND ('ethical guidelines' OR 'ethical standards')”.

W toku wyszukiwania zidentyfikowano liczne dokumenty zawierające rekomendacje i wytyczne dla twórców polityk publicznych w zakresie regulacji sztucznej inteligencji. Choć wartościowe, wykraczają one poza zakres niniejszego przeglądu, który skupia się na wytycznych dotyczących korzystania z gen AI przez pracowników organizacji. Dokumenty odpowiadające kryteriom regulacji organizacyjnych znaleziono w przypadku trzech znaczących organizacji międzynarodowych:

- Organizacji Narodów Zjednoczonych,
- Organizacji Współpracy Gospodarczej i Rozwoju,
- Komisji Europejskiej przy Unii Eurpejskiej.

Państwa

W przypadku przeglądu wytycznych dotyczących korzystania z gen AI przez rządy, przyjęto nieco inną strategię wyszukiwania.

- Wyszukiwanie za pomocą Perplexity.AI służyło temu, by „wpaść na trop” poprzez zidentyfikowanie stron internetowych opisujących interesujące nas praktyki rządów.

Zapytanie: „Which governments have developed guidelines on using generative AI for their staff?”.

- Wyszukiwarki Google i Bing służyły temu, by znaleźć konkretne dokumenty wspomniane w tekstach zidentyfikowanych w pierwszym kroku. Zapytanie:

  „[GOVERNMENT NAME] AND 'guidelines on using generative AI'”.

Wyszukiwanie doprowadziło do odkrycia dokumentów opracowanych przez:

- Rząd Wielkiej Brytanii,
- Rząd Kanady,
- Rząd Australii,
- Urząd Stanów Zjednoczonych ds. Zarządzania Kadrami (US Office of Personnel Management).

Ponadto okazało się, że w szeregu amerykańskich stanów opracowano i opublikowano dokumenty odnoszące się do interesującej nas tematyki (niektóre z nich mają charakter dekretów – *executive orders*):

- Kalifornia,
- Kansas,
- Maine,
- New Jersey,
- Oklahoma,
- Oregon,
- Pennsylvania,
- Virginia,
- Wisconsin.

Miasta

W przypadku przeglądu wytycznych dotyczących korzystania z generatywnej sztucznej inteligencji opracowanych przez miasta, zastosowano podobną strategię jak w przypadku rządów.

- Wyszukiwanie za pomocą Perplexity.AI służyło temu, by „wpaść na trop” poprzez zidentyfikowanie różnych stron internetowych opisujących praktyki różnych miast.

Zapytanie: „Which cities have developed guidelines on using generative AI for their staff?”.

- Wyszukiwarki Google i Bing służyły temu, by znaleźć konkretne dokumenty wspomniane w tekstach zidentyfikowanych w pierwszym kroku. Zapytanie: „[CITY NAME] AND 'guidelines on using generative AI'”.

Dodatkowo przeszukano bazę danych Atlas of Urban AI, która zbiera informacje o inicjatywach miejskich w zakresie sztucznej inteligencji. Większość inicjatyw miejskich opisanych w tej bazie koncentruje się na wymiarze regulacyjnym zarządzania AI.

Ostatecznie, na liście znalazły się następujące miasta:

- Baltimore,
- Boston,
- Helsinki,
- Massachusetts,
- Nowy Jork,
- San Francisco,
- San Jose,
- Seattle,
- Tempe,
- Wiedeń.

Duże organizacje pozarządowe

Przeszukując międzynarodowe organizacje pozarządowe, które pełnią funkcje społeczne, skoncentrowaliśmy się na 15 największych (według portalu Human Rights Careers) międzynarodowych humanitarnych NGO. Wyszukiwanie przeprowadzono za pomocą wyszukiwarek Google i Bing, używając zapytania: „[NGO NAME] AND 'guidelines on using generative AI'”.

W tym przypadku wyszukiwanie nie przyniosło rezultatów. Żadna z przeszukiwanych międzynarodowych humanitarnych NGO nie udostępniła publicznie wytycznych dotyczących korzystania z gen AI przez swoich pracowników.

Metoda analizy pozyskanych źródeł

Analiza zidentyfikowanych dokumentów została przeprowadzona za pomocą oprogramowania MaxQDA 24. Jest to zaawansowane narzędzie do analizy danych jakościowych i mieszanych, które umożliwia systematyczne kodowanie (przypisywanie „etykiet" wybranym fragmentom materiału) i kategoryzowanie fragmentów tekstu. Pozwala to na odkrywanie wzorców w złożonym materiale badawczym.

Proces analizy przeprowadzono w dwóch etapach. W pierwszym skupiono się na identyfikacji kluczowych elementów w zebranych dokumentach nt. korzystania z generatywnej sztucznej inteligencji. Uwagę skoncentrowano na następujących aspektach:

- Preambuła/uzasadnienie – wstępne części dokumentów, w których zazwyczaj przedstawiono cel i kontekst tworzenia wytycznych.
- Identyfikacja szans i zagrożeń – wzmianki o potencjalnych korzyściach i ryzykach związanych z wykorzystaniem gen AI.
- Zasady przewodnie – można określić jako swego rodzaju principia etycznego i odpowiedzialnego korzystania z gen AI, sformułowane jako hasła mające przyświecać korzystaniu z gen AI.
- Wytyczne operacyjne – mniej lub bardziej szczegółowe instrukcje i procedury dotyczące praktycznego wdrażania i korzystania z gen AI w codziennej działalności organizacji.
- Praktyczne wskazówki – przedstawione wprost przykłady promptowania lub opisane egzemplifikacje stosowania gen AI w praktyce.

Z kolei w drugim etapie analizy dokonano szczegółowego przeglądu zidentyfikowanych zasad, wytycznych i wskazówek w analizowanych dokumentach. Celem było sklasyfikowanie ich według powtarzających się wątków. Wynikiem całego procesu jest synteza przedstawiona w rozdziale „Gen AI w organizacjach publicznych – regulacje i *codes of conducts*".

Publikacja powstała w ramach projektu współfinansowanego z Europejskiego Funduszu Rozwoju Regionalnego w ramach programu Fundusze Europejskie dla Nowoczesnej Gospodarki, 2021-2027.