

# Wykorzystanie dużych modeli językowych (LLM) w audycie jako narzędzia redukcji asymetrii informacji i ryzyka nadużyć w relacji udziałowcy – zarząd

---

MARTA GRACZYK

ORCID: 0009-0006-8745-4188

WALDEMAR MAJEK

ORCID: 0009-0001-4257-9300

PIOTR MODZELEWSKI

ORCID: 0000-0003-2817-9885

## Streszczenie

**Cel:** Celem opracowania jest zbadanie możliwości wykorzystania dużych modeli językowych (LLM) w przeprowadzaniu testów audytowych oraz ocena ich potencjału w redukcji asymetrii informacji i ryzyka nadużyć w relacji firma audytorska (zaangażowana przez udziałowców badanej spółki) – zarząd odpowiedzialny za sporządzenie badanego sprawozdania finansowego. W artykule zostały zweryfikowane hipotezy badawcze dotyczące wpływu generatywnej sztucznej inteligencji (Gen AI) na dokładność i skuteczność wykrywania nieprawidłowości finansowych w porównaniu do metod tradycyjnych.

**Metodyka / Podejście badawcze:** W artykule zastosowano triangulację metod badawczych: analizę literatury, konceptualizację problemu w oparciu o teorię agencji i koncepcję kosztów agencyjnych oraz badanie empiryczne typu symulacyjnego. Badanie polegało na przeprowadzeniu ośmiu rzeczywistych testów audytowych na próbie 167 pozycji (umów, faktur, zapisów z systemów księgowych i innych dokumentów finansowych) pochodzących ze studium przypadku w postaci międzynarodowej grupy kapitałowej. Do analizy wykorzystano pięć modeli językowych powszechnie uznawanych za kluczowe w rozwoju technologii LLM: ChatGPT, Gemini, Microsoft Copilot, Llama oraz Claude (Benaich, 2024; Maslej et al., 2025). Zastosowano dwuetapową procedurę badawczą: fazę wstępną z użyciem podstawowych zapytań oraz fazę optymalizacji z wykorzystaniem zaawansowanej inżynierii zapytań (ang. prompt engineering), dokonując szczegółowej analizy porównawczej wyników.

**Wyniki i rekomendacje:** Wyniki badania wskazują, że „surowe” modele LLM w fazie wstępnej wykazują zróżnicowaną skuteczność (od 74 do 96 proc.), przy czym ich działanie jest obciążone istotnym ryzykiem halucynacji (generowania treści poprawnych gramatycznie i logicznie, lecz niezgodnych z faktami) i błędów interpretacyjnych przy analizie złożonych klauzul prawnych. Halucynacje w tym kontekście różnią się od klasycznych błędów pierwszego i drugiego rodzaju, które występują w audycie: LLM mogą generować treści fałszywe mimo pozornej poprawności, co niekoniecznie odpowiada błędnej identyfikacji nieprawidłowości lub przeoczeniu w sensie statystycznym. Jednakże po zastosowaniu optymalizacji zapytań i doprecyzowaniu kontekstu, skuteczność modeli znacząco wzrosła, przy czym model Gemini osiągnął 100 proc. poprawności w badanej próbie. Technologia ta pozwala na szybką analizę 100 proc. populacji danych nieustrukturyzowanych, co sprzyja wykrywaniu potencjalnych nieprawidłowości niewidocznych przy tradycyjnym próbkowaniu.

**Ograniczenia / Implikacje badawcze:** Badanie przeprowadzono na zanonimizowanych danych jednej grupy kapitałowej z branży AGD, co może ograniczać uniwersalność wniosków dla specyficznych sektorów (np. finansowego). Główne ograniczenia technologii to limity wielkości plików, kwestie poufności danych oraz ryzyko nadmiernego zaufania do wyników generowanych przez AI.

**Oryginalność / wartość:** Opracowanie wypełnia lukę badawczą w zakresie praktycznego zastosowania GenAI w konkretnych procedurach audytowych na danych rzeczywistych. Prezentuje dowody empiryczne na skuteczność inżynierii zapytań w eliminowaniu błędów modeli językowych w audycie, stanowiąc wkład w dyskusję nad przyszłością zawodu biegłego rewidenta.

**Słowa kluczowe:** duże modele językowe (LLM), audyt, asymetria informacji, ryzyko nadużyć, generatywna sztuczna inteligencja (Gen AI), teoria agencji.

## Wprowadzenie

Współczesny rynek usług finansowych i kapitałowych charakteryzuje się rosnącą złożonością relacji między interesariuszami oraz dynamicznym przyrostem danych generowanych przez podmioty gospodarcze. W gospodarce opartej na danych, podstawą zaufania udziałowców i inwestorów do zarządów spółek jest rzetelność i wiarygodność prezentowanych sprawozdań finansowych. Gospodarka oparta na danych stanowi rozwinięcie koncepcji gospodarki opartej na wiedzy, w której kluczowym zasobem stają się nie tylko wiedza i kapitał intelektualny, lecz również dane jako podstawowy czynnik tworzenia wartości (OECD, 1996; European Commission, 2020). Jednakże, naturalnie występująca asymetria informacji między przedsiębiorstwem, posiadającym pełną wiedzę o swojej kondycji, a interesariuszami zewnętrznymi, stwarza strukturalne warunki do występowania błędnych decyzji alokacyjnych oraz nadużyć finansowych.

Jednym z mechanizmów systemowych, mającym na celu redukcję tej asymetrii i uwiarygodnienie danych, jest niezależny audyt finansowy. Mimo to, tradycyjne metody audytu, oparte w dużej mierze na manualnym przeglądzie dokumentacji oraz wnioskowaniu na podstawie ograniczonej próby badawczej, w obliczu ery Big Data stają się coraz mniej skuteczne i efektywne. Są one czasochłonne, podatne na błędy ludzkie (zmęczenie, przeoczenie) oraz subiektywizm osądu. Dodatkowo, presja rynkowa na szybką realizację badania i redukcję jego kosztów może skutkować obniżeniem jakości usług atestacyjnych, co negatywnie wpływa na zdolność wykrywania oszustw księgowych (Knechel et al., 2013).

Ostatnie lata przyniosły dynamiczny rozwój generatywnej sztucznej inteligencji (GenAI), a w szczególności dużych modeli językowych (LLM – Large Language Models). Modele te, trenowane na bardzo dużych zbiorach danych tekstowych, posiadają zdolność przetwarzania języka naturalnego, rozumienia kontekstu, analizy sentymentu oraz wyciągania wniosków logicznych. W odniesieniu do audytu, LLM mogą dokonać rewolucji w zakresie analizy danych nieustrukturyzowanych – umów, faktur, notatek z posiedzeń zarządu, czy korespondencji mailowej – które stanowią znaczną część dowodów badania, a dotychczas trudno było zautomatyzować ich analizę.

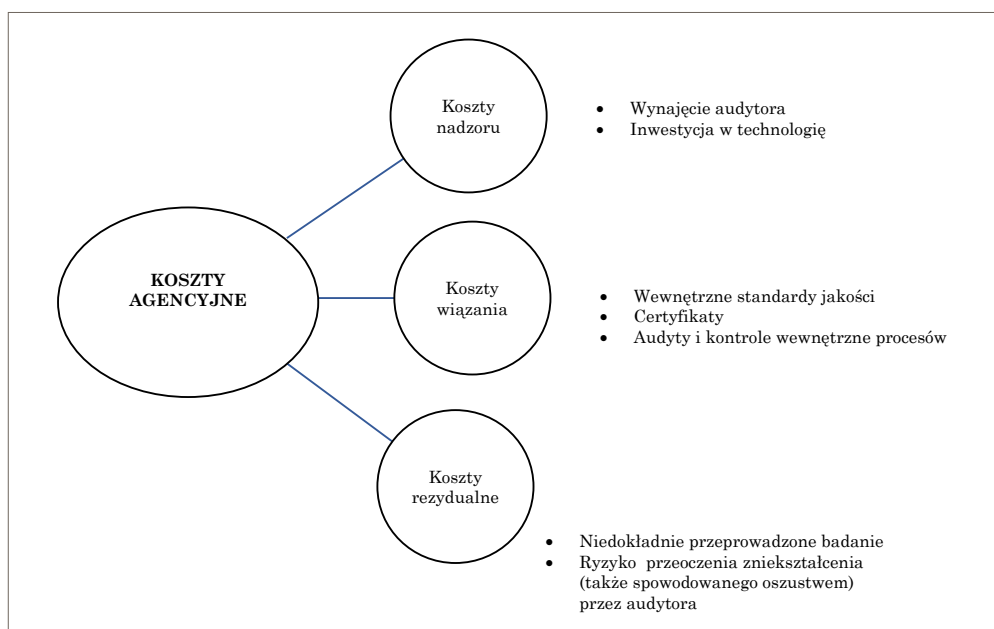
W oparciu o badanie empiryczne symulujące pracę audytora, w niniejszym artykule podjęto próbę odpowiedzi na pytanie, czy GenAI może stać się skutecznym narzędziem wspierającym biegłego rewidenta, zwiększając precyzję i zakres badania oraz finalnie redukując asymetrię informacji między zarządem, a udziałowcami.

## 1. Teoretyczne podstawy badania: Teoria agencji i ryzyko nadużyć

Teoria agencji, której fundamenty stworzyli m.in. Jensen i Meckling, opisuje relacje, w których jedna strona (pryncypał) powierza drugiej (agentowi) wykonanie zadania, przekazując jej część uprawnień decyzyjnych. Na rynku audytorskim występuje specyficzna, trójstronna konfiguracja tej relacji. Udziałowcy (pryncypał) powierzają kapitał zarządowi (agentowi), angażując firmę audytorską mającą weryfikować jego działania (Jensen & Meckling, 1976).

Kluczowym problemem w tej strukturze jest fakt, że firma audytorska nie współpracuje bezpośrednio z udziałowcami, lecz z zarządem spółki, co rodzi konflikt interesów i ryzyko działania na niekorzyść inwestorów w celu utrzymania klienta (Schaefer, 2023). Prowadzi to do powstawania kosztów agencyjnych, które obejmują koszty nadzoru (np. wynagrodzenie audytora), koszty wiązania (budowanie reputacji) oraz koszty rezydualne wynikające z rozbieżności interesów (Watts et al., 1986).

Rys. 1. Struktura kosztów agencyjnych w relacji udziałowcy – firma audytorska



Źródło: Opracowanie na podstawie: Jensen, 1976; Damodaran, 2010; Solomon, 2020.

W ramach teorii agencji identyfikuje się dwa główne problemy wpływające na jakość audytu:

- 1) Selekcja negatywna (adverse selection): sytuacja, w której audytorzy nie posiadają pełnej wiedzy o dokumentach dostarczonych przez jednostkę badaną, co obniża jakość weryfikacji (Jensen & Meckling, 1976; Eisenhardt, 1989).
- 2) Pokusa nadużycia (ang. moral hazard): zjawisko polegające na przedkładaniu przez audytora własnych interesów nad interesy udziałowców, np. poprzez nieujawnianie wykrytych nieprawidłowości w celu zachowania kontraktu (Jensen & Meckling, 1976; Eisenhardt, 1989).

Mimo istnienia regulacji prawnych i standardów etycznych, rynek audytorski charakteryzuje się systemowymi źródłami konfliktów interesów. Do najważniejszych należą:

- finansowa zależność od zarządu spółki (audytor jest opłacany przez podmiot, w którym przeprowadza badanie),
- świadczenie usług doradczych równoległe z audytem,
- presja zarządu na akceptację agresywnej księgowości

oraz

- zjawisko „familiarity threat”, czyli zbyt bliskich relacji z klientem (International Ethics Standards Board for Accountants, 2018; DeAngelo, 1981; Frankel, Johnson & Nelson, 2002).

Historia gospodarcza dostarcza dowodów na negatywną materializację tych ryzyk. Upadek Enronu i firmy audytorskiej Arthur Andersen uwypuklił problem braku niezależności spowodowany świadczeniem usług doradczych i ukrywaniem kreatywnej księgowości. Podobne mechanizmy zadziałały w przypadku WorldCom oraz niemieckiej spółki Wirecard, gdzie audytorzy przez lata ignorowali sygnały o manipulacjach finansowych. Na gruncie polskim przykładem jest sprawa GetBack, gdzie podważono skuteczność działań audytora w zakresie weryfikacji sprawozdania finansowego spółki (McLean & Elkind, 2003; Coates, 2007; Pulliam & Solomon, 2002; McCrum & Jones, 2020; Business Insider Polska, 2020).

Podstawowym celem audytu jest dostarczenie niezależnej opinii o sprawozdaniu finansowym, co ma na celu uwiarygodnienie danych i redukcję luki informacyjnej między zarządem a udziałowcami. Mechanizmy takie jak niezależność weryfikacji, profesjonalny sceptycyzm oraz zgodność ze standardami mają gwarantować jakość badania (Watts, 1977).

Jednakże, tradycyjny audyt posiada istotne ograniczenia, które uniemożliwiają całkowitą eliminację ryzyka przeoczenia zniekształceń oraz ryzyka nadużyć.

Do głównych ograniczeń należą:

- testowy charakter badania (brak weryfikacji 100 proc. transakcji);
- opóźnienie informacji w stosunku do zdarzeń gospodarczych;
- ograniczona pewność;
- subiektywność ocen audytora;
- inherentne ryzyko niewykrycia oszustw.

Ponadto, zjawisko pokusy nadużycia może prowadzić do nadmiernej ufności udziałowców w opinię audytora, co paradoksalnie usypia ich czujność (Dopierała, 2012; DeFond, 2010). Przejrzystość badania jest kluczową wartością dla inwestorów, pozwalającą na dokładniejsze szacowanie ryzyka i egzekwowanie odpowiedzialności zarządu. Wobec wyzwań związanych z tradycyjnymi metodami testowania, rośnie zapotrzebowanie na wdrożenie nowoczesnych narzędzi, takich jak sztuczna inteligencja.

Analiza literatury i historycznych przypadków nadużyć wskazuje, że tradycyjne metody audytowe w kontekście teorii agencji bywają zawodne i podatne na wpływ czynnika ludzkiego oraz presję biznesową. Istnieje zatem pilna potrzeba implementacji rozwiązań, które zautomatyzują proces analizy dokumentacji i zwiększą niezależność badania. W odpowiedzi na te wyzwania sformułowano hipotezy badawcze dotyczące wpływu generatywnej sztucznej inteligencji (Gen AI) na dokładność i skuteczność wykrywania nieprawidłowości finansowych w porównaniu do metod tradycyjnych.

## 2. Charakterystyka i potencjał dużych modeli językowych (LLM)

Dynamiczny rozwój technologii sztucznej inteligencji, a w szczególności dużych modeli językowych (LLM – ang. Large Language Models), rewolucjonizuje metodykę pracy w sektorach opartych na analizie danych, w tym w audycie finansowym. W odróżnieniu od tradycyjnych metod analitycznych, które koncentrują się głównie na danych ustrukturyzowanych (liczbowych), LLM oferują możliwość efektywnego przetwarzania ogromnych zbiorów danych nieustrukturyzowanych, takich jak umowy, korespondencja czy raporty narracyjne. Niniejsza sekcja przybliży charakterystykę tej technologii oraz jej zastosowanie w wykrywaniu anomalii i redukcji asymetrii informacji (Becker et al., 2024).

Duże modele językowe to zaawansowane systemy oparte na sieciach neuronowych, trenowane na wielkich zbiorach danych tekstowych, zdolne do generowania i analizy języka naturalnego na poziomie zbliżonym do ludzkiego. Fundamentem ich działania jest architektura uczenia głębokiego opracowanego przez Google (architektura Transformer) oraz zdolność rozpoznawania kontekstu dzięki mechanizmowi uwagi (ang. attention mechanism) – oba mechanizmy zostały zaprezentowane przez Vaswaniego i współautorów w 2017 roku w publikacji „Uwaga to wszystko, czego potrzebujesz” (ang. „Attention is all you need”). Mechanizm uwagi pozwala modelowi na identyfikację zależności między słowami w zdaniu nawet w rozbudowanych fragmentach tekstu, co umożliwia głębokie zrozumienie kontekstu, niuansów językowych oraz relacji przyczynowo-skutkowych (Vaswani et al., 2017).

Proces uczenia LLM przebiega dwuetapowo:

- 1) faza uczenia wstępnego (ang. pre-training), w której model nabywa ogólną wiedzę językową poprzez naukę na olbrzymich zbiorach danych tekstowych;
- 2) faza dostrajania (ang. fine-tuning) adaptująca model do specyficznych zadań (Raffel et al., 2020).

Istotną cechą nowoczesnych modeli jest zdolność do tzw. few-shot learning, czyli wykonywania nowych zadań na podstawie zaledwie kilku przykładów zawartych w zapytaniu (ang. prompt), co znacząco obniża bariery wdrożenia tej technologii w specyficznych procedurach audytowych (Brown et al., 2020). Mimo imponujących możliwości, technologia ta obarczona jest ryzykiem tzw. halucynacji – generowania treści poprawnych gramatycznie i logicznie, lecz niezgodnych z faktami (Ji et al., 2023). Wymaga to od audytorów zachowania krytycznego podejścia i weryfikacji wyników generowanych przez AI (Mökander i Floridi, 2023).

Z punktu widzenia audytu, niezwykle istotna jest umiejętność analizy danych tekstowych (często w kilku językach) w formie nieustrukturyzowanej, takich jak:

umowy, korespondencja mailowa, raporty zdarzeń czy protokoły i notatki ze spotkań. Dotychczasowe systemy analityczne pozwalały głównie na pomoc w analizie danych liczbowych, pomijając co do zasady dane tekstowe. Modele LLM rozszerzają te możliwości, dzięki czemu mogą przyczynić się do zwiększenia dokładności badania audytowego (Becker et al., 2024). Współczesny audyt, oprócz analizy danych finansowych, w znacznym stopniu wymaga również analizy rosnącej liczby dokumentów tekstowych. LLM dzięki zdolności do analizy sentymentu (umiejętność polegająca na rozumieniu intencji i tonu wypowiedzi), mogą zautomatyzować i uszczelnić ten proces w obszarach, które dotychczas wymagały pracochłonnej weryfikacji manualnej.

Do kluczowych funkcjonalności wspierających audyt w analizie jakościowej należą:

- 1) Ekstrakcja informacji – automatyczne wydobywanie kluczowych danych (daty, kwoty, klauzule kar, podpisy) z dokumentów o zróżnicowanej strukturze, takich jak faktury czy umowy. Badania wskazują, że modele LLM potrafią przetwarzać wielojęzyczną dokumentację znacznie szybciej i z wyższą precyzją niż zespoły ludzkie (Dennis, 2024);
- 2) Klasyfikacja dokumentów – przypisywanie dokumentów źródłowych do odpowiednich kategorii badawczych, co usprawnia organizację dowodów badania (Becker et al., 2024);
- 3) Analiza sentymentu – ocena tonu wypowiedzi w korespondencji wewnętrznej lub notatkach zarządu. Wykrycie negatywnego nacechowania emocjonalnego, niepewności lub presji w komunikacji może stanowić sygnał ostrzegawczy (ang. red flag) wskazujący na ryzyko nadużyć lub problemów z kontynuacją działalności (Srivastava et al., 2023);
- 4) Inżynieria zapytań (promptowanie, ang. prompt engineering) – możliwość tworzenia specjalnych instrukcji (promptów) dla konkretnych testów audytowych, np. weryfikacji zgodności faktur z zamówieniami, co pozwala na standaryzację procedur (Mökander et al., 2023).

Zastosowanie tych narzędzi pozwala na przejście od badania wyrywkowego (próbki) do analizy pełnej populacji dokumentów, co bezpośrednio wpływa na redukcję ryzyka przeoczenia istotnych zniekształceń.

Jednym z najważniejszych zastosowań LLM w audycie jest identyfikacja anomalii, rozumianych nie tylko jako zniekształcenia liczbowe, ale przede wszystkim jako niespójności semantyczne i narracyjne. Tradycyjne metody analityczne często pomijają ten aspekt. Modele językowe są w stanie wykryć subtelne sygnały świadczące o manipulacji, takie jak:

- sprzeczności między danymi finansowymi a częścią opisową raportu rocznego (tzw. oszustwa narracyjne);
- nagle zmiany stylu lub tonu wypowiedzi w raportach zarządu;
- nietypowe klauzule w umowach lub brak standardowych zabezpieczeń prawnych;
- pojawienie się słów kluczowych sugerujących ukrywanie informacji, np. „niezidentyfikowane”, „utajnione” (Crawford, 2020).

Badania Rahamana i in. (2024) dowodzą, że modele LLM potrafią zidentyfikować niespójności w narracji korporacyjnej z większą dokładnością i wyprzedzeniem niż analitycy posługujący się tradycyjnymi metodami. Poniższa tabela przedstawia porównanie skuteczności różnych metod w wykrywaniu anomalii.

**Tabela 1. Porównanie skuteczności wykrywania anomalii finansowych**

Metoda wykrywania	Typ anomalii	Skuteczność szacowana <sup>1</sup>	Uwagi
przegląd manualny	sprzeczne stwierdzenia	niska – zależna od osoby	czasochłonna, subiektywna
tradycyjne narzędzia <sup>2</sup>	frazy kluczowe	średnia	ograniczona do prostych schematów
modele LLM	zmiana tonu, niespójność narracji	wysoka (70-90%)	wymaga nadzoru człowieka

Źródło: Opracowanie na podstawie: Becker et al., 2024; Mökander et al., 2023.

Warto podkreślić, iż wykrywanie anomalii i nadużyć będzie najskuteczniejsze w połączeniu z tradycyjnymi metodami audytorskimi. Połączenie tych dwóch podejść może prowadzić do kompleksowego oraz dokładnego zobrazowania potencjalnych nieprawidłowości.

Podsumowując, integracja dużych modeli językowych z warsztatem biegłego rewidenta stwarza szansę na znaczące podniesienie jakości usług atestacyjnych. Zdolność do szybkiej analizy kontekstowej i wykrywania anomalii w danych nieustrukturyzowanych wypełnia luki, które pozostawiają metody tradycyjne. Weryfikacja tej tezy wymaga jednak dowodów empirycznych.

<sup>1</sup> Szacunki oparte na literaturze: Becker et al., 2024; Mökander et al., 2023.

<sup>2</sup> Tradycyjne narzędzia analizy danych, służące do wyszukiwania i filtrowania fragmentów tekstu oraz dopasowywania wzorców, takie jak Microsoft Excel i CaseWare IDEA.

### **3. Model agencyjny na linii udziałowcy – zarząd spółki: badanie empiryczne**

Aby zweryfikować potencjał LLM w praktyce audytowej, zaprojektowano i przeprowadzono badanie symulacyjne na rzeczywistych danych. Dane wykorzystane w badaniu empirycznym do tej pracy zostały pozyskane od spółki działającej w branży AGD. Spółka ta, w roku 2024, przeszła proces akwizycji, stając się częścią większej korporacji. Obecnie firma prowadzi działalność w ponad 100 krajach, posiada 18 zakładów produkcyjnych, w tym 11 w Europie oraz jest jednym z liderów europejskiego rynku zaopatrującego gospodarstwa domowe w sprzęt AGD. Przedsiębiorstwo posiada rozbudowaną strukturę organizacyjną w obszarze EMEA (Europa, Bliski Wschód oraz Afryka), w tym liczne lokalne jednostki, które są odpowiedzialne za sprzedaż, logistykę, marketing, obsługę klienta oraz serwis urządzeń. Działalność spółki charakteryzuje się złożoną strukturą organizacyjną oraz rozbudowanymi procesami biznesowymi i finansowo-księgowymi, które każdego roku generują ogromne ilości danych finansowych, jak i niefinansowych. Ze względu na spełnianie kryteriów określonych w ustawie o rachunkowości, dotyczących wielkości spółki mierzonej poziomem zatrudnienia oraz osiąganych przychodów, spółka podlega obowiązkowemu corocznemu badaniu sprawozdania finansowego.

Współpraca z firmą audytorską obejmuje coroczne badanie sprawozdania finansowego oraz cykliczne testy audytowe, przeprowadzane w różnych krajach, w ramach zgodności z regulacjami SOX – Sarbanes-Oxley Act 2002. Na początku warto podkreślić, iż wszelkie dokumenty użyte w testach audytowych zostały w pełni zanonimizowane, nie ujawniając tym samym wrażliwych informacji finansowych przedsiębiorstwa oraz danych kontrahentów, które posłużyły za źródło do badania empirycznego. Dane do testów audytowych zostały wyselekcjonowane z archiwum badań przeprowadzonych w regionie EMEA. W ramach tego regionu, każdy kraj reprezentowany jest przez odrębną jednostkę, co zapewnia zróżnicowanie badanej populacji. Zestaw danych został przygotowany wyłącznie w celach badawczych, przy zachowaniu zasad bezpieczeństwa informacji i poufności.

W celu uzyskania kompleksowej oceny możliwości dużych modeli językowych, do próby badawczej wybrano dokumentację testową w taki sposób, aby pokryła kluczowe obszary działalności spółki oraz zawierała typowe audytowe procedury weryfikacyjne. Jednocześnie próbę świadomie zróżnicowano, aby zawierała różnorodne dane zarówno w formie jak i treści. Analizie poddano dokumenty pochodzące z 8 różnych testów audytowych z 4 różnorodnych audytów finansowych oraz z 4 kontroli SOX, które różnią się między sobą:

- rodzajem oraz zakresem tematycznym – przetestowana próba zawiera różnorodną dokumentację w postaci umów, faktur, danych HR, polityk wewnętrznych;
- źródłowym językiem dokumentów – wybrana próba zawiera dokumentację w 4 różnych językach;

- krajem pochodzenia – dobrana do próby dokumentacja pochodzi z 5 różnych krajów;
- rodzajem audytu – uzyskana dokumentacja pochodzi z kontroli SOX, audytów finansowych, audytu zgodności oraz audytu antykorupcyjnego;
- formatem danych, próba jest zróżnicowana również pod kątem rodzaju dokumentów i zawiera PDF, e-maile, faktury, formularze, umowy, screeny z systemu księgowego oraz tabelaryczne zestawienia.

Ta świadomie zróżnicowana próba ma na celu zbadanie, jak modele LLM radzą sobie z różnymi zadaniami audytowymi, językami oraz rodzajami dokumentów. W celu zapewnienia reprezentatywności próby oraz możliwości uogólnienia wyników, każdy test audytowy został wybrany z odrębnego badania, przeprowadzonego w kraju regionu EMEA, z wyjątkiem jednego badania zawierającego audyt dwóch państw jednocześnie (wybrano tutaj dwa testy, a każdy z nich dotyczył innego kraju). Taka metodologia pozwala na uwzględnienie różnic kulturowych, regulacyjnych i operacyjnych, występujących w poszczególnych krajach. Badanie empiryczne zawiera testy audytowe, których szczegółową charakterystykę opisuje Tabela nr 2.

Tabela 2. Charakterystyka testów audytowych wybranych do badania

Numer testu i zakres	Audytowany obszar i rok	Języki dokumentacji	Rodzaj dokumentów/ danych	Cel testu
Test 1 – weryfikacja próby kontraktów	2023 EMEA – Niemcy i Włochy: sprzedaż bezpośrednia do konsumenta	angielski, włoski	umowy sprzedaży, klauzule prawne	weryfikacja czy umowa została zawarta i była ważna w badanym okresie
Test 2 – weryfikacja próby dokumentów przewozowych	2021 EMEA – Jednostka sprzedażowa w Wielkiej Brytanii	angielski	listy przewozowe, potwierdzenia dostawy, daty dostawy	poprawność rozpoznania przychodu
Test 3 – weryfikacja próby faktur kosztowych	SOX 2023 – EMEA: Kontrola inwentaryzacyjna RTP (Inv5) w Centrum Usług Wspólnych w Łodzi	angielski	faktury od dostawców, opisy usług lub towarów, kwoty, daty	weryfikuje czy dla każdej faktury o wartości powyżej 10 000 euro został przeprowadzony proces ofertowania, bądź istnieje ważny kontrakt z dostawcą

Numer testu i zakres	Audytowany obszar i rok	Języki dokumentacji	Rodzaj dokumentów/danych	Cel testu
Test 4 – weryfikacja próby faktur korygujących	2023 EMEA – Niemcy i Włochy: sprzedaż bezpośrednia do konsumenta	angielski, niemiecki	noty korygujące, faktury pierwotne, przyczyny korekt	analiza przyczyn i poprawności wystawionych korekt, identyfikacja potencjalnych nieprawidłowości w rozliczeniach
Test 5 – weryfikacja akceptacji finansowych zgodnie z procedurą akceptacji	2023 EMEA – System zachęt dla partnerów handlowych	angielski	dane z systemu, workflow akceptacji, matryca akceptacji, kwoty wypłat, daty akceptacji i księgowania	ocena zgodności procesu akceptacji wypłat z obowiązującą procedurą, weryfikacja uprawnień osób akceptujących
Test 6 -weryfikacja próbki zmian limitów kredytowych oraz terminów płatności klientów	SOX 2023 – EMEA: Kontrola rozliczenia należności OTC (AR4) we Włoszech	angielski, włoski	wnioski o zmianę limitu lub terminu, zatwierdzenia, dane klientów	ocena zasadności i prawidłowości zmian limitów kredytowych oraz terminów płatności, ocena zgodności z polityką kredytową spółki
Test 7 -weryfikacja próbki poprawności założenia nowego klienta	SOX 2023 – EMEA: Kontrola rozliczenia należności OTC (AR1) we Włoszech	angielski, włoski	formularze założenia klienta, dokumenty rejestracyjne, dane w systemie CRM	ocena kompletności i poprawności procesu założenia profilu nowego klienta, weryfikacja wymaganych dokumentów i informacji
Test 8 – weryfikacja próbki zmian płac dla pracowników	2023 EMEA – Audyt przeciwdziałania korupcji i łapownictwu	niemiecki, francuski	umowy o pracę, aneksy, dokumenty potwierdzające zmiany płac, daty obowiązywania zmian	ocena prawidłowości udokumentowania i wdrożenia zmian w wynagrodzeniach pracowników, ocena zgodności z umowami i regulacjami wewnętrznymi

Źródło: Opracowanie własne na podstawie zebranych danych empirycznych.

Wybór powyższych testów podyktowany był chęcią reprezentacji kluczowych obszarów audytu, skupiono się na najbardziej czasochłonnych zadaniach (długie umowy, zrzuty ekranów) i odrzucono duże arkusze kalkulacyjne, z którymi obecne modele GenAI sobie nie radzą. Celowo dobrano próbę tak, aby przetestować dokumenty związane zarówno ze sprzedażą, jak i kosztami spółki.

Do analizy porównawczej zawartej w badaniu empirycznym wykorzystano pięć dużych modeli językowych w płatnych wersjach, reprezentujących czołowych dostawców technologii sztucznej inteligencji:

- Chat GPT od OpenAI – oparty na architekturze GPT-4 jeden z pionierskich modeli LLM charakteryzuje się wysoką zdolnością rozumienia i przetwarzania języka naturalnego. Model oferuje odpowiedzi na złożone pytania, interpretację dokumentów i wyszukiwania anomalii logicznych.
- Gemini od Google – znany z wysokiej jakości analizy językowej i interpretacyjnej, dobrze radzi sobie z interpretacją i analizą informacji zawartych w tekście, obrazie oraz kodzie.
- Microsoft Copilot – narzędzie ściśle zintegrowane ze środowiskiem Microsoftu i wykorzystujące elementy czatu GPT w swojej architekturze. Oferuje zaawansowane możliwości analizy tekstu i danych również z dokumentów Word.
- Claude od Anthropic – model stworzony z myślą o bezpieczeństwie danych, znany ze zdolności do prowadzenia długich i złożonych konwersacji oraz generowania spójnych tekstów.
- Llama od Mety to model wykorzystywany do analiz porównawczych również w mniej popularnych na świecie językach. Ze względu na fakt, że spółka Meta nie udostępniła jeszcze modelu Llama na rynku polskim, w testowaniu wykorzystano platformę Abacus AI, za pomocą której można pracować na tym modelu.

Tak zróżnicowany zestaw testów i narzędzi pozwolił na rzetelną ocenę potencjału wykorzystania dużych modeli językowych w testach audytowych. Dzięki różnorodności formatów, języków, typów danych oraz krajów pochodzenia dokumentacji, przeprowadzona w tej pracy analiza stanowi solidną podstawę do wyciągania wniosków i możliwości uogólniania wyników. Wnioski z niej wypływające pomagają w lepszym zrozumieniu potencjału i ograniczeń technologii GenAI w sektorze audytu finansowego oraz dostarczają praktycznych wskazówek na temat efektywnego wykorzystania narzędzi AI w przyszłości.

Badanie empiryczne przeprowadzono w dwóch fazach:

- 1) Faza wstępna: modele otrzymały podstawowe zapytania (prompty) opisujące cel testu i kroki testowe, bez dodatkowych wyjaśnień kontekstowych;

- 2) Faza optymalizacji (prompt engineering): na podstawie błędów z fazy pierwszej, opracowano udoskonalone prompty, zawierające precyzyjne instrukcje interpretacyjne i ponownie przeprowadzono testy.

Wyniki generowane przez AI (status „zaliczone”/„niezaliczone” wraz z uzasadnieniem) porównywano z wynikami uzyskanymi tradycyjnie przez zespół audytowy, przyjmując te drugie za wzorzec poprawności (ang. ground truth).

## 4. Wyniki badań – Faza wstępna: analiza błędów i ograniczeń

W pierwszej fazie testowania modele zostały skonfrontowane z „surowym” zadaniem audytowym. Wyniki tej fazy ujawniły zarówno duży potencjał technologii, jak i istotne ryzyka związane z jej bezkrytycznym stosowaniem.

**Tabela 3. Podsumowanie wyników analizy testów audytowych**

Model	Testy zadane	Testy odrzucone <sup>3</sup>	Testy wykonane	Poprawne odpowiedzi	Stosunek odpowiedzi poprawnych do wszystkich zadanych testów (w proc.)	Stosunek odpowiedzi poprawnych do wykonanych testów (w proc.)
Chat GPT	167	10	157	123	74	78
Gemini	167	-	167	161	96	96
Microsoft Copilot	167	6	161	153	92	95
Llama	167	40	127	124	74	98
Claude	167	10	157	153	92	97

Wstępna analiza wykazała znaczące różnice w skuteczności i możliwościach technicznych poszczególnych modeli. Zestawienie wyników dla wszystkich 167 próbek prezentuje się następująco:

- Gemini – model ten wykazał najwyższą skuteczność techniczną i merytoryczną. Przetworzył 100 proc. zadanych próbek (167/167) i udzielił poprawnej odpowiedzi w 161 przypadkach, co daje skuteczność na poziomie 96 proc. Nie odrzucił żadnego pliku z powodu formatu czy wielkości.
- Microsoft Copilot – osiągnął wysoką skuteczność merytoryczną (92 proc.), poprawnie rozwiązując 153 próbki. Napotkał jednak bariery techniczne – odrzucił 6 próbek w teście pierwszym z powodu przekroczenia limitu wielkości pliku (1 MB), co w praktyce audytowej jest istotnym ograniczeniem, ponieważ używane pliki mają zazwyczaj większy rozmiar.
- Claude – podobnie jak Copilot, osiągnął 92 proc. skuteczności merytorycznej (153 poprawne odpowiedzi), ale odrzucił 10 próbek z przyczyn technicznych.
- Llama – model ten napotkał największe trudności z analizą plików w formacie .xls odrzucając łącznie 40 próbek. Jednak w przypadku dokumentów

<sup>3</sup> Liczba próbek odrzuconych ze względów technicznych.

tekstowych w formacie .pdf, które był w stanie przetworzyć, jego skuteczność była imponująca – 98 proc.

- Chat GPT – uzyskał najniższy wynik ogólny – 74 proc. poprawności (123 poprawne odpowiedzi). Model ten najczęściej popełniał błędy merytoryczne i interpretacyjne.

Analiza przyczyn błędów w fazie wstępnej pozwoliła zidentyfikować kluczowe słabości „surowych” modeli LLM w zastosowaniach audytowych. Najwięcej problemów sprawił test pierwszy (weryfikacja kontraktów), który wymagał analizy prawnej. Do kluczowych słabości można zaliczyć:

### 1) Problem interpretacji klauzul

W teście pierwszym, modele miały problem z interpretacją klauzul o automatycznym odnowieniu umowy.

- Gemini w trzech przypadkach uznało test za negatywny, argumentując, że brak jest dowodu na to, że umowa nie została wypowiedziana, mimo że sama umowa zawierała klauzulę o automatycznym przedłużeniu. Jest to przykład nadmiernej ostrożności, która generuje fałszywe alarmy (wynik fałszywie dodatni w kontekście wykrycia błędu).
- Chat GPT w pięciu przypadkach popełnił błędy. W próbie 24 uznał umowę za ważną („wynik zaliczony”), mimo że nie pokrywała ona całego badanego okresu. Model halucynował, uznając, że częściowe pokrycie okresu jest wystarczające, co jest fundamentalnym błędem w audycie.
- Claude również w 3 przypadkach błędnie ocenił ważność umów, które zakładały ich automatyczne odnowienie.

### 2) Problem dowodów alternatywnych:

W teście drugim (dotyczącym dokumentów przewozowych) zadaniem było potwierdzenie dostawy. Część dokumentów nie miała podpisu klienta (co mogłoby zostać uznane za błąd), ale audytorzy dysponowali alternatywnym dowodem w postaci zrzutu ekranu z systemu logistycznego klienta. Modele w fazie wstępnej często ignorowały te alternatywne dowody, sztywno trzymając się instrukcji o „podpisie na dokumencie”, co skutkowało błędną oceną „niezaliczone” dla poprawnych transakcji.

### 3) Ryzyko wyniku fałszywie dodatniego (błąd I rodzaju):

Najbardziej niebezpiecznym zjawiskiem zaobserwowanym w fazie wstępnej było uznawanie przez modele próbek wadliwych za poprawne. Chat GPT w całej próbie uznał 19 próbek niepoprawnych za poprawne. W audycie oznacza to ryzyko niewykrycia istotnego zniekształcenia (ryzyko przeoczenia), co jest znacznie groźniejsze niż fałszywy alarm wymagający dodatkowej weryfikacji.

Powyższa analiza wyników testów prowadzi do wniosku, że wykorzystanie modeli LLM w audycie wymaga nadzoru człowieka i walidacji przeprowadzonych wyników oraz, że modele nie mogą w pełni zastąpić profesjonalnego osądu doświadczonego audytora. Konieczna była też kolejna faza testowania – optymalizacja.

## 5. Faza optymalizacji i wyniki końcowe – rola inżynierii zapytań

Na podstawie analizy błędów z fazy pierwszej, przeprowadzono drugą część badania empirycznego, wykorzystując techniki inżynierii zapytań (prompt engineering). Opracowano zestaw doprecyzowanych instrukcji, które miały na celu wyeliminowanie niejednoznaczności i narzucenie modelom logiki audytorskiej.

**Tabela 4. Zestawienie sformułowań służące optymalizacji zapytań audytowych**

Numer	Przyczyna	Propozycja optymalizacji
1	Każde z narzędzi wskazało inne próbki jako negatywne.	Wyjaśnij dla wyników negatywnych tok rozumowania.
2	Chat GPT oraz Claude dla umów z automatycznym odnowieniem uznały, że nie mamy pewności czy umowa nie została zerwana wcześniej.	Dla umów z automatycznym przedłużeniem uznaj, że umowy są ważne.
3	Chat GPT oraz Llama uznały, że umowa przedłużona mieści się w okresie badanym gdzie w praktyce nie obejmowała go całego.	Umowa musi pokrywać cały okres badany.
4	Chat GPT i Llama uznały podpis dostawcy za podpis klienta.	Podpis klienta znajduje się w dokumencie w pozycji 24.
5	Gemini nie uznało potwierdzenia z systemu wewnętrznego klienta za wystarczający dowód.	Jeśli mamy potwierdzenie z systemu wewnętrznego klienta i zawiera ono numer i datę dostawy zgodne z danymi w dokumentach spółki uznaj, że jest to prawidłowe.
6	Copilot nie uznał korekty sprzedaży na podstawie zrzutów ekranu z SAP za wystarczający dowód.	Jeśli mamy potwierdzenie korekty sprzedaży na zrzucie ekranu z SAP, które dotyczy tego samego albo następującego miesiąca uznaj, że jest to prawidłowe.
7	Chat GPT, Gemini oraz Llama dla próbki, która dotyczyła postępowania sądowego uznały, że test jest niezaliczony.	W przypadku kosztów sądowych uznaj test jako niedotyczący (N/A).

Źródło: Opracowanie własne na podstawie: wyników testów modeli LLM.

Zastosowanie zoptymalizowanych zapytań przyniosło skokową poprawę wyników, szczególnie w testach, które wcześniej sprawiały trudności.

**Tabela 5. Podsumowanie wyników analizy testów audytowych po optymalizacji zapytań<sup>4</sup>**

Model	Testy zadane	Testy odrzucone <sup>5</sup>	Testy wykonane	Poprawne odpowiedzi	Stosunek odpowiedzi poprawnych do wszystkich zadanych testów (w proc.)	Stosunek odpowiedzi poprawnych do wykonanych testów (w proc.)
Chat GPT	167	10	157	127	76	81
Gemini	167	-	167	167	100	100
Microsoft Copilot	167	6	161	160	96	99
Llama	167	40	127	127	76	100
Claude	167	10	157	157	94	100

Źródło: Opracowanie własne na podstawie zebranych danych empirycznych

Oto, jak w fazie optymalizacji radziły sobie poszczególne modele:

- Gemini – pełna skuteczność. Po optymalizacji model Gemini osiągnął 100 proc. skuteczności. Poprawnie zweryfikował wszystkie 167 próbek we wszystkich 8 testach. Wyeliminowano wszystkie błędy interpretacyjne dotyczące umów i dowodów alternatywnych. Model ten wykazał się największą elastycznością adaptacji do nowych instrukcji.
- Claude i Llama – modele te również osiągnęły 100 proc. poprawności w próbkach, które były w stanie technicznie przetworzyć (odpowiednio 157 i 127 próbek). Inżynieria zapytań skutecznie wyeliminowała ich wcześniejsze błędy merytoryczne.
- Microsoft Copilot – osiągnął 99 proc. skuteczności (160/161 próbek), popełniając tylko jeden błąd. Nadal jednak borykał się z ograniczeniem wielkości plików.
- Chat GPT – mimo poprawy, model ten nadal wykazywał pewne problemy. Jego skuteczność wzrosła z 74 do 89 proc. (w teście pierwszym), ale w skali wszystkich testów nadal popełniał błędy, wykazując tendencję do halucynowania przy bardzo złożonych dokumentach prawnych.

<sup>4</sup> Podsumowanie przedstawiono w oparciu o wszystkie testy od 1 do 8.

<sup>5</sup> Liczba próbek odrzuconych ze względów technicznych.

Symulacje pokazują, że modele LLM mogą być wartościowym narzędziem wspierającym audytora, szczególnie w analizie dokumentów tekstowych. Ich skuteczność zależy jednak od jakości i sposobu sformułowanego zapytania.

W podsumowaniu analizy warto również zaznaczyć, że uznanie przez modele LLM próbki niepoprawnej za poprawną stwarzało ryzyko przeoczenia błędu przez audytora analizującego wyniki testu. Sytuację tę podsumowuje tabela nr 6.

**Tabela 6. Podsumowanie rzetelności weryfikacji modeli LLM**

Model	Testy zadane	Uznanie próbki niepoprawnej za poprawną	Rzetelność weryfikacji (w proc.)
Chat GPT	167	19	89
Gemini	167	-	100
Microsoft Copilot	167	1	99
Llama	167	1	99
Claude	167	2	99

Źródło: Opracowanie własne na podstawie: wyników testów modeli LLM.

## 6. Dyskusja: Szanse i wyzwania dla profesji

Wyniki badania empirycznego prowadzą do ważnych wniosków dotyczących przyszłości audytu w kontekście wykorzystania sztucznej inteligencji. Szanse i wyzwania można przyporządkować do trzech kategorii.

### 1. Redukcja asymetrii informacji.

Zdolność modelu Gemini do bezbłędnej analizy 167 dokumentów w czasie nieporównywalnie krótszym niż praca człowieka (analiza dokonana przez modele językowe zajmuje kilka minut, analogiczna analiza dokonana przez biegłego rewidenta – kilka godzin lub nawet dni) wskazuje na nadchodzącą, potencjalną zmianę paradygmatu audytu. Zamiast badać próbę obejmującą 5-10 proc. dokumentów, audytorzy wyposażeni w LLM będą mogli weryfikować 100 proc. transakcji. To drastycznie zmniejsza pole do ukrywania nadużyć przez zarządy (agentów) i dostarcza udziałowcom (pryncypałom) znacznie pewniejszą informację o stanie spółki.

### 2. Wykrywanie anomalii jakościowych.

Badanie potwierdziło, że LLM są szczególnie skuteczne w wykrywaniu anomalii w danych tekstowych (np. brakujące klauzule RODO, zmiany w szablonach umów), które są trudne do wychwycenia przy masowym, manualnym przeglądzie, zwłaszcza w kontekście ograniczonych możliwości utrzymywania uwagi przez człowieka.

### 3. Wyzwania implementacyjne.

- I) Zależność od jakości zapytania. Badanie dowiodło, że jakość pracy AI jest bezpośrednią pochodną jakości promptu. Biegły rewident musi stać się ekspertem w komunikacji z modelem („inżynierem promptów”), precyzyjnie definiując kryteria badania.
- II) Ryzyko halucynacji. Przypadek Chat GPT, który „zmyślał” poprawność umów, pokazuje, że AI nie może działać bez nadzoru. Konieczna jest weryfikacja wyników przez człowieka i jego udział w procesie (konceptcja „human-in-the-loop”), szczególnie w bardziej złożonych przypadkach, które mogą budzić wątpliwości.
- III) Kwestie techniczne i prawne. Ograniczenia wielkości plików (Copilot) oraz problemy z analizą formatu xls (Llama) wskazują, że obecne narzędzia wymagają jeszcze rozwoju lub integracji ze specjalnymi systemami audytowymi. Ponadto, przesyłanie poufnych danych klientów do publicznych modeli rodzi pytania o bezpieczeństwo danych i zgodność z RODO i artykułem 54 Rozporządzenia Parlamentu Europejskiego i Rady (UE) 2016/679 (GDPR).

## Podsumowanie

Wdrożenie dużych modeli językowych do warsztatu narzędzi biegłego rewidenta wydaje się nieuniknionym kierunkiem rozwoju profesji. Technologia ta, jak wykazało badanie na przykładzie modelu Gemini, osiągnęła już poziom dojrzałości pozwalający na wspieranie efektywności człowieka, a w niektórych obszarach nawet jej przewyższanie (masowa analiza dokumentów).

LLM oferują realną szansę na redukcję luki oczekiwań (expectation gap) i asymetrii informacji na rynku. Pozwalają na przesunięcie środka ciężkości pracy audytora z pracochłonnej, manualnej weryfikacji dokumentów na analizę ryzyk, ocenę osądów kierownictwa i interpretację wyników dostarczanych przez systemy AI. Warunkiem sukcesu jest jednak świadome wdrażanie tych narzędzi, z uwzględnieniem konieczności optymalizacji zapytań oraz zachowania krytycznego nadzoru nad wynikami ich pracy.

## Literatura

- Amirizianiani M., Martin E., Roosta T., Chadha A., Shah Ch. *A Tool for Auditing Large Language Models Using Multiprobe Approach*. 2024.
- Bazerman M. H., Loewenstein G., & Moore D. A. „Why good accountants do bad audits”. *Harvard Business Review*. 2002.
- Becker K., Günther T., & Neuhüttler J. „An investigation into how generative AI can improve auditors’ understanding of audit evidence”. 2024.
- Benaich, N., & Hogarth, I. *State of AI Report 2024*. Air Street Capital. 2024.
- Brown T. B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., & Amodei D. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*. 2022.
- Business Insider Polska. „NIK: państwo nie zadziałało ws. GetBack. Zawiódł też audyt”. 2020.
- Cockcroft S., & Russell M. „Artificial intelligence: perspectives from auditing academics and practitioners”. *Accounting and Finance*. 2018.
- Coffee J.C. *Gatekeepers: The Professions and Corporate Governance*. Oxford University Press. 2006.
- Damodaran A. *Corporate Finance: Theory and Practice*. 2010.
- DeAngelo L. E. „Auditor size and audit quality”. *Journal of Accounting and Economics*. 1981.
- Devlin J., Chang M. W., Lee K., & Toutanova K. „BERT: Pre-training of deep bidirectional transformers for language understanding”. In: *Proceedings of NAACL-HLT*. 2019.
- Dopierała W. „Asymetria informacji a rynek usług audytorskich”. *Zeszyty Teoretyczne Rachunkowości*, 66 (122), 79-94. 2012.
- Eimers M. E., & Stagno M. C. „The impact of artificial intelligence on the accounting profession”. *Journal of Emerging Technologies in Accounting*. 2021.
- Hay D., Knechel W.R., Wong N. „Audit Fees: A Meta-analysis of the Effect of Supply and Demand Attributes”. *Contemporary Accounting Research*. 2006.
- Jensen M.C., & Meckling W.H. „Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure”. *Journal of Financial Economics*. 1976.
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. Survey of Hallucination in Natural Language Generation. 2023
- Knechel W. R., Krishnan G. V., Pevzner M., Shefchik L. B., & Snaith M. S. „Audit quality: Insights from the academic literature”. *Auditing: A Journal of Practice & Theory*, 32 (1). 2013.
- Maslej, N., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Kariuki, N., Capstick, E., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., Walsh, T., Hamrah, A., Santarlasci, L., Lotufo, J. B., Rome, A., & Shi, A. *Artificial Intelligence Index Report 2025*. Stanford Institute for Human-Centered Artificial Intelligence (HAI), Stanford University. 2025.

- McCrum D. & Jones S. „The €1.9bn that vanished: Inside Wirecard’s missing millions”. *Financial Times*. 2020.
- Mökander J., Schuett J., Kirk H. R., & Floridi L. „Auditing large language models: A three-layered approach”. 2023.
- Najwyższa Izba Kontroli. *Działania organów państwa wobec GetBack S.A.* 2020.
- O’Dwyer B., Owen D., & Humphrey C. Assurance beyond accounting: Towards a contemporary agenda. *Accounting and Business Research*, 41 (3), 205-231. 2011.
- Porter B. A. „The audit expectation gap: Underlying causes and modifying factors”. *Accounting and Business Research*. 1993.
- Pulliam S., & Solomon D. „WorldCom Overstated Profit By \$3.8 Billion Through Accounting Error”. *The Wall Street Journal*. 2002.
- Radford A., Wu J., Child R., Luan D., Amodei D., & Sutskever I. „Language models are unsupervised multitask learners”. *OpenAI Blog*. 2019.
- Raffel C., Shazeer N., Roberts A., Lee K., Narang S., Matena M.,... & Liu, P. J. „Exploring the limits of transfer learning with a unified text-to-text transformer”. *Journal of Machine Learning Research*. 2020.
- Wallace W. A. „The economic role of the audit in free and regulated markets: A review”. *Research in Accounting Regulation*. 1980.
- Watts R. L. „Corporate financial statements, a product of the market and the state”. *Accounting Review*. 1977.
- Watts R. L., & Zimmerman J. L. *Positive accounting theory*. Prentice Hall. 1986.
- International Ethics Standards Board for Accountants. *International Code of Ethics for Professional Accountants*. 2020.
- Międzynarodowy Standard Badania 260. *Komunikowanie się z osobami sprawującymi nadzór nad jednostką*. 2023.
- Ustawa o rachunkowości z dnia 29 września 1994 r, art. 66 ust. 5.
- Ustawa o rachunkowości z dnia 29 września 1994 r, art. 64 ust. 1.
- Ustawa z dnia 11 maja 2017 r. o biegłych rewidentach, firmach audytorskich oraz nadzorze publicznym (t.j. Dz. U. z 2017 r. poz. 1089).
- Anthropic. *Claude 3 Technical Overview*. 2024. Online. Dostęp 19 maja 2025. <https://www.anthropic.com/news/claude-3-family>.
- Center for Audit Quality. *Auditing in the age of generative AI*. 2024. Online. Dostęp 8 maja 2025. <https://www.thecaq.org/auditing-in-the-age-of-generative-ai>.
- Google. *Introducing Gemini: our most capable AI model*. 2024. Online. Dostęp 19 maja 2025. <https://blog.google/technology/ai/google-gemini-ai>.
- Grand View Research. *Wearable AI Market Size, Share & Trends Analysis Report By Type (Smartwatches, Smart Eyewear, Smart Earwear), By Application, By Operations, By Component, By Region, And Segment Forecasts, 2023 – 2030*. 2023. Online. Dostęp 13 maja 2025. <https://www.grandviewresearch.com/industry-analysis/wearable-ai-market-report>.
- HAI Stanford. *Artificial Intelligence Index Report 2025*. Stanford University Human-Centered Artificial Intelligence. 2025. Online. Dostęp 7 maja 2025. <https://hai.stanford.edu/ai-index/2025-ai-index-report>.

Meta AI. *Meta introduces LLaMA 3: Open foundation models*. 2023. Online. Dostęp 19 maja 2025. <https://ai.meta.com/blog/meta-llama-3>.

Microsoft. *Microsoft 365 Copilot Overview*. 2023. Online. Dostęp 19 maja 2025. <https://learn.microsoft.com/en-us/microsoft-365/copilot/overview>.

OpenAI. *GPT-4 Technical Report*. 2023. Online. Dostęp 8 maja 2025. <https://openai.com/index/gpt-4-research>.

