

Use of large language models (LLMs) in auditing as a tool to reduce information asymmetry and fraud risk in the shareholder-management relationship

MARTA GRACZYK
ORCID: 0009-0006-8745-4188

WALDEMAR MAJEK
ORCID: 0009-0001-4257-9300

PIOTR MODZELEWSKI
ORCID: 0000-0003-2817-9885

Astract

Aim: The purpose of the paper is to explore the use of Large Language Models (LLMs) in audit testing and to assess their potential in reducing information asymmetry and fraud risk in the relationship between the audit firm (engaged by the audited company's shareholders) and the management responsible for the preparation of the audited financial statements. The paper verifies the research hypotheses regarding the impact of generative artificial intelligence (Gen AI) on the accuracy and efficiency of financial irregularity detection as compared to traditional methods.

Methodology / Research approach: The paper employs triangulation of research methods: an analysis of the literature, the conceptualization of the problem based on the theory of agency and the concept of agency costs, and a simulation-type empirical study. The study consisted of eight actual audit tests on a sample of 167 items (contracts, invoices, entries from accounting systems, and other financial documents) sourced from a case study involving an international capital group. Fivelanguage models widely recognized as key to the development of LLM technology were used in the analysis: ChatGPT, Gemini, Microsoft Copilot, Llama and Claude (Benaich, 2024; Maslej at al, 2025). A two-stage research procedure was employed: an initial phase using basic prompts and an optimisation phase using advanced prompt engineering, with a detailed comparative analysis of the results.

Results and recommendations: The results of the study indicate that "crude" LLMs in the initial phase show varying degrees of efficiency (between 74 and 96 per cent), however, their use carries a significant risk of hallucinations (the generation of content that is grammatically and logically correct but inconsistent with the facts)



and interpretation errors when analysing complex legal clauses. Hallucinations in this context differ from classic type I and type II errors that occur in auditing: LLMs can generate false content that may have the appearance of correctness, although this does not necessarily equate to a misidentification of irregularities or an oversight in a statistical sense. However, after applying prompt optimisation and refining the context, the models' efficiency increased significantly, with the Gemini model achieving 100 per cent correctness in the test sample. This technology allows for a rapid analysis of 100 per cent of the population of unstructured data, which is conducive to detecting potential irregularities that are not visible through traditional sampling.

Research limitations / implications: The study was carried out on anonymised data from a single capital group operating in the household appliances sector, which may limit the universality of the conclusions for specific sectors (e.g. finance). The main limitations of the technology include the file size limits, data confidentiality issues and the risk of overconfidence in the results generated by AI.

Originality / value: The paper fills a research gap in the practical application of GenAI in specific audit procedures on real data. It presents empirical evidence of the effectiveness of prompt engineering in eliminating language model errors in auditing, contributing to the discussion on the future of the statutory auditor profession.

Keywords: Large Language Models (LLMs), audit, information asymmetry, fraud risk, generative artificial intelligence (GenAI), theory of agency.

Introduction

The present-day financial and capital services market is characterised by increasingly complex relationships between stakeholders and a dynamic increase in the amount of data generated by business entities. In a data-driven economy, the shareholder and investor confidence in company management boards is based on the integrity and reliability of the presented financial statements. The data-driven economy is an extension of the concept of the knowledge-based economy, in which not only knowledge and intellectual capital become key resources, but also data as a fundamental factor in value creation (OECD, 1996; European Commission, 2020). However, the naturally occurring asymmetry of information between a company, which has full knowledge of its condition, and external stakeholders creates structural conditions for the occurrence of erroneous allocation decisions and financial fraud.

One systemic mechanism to reduce this asymmetry and make the data more reliable is an independent financial audit. Nevertheless, traditional audit methods, based largely on manual review of documentation and inference based on limited test samples, are becoming increasingly ineffective and inefficient in the era of Big Data. They are time-consuming, prone to human error (fatigue, omission) and subjectivity of judgement. In addition, market pressure to deliver the audit quickly and reduce its costs may result in a reduction in the quality of assurance services, which has an adverse effect on the ability to detect accounting fraud (Knechel et al., 2013).

Recent years have seen rapid development of generative artificial intelligence (GenAI) and, in particular, Large Language Models (LLMs). Trained on very large text datasets, these models have the ability to process natural language, understand context, analyse sentiment and draw logical conclusions. In terms of auditing, LLMs can revolutionise the analysis of unstructured data – contracts, invoices, memos from management board meetings and email correspondence – which make up a significant portion of audit evidence and the analysis of which has until recently been difficult to automate.

Based on empirical studies simulating the work of an auditor, this paper attempts to answer the question of whether GenAI can become an effective tool to support statutory auditors, enhancing the precision and expanding the scope of audits and ultimately reducing information asymmetry between management and shareholders.

1. Theoretical basis of the research: Theory of agency and the risk of fraud

The theory of agency, the foundations of which were laid by, among others, Jensen and Meckling, describes a relationship in which one party (the principal) entrusts another (the agent) with a task, delegating some of its decision-making powers to that party. There is a specific three-way configuration of this relationship in the audit market. The shareholders (the principal) entrust capital to the management (the agent), engaging an audit firm to verify the latter's actions (Jensen & Meckling, 1976).

A key problem with this structure is that the audit firm does not work directly with the shareholders, but with the company's management, which creates a conflict of interest and the risk of disadvantaging investors in order to retain the client (Schaefer, 2023). This leads to agency costs, which include oversight costs (e.g. the auditor's fee), bonding costs (reputation building) and residual costs arising from the divergence of interests. (Watts et al., 1986).

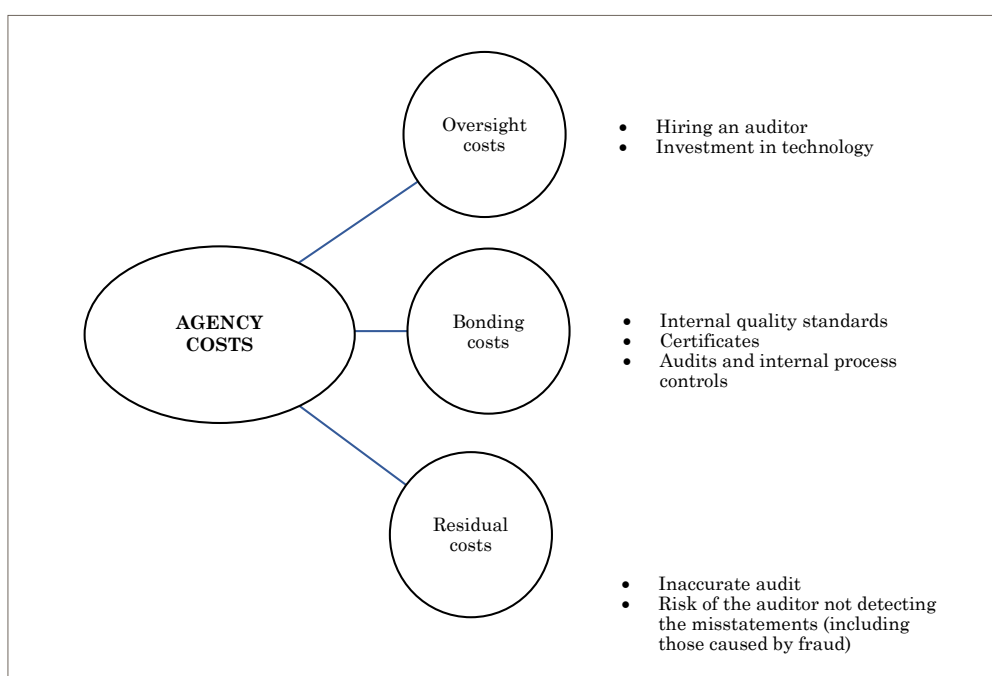


Fig. 1. The structure of agency cost in the relationship between shareholder and the audit firm

Source: Own study based on: Jensen. 1976, Damodaran. 2010, Solomon. 2020).

Within the theory of agency there are two main issues affecting audit quality:

1. Adverse selection: a situation where the auditors do not have full knowledge of the documents provided by the audited entity, which reduces the quality of the verification (Jensen & Meckling. 1976, Eisenhardt K.M. 1989).
2. Moral hazard: the phenomenon of an auditor putting their own interests above those of shareholders, e.g. by not disclosing detected irregularities in order to retain the contract (Jensen & Meckling. 1976, Eisenhardt K.M. 1989).

Despite the existence of legal regulations and ethical standards, the audit market is characterised by systemic sources of conflicts of interest. The most important of these include:

- financial dependence on the company's management (the auditor is paid by the entity whose accounts they are auditing),
- provision of advisory services in parallel with the audit,
- pressure from the management board to accept aggressive accounting practices,

and

- “familiarity threat”, i.e. an overly close relationship with the client.

(International Ethics Standards Board for Accountants. 2018, DeAngelo, 1981, Frankel, Johnson, & Nelson, 2002).

Economic history provides evidence of the negative materialisation of these risks. The collapse of Enron and the audit firm Arthur Andersen highlighted the problem of lack of independence caused by the provision of advisory services and the concealment of creative accounting. Similar mechanisms were at work in the case of WorldCom and Germany's Wirecard, where auditors ignored signs of financial manipulations for years. In Poland, one example is the GetBack case where the effectiveness of the auditor's actions in verifying the company's financial statements was called into question (McLean & Elkind, 2003, Coates, 2007, Pulliam & Solomon, 2002, McCrum & Jones, 2020, Business Insider Polska, 2020).

The primary objective of an audit is to provide an independent opinion on the financial statements in order to give credibility to the data and reduce the information gap between the management and shareholders. The quality of audit is meant to be ensured by such mechanisms as independent verification, professional scepticism and compliance with standards (Watts, 1977).

However, traditional auditing has significant limitations that prevent complete eradication of the risk of overlooking misstatements and the risk of fraud. The main limitations include:

- the test nature of audits (less than 100 per cent of transactions are verified);

- delay of information in relation to economic events;
- limited assurance;
- subjectivity of the auditor's assessments;
- inherent risk of failure to detect fraud.

Furthermore, the phenomenon of moral hazard can lead to shareholders' overconfidence in the auditor's opinion, which paradoxically lulls them into a false sense of security (Dopierała, 2012, DeFond, 2010).

The transparency of an audit is a key value for investors, allowing them to estimate risk more accurately and to enforce management accountability. In the face of the challenges of traditional testing methods, there is a growing demand for the implementation of modern tools such as artificial intelligence.

An analysis of the literature and historical cases of fraud indicates that traditional audit methods in the context of the theory of agency are sometimes unreliable and vulnerable to the human factor and business pressures. There is therefore an urgent need to implement solutions that automate the document analysis process and increase the independence of auditing. In response to these challenges, research hypotheses were formulated regarding the impact of generative artificial intelligence (Gen AI) on the accuracy and efficiency of financial irregularity detection as compared to traditional methods.

2. Characteristics and potential of Large Language Models (LLMs)

The dynamic development of the artificial intelligence technology, in particular Large Language Models (LLMs), is revolutionising working methods in data-driven sectors, including financial auditing. Unlike traditional analytical methods, which focus mainly on structured (numerical) data, LLMs offer the ability to efficiently process large sets of unstructured data, such as contracts, correspondence or narrative reports. This section takes a closer look at the characteristics of this technology and its application detecting anomalies and reducing information asymmetry (Becker et al., 2024).

Large Language Models are sophisticated neural network-based systems trained on large text datasets and capable of generating and analysing natural language at a human-like level. They are underpinned by the deep learning architecture developed by Google (the Transformer architecture) and the ability to recognise context through the attention mechanism – both mechanisms were presented by Vaswani et al. in 2017 in the publication “Attention is all you need”. The attention mechanism allows the model to identify relationships between words in a sentence even in complex passages of text, enabling a deep understanding of context, linguistic nuances and cause-and-effect relationships (Vaswani et al., 2017).

The LLM learning process takes place in two stages:

1. the pre-training phase, in which the model acquires general linguistic knowledge by learning from huge text data sets;
2. the fine-tuning phase, which adapts the model to specific tasks (Raffel et al., 2020).

An important feature of modern models is the ability to perform what is referred to as few-shot learning, i.e. to perform new tasks based on just a few examples contained in a prompt, which significantly lowers the barriers to implementing this technology in specific audit procedures (Brown et al., 2020). Despite its impressive capabilities, the technology carries the risk of so-called hallucinations, i.e. generation of content that is grammatically and logically correct, but factually wrong (Ji et al., 2023). This requires auditors to maintain a critical approach and verify the results generated by AI (Mökander and Floridi, 2023).

From an audit perspective, the ability to analyse textual data (often in several languages) in unstructured form, such as contracts, email correspondence, event reports or minutes and memos from meetings, is extremely important. Up to now, analytical systems have mainly been able to help analyse numerical data, disregarding textual data as a rule. LLMs extend these capabilities and can therefore

contribute to increasing the accuracy of audits (Becker et al., 2024). In addition to analysing financial data, modern auditing largely requires the analysis of a growing number of text documents. With their ability to analyse sentiment (a skill that involves understanding the intent and tone of speech), LLMs can automate and seal the process in areas that previously required labour-intensive manual verification.

Key functionalities that support auditing in qualitative analysis include:

1. Extraction of information – automatic extraction of key data (dates, amounts, penalty clauses, signatures) from documents of varying structure, such as invoices or contracts. Studies indicate that LLMs can process multilingual documentation much faster and with higher precision than human teams (Dennis, 2024).
2. Classification of documents – assigning source documents to appropriate audit categories, which improves the organisation of audit evidence (Becker et al., 2024).
3. Analysis of sentiment – assessing the tone of speech in internal correspondence or management memos. Detecting negative emotionality, uncertainty or pressure in communication can be a red flag that indicates the risk of fraud or business continuity problems (Srivastava et al., 2023).
4. Prompt engineering – the ability to create special instructions (prompts) for specific audit tests, e.g. verification of the compliance of invoices with purchase orders, thus providing for standardisation of procedures (Mökander et al., 2023).

The use of these tools makes it possible to progress from random testing (sampling) to an analysis of the full population of documents, which directly reduces the risk of overlooking material misstatements.

One of the most important applications of LLMs in auditing is the identification of anomalies, understood not only as numerical distortions, but above all as semantic and narrative inconsistencies. Traditional analytical methods often overlook this aspect. Language models are able to detect subtle signals indicative of manipulation, such as:

- contradictions between the financial data and the narrative part of the annual report (known as narrative fraud);
- sudden changes of style or tone in management reports;
- unusual clauses in contracts or lack of standard legal safeguards;
- occurrence of keywords suggesting concealment of information, e.g. “unidentified”, “classified” (Crawford, 2020).

The study by Rahaman et al. (2024) demonstrates that LLMs can identify inconsistencies in corporate narratives with greater accuracy and faster than analysts who use traditional methods. The table below compares the effectiveness of different methods in detecting anomalies.

Table 1. Comparison of the effectiveness of financial anomaly detection

Detection method	Type of anomaly	Estimated effectiveness ¹	Notes
manual review	contradictory statements	low – depending on the person	time-consuming, subjective
traditional tools ²	key phrases	medium	limited to simple patterns
LLMs	change in tone, inconsistency in narrative	high (70-90 %)	requires human supervision

Source: Prepared based on: Becker et al. 2024, Mökander et al. 2023).

It is worth noting that the detection of anomalies and fraud will be most effective when combined with traditional auditing methods. Combining these two approaches can lead to a comprehensive and accurate depiction of potential anomalies.

In summary, integration of Large Language Models into the statutory auditor's workshop provides an opportunity to significantly improve the quality of attestation services. The ability to quickly analyse context and detect anomalies in unstructured data fills the gaps left by traditional methods. However, verification of this thesis requires empirical evidence.

¹ Estimates based on the literature: Becker et al. 2024, Mökander et al. 2023.

² Traditional data analysis tools for searching and filtering text fragments and pattern matching such as excel, IDEA.

3. The agency model in the shareholder-management relationship: an empirical examination

To verify the potential of LLMs in audit practice, a simulation audit was designed and conducted on real data. The data used in the empirical study for this paper was obtained from a company operating in the household appliances sector. In 2024, the company went through an acquisition process and became part of a larger corporation. Today, the company operates in more than 100 countries, has 18 production sites, including 11 in Europe, and is one of the leaders in the European household appliances supply market. The company has an extensive organisational structure in EMEA (Europe, Middle East and Africa), including numerous local units that are responsible for sales, logistics, marketing, customer service and equipment maintenance. The company's operations are characterised by a complex organisational structure and extensive business, financial and accounting processes which generate enormous volumes of financial as well as non-financial data each year. As it meets the criteria set out in the Accounting Act regarding the size of the company as measured by the level of employment and revenues generated, the company is subject to mandatory annual audits of its financial statements.

Cooperation with an audit firm includes an annual audit of the company's financial statements and cyclical audit tests carried out in various countries, as part of compliance with SOX regulations (Sarbanes-Oxley Act 2002). To begin with, it is worth emphasising that all documents used in the audit tests have been fully anonymised, thus not revealing the company's sensitive financial information and counterparty data that served as the source for the empirical examination. The data for the audit tests was selected from an archive of audits conducted in the EMEA region. Within this region, each country is represented by a separate unit, ensuring the diversity of the test population. The dataset has been prepared for research purposes only, respecting the principles of information security and confidentiality.

In order to obtain a comprehensive assessment of the capabilities of Large Language Models, test documentation was selected for the test sample in such a way that it covered the company's key business areas and included typical audit verification procedures. At the same time, the sample was intentionally differentiated to include a variety of data both in terms of form and content. Documents from eight different audit tests from four different financial audits and four SOX audits were analysed; they differed from one another in terms of:

- type and thematic scope – the sample tested contains a variety of documentation in the form of contracts, invoices, HR data, internal policies;
- source language of documents – the selected sample contains documentation in four different languages;
- country of origin – the documentation selected for the sample comes from five different countries;

- type of audit – the documentation obtained comes from SOX audits, financial audits, compliance audits and anti-corruption audits;
- data format – the sample is also diversified in terms of the type of documents and includes PDFs, emails, invoices, forms, contracts, screenshots from the accounting system and tabular summaries.

The aim of this intentionally diverse sample is to explore how LLMs cope with different audit tasks, languages and document types. To ensure that the sample is representative and the results can be generalised, each audit test was selected from a separate audit, conducted in an EMEA country, with the exception of one containing the audit of two countries at the same time (two tests were selected here, each covering a different country). Such a methodology allows for taking into account the cultural, regulatory and operational differences encountered in each country. The empirical study includes audit tests which are characterised in detail in Table 2.

Table 2. Characteristics of the audit tests selected for the study

Test number and range	Audited area and year	Documentation languages	Type of documents/ data	Purpose of the test
Test 1 - verification of a sample of contracts	2023 EMEA - Germany and Italy: direct-to-consumer sales	English, Italian	sales contracts, legal clauses	verification whether the contract was concluded and valid during the period under examination
Test 2 - verification of a sample of transport documents	2021 EMEA - UK-based sales unit	English	waybills, delivery notes, delivery dates	correctness of revenue recognition
Test 3 - verification of a sample of cost invoices	SOX 2023 - EMEA: RTP inventory control (Inv5) at the Shared Services Centre in Łódź	English	invoices from suppliers, descriptions of services or goods, amounts, dates	verifies whether a bidding process has been carried out for each invoice above EUR 10,000 or whether a valid supplier contract exists
Test 4 - verification of a sample of correction invoices	2023 EMEA - Germany and Italy: direct-to-consumer sales	English, German	correction notes, original invoices, reasons for correction	analysis of the reasons for and correctness of the corrections issued, identification of potential irregularities in the accounts

Test number and range	Audited area and year	Documentation languages	Type of documents/ data	Purpose of the test
Test 5 - verification of financial acceptances according to the acceptance procedure	2023 EMEA – a trading partner incentive scheme	English	system data, acceptance workflow, acceptance matrix, payment amounts, acceptance and posting dates	assessment of the conformity of the payment acceptance process with the applicable procedure, verification of the authorisations of the accepting officers
Test 6 - verification of a sample of changes to credit limits and customer payment terms	SOX 2023 – EMEA: review of the settlement of OTC receivables (AR4) in Italy	English, Italian	requests for limit or deadline changes, approvals, customer data	assessment of the relevance and appropriateness of changes to credit limits and payment terms, assessment of compliance with the company's credit policy
Test 7 - verification of a sample of new client creation	SOX 2023 – EMEA: review of the settlement of OTC receivables (AR1) in Italy	English, Italian	customer creation forms, registration documents, CRM system data	assessment of the completeness and correctness of the process of creating a new client profile, verification of the required documents and information
Test 8 - verification of a sample of employee wage changes	2023 EMEA – anti-corruption and bribery audit	German, French	employment contracts, annexes, documents evidencing wage changes, change effective dates	assessment of the correct documentation and implementation of changes in employee wages, assessment of compliance with contracts and internal regulations

Source: Prepared by the authors based on collected empirical data.

The choice of the above tests was dictated by the desire to represent key audit areas, with the focus on the most time-consuming tasks (long contracts, screenshots) and exclusion of large spreadsheets, which current GenAI models cannot cope with. The sample was deliberately selected to test documents related to both sales and company costs. Five Large Language Models (paid versions), representing leading AI technology providers, were used for the comparative analysis included in the empirical study:

- GPT Chat from OpenAI – based on the GPT-4 architecture, one of the pioneering LLMs, characterised by high capacity for understanding and processing

natural language. The model offers answers to complex questions, interpretation of documents and search for logical anomalies.

- Gemini from Google – known for its high quality linguistic and interpretive analysis, it copes well with interpreting and analysing information contained in text, images and code.
- Microsoft Copilot – a tool that is tightly integrated into the Microsoft environment and uses GPT chat elements in its architecture. It offers advanced analysis of text and data, also from Word documents.
- Claude from Anthropic – a model designed with data security in mind, known for its ability to hold long and complex conversations and generate coherent texts.
- Llama from Meta is a model used for comparative analyses in also less common languages. Due to the fact that Meta has not yet made the Llama model available on the Polish market, the Abacus AI platform was used in the tests.

Such a diverse set of tests and tools made it possible to reliably assess the potential for using Large Language Models in audit tests. Owing to the variety of formats, languages, data types and countries of origin of the documentation, the analysis carried out for this paper provides a solid basis for drawing conclusions as well as the possibility of generalising the results. Its conclusions help to better understand the potential and limitations of the GenAI technology in the financial audit sector and provide practical guidance on the effective use of AI tools in the future.

The empirical examination was conducted in two phases:

1. The preliminary phase: the models were given basic prompts describing the purpose and steps of the test, without any additional contextual explanations;
2. The optimisation phase (prompt engineering): based on the errors from phase one, improved prompts were developed with precise interpretative instructions, and tests were run again.

The results generated by AI (pass/fail status with reasoning) were compared with those traditionally obtained by the audit team, taking the latter as the ground truth.

4. Test results – The preliminary phase: analysis of errors and limitations

In the first testing phase, the models were confronted with a “raw” audit task. The results of this phase revealed both the great potential of the technology and the significant risks associated with its indiscriminate use.

Table 3. A summary of the results of the audit test analysis

Model	Assigned tests	Rejected tests ³	Executed tests	Correct answers	Ratio of correct answers to all tests assigned (per cent)	Ratio of correct answers to tests executed (per cent)
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

Source: Prepared by the authors based on the collected empirical data.

The preliminary analysis showed significant differences in the effectiveness and technical capabilities of the different models. A summary of the results for all 167 samples is as follows:

- Gemini – this model demonstrated the highest technical and substantive performance. It processed 100 per cent of the samples assigned (167/167) and gave the correct answers in 161 cases, a success rate of 96 per cent. It did not reject any file due to format or size.
- Microsoft Copilot – it achieved a high substantive efficiency (92 per cent), correctly solving 153 samples. However, it encountered technical barriers – it rejected 6 samples in test one due to exceeding the file size limit (1 MB), which is a significant limitation in audit practice as the files used are usually larger in size.
- Claude - similarly to Copilot, it achieved 92 per cent substantive efficiency (153 correct answers), but rejected 10 samples due to technical reasons.

³ The number of samples rejected on technical grounds.

- Llama - this model encountered the greatest difficulties in analysing .xls files, rejecting a total of 40 samples. However, for text documents in .pdf format, which it was able to process, its efficiency was impressive – 98 per cent.
- Chat GPT - it received the lowest overall score with 74 per cent correctness (123 correct answers). This model most often made substantive and interpretation errors.

An analysis of the causes of errors in the preliminary phase led to the identification of key weaknesses of “raw” LLMs in audit applications. The most problematic test was test one (verification of contracts), which required legal analysis. Key weaknesses include:

1. The problem with interpretation of contractual clauses

In test one, the models had a problem with interpreting automatic contract renewal clauses.

- Gemini found the test negative in three cases, arguing that there was no evidence that the contract had not been terminated, even though the contract itself contained an automatic renewal clause. This is an example of excessive caution which generates false alarms (false positives in the context of error detection).
- Chat GPT made errors in five cases. In sample no. 24, the model considered the contract to be valid (“score passed”), even though it did not cover the entire period under examination. The model hallucinated, deeming partial period coverage sufficient, which is a fundamental audit error.
- Claude also made errors in assessing the validity of contracts in three cases which provided for their automatic renewal.

2. The problem of alternative evidence:

In test two (regarding transport documents), the task was to confirm delivery. Some of the documents did not have the client’s signature (which could be considered an error), but the auditors had alternative evidence in the form of a screenshot from the client’s logistics system.

In the preliminary phase, the models often ignored such alternative evidence, rigidly adhering to the “signature on the document” prompt, which resulted in an erroneous “failed” assessment for correct transactions.

3. The risk of false positives (type I error):

The most dangerous phenomenon observed at the preliminary phase was that the models considered defective samples to be correct. In the entire test sample, Chat GPT considered 19 incorrect samples as correct. In auditing, this means the risk

of not detecting a material misstatement (risk of omission), which is much more dangerous than a false alarm requiring additional verification.

The above analysis of the test results leads to the conclusion that the use of LLMs in auditing requires human supervision and validation of the results, and that the models cannot fully replace the professional judgement of an experienced auditor. Another testing phase, i.e. optimisation, was also necessary.

5. The optimisation phase and final results – the role of prompt engineering

Based on the error analysis from phase one, the second part of the empirical examination was carried out using prompt engineering techniques. A set of clarified instructions was developed to eliminate ambiguities and impose audit logic on the models.

Table 4. A summary of the wording for optimisation of audit prompts

No.	Cause	Proposed optimisation
1	Each tool indicated different samples as negative.	Explain the reasoning for negative results.
2	For contracts with automatic renewal, Chat GPT and Claude found that there is no certainty whether the contract has not been terminated before.	For contracts with automatic renewal, recognise the contracts as valid.
3	Chat GPT and Llama found that a renewed contract fell within the period under examination, where in fact it did not cover the entire period.	The contract must cover the entire period under examination.
4	Chat GPT and Llama recognised the supplier's signature as the client's signature.	The client's signature can be found in the document under item 24.
5	Gemini did not recognise a confirmation from the client's internal system as sufficient evidence.	If there is a confirmation from the client's internal system and it contains a delivery number and date that match the data in the company's documents, recognise it as correct.
6	Copilot did not recognise a sales adjustment based on SAP screenshots as sufficient evidence.	If there is a confirmation of a sales adjustment in a screenshot from SAP that relates to the same or the following month, recognise it as correct.
7	In the case of a sample that related to court proceedings, Chat GPT, Gemini and Llama recognised the test as a fail.	For court costs, consider the test as not applicable (N/A).

Source: Prepared by the authors based on the results of LLM tests.

The use of optimised prompts brought a sharp improvement in performance, particularly in tests that previously caused difficulties.

Table 5. A summary of the results of the audit test analysis after prompt optimisation⁴

Model	Assigned tests	Rejected tests ⁵	Executed tests	Correct answers	Ratio of correct answers to all tests assigned (per cent)	Ratio of correct answers to tests executed (per cent)
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

Source: Prepared by the authors based on the collected empirical data

Below is how each model coped during the optimisation phase:

- Gemini – full effectiveness. After optimisation, the Gemini model achieved 100 per cent efficiency. It correctly verified all 167 samples in all eight tests. All errors of interpretation regarding contracts and alternative evidence were eliminated. This model showed the greatest flexibility in adapting to new prompts.
- Claude and Llama - these models also achieved 100 per cent correctness in the samples which they were technically able to process (157 and 127 samples respectively). Prompt engineering effectively eliminated their previous substantive errors.
- Microsoft Copilot - it achieved 99 per cent efficiency (160/161 samples), making only one error. However, it still struggled with the file size limitation.
- Chat GPT - despite improvements, this model still showed some problems. Its efficiency increased from 74 to 89 per cent (in test one), but across all the tests it still made errors, showing a tendency to hallucinate when dealing with very complex legal documents.

⁴ The summary is based on all tests from 1 to 8.

⁵ The number of samples rejected on technical grounds.

The simulations show that LLMs can be a valuable tool in supporting auditors, especially in analysing text documents. However, their effectiveness depends on the quality and formulation of the prompt.

In summary to the analysis, it is also worth noting that the recognition by the LLMs of an incorrect sample as correct posed a risk of error being overlooked by the auditor analysing the test results. This situation is summarised in Table 6.

Table 6. A summary of the reliability of verification by LLMs

Model	Assigned tests	Recognition of an incorrect sample as correct	Reliability of verification (per cent)
Chat GPT	167	19	89
Gemini	167	-	100
Microsoft Copilot	167	1	99
Llama	167	1	99
Claude	167	2	99

Source: Prepared by the authors based on LLM test results.

6. Debate: Opportunities and challenges for the profession

The results of the empirical study lead to important conclusions about the future of auditing in the context of the use of artificial intelligence. Opportunities and challenges can be assigned to three categories.

1. Reduction of information asymmetry.

The Gemini model's ability to flawlessly analyse 167 documents in a time that is incomparably shorter than human work (an analysis by language models takes a few minutes, a similar analysis by an auditor takes a few hours or even days) shows that a potential shift in the audit paradigm is on the horizon. Rather than examining a sample of 5-10 per cent of documents, auditors equipped with an LLM will be able to verify 100 per cent of transactions. This dramatically reduces the scope for management boards (agents) to conceal fraud and provides shareholders (principals) with much more certain information about the state of the company.

2. Detection of quality anomalies.

The examination confirmed that LLMs are particularly effective at detecting anomalies in textual data (e.g. missing GDPR clauses, changes to contract templates) which are difficult to capture with mass manual review, especially in the context of limited human capacity to maintain attention.

3. Implementation challenges.

I. Dependence on the prompt quality. The examination found that the quality of AI performance is a direct corollary of prompting quality. The auditor must become an expert in communicating with the model ("a prompt engineer") by precisely defining the audit criteria.

II. Risk of hallucinations. The case of Chat GPT, which "made up" the correctness of contracts, shows that AI cannot operate without supervision. Human verification of results and human participation in the process is necessary (the human-in-the-loop concept), particularly in more complex cases that may raise doubts.

III. Technical and legal issues. File size limitations (Copilot) and problems with analysing .xls format (Llama) show that the current tools still require developing or integrating into special audit systems. Furthermore, sending confidential client data to public models raises questions about data security and compliance with GDPR and Article 54 of Regulation (EU) 2016/679 of the European Parliament and of the Council.

Conclusions

The implementation of Large Language Models into the auditor's toolkit seems an inevitable direction the profession is heading in. This technology, as shown on the example of Gemini, has already reached a level of maturity to support human performance and even surpass it in some areas (mass document analysis).

LLMs offer a real opportunity to reduce the expectation gap and information asymmetry in the market. They allow the centre of gravity of the auditor's work to shift from labour-intensive, manual document verification to analysing risks, assessing management judgements and interpreting the results provided by AI systems. However, a prerequisite for success is the conscious implementation of these tools, taking into account the need to optimise prompts and maintain critical supervision of their performance.

References

- Amirizani 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).
- International Standard on Auditing 260. *Communications with those charged with governance*. 2023).
- Act of 29 September 1994 on accounting, Article 66(5).
- Act of 29 September 1994 on accounting, Article 64(1).
- Act of 11 May 2017 on statutory auditors, audit firms and public oversight (consolidated text: Polish Journal of Laws of 2017, item 1089).
- Anthropic. *Claude 3 Technical Overview*. 2024) Online. Accessed 19 May 2025. <https://www.anthropic.com/news/claude-3-family>.
- Centre for Audit Quality. *Auditing in the age of generative AI*. 2024) Online. Accessed 8 May 2025. <https://www.thecaq.org/auditing-in-the-age-of-generative-ai>
- Google. *Introducing Gemini: our most capable AI model*. 2024. Online. Accessed 19 May 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. Accessed 13 May 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. Accessed 7 May 2025. <https://hai.stanford.edu/ai-index/2025-ai-index-report>.

- Meta AI. *Meta introduces LLaMA 3: Open foundation models*. 2023) Online. Accessed 19 May 2025. <https://ai.meta.com/blog/meta-llama-3>.
- Microsoft. *Microsoft 365 Copilot Overview*. 2023. Online. Accessed 19 May 2025. <https://learn.microsoft.com/en-us/microsoft-365/copilot/overview>.
- OpenAI. *GPT-4 Technical Report*. 2023. Online. Accessed 8 May 2025. <https://openai.com/index/gpt-4-research>.