Preferowany sposób cytowania:
Olejniczak, K., Kupiec, T. i Wojtowicz, D. (2025). Przeglądy źródeł wspierane gen AI. W: K. Olejniczak, D. Batorski i J. Pokorski (red.), Generatywna AI w badaniach. Praktyczne zastosowania w ewaluacji polityk publicznych, s. 154–184. Warszawa: Polska Agencja Rozwoju Przedsiębiorczości.

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).