SILESIAN UNIVERSITY OF TECHNOLOGY PUBLISHING HOUSE

SCIENTIFIC PAPERS OF SILESIAN UNIVERSITY OF TECHNOLOGY 2025
ORGANIZATION AND MANAGEMENT SERIES NO. 239

ASSESSING LARGE LANGUAGE MODELS IN IT SYSTEMS AUDITING

Mariusz ŻYTNIEWSKI[1*], Małgorzata PAŃKOWSKA[2]

[1] University of Economics in Katowice; mariusz.zytniewski@ue.katowice.pl, ORCID: 0000-0003-2170-1191
[2] University of Economics in Katowice; malgorzata.pankowska@ue.katowice.pl,
ORCID: 0000-0001-8660-606X
* Correspondence author

Purpose: This study examines the potential of large language models (LLMs) in the field of information systems auditing (ISA). Exploration of LLMs' internal knowledge is one of the most interesting research directions and concerns the range of information, knowledge such agents store without any access to external sources. The foundation of IS audit and control has evolved through the recognition of the need for strong information technology (IT) management by professional organizations, e.g., ISACA, business and government institutions. Other major areas for IS auditors are the issues of security and privacy of IT. The demand by the public and governments have generated concern for the protection of individual rights and business information, as well as the need to supervise the whole organization to ensure effective and efficient management.
Design/methodology/approach: The research applied both quantitative and qualitative evaluation methods using 146 official CISA certification questions to assess the auditing knowledge of seven large language models. The analysis measured model accuracy, inter-agent agreement, and A-bias tendencies, while qualitative examination identified reasoning errors and interpretation patterns in incorrectly answered questions.
Findings: The study found that large language models achieved an average accuracy of 81,8% on CISA certification questions, with Claude Sonnet 4 and Chat GPT-4.1 performing best at 88.4%. Some models showed certain reasoning limitations, such as bias toward specific answer options.
Research limitations/implications: The main limitation of this study is that the evaluation was based solely on multiple-choice CISA certification questions, which may not fully capture the complexity of real-world auditing scenarios.
Practical implications: The study indicates that large language models can enhance the efficiency of audit preparation and training by automating knowledge-intensive tasks.
Originality/value: The paper highlights the growing role of generative artificial intelligence in professional certification and auditing practice.
Keywords: Large Language Models, Information Systems Auditing, Generative AI, Audit Automation, AI Reliability.
Category of the paper: Research paper.

1. Introduction

The growing popularity of generative artificial intelligence, particularly large language models (LLMs), is having a significant effect on the interaction between information systems, users and specialists. According to studies, it has been found that the latest language models can perform scientific literature selection at a human-level, but are still prone to making mistakes (Khraisha et al., 2024). Exploration of LLMs' internal knowledge is one of the most interesting research directions and concerns the range of information, knowledge such agents store without any access to external sources. This type of study reveals the depth of an agent's knowledge and the possible extent of its hallucinations (Huang et al., 2025). Within this area are standards like BELIEF(ICL), which assess model memory (Zhao, Yoshinaga, Oba, 2024), and tools that not only measure model memory, but also tests of factual consistency like TeCFaP. TeCFaP (Temporally Consistent Factuality Probing) adds a temporal dimension to correctness evaluations and expose significant gaps in LLM recall (Bajpai, Goyal, Anwer, Chakraborty, 2024). In addition, the literature includes domain specific knowledge studies such as medical testing evaluations (Huang, Hsiao, Yeh, Wu, Kao, 2024), (Herrmann-Werner et al., 2024). The research presented in this article focuses on assessing agents' knowledge in the field of information systems auditing.

Information systems auditing (ISA) involves a systematic evaluation of the security, integrity and effectiveness of an organisation's IT solutions. Frameworks such as COBIT 2019 and the updated ISO/IEC 27001 2022 standard define catalogues of processes and controls ranging from governance to technical aspects. The benefits of implementing COBIT 2019 have been benchmarked and analysed for their impact in relation to risk management, resource optimisation and operational efficiency (Deagama Antariksa, Perangin Angin, Widodo, 2025). Literature reviews of ISO/IEC 27001 identify five main research themes: relationships with related standards, organisational motivations, implementation challenges, outcomes achieved and contextual factors (Culot, Nassimbeni, Podrecca, Sartor, 2021). There is still a lack of systematic studies examining whether LLMs can accurately handle tasks related to IT auditing, like business continuity control tests and compliance assessments against ISO/IEC 27001. As artificial intelligence permeates audit processes, interest is growing in whether LLMs might support auditors in competencies validated by the Certified Information Systems Auditor (CISA) examination. CISA, administered by ISACA, is the global standard confirming proficiency in five key domains:

- The information systems audit process,
- IT governance and management,
- information systems acquisition, development and implementation,
- information systems operations and business resilience,
- information asset protection.

In this study, CISA certification questions formed the basis of the research methodology applied to selected LLMs, allowing for a rigorous evaluation of their knowledge. The aim of this article is to provide an empirical assessment of the level of understanding demonstrated by these LLMs in the field of IT systems audit. The results are intended to offer practical guidance for auditors, regulators and LLM developers, and to contribute to the development of methods for assessing the reliability of generative AI in specialised technological risk management processes. In the following sections, authors will introduce the theoretical framework for evaluating LLMs, describe the research methodology adopted, present the results obtained across seven LLMs, and discuss their implications.

2. Background and Related Work

Research into the knowledge of generative AI agents focuses on multifaceted benchmarks that assess LLMs' ability to store and reproduce factual information. Zhao et al. (2024) proposed the BELIEF benchmark. BELIEF (Benchmark to Evaluate Language model Internal factual knowledgE and Faithfulness) uses diverse prompting schemes to measure the accuracy, consistency and reliability of fact recall by language models (Zhao, Yoshinaga, Oba, 2024). Their findings indicated that, although LLMs retain large quantities of information, they may struggle to retrieve it compared with human performance. At the same time, larger models may store more information but often exhibit lower response reliability than humans. The authors also showed that including examples of questions and answers within the prompt can yield up to twice as many correct responses as simply posing the question.

Lin et al. (2022) used the TruthfulQA benchmark to evaluate how readily models mimic human false beliefs. They found (across 817 questions) that the models achieved only 58% correct answers compared with 94% for human participants (Lin, Hilton, Evans, 2022). Their results suggest that increasing an LLM's parameter count leads to only small improvements in truthfulness. This means that smaller, less resource-heavy models might perform almost equally well. Bajpai et al. (2024) proposed the time oriented TC Probe methodology (Temporally Consistent Factuality Probe). This methodology uses a range of time-based questions to identify information, knowledge gaps in large language models (LLMs). In line with previous findings, they noted that embedding example answers in the prompt improves the quality of the model's outputs.

In addition to examining the general knowledge of large language models, researchers have also focused on their understanding of specific domains. Huang et al. (2024) tested GPT-4 based on medical licensing exam questions (in Taiwan). During this analysis GPT-4 achieved a general accuracy of 87.8%. Their study indicated that LLMs have potential for theoretical questions but researchers recognised LLMs' problems with more detailed questions.

This analysis demonstrated strong potential in multiple-choice settings, where LLMs must select one or more preset answers.

Herrmann-Werner et al. (2024) applied Bloom's Taxonomy (a classification of cognitive skills) to examine how well GPT-4 can respond to medical questions. The authors classified each question based on its cognitive level and recognised patterns indicating that most mistakes occurred at the lowest cognitive level questions. While GPT-4 delivered very high overall scores, the authors observed that it still tended to omit critical details in complex subject matter.

In the field of IT auditing, research remains sporadic. Fotoh and Mugwira (2025) conducted preliminary analyses of ChatGPT's role in external audits. Researchers highlighted benefits in automating routine tasks alongside risks from hallucinations and incorrect outputs. They showed that LLMs may falter in providing sufficient depth of audit-standard interpretation, may rely on outdated information, and may hallucinate, all of which can undermine audit assessments. They further emphasised challenges related to preserving auditor objectivity and professionalism, ensuring client confidentiality and data security, and maintaining clear lines of accountability.

Based on the conducted literature analysis, the authors decided to examine various language models, including models of different sizes. To determine the usefulness of LLMs in the selected area, the agents were evaluated in terms of their information, knowledge of information systems auditing. The study combined both quantitative and qualitative approaches, focusing in particular on identifying the most frequent types of agent hallucinations.

3. ISACA Credentials and AI usage

In business organizations, the role of information technology control and audit has become a critical mechanism for ensuring the integrity of information systems and the reporting of organizational finances for managerial decisions. Traditionally, auditing contributes knowledge on internal control practices and the overall control philosophy. Henri Fayol identified five functions of management, i.e., planning, organizing, leading, coordinating, and control. These functions constitute a framework for managers for their decision making on the material resources, processes, employees, and their performance. The control focuses on setting the business goals, monitoring the goals' achievement, and taking corrective actions in situations of discrepancies among goals and activities for their achievements (Senft, Gallegos, 2009).

The information system auditing is an integral part of the audit functions because it supports the auditor's judgement on the quality of the information processes supported by computer systems. The information system auditor's role has evolved to provide assurance that adequate and appropriate controls are in organizations. An information system auditor as a counselor must take an active role in developing policies on auditability, control, testing, and standards.

Auditors must convince end-users and IT personnel about the need for a controlled IT environment. IS auditors are partners of top management. Decisions concerning the need for a system traditionally belong to top management, however, when allocating funds for new technology, management has to rely on the computer personnel judgement verified by the IS auditor. Top management needs the support of a skilled computer staff that understands the organization's requirements, and IS auditors are in such a position to provide that information. The IS auditors can verify suitability of all alternatives for a given IT project. All the risks are to be accurately assessed, the technical hardware and software solutions are to be assured as correct. Beyond that, the IS auditors can work as an investigator, because the knowledge, awareness and use of computer-assisted tools and techniques in performing forensic support work have provided new opportunities for the IS auditors and IT security managers. Nowadays, the auditing data analytics and LLMs support auditing processes, by providing process data for decision making as well as by simplification of the knowledge management process.

In the information-based business environment, business professionals, who are technically competent in IT or IT specialists, who understand the accounting, management, and financial operation are highly demanded for IS auditing careers. IS auditors must continuously receive education to upgrade their knowledge, skills, and abilities. IS auditors' competences are required in the following fields: legal issues and their impact on IT, computer-assisted audit tools and techniques, ethical issues and professional standards, IT governance, management of processes and architectures, risk management, quality management, financial management, IT project management, software development and implementation, IT sourcing, application control and maintenance, change management, service management, security and service management, e-business and enterprise resource planning (ERP) systems. This set of technologies is not finished and lately the business organization digital transformation forces IT auditors to learn the new technologies, e.g., data science, LLMs, or blockchain solutions. The rapid development of artificial intelligence creates the requirement to provide explainable solutions to business, to manage artificial intelligence (AI) supported business models as well as AI risk.

Information Systems Audit and Control Association (ISACA) (www.isaca.org) is a community of information systems and information technology (IS/IT) professionals, involved in creating information systems in safe, secure, and accessible ecosystems. ISACA continuously recognizes the technology challenges and provides lifetime learning and career development opportunities. ISACA is a leader in career development credentials. The credentialing is a particular way to validate professional knowledge and information communication technology (ICT) skills. ISACA offers the following certificates:

- The Certified Information Systems Auditor (CISA), recognized as a standard of achievement for people who audit, control, monitor, and assess business information technology and business information systems.

- The Advanced in AI Audit (AAIA), which permits IT professionals to recognize the complexities of AI, assess risks, identify opportunities, and ensure compliance with standards. This certification confirms competencies in conducting AI-focused audits, AI integration, and enhancing AI-driven audit processes.
- The Certified Information Security Manager (CISM), which confirms competencies in information security governance, incident management and risk management. This certificate is particularly suitable for senior managers in IT security and control.
- The Advanced in AI Security Management (AAISM), which acknowledges the experience and knowledge of CISM holders concerning AI specific security issues.
- The Certified in the Governance of Enterprise IT (CGEIT), which is to confirm the professional mindset to assess, design, implement, and manage enterprise IT governance systems aligned with general business strategy.
- The Certified in Risk and Information Systems Control (CRISC), which confirms expertise in identification and management of business – IT risk, and implementation and maintaining the information system control. This certificate is suitable for managers focused on IT and cyber risk and control.
- The Certified Data Privacy Solutions Engineer (CDPSE), which is required for successful implementation of privacy by design and by default into IT systems, networks and applications. The privacy control managers are working with software engineers, system and networks engineers, application and databases administrators, and with the project managers.
- The Certified Cybersecurity Operations Analyst (CCOA), which is provided for development of technical skills to evaluate threats, identify vulnerabilities, and prevent cyber incidents.
- The Cybersecurity Practitioner Certification (CSX-P) certification for confirming one's ability to perform globally validated cybersecurity skills covering security functions, i.e., Identify, Protect, Detect, Respond, and Recover derived from the NIST Cybersecurity Framework.

The CISA certification is one of the most sought-after and highest-paying IT certifications. This certification is strongly required by the professionals in the banking sector. The CISA certification addresses innovations like AI and blockchain, and as such is an evidence of human competencies to apply a risk-based approach to audit engagements. The CISA certificate is to confirm competencies to cope with challenges in the following domains: information systems auditing process, governance and management of information technology, governance and management of information technology, information systems acquisition, development, and recommendation, information systems operations and business resilience, and protection of information assets. To the certification, ISACA provides all necessary materials, with the additional guidance and expert instruction. ISACA distributes the online courses,

on-demand instruction and in-depth exam preparation, as well as exam questions, answers, and explanations. ISACA provides a comprehensive reference guide that helps candidates prepare for the CISA exam and permits them to understand their roles and responsibilities as information systems auditors.

The exam course materials are elaborated by subject matter experts and volunteers to align with actual practices and technologies. ISACA distributes course materials in printed and electronic versions. Beyond that, the candidates have opportunities to participate in various training online and in place. Learning based on ISACA course materials is absolutely necessary to pass the exams and receive certificates. Actually, ISACA does not provide AI-supported teaching nor examinations. Exams are conducted online or in place. However, according to ISACA requirements, the CISA certification requires not only passing the CISA exam, but next applying for certification within the five year period of time, five years of IS audit, control, assurance or security work experience, experience in at least one of the five CISA job practice domain areas, and verification of work experience.

Responding to the future needs concerning AI technology usage and learning, an increasing number of professional institutions have recently implemented a model of certifications. The ICT companies as well as professional associations are interested in open-ended and real-life teaching. The market of vocational education institutions is developing, although one of the primary objectives of higher education is to produce specialized human resources with the necessary and actual competencies for the challenges in a complex business environment. Nowadays, the educational processes are supported by AI solutions. The particular place belongs to GenAI, which primarily comprises technologies, such as deep learning, natural language processing (NLP), generative adversarial networks (GANs), and transformer architectures (Al-Thani, Ahmad, 2025).

Mostly, at higher educational institutions (HEIs), GenAI supports administrative work, however, GenAI is increasingly popular in supporting examining processes. For instance, ChatGPT-3.5 and ChatGPT-4o are utilised in the Japanese National Dental Examination, to assess the clinical reasoning skills and dental knowledge to determine their practical usefulness in dental education (Uehara et al., 2025). GenAI is expected to revolutionize medical education. Salman et al. (2025) admit that AI-generated answers exam questions were reviewed by pharmacology experts. These authors admitted that ChatGPT, Copilot, and Gemini demonstrate high accuracy scores for questions on intermediate levels. Yoon et al. (2024) argue that ChatGPT has been tested in health care, including the US Medical Licensing Examination and specialty exams, showing near-passing results. However, these authors emphasize the need for cautious application and further refinement, particularly in non-English medical contexts. Although the AI results are promising, their careful evaluation by human beings is required to ensure acceptable performance. Similar conclusions were formulated by Liu et al. (2024), who revealed that over 2 years, researchers have used various medical licensing examinations to test whether ChatGPT possesses proper medical knowledge. Unfortunately, they concluded

that due to the ChatGPT insufficient accuracy and inconsistent performance, the ChatGPT-4 is not yet suitable for use in medical education. Similar arguments have been provided by Hong et al. (2024), who said that while ChatGPT did not yet meet the passing criteria for the Intermediate Professional Technical Qualification Examination in Ultrasound Medicine, it has potential as a supplementary tool in medical education. Wang et al. (2023) add that although ChatGPT cannot pass the examination, it can be improved quickly through deep learning.

GenAI is expected to be used in various ways in education, e.g., to assess students (Chen et al., 2020; Baidoo-Anu, Ansah, 2023), to provide feedback to students and learners (Chen et al., 2020), for personalized intelligent teaching (Wang et al., 2021), for learning analytics, language correction, for assistance of students with special educational needs (Kalnina et al., 2024). Generally, the GenAI convergence education methods are systematically researched and they will be implemented in the future. Farber (2024) indicated that GenAI complements traditional teaching and examination methods, even providing a personalized and practical learning experience. Fotoh and Mugwira (2025) highlight that ChatGPT can generate coherent and contextually relevant prompts on a wide scope of topics, but it possesses limitations, weaknesses, and risks in terms of security, bias, hallucinations, and accuracy. The weaknesses of the GenAI cover outdated information and insufficient depths in interpretations of the auditing standards (Munoko et al., 2020).

4. Research Methodology

The methodological framework of this study is based on the fact-probing approach introduced by Zhao et al. (2024) and on experiments evaluating LLMs in specialised tests (Huang et al., 2025).

The first step involved obtaining the CISA certification questions in PDF format. One hundred and forty-six test questions (from the eleventh test edition, all in English) were verified for completeness. Their content was then extracted into a JSON structure that captured each question, its possible answers and the correct option. This enabled automated validation of the LLMs' outputs. Custom Python scripts were developed to query seven selected LLMs via the API. These were:

- Chat GPT 4.1 and GPT 3.5 Turbo – developed by OpenAI. GPT 4.1 was released in April 2025 and has been compared with GPT 3.5, which was introduced in 2022.
- Claude Sonnet 4 – created by Anthropic and released in May 2025. It falls into the medium-sized model category, with approximately 50-100 billion parameters (unverified data).

- DeepSeek V3 – developed by the Chinese company DeepSeek AI and first presented in December 2024. It has 671 billion parameters.
- Llama 3 (70B-Instruct) – created by Meta Platforms Inc. (Meta AI). It has 70 billion parameters.
- Gemma 3 (27b-it) – created by Google in March 2025. This model comprise 27 billion parameters.
- Phi 4 – developped by Microsoft in December 2024. It has 14 billion parameters.

For each model, a uniform prompt template containing the question and answer was used. API parameters (temperature, max_tokens and top_p) were chosen based on vendors' recommendations and generation stability literature (Bajpai, Goyal, Anwer, Chakraborty, 2024). To eliminate randomness, each experiment was run five times for each model and the most frequently given answer was taken as the final response.

From the perspective of inter-system communication, access through application programming interfaces (APIs) to appropriate versions of LLM models is a significant characteristic. It is particularly important for the use of LLMs in the cloud computing model. In the initial version of the article, the authors intended to analyse eight models, including two versions of the Gemma model: Gemma 3 (27b-it) and Gemma 2 (27b-it). The results obtained for both versions of the model were identical. Information received from the provider indicated that, rather than disabling access to the deprecated Gemma 2 (27b-it) model, its API was redirecting queries to the Gemma 3 (27b-it) model.

5. Analysis and Results

To examine the selected models, the following quantitative analyses were conducted:

- LLM Accuracy – this analysis evaluates LLMs' accuracy in answering the test questions. The benchmark consisted of 146 single-choice questions, and the average accuracy across the analysed dataset was 81.8%.
- Inter-Agent Agreement Heatmap – this analysis assesses the similarity of responses across different agents.
- A-Bias vs. Accuracy – this analysis examines the relationship between A-Bias and overall accuracy.
- A-Bias in Errors – this analysis measures A-Bias considering only the cases where the LLMs answered incorrectly.

Based on the analysis, the most accurate LLMs were Claude Sonnet 4 and GPT-4.1, both achieving 88.4% correct answers. The lowest performance was observed for GPT-3.5, with 70.3% correct answers. The full results are presented in Figure 1.

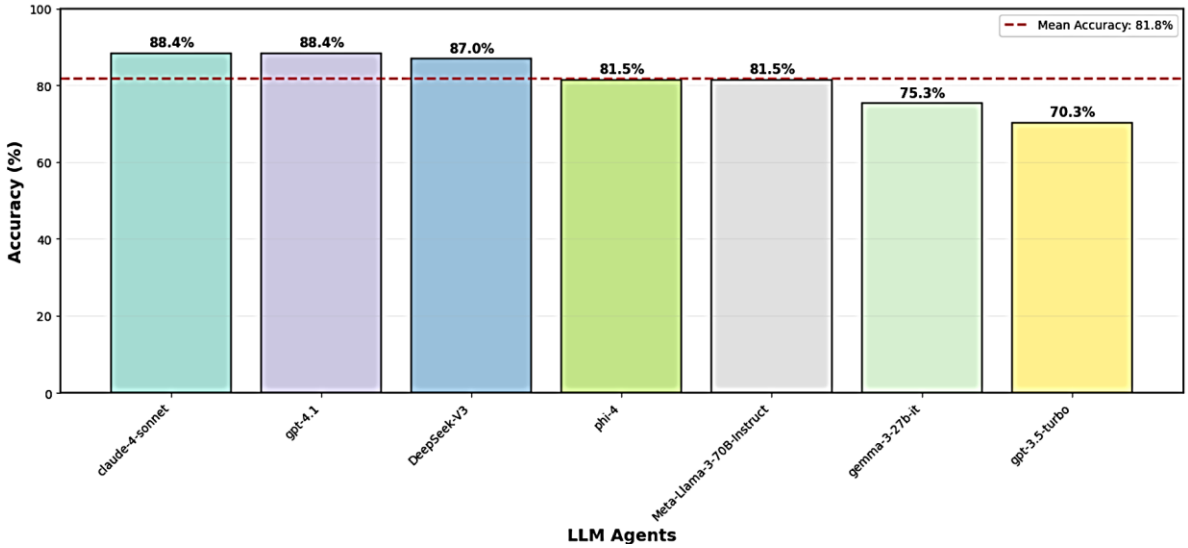

Figure 1. LLM Agents Accuracy Comparison.

Source: own elaboration.

The second analysis was conducted to evaluate the similarity of answers generated by the analysed agents. Figure 2 presents a heatmap illustrating the degree of similarity. The highest similarity scores were observed between Claude Sonnet 4 and DeepSeek V3 (88%), and between Llama 3 70B-Instruct and Claude Sonnet 4 (87%). In contrast, GPT-4.1 and GPT-3.5 showed only 70% similarity.

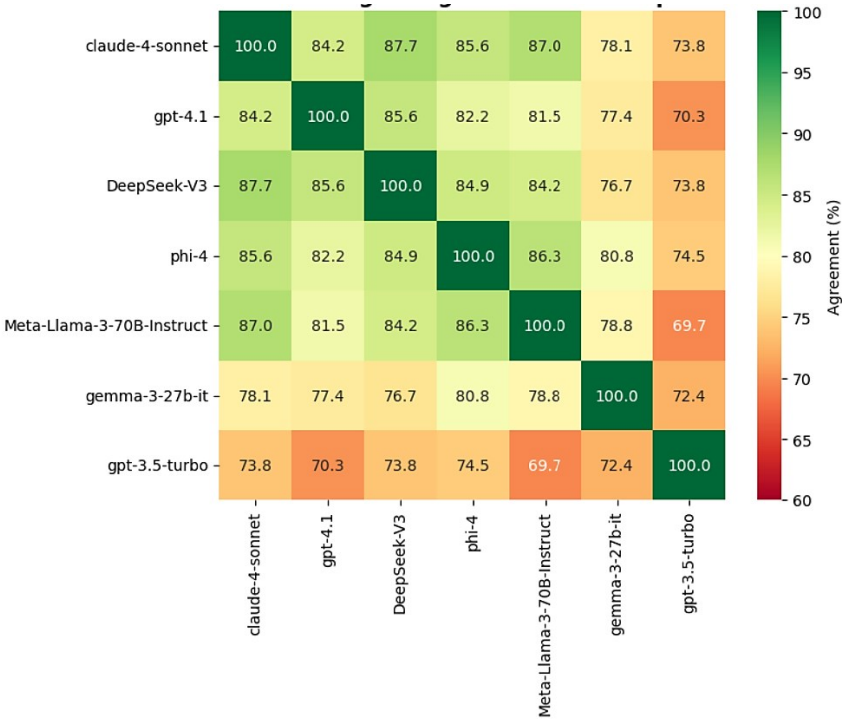

Figure 2. Inter-Agent Agreement Heatmap.

Source: own elaboration.

The next analysis focused on the A-Bias effect. LLMs tend to select option A more frequently when answering multiple-choice questions, especially in cases where they are uncertain about the correct answer. Figure 3 presents a comparison between A-Bias (measured as the deviation from the expected baseline frequency of option A) and model accuracy. For reference, the distribution of correct answers in the analysed dataset was: A – 25.3%, B – 24.0%, C – 29.5%, and D – 21.2%, which serves as the baseline for bias estimation. As shown in Figure 3, Claude Sonnet 4 achieved the highest accuracy with low bias, while Phi-4 demonstrated the strongest A-Bias combined with moderate accuracy.

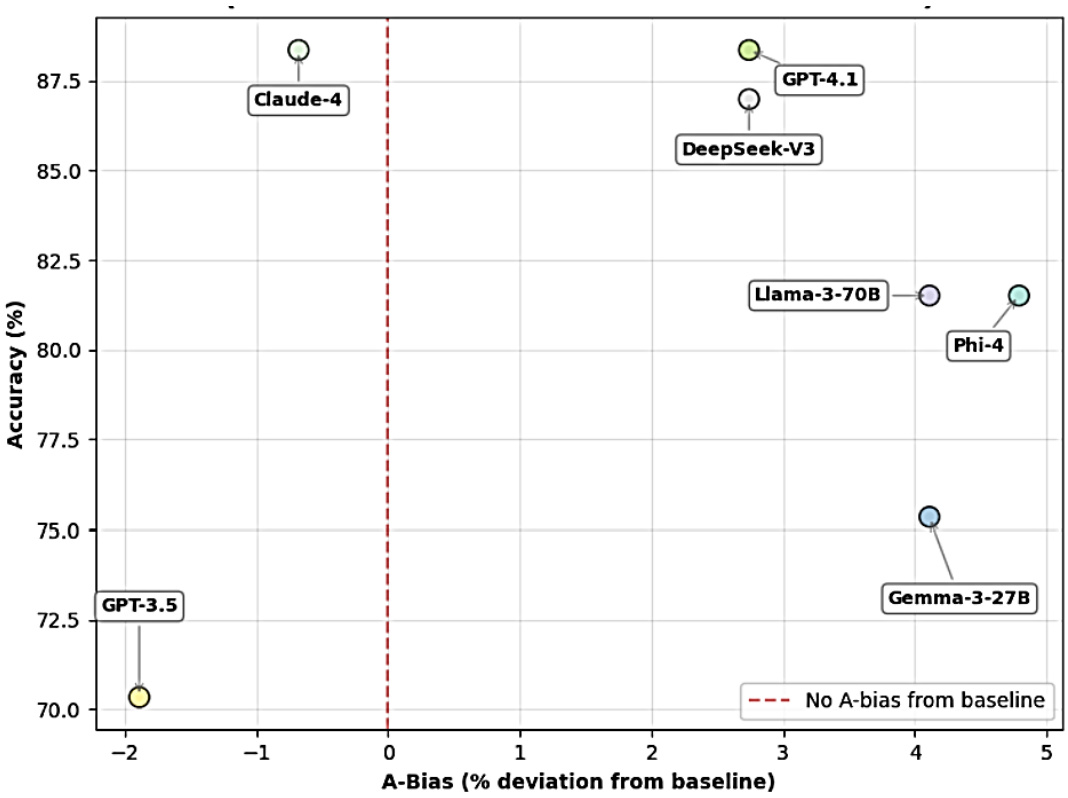

Figure 3. A-Bias vs. Accuracy.

Source: own elaboration.

To better understand the A-Bias problem in the selected LLMs, an additional analysis was performed where A-Bias was calculated only for the incorrect answers (Figure 4).

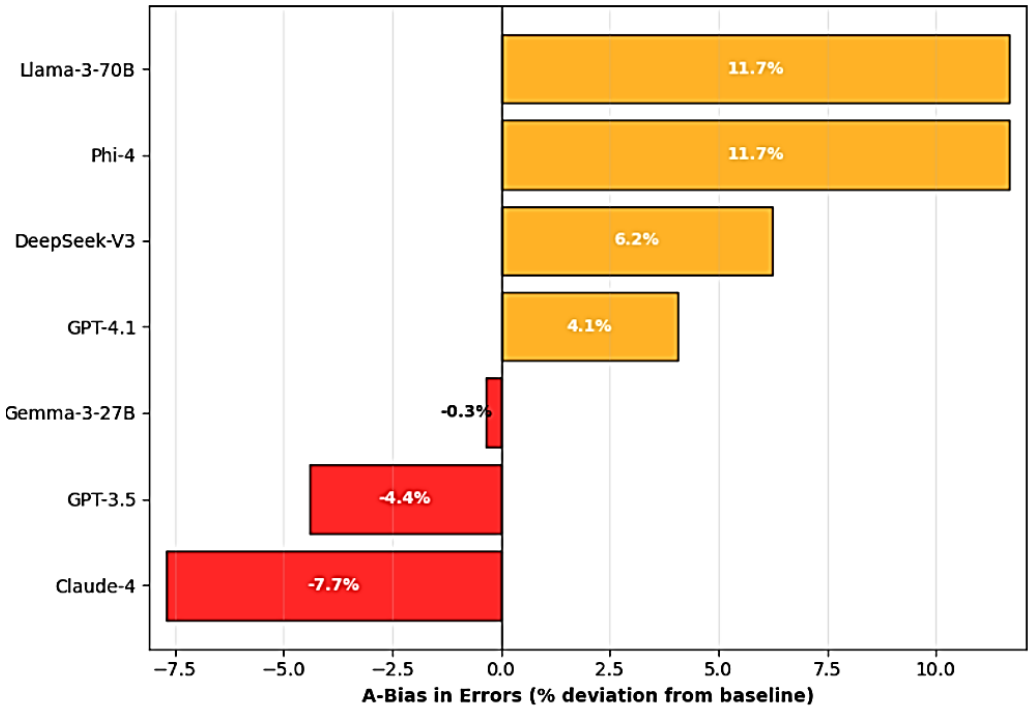

Figure 4. A-Bias in Errors.

Source: own elaboration.

This analysis explains how often agents tend to choose option A when they do not know the correct answer. In this case, Llama 3 70B-Instruct and Phi 4 showed the highest bias values (11.7%). Claude Sonnet 4 and GPT-3.5 exhibit negative A-Bias in errors, indicating that in their incorrect responses, they selected option A less frequently than expected from the baseline distribution.

6. Discussion

Summarising the analyses, 3 out of 7 LLMs achieved accuracy above the overall average. The best-performing models were Claude Sonnet 4 and GPT-4.1. Nevertheless, their answers were similar in 84% of the cases. GPT-4.1 clearly outperformed its predecessor GPT-3.5 (88.4% vs. 70.3%). The analysis further demonstrated that some questions were not answered correctly by any of the LLMs. In total, four such questions were identified:

- Question 38 – "Which of the following forms of evidence would an IS auditor consider the MOST reliable?"
- Question 75 – "Which of the following would an IS auditor use to determine if unauthorised modifications were made to production programs?"
- Question 92 – "Which of the following will MOST successfully identify overlapping key controls in business application systems?"

- Question 126 – "During an IS audit, which is the BEST method for an IS auditor to evaluate the implementation of segregation of duties within an IT department?"

In question 38, all agents answered D – "A confirmation letter received from an outside source," whereas the correct answer was B – "The results of a test performed by an external IS auditor". GPT-4.1 was asked for an explanation, and the authors received the following response:

B. The results of a test performed by an external IS auditor: This is reliable, but not as strong as direct confirmation from an independent third party. D. A confirmation letter received from an outside source: This is the most reliable because it is independent, objective, and comes from a third party. External, independent evidence is always considered the most reliable in auditing.

This outcome indicates that the LLM assigns greater reliability to a confirmation letter than to the results of procedures performed by an external auditor.

In question 75, all agents answered A – "System log analysis", whereas the correct answer was B – "Compliance testing". GPT-4.1 was asked again for an explanation, and the authors received the following response:

System log analysis involves reviewing logs that record changes to production programs, such as version control logs, change management logs, or audit trails. These logs can show who made changes, when, and what was changed, helping the auditor identify unauthorized modifications. Compliance testing checks if controls and procedures are being followed, but does not directly identify unauthorized changes.

The incorrect answer may have resulted from the models' tendency to focus on technical solutions rather than on audit procedures.

In question 92, all agents answered A – "Reviewing system functionalities that are attached to complex business processes", whereas the correct answer was C – "Replacing manual monitoring with an automated auditing solution". GPT-4.1 was subsequently asked again for an explanation, and the authors received the following response:

Reviewing system functionalities that are attached to complex business processes. This approach allows you to map out where controls exist within the business application and identify if multiple controls are addressing the same risk (i.e., overlapping controls). By understanding the functionalities and their relationship to business processes, you can see where controls may duplicate or overlap. C. Replacing manual monitoring with an automated auditing solution This is a method of monitoring, not of identifying overlapping controls. The most successful way to identify overlapping key controls is to review system functionalities attached to complex business processes (Option A), as this provides a holistic view of where controls are implemented and if there is redundancy.

The incorrect answers were likely caused by excessive keyword matching and overreliance on semantic similarity between the question and the answer.

In question 126, two agents answered D – "Evaluate the organizational structure" (GPT-3.5 and Claude Sonnet 4) and five answered B – „Review the job descriptions of the IT functions" whereas the correct answer was A – "Discuss it with the IT managers". GPT-4.1 was re-prompted for an explanation, and the authors received the following response:

Reviewing job descriptions allows the auditor to see the specific roles, responsibilities, and access levels assigned to each position. This directly reveals whether critical duties are appropriately segregated to prevent conflicts of interest, fraud, or error. Discussing with IT managers (A) may provide insights, but it is subjective and may not reflect actual practices or documented responsibilities.

The incorrect answer likely resulted from the model's excessive reliance on textual information sources and its lack of deeper understanding of the audit context.

The application of LLMs in examination processes is mostly noticed at medical science schools. Medical science teaching and learning require systematic studying, therefore maybe other disciplines, i.e., chemistry, food production, or biology will be soon supported by AI in the examination process. Auditing is developed in management and computer science studies, but in these disciplines application of AI in examination is not popular. The ISACA association is observed as a rather traditionally thinking people society. During the ISACA conferences and summits in Europe, in other continents' countries, as well as in the ISACA Headquarters in Chicago, USA, auditors discuss the impact of AI on the auditing processes as well as on other business processes. They promote rational actions and regulative procedures to manage risks and challenges connected with an implementation of AI and LLMs in and for business companies. The ISACA experts work on management of models and reports generated by the GenAI solutions. These issues are included in the certificated exam questions. However, on the other side, ISACA does not offer AI-supported courses. The teaching and examinations are conducted in a traditional way. Beyond the courses based on the ISACA original course materials, there are plenty of small business enterprises (SMEs), who offer materials, lecturing, and exam preparation. That business units are usually located in the cities, where ISACA chapter boards are placed. Hence, the provided course teaching is a source of additional support to the local chapter budget. Some materials needed to take the exam, i.e., CISA, CISM etc. are available for ISACA members on the Internet for free. Availability of course questions and answers is always an opportunity for the Internet bots to collect data and use in the LLMs. From one side, it is very good for knowledge popularization and making it easier to pass the exams and receive the certificates, however, from the other side the knowledge is not hidden as the internal knowledge of the ISACA chapters. Therefore, there is a risk that this inter-society developed knowledge will soon be freely revealed. The interpretation of the certificate exam answers also belongs to the ISACA experts, but also here, there is a risk that this information will soon be available on the Internet. The ISACA exam passing is not a trivial issue, therefore the learners are constantly looking for available materials to prepare for exams. In their own interest, the ISACA can soon implement their own GenAI bots supporting the auditing and security knowledge and information about the learners capabilities and competencies.

7. Conclusions

The obtained results confirmed the high level of knowledge demonstrated by the examined large language models (LLMs) in the field of information systems auditing. The performance of smaller models, such as Phi 4, did not differ significantly from that of the larger ones. In the case of GPT, a clear improvement in answer quality was observed between versions 4.1 and 3.5. Most LLMs exhibited the A-bias effect reported in the literature. A detailed analysis of the responses revealed four questions that require further investigation. The next stage of the presented research may involve building a society of agents, in which the system's response would be generated based on the cooperation of several models simultaneously.

This research revealed some issues important from the point of view of AI usage for teaching auditors. There is the opportunity to compare publicly available agents' answering results as well as to develop a system for various agents' deliberation and concluding. On the other side, there is still an open question as to how to provide ISACA courses and how to distribute learning materials. For sure that association needs to monitor the risk of unprotected distribution of course materials. Successfully, ICT development encourages ISACA to constantly renew the courses' materials, so questions and answers (Qs&As) are changeable. Therefore, there is the question of revealing or not revealing the questions and answers. There are business organizations, which enable learners access to the exam answers, because they believe that the final exam is the rational verification of the acquired competences. The exam for driving licence is the best example in this case. However, there are also organizations, which argue that new sets of questions and answers are to be better and better protected, because the association wants to control the examination process.

References

1. Al-Thani, N.J., Ahmad, Z. (2025). Learning through "research cognitive theory": A new framework for developing 21st-century research skills in secondary school students. *Heliyon, 11(2),* e41950. https://doi.org/10.1016/j.heliyon.2025.e41950
2. Baidoo-Anu, D., Ansah, L.O. (2023). Education in the era of generative artificial intelligence (AI): understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI, 7*, 52-62, doi: 10.61969/jai.1337500
3. Bajpai, A., Goyal, A., Anwer, A., Chakraborty, T. (2024). *Temporally consistent factuality probing for large language models.* Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), 15864-15881. https://doi.org/10.18653/v1/2024.emnlp-main.887

4. Chen, L., Chen, P., Lin, Z. (2020). Artificial intelligence in education: a review. *IEEE Access, 8*, 75264-75278, doi: 10.1109/ACCESS.2020.2988510
5. Culot, G., Nassimbeni, G., Podrecca, M., Sartor, M. (2021). The ISO/IEC 27001 information security management standard: Literature review and theory-based research agenda. *The TQM Journal, 33(7),* 76-105. https://doi.org/10.1108/TQM-09-2020-0202
6. Deagama Antariksa, M., Perangin Angin, M., Widodo, A.P. (2025). COBIT 2019 framework in IT governance: A systematic literature review of implementation challenges and benefits across various industry sectors. *Journal of Renewable Energy, Electrical, and Computer Engineering, 5(1),* 99-105. https://doi.org/10.29103/jreece.v5i1.19501
7. Farber, S. (2024). Harmonizing AI and human instruction in legal education: a case study from Israel on training future legal professionals. *International Journal of the Legal Profession, 31(3)*, pp. 349-363, https://doi.org/10.1080/09695958.2024.2430018
8. Fotoh L.E., Mugwira, T. (2025). Exploring Large Language Models in external audits: Implications and ethical considerations. *International Journal of Accounting Information Systems, Vol. 56,* 100748, https://doi.org/10.1016/j.accinf.2025.100748
9. Herrmann-Werner, A., Festl-Wietek, T., Holderried, F., Herschbach, L., Griewatz, J., Masters, K., Zipfel, S., Mahling, M. (2024). Assessing ChatGPT's mastery of Bloom's taxonomy using psychosomatic medicine exam questions: Mixed-methods study. *Journal of Medical Internet Research, 26(1),* e52113. https://doi.org/10.2196/52113
10. Hong, D.-R., Huang, C.-Y. (2024). The performance of AI in medical examinations: an exploration of ChatGPT in ultrasound medical education. *Frontiers in Medicine, 11,* 1472006, doi: 10.3389/fmed.2024.1472006
11. Huang, C.-H., Hsiao, H.-J., Yeh, P.-C., Wu, K.-C., Kao, C.-H. (2024). Performance of ChatGPT on Stage 1 of the Taiwanese medical licensing exam. *Digital Health, 10.* https://doi.org/10.1177/20552076241233144
12. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems, 43(2), Article 42*. https://doi.org/10.1145/3703155
13. Kalnina, D., Nimante, D., Baranova, S. (2024). Artificial intelligence for higher education: benefits and challenges for pre-service teachers. *Frontiers in Education, 9*, 1501819. doi: 10.3389/feduc.2024.1501819
14. Khraisha, Q., Put, S., Kappenberg, J., Warraitch, A., Hadfield, K. (2024). Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Research Synthesis Methods, 15(4),* 616-626. https://doi.org/10.1002/jrsm.1715
15. Lin, S., Hilton, J., Evans, O. (2022). *TruthfulQA: Measuring How Models Mimic Human Falsehoods.* Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022), 3211-3225. https://doi.org/10.18653/v1/2022.acl-long.229

16. Liu, M., Okuhara, T., Chang, X., Shirabe, R., Nishiie, Y., Okada, H., Kiuchi T. (2024): Performance of ChatGPT Across Different Versions in Medical Licensing Examinations Worldwide: Systematic Review and Meta-Analysis. *Journal of Medical Internet Research, vol. 26,* e60807, doi: 10.2196/60807
17. Munoko, I., Brown-Liburd, H.L., Vasarhelyi, M. (2020). The ethical implications of using artificial intelligence in auditing. *Journal of Business Ethics 167(2),* 209-234. https://doi.org/10.1007/s10551-019-04407-1
18. Salman, I.M., Ameer, O.Z., Khanfar, M.A., Hsieh, Y.-H. (2025). Artificial intelligence in healthcare education: evaluating the accuracy of ChatGPT, Copilot, and Google Gemini in cardiovascular pharmacology. *Frontiers in Medicine, 12,* 1495378, https://doi.org/10.3389/fmed.2025.1495378
19. Senft, S., Gallegos, F. (2009). *Information Technology Control and Audit*. Boca Raton: CRC Press, Taylor and Francis.
20. Uehara, O., Morikawa, T., Harada, F., Sugiyama, N., Matsuki, Y., Hiraki, D., Sakurai, H., Kado, T., Yoshida, K., Murata, Y., Matsuoka, H., Nagasawa, T., Furuichi, Y., Abiko, Y., Miura, H. (2025). Performance of ChatGPT-3.5 and ChatGPT-4o in the Japanese National Dental Examination. *Journal of Dental Education, 89(4),* pp. 459-466, doi: 10.1002/jdd.13766
21. Wang, S., Wang, H., Jiang, Y., Li, P., Yang, W. (2021). Understanding students' participation of intelligent teaching: an empirical study considering artificial intelligence usefulness, interactive reward, satisfaction, university support and enjoyment. *Interactive Learning Environments, 31,* pp. 5633-5649.
22. Wang, Y.-M., Shen, H.-W., Chen, T.-J. (2023). Performance of ChatGPT on the pharmacist licensing examination in Taiwan. *Journal of the Chinese Medical Association, 86(7),* pp. 653-658.
23. Yoon, S.-H., Oh, S.K., Lim, B.G., Lee, H.-J. (2024). Performance of ChatGPT in the In-Training Examination for Anesthesiology and Pain Medicine Residents in South Korea: Observational Study. *Jmir Medical Education, 10*, e56859, doi: 10.2196/56859
24. Zhao, X., Yoshinaga, N., Oba, D. (2024). *What matters in memorizing and recalling facts? Multifaceted benchmarks for knowledge probing in language models*. Findings of the Association for Computational Linguistics: EMNLP 2024, 13186-13214. https://doi.org/10.18653/v1/2024.findings-emnlp.771