South African Journal of Economic and Management Sciences
ISSN: (Online) 2222-3436, (Print) 1015-8812

Examining different artificial intelligence models' ability to pass Certificate of Theory in Accountancy-level tax questions

Authors:
Asheer J. Ram[1]
Wayne van Zijl[1]

Affiliations:
[1]Margo Steele School of Accountancy, Faculty of Commerce, Law and Management, University of the Witwatersrand, Johannesburg, South Africa

Corresponding author:
Asheer Jaywant Ram,
asheer.ram@wits.ac.za

Dates:
Received: 12 June 2025
Accepted: 20 Nov. 2025
Published: 23 Jan. 2026

How to cite this article:
Ram, A.J. & Van Zijl, W., 2026, 'Examining different artificial intelligence models' ability to pass Certificate of Theory in Accountancy-level tax questions', *South African Journal of Economic and Management Sciences* 29(1), a6348. https://doi.org/10.4102/sajems.v29i1.6348

As Artificial Intelligence (AI) models become more sophisticated and entrenched in accountancy professions, this raises questions about their ability to outperform humans. This article is one of the first to examine the ability of five different AI models to pass professional tax examinations.

Contribution: This article provides evidence about AI's current ability to support or replace tax practitioners. It provides a baseline to track the progress of different AI models as they evolve. Only Grok passed, while ChatGPT, Claude, CoPilot, and Gemini failed. Notably, the AI models provided persuasive answers despite being incorrect, negating their ability to replace tax practitioners.

Keywords: artificial intelligence; education; ChatGPT; Claude; Copilot; Gemini; Grok; taxation.

Introduction

Artificial intelligence (AI) is rapidly becoming commonplace in our lives. As part of this, different professions are grappling with the extent to which AI supports or replaces tasks historically carried out by their professionals. The accounting profession is no different, and a key area of professional advice stems from the tax practitioner community. Tax practitioners are often a necessity for businesses and individuals alike because of the complexity and constantly changing nature of tax. It is currently unclear to what extent different AI models can support or replace tax practitioners. This article addresses this gap by evaluating different AI models' ability to cope with tax queries through a qualitative analysis.

Tax examination questions written by final-year Certificate of Theory in Accountancy (CTA) students studying for their Chartered Accountant (South Africa) (CA[SA]) designation were given to five different AI models. Chartered Accountant (South Africa) is a designation conferred by the South African Institute of Chartered Accountants (SAICA). The SAICA is regularly lauded as the number one chartered accountant institute in the world (SAICA 2023). Students who seek to become SAICA members need to complete an approved undergraduate degree and a postgraduate degree (commonly referred to as CTA) (Van Wyk 2011). As part of this, they need to complete advanced taxation studies to complete their CTA. The CTA examinations do not have any multiple-choice questions (MCQs) or true or false questions, as they focus on detailed calculations and interpretive-essay-type questions. This is to simulate what work one would perform as a tax practitioner.

To be a tax practitioner in South Africa, one must belong to a professional body that is a recognised controlling body by the South African Revenue Service (SARS 2024). The SAICA is one such body. As a result, there is an important link between tax practitioners and their education and one would expect that tax practitioners are able to appropriately handle CTA tax questions. Tax practitioners support businesses and natural persons in ensuring that they comply with tax laws and regulations. Any tax practitioner would also need to comply with relevant levels of knowledge and due care as per the professional requirements of their recognised controlling bodies.[1] This article asks the question: *Can current AI models reliably provide tax advice at the level expected of professional tax practitioners?*

This article makes multiple important contributions. Firstly, the article evaluates how well different AI models are able to pass a real-world tax examination that requires different skill

1.The SAICA code of Professional Conduct (SAICA 2024) and the South African Institute of Taxation (SAIT) Code of Conduct (SAIT 2025) require professional competence and due care of members.

sets in comparison to students studying to become CA(SA)s. Secondly, the article sets an important benchmark that can be used in the future to evaluate the progress of AI models in providing tax-related business advice. Finally, the article provides important insights from running the experiment and builds on the limited non-United States (US) research evaluating AI's ability to support and/or replace accounting professionals.

Literature review

Artificial intelligence models

Large language models (LLMs) are designed to simulate conversations, with users entering natural-language prompts and questions to receive human-like responses (Borger et al. 2023; Teubner et al. 2023). However, the response is not generated by a free-thinking human. It is generated using the probabilities and patterns inherent in the data used to train the respective AI models, reflecting the importance and impact of the training data provided. Different models use different approaches. For example, some focus on large quantities of unspecified data and often use the internet as a primary source (Roberts, Baker & Andrew 2024). Others may follow a more focused strategy and selectively feed it information that facilitates its knowledge base in a particular field, discipline, or function. Regardless of the strategy, extensive neural networks act as a system for working through and making sense of the training data (Borger et al. 2023; Roberts et al. 2024).

An AI model's training will dictate its strengths and weaknesses. Artificial intelligence models trained in specific tasks will be better at those tasks. This comes at the expense of handling more general, perhaps conversational tasks. Other models are trained on a wide variety of information from different disciplines and functions. These models may perform better in what is known as zero-shot performance. This refers to an AI model's ability to be given an entirely new task in which it has not been specifically trained (Eulerich et al. 2024).

Considering the models used in this article, ChatGPT, being trained on OpenAI's proprietary data and web data, is considered more of a conversational AI well-suited for content creation. Gemini, being trained on Google-scale datasets (text, images, audio, video), appears to be aimed towards multimedia content creation, cross-modal analysis, and multimodal AI tasks. Claude, trained in public web content, open-source code, documentation, and books, is suitable for programming, in-depth scientific studies and lengthy essays. Copilot, trained on a vast amount of code from public repositories, is useful for generating efficient solutions to algorithmic problems (Jabbar, Ul Islam & Boudjadar 2025). Grok, trained on X.com (formerly Twitter) data, is suited to analysing trends and interactions on social networks (De Carvalho Souza & Weigang 2024).

Artificial intelligence models suffer from four key concerns. Firstly, the extent to which AI models hallucinate is considered. When AI models provide an answer despite lacking the requisite knowledge and capabilities to form one, the output may be a hallucination. This is a difficult area, as there is a trade-off. On the one hand, allowing AI models to be creative enhances their ability to solve problems in new ways, thereby leveraging the benefits AI can offer. On the other hand, creativity increases the likelihood that AI models will provide false answers, with no indication to the user that the response may be a hallucination or its degree of uncertainty or confidence in its answer (Dahl et al. 2024).

Secondly, AI models typically lack the ability to discern truth from untruth and bias (Roberts et al. 2024). The training material given to a model is, consequently, a significant factor in determining the credibility and potential bias of the AI model's output. Using general internet data as training material may, accordingly, have negative implications for a model's ability to provide credible and reliable solutions to queries in, for example, highly regulated disciplines. Thirdly, Roberts et al. (2024) also raised the idea of AI models being 'overconfident'. Said differently, AI models may *want* to provide you with an answer, whether or not they can do so. This may result in hallucination to satisfy the user's request. Finally, Borger et al. (2023) and Roberts et al. (2024) raise important concerns about an AI model's likelihood of reinforcing stereotypes, especially when trained on biased data and biased users accept these outputs.

Research on the use of artificial intelligence models to answer assessments

There is limited research into AI models' ability to answer formal assessments. Many focus only on MCQs or on a single AI model, leaving two major gaps: Firstly, AI's ability to address essay and discipline-specific assessments and secondly, there is no comparison of different AI models' ability to handle the same assessments that evaluate different types of assessment components.

In one of the first articles to consider the efficacy of AI in answering accounting assessment questions, Bommarito et al. (2023) evaluated the performance of OpenAI's earlier versions of ChatGPT and Text-davinci-003 on the uniform CPA examination written in the US. Text-davinci-003 achieved a correct rate of 14.4% on Regulation examination questions, underperforming human capabilities on quantitative reasoning in zero-shot prompts. The model achieved human-level performance in remembering, understanding, and application, answering 57.6% of questions correctly. ChatGPT 3 outperformed the text-davinci models.

Wood et al. (2023), also with a US focus, evaluated ChatGPT 3's performance on accounting questions in comparison to students' performance. ChatGPT achieved an average of 56.5% with partial credit[2], significantly underperforming in comparison to students' average score of 76.7%. ChatGPT

2.Partial credit means that points are assigned to very close and moderately close answers in MCQ questions, as opposed to only awarding points to a fully correct answer (Schneid, Armour & Brandl 2025).

performed better on true or false (68.7%) and MCQs (59.5%) but struggled with calculation (28.7%) and short-answer questions (39.1%). ChatGPT performed relatively well in accounting information systems and auditing assessments but had lower accuracy in taxation, financial, and managerial accounting questions. Higher-order learning questions posed a challenge to ChatGPT. Notably, 'ChatGPT struggled to handle long, written questions with multiple parts, even when allowing for "carry over" mistakes' (Wood et al. 2023:15). ChatGPT often 'made up' facts and provided descriptive explanations for its answers, even if incorrect, which can easily, but incorrectly, convince AI users of AI's correctness.

Atanasovski et al. (2023) explored the effectiveness of ChatGPT 3.5 in answering examination questions in accounting and auditing in North Macedonia. The research involved 11 subject examinations with a total of 401 questions. ChatGPT 3.5 successfully passed 8 out of the 11 subjects, achieving a pass rate of 73%. For true or false questions, ChatGPT had a 65% correct response rate. For MCQs with a single correct answer, ChatGPT 3.5 achieved a 72% correct response rate, while for MCQs with multiple correct answers, ChatGPT 3.5 only achieved a 48% correct response rate. This is similar to a study in Portugal where ChatGPT failed to pass the Portuguese Order of Chartered Accountants examination (Albuquerque & Dos Santos 2024). In open-ended short questions, a strong performance was demonstrated with a 78% correct response rate. However, for essay questions, ChatGPT 3.5 earned a 55% score.

ChatGPT 3.5 excelled in subjects such as Principles of Accounting, Auditing, and Internal Auditing but struggled with Management Accounting II, Governmental Accounting, and International Accounting. The study concludes that while ChatGPT is proficient in qualitative questions and simpler MCQs, it faces challenges with quantitative calculations, complex MCQs and essay questions (Atanasovski et al. 2023). Given the nature of the CTA examination as discussed in the introduction, one would expect ChatGPT to struggle with passing.

Later, Eulerich et al. (2024) examined ChatGPT's performance in the US CPA and other accounting certification examination (CMA, CIA, EA). They find that ChatGPT 3.5 scored an average of 53.1% across all examinations, failing to pass any[3]. However, ChatGPT 4 improved its scores by 16.5% after 10-shot training[4]; it achieved an average score of 85.1%, passing all the areas. Their work concurs with that of Bommarito et al. (2023), evidencing that more recent versions of ChatGPT perform better than their older counterparts. However, zero-shot attempts still showed issues with passing examinations.

Following a legal perspective, Katz et al. (2024) found that ChatGPT 4 substantially outperformed human students and previous ChatGPT models in the US Bar examinations. ChatGPT achieves roughly 297 points, well above the passing threshold for all universal Bar jurisdictions. On the multistate Bar examination, which consists solely of MCQs, ChatGPT 4 achieves a 75.7% accuracy rate, outperforming the average human test-taker by more than 7% and demonstrating a 26% increase over previous ChatGPT versions. On the multistate essay examination, ChatGPT 4 scored 4.2 out of 6 points, with ChatGPT 3 scoring 3 out of 6 points. A passing grade is considered 4 out of 6 points, which indicates that ChatGPT 4 was able to marginally pass the multistate essay examination. The decrease in its score from the MCQs is noticeable, indicating once more that essay- and discussion-type questions prove more challenging for the AI model.

Cheng et al. (2024) explored the capabilities of ChatGPT 3.5 and 4 to answer educational accounting cases in the US. They found that ChatGPT 4 performs better than ChatGPT 3.5, especially in tasks requiring explanation, application of rules, and ethical evaluation. However, both ChatGPT 3.5 and ChatGPT 4 struggled with tasks requiring financial statement creation, journal entries, or software use. As indicated in the previous studies, much of the research has been centred in the US (Bommarito et al. 2023; Eulerich et al. 2023; Katz et al. 2024; Wood et al. 2023).

Pinto et al. (2024) compare and contrast different AI models. They evaluate the performance of ChatGPT 3.5, ChatGPT 4, and Gemini in the Portuguese Chartered Accountant Examination. This examination consists solely of MCQs. With an average accuracy for tax questions of 48%, ChatGPT 4 outperformed Gemini (38%) and ChatGPT 3.5 (36%). All these models failed, reinforcing the findings of Albuquerque and Dos Santos (2024), who found that ChatGPT struggled with tax questions where judgement was required. The AI models struggled most with management and financial accounting questions, but performed better in taxation and ethics. Interestingly, this result in respect of taxation is contrary to Wood et al. (2023). In further contrast, the USA's tax system (22nd most complex out of 64 countries) is regarded as being less complex than that of Portugal (17th most complex out of 64 countries) (Tax Complexity Index 2022). Consequently, one would expect the AI models to perform better in the US than in Portugal. However, the Portuguese examination, consisting only of MCQs (Albuquerque & Dos Santos 2024; Pinto et al. 2024), demonstrates, again, that ChatGPT performed better in the MCQs compared to the written question considered by Wood et al. (2023). As the South African CTA tax examinations do not include any MCQs, it is expected that these models will struggle with the written and complex calculations required. These studies clearly demonstrate that, while AI models show potential, they are not yet capable of consistently passing rigorous professional examinations without further training and improvement.

Most existing studies have focused on MCQ questions or assessments that include both essays and MCQs. The South African approach to CTA tax is very different, with no MCQs used at all. Instead, higher-order thinking and critical

3.These examinations do not use a 50% pass mark. Rather, they use thresholds for various sections and parts, and the 53.1% does not meet the passing threshold in any section.

4.Ten-shot training means that the model is trained ten times on a small dataset containing only a few examples. The aim is to assist the model in applying what it has learnt to more generalisable, unseen examples.

evaluation skills via complex calculation and application-type questions are used. Accordingly, this article contributes to the existing literature and evaluates different AI models' ability to handle different types of questions. How this was achieved is discussed next.

Method

The article takes an exploratory descriptive empirical research approach to assessing the tax capabilities of the AI models. The article posed one discussion theory-based tax question and one numerical tax question to five different AI models. The purpose was to assess their ability to pass a final-year exit-level examination at a SAICA-accredited South African university with a zero-shot prompt. Because of the lengthy limitations of most free AI models (AI for Education 2025), the paid versions of ChatGPT 4, Claude, Copilot, Gemini, and Grok were used. This also ensures that the AI models provide their best possible answers to examination questions. In addition, the latest AI models as of June 2025 were used. The article did not opt to use older models that have been assessed in prior articles, as the purpose is not to assess any particular version but rather the latest and best AI capabilities at present.

The only prompts given to each AI were the purpose of the study, the scenario, and the two tasks required to be completed (as given to 2024 students). The question paper with the scenario information and the Table 1 tasks was uploaded to each AI model. There was no attempt to alter or amend the information or tasks to possibly assist readability by the AI models. This is a first zero-shot look at the ability of the AI models to respond to these tax questions with no training or prior preparation, to better simulate how one may use an AI model to replace a tax practitioner.

Given the localised nature of the tax examination information and tasks provided (Table 1) and that this was an in-person written examination, along with the fact that this examination was set new and never used previously, it was not considered a risk that any of the data had leaked into the training data used in the AI models. Considering this, no contamination checks were run on the AI models (see Katz et al. 2024).

TABLE 1: Tasks required to be performed provided to the artificial intelligence models.

Number	Description	Marks
1.	**Refer to the section titled 'The case of Melusi Gwabe'.** You are required to assist O'lerato Gwabe with the determination of the Estate Duty payable amount in respect of her late father, Melusi Gwabe. Using only the information provided in the scenario, draft a list of questions which you would need to ask O'lerato Gwabe to accurately and completely assist with completing the Estate Duty payable calculation. For each question, provide a reason as to why you are asking that question. Supporting calculations are not required. References to any relevant legislation are not required.	12
2.	**Refer to the section titled 'The case of Melusi Gwabe'.** Assume that the bank account had a balance of R428 000 invested, the BMW 1-series had a market value of R380 000 and the shares in the unlisted company had a market value of R240 000 on the date of Melusi's death. Furthermore, assume the following: • The funds in the bank account and the BMW 1-series were bequeathed to Zinhle Gwabe, • The unlisted company shares were bequeathed to O'lerato Gwabe, • All the transfers took place in February 2024. Calculate the taxable income of the deceased estate of the late Melusi Gwabe for the 2024 year of assessment. Provide reasons for nil amounts.	11

The same solution as that used for the 2024 students was used, and the same marker of the 2024 students graded each AI model's answer. This enhances the validity of the comparison between the AI's marks and the marks of the students. The specific tasks provided to each AI model are detailed in Table 1.

A qualitative approach is taken to assess the quantitative results of the AI models in responding to Table 1 tax tasks. This research approach is consistent with other AI studies in this area (Atanasovski et al. 2023; Bommarito et al. 2023; Eulerich et al. 2023; Wood et al. 2023). As an exploratory study, the aim is not to present a generalisable positivist conclusion, but to provide insight into the real-world ability of AI models to handle tax questions.

Results

Table 2 presents the results. Notably, four of the five models failed overall[5], with Grok standing out by passing with 56.52% (and exceeding the next best mark by 17.39% points). As most AI models are LLMs, the intuitive expectation would be that the AI models would perform better in the discussion questions. This was not generally the case and supports the findings of Wood et al. (2023) and Atanasovski et al. (2023). Our results reflect mixed outcomes, with three models performing better in the calculation-style questions and the other two achieving better discussion question results.

The students' pass rates of 75.86% and 72.80%, respectively, for the discussion (question 1) and calculation (question 2) questions are high. This suggests that your average entry-level tax practitioner should be able to cope with them relatively easily. Similarly, the students' average mark for both questions was approximately 63%. Given the intensive training and knowledge required for tax practitioners, as discussed in the introduction, this reinforces that the questions posed to the AI models were reasonable and not so complex as to indicate that only very experienced tax practitioners would be able to cope with them.

In the discussion component, nothing stood out between the AI models' quality of language and that of the students. Both were fairly professional and on topic. A key insight gained is that where Grok simply supplied its answer, all other AI models sought to convince the reader that their answer was justified and correct, supporting the findings of Wood et al. (2023). The important message this finding sends is that these AI models make it particularly difficult for users to gauge the credibility of their advice and solutions. In contrast, a critical distinction exists in professional practice. While tax practitioners may provide exploratory explanations during initial client consultations,

5.Failed, in a South African university context, refers to where a mark of less than 50% overall was achieved.

TABLE 2: Results from the artificial intelligence models.

AI model (in alphabetical order)	Discussion (question 1) score achieved (%)	Calculation (question 2) score achieved (%)	Total for Table 1 tasks (%)	Deviation of the total percentage from the student average
ChatGPT 4	16.67	27.27†	21.74	-41.26
Claude	50.00†	27.27	39.13	-23.87
Copilot	16.67	18.18†	17.39	-45.61
Gemini	33.33†	18.18	26.09	-6.91
Grok	50.00	63.64†	56.52	-6.48
Student average	63.06	62.94	63.00	-
Student pass rate	75.86	72.80	74.33	-

AI, artificial intelligence.

†, Values show the highest component (discussion or calculation) for each AI model.

professional conduct codes (as discussed in the introduction) require that formal tax advice (such as tax return positions or formal opinions) only be provided when the practitioner has sufficient knowledge and certainty. Artificial intelligence models do not make this contextual distinction, providing equally persuasive answers regardless of whether they represent exploratory discussion or formal advice. Said differently, it would take an up-to-date expert to determine whether these AI models' answers are correct, negating their value as providers of tax support and advice.

A noticeable issue between AI models' responses and those of students is that, typically, students who do not know the answer will have short, if any, solutions provided. Artificial intelligence models provided full solutions, despite the marks clearly indicating that they did not have the requisite knowledge and capabilities to correctly provide solutions. This can be linked to current concerns about AI models' hallucinations, reducing their value for tasks where correct answers, without any creativity on the AI model's part, are required (Dahl et al. 2024; Roberts et al. 2024). This may suggest that, should businesses and individuals want AI models that can support or replace tax advice, AI models with limited scope for creativity may be preferable.

The results, in a tax sense, align with Wood et al. (2023) but are contrary to Pinto et al. (2024). This raises the point that the AI models may be able to do better with respect to tax in certain jurisdictions compared with others. South Africa's tax system is ranked as the 45th most complex out of 64 countries[6] (Tax Complexity Index 2022), which is a lower complexity than the tax systems of Portugal (17th most complex out of 64 countries) (Tax Complexity Index 2022) (see Pinto et al. 2024) and the US (22nd most complex out of 64 countries) (Tax Complexity Index 2022) (see Wood et al. 2023). Accordingly, one would have expected the AI models to perform better given the lower SA tax complexity, but this does not seem to be the case, suggesting jurisdiction-specific training data used in the AI models or complexity interpretation issues. Furthermore, Pinto et al. (2024) showed that ChatGPT outperformed Gemini, contrary to this study, which found that Gemini outperforms ChatGPT by 4.35% points. Given the mixed nature of these findings, what is clear is that no AI model can confidently be used to support or replace tax practitioners at present.

6.Where a country ranked 1st represents the most complex tax system.

Conclusion

Overall, the results, especially in the light of other studies' findings, point to the fact that AI models are still untrustworthy as far as providing tax support and advice is concerned. They may be useful when used as a type of 'sounding board' to facilitate knowledgeable tax practitioners' own reasoning and decision-making rather than prescribing absolute solutions. But they cannot replace or be used in place of tax practitioners and consultants.

Interestingly, Grok was at least able to pass both questions, indicating that AI has the potential to play a significant role in tax practitioners' work. Coupled with the rapid advancement of AI (Teubner et al. 2023), this study should be repeated regularly to track the progress of AI models and to provide empirical, comparable observations about the ability of AI models to perform as tax practitioners. This information should inform regulators, universities, and students' planning so that they stay relevant and efficient while maintaining high standards of tax advice and support.

This article finds that AI models' persuasiveness and creativity are key risks and areas that require more research. In addition, regulators and universities need to track AI's performance to design acceptable uses for AI that take advantage of its efficiencies yet protect the integrity of the profession. Research is also required to better understand the impact of different countries' tax complexity on AI's capabilities.

This study has some limitations. A sample size of only two tasks was used, and a single marker was used (although it was consistent with who marked the students). The different AI models were used at a specific time, and there may be changes to these over time. As an exploratory study grounded in a South African study, there are limitations to the generalisability of these results.

Acknowledgements

Competing interests

The authors declare that they have no financial or personal relationships that may have inappropriately influenced them in writing this article. The author, Wayne van Zijl, serves as

an editorial board member of this journal. Wayne van Zijl has no other competing interests to declare.

CRediT authorship contribution

Asheer J. Ram: Methodology, formal analysis, investigation, Writing-original draft, resources and writing – review & editing. Wayne van Zijl: Conceptualisation, formal analysis, investigation, Writing-original draft, visualisation and writing – review & editing. All authors reviewed the article, contributed to the discussion of results, approved the final version for submission and publication, and take responsibility for the integrity of its findings.

Funding information

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Data availability

The authors declare that all data that support this research article and findings are available in the article and its references.

Disclaimer

The views and opinions expressed in this article are those of the authors and are the product of professional research. They do not necessarily reflect the official policy or position of any affiliated institution, funder, agency or that of the publisher. The authors are responsible for this article's results, findings and content.

References

AI for Education, 2025, *AI model comparison: Free vs Paid tiers*, viewed 26 October 2025, from https://www.aiforeducation.io/ai-resources/ai-model-comparison-free-vs-paid-tiers.

Albuquerque, F. & Dos Santos, P.G., 2024, 'Can ChatGPT be a certified accountant? Assessing the responses of ChatGPT for the professional access exam in Portugal', *Administrative Sciences* 14(7), 1–15. https://doi.org/10.3390/admsci14070152

Atanasovski, A., Tocev, T., Dionisijev, I., Minovski, Z. & Jovevski, D., 2023, 'Evaluating the performance of ChatGPT in accounting and auditing exams: An experimental study in North Macedonia', in M. Trpeska (ed.), *4th international scientific conference: Economic and Business Trends Shaping the Future, online conference proceedings*, Skopje, North Macedonia, November 9–10, 2023, pp. 40–50, viewed n.d., from http://hdl.handle.net/20.500.12188/28871.

Bommarito, J., Bommarito, M., Katz, D.M. & Katz, J., 2023, 'GPT as knowledge worker: A zero-shot evaluation of (AI) CPA capabilities', Report, arXiv. https://doi.org/10.48550/arXiv.2301.04408

Borger, J.G., Ng, A.P., Anderton, H., Ashdown, G.W., Auld, M., Blewitt, M.E. et al., 2023, 'Artificial intelligence takes center stage: Exploring the capabilities and implications of ChatGPT and other AI-assisted technologies in scientific research and education', *Immunology & Cell Biology* 101(10), 923–935. https://doi.org/10.1111/imcb.12689

Cheng, X., Dunn, R., Holt, T., Inger, K., Jenkins, J.G., Jones, J. et al., 2024, 'Artificial intelligence's capabilities, limitations, and impact on accounting education: Investigating ChatGPT's performance on educational accounting cases', *Issues in Accounting Education* 39(2), 23–47. https://doi.org/10.2308/ISSUES-2023-032

Dahl, M., Magesh, V., Suzgun, M. & Ho, D., 2024, 'Large legal fictions: Profiling legal hallucinations in large language models', *Journal of Legal Analysis* 16(1), 64–93. https://doi.org/10.1093/jla/laae003

de Carvalho Souza, M.E. & Weigang, L., 2025, 'Grok, Gemini, ChatGPT and DeepSeek: Comparison and applications in conversational artificial intelligence', *Report, Dept. of Computer Science*, University of Brasília. https://doi.org/10.5281/zenodo.14885243

Eulerich, M., Sanatizadeh, A., Vakilzadeh, H. & Wood, D.A., 2024, 'Is it all hype? ChatGPT's performance and disruptive potential in the accounting and auditing industries', *Review of Accounting Studies* 29(3), 2318–2349.

Jabbar, A., Ul Islam, S. & Boudjadar, J., 2025, 'A comparative review of LLM-based conversational systems: Insights from DeepSeek, ChatGPT, Gemini, Claude, and Copilot', in Institution of Engineering and Technology (ed.), *International Conference on AI and the Digital Economy (CADE 2025)*, Hybrid Conference, Venice, Italy, July 14–16, 2025, pp. 167–173.

Katz, D.M., Bommarito, M.J., Gao, S. & Arredondo, P., 2024, 'GPT-4 passes the bar exam', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 382(2270), 1–17. https://doi.org/10.1098/rsta.2023.0254

Pinto, A.S., Abreu, A., Costa, E. & Paiva, J., 2024, 'AI in accounting: Can AI models like ChatGPT and Gemini successfully pass the Portuguese chartered accountant exam?', in A. Abreu, J.V. Carvalho, A. Mesquita, A. Sousa Pinto & M. Mendonça Teixeira (eds.), *BT – Perspectives and trends in education and technology*, pp. 429–438, Springer Nature Switzerland, Cham.

Roberts, J., Baker, M. & Andrew, J., 2024, 'Artificial intelligence and qualitative research: The promise and perils of large language model (LLM) "assistance"', *Critical Perspectives on Accounting* 99, 102722. https://doi.org/10.1016/j.cpa.2024.102722

South African Institute of Chartered Accountants (SAICA), 2023, *CA(SA) and SAICA are back to number 1 in the world*, viewed 06 May 2025, from https://www.saica.org.za/news/south-african-chartered-accountants-lead-in-global-trustworthiness.

South African Institute of Chartered Accountants (SAICA), 2024, *SAICA code of conduct*, viewed 31 May 2025, from https://www.saica.org.za/about/general/ethics/saica-code-of-conduct.

South African Institute of Taxation (SAIT), 2025, *SAIT member code of conduct*, viewed 31 May 2025, from https://thesait.org.za/wp-content/uploads/2025/02/SAIT_code_of_ethics7.pdf.

South African Revenue Service (SARS), 2024, *Register as a tax practitioner*, viewed 26 October 2025, from https://www.sars.gov.za/tax-practitioners/register-as-a-tax-practitioner/.

Schneid, S.D., Armour, C. & Brandl, K., 2025, 'Beyond right or wrong: How partial credit scoring on multiple-choice questions improves student performance and assessment perceptions', *British Journal of Clinical Pharmacol*ogy, 1–7. https://doi.org/10.1002/bcp.70127

Tax Complexity Index, 2022, *Tax complexity index*, viewed 08 May 2025, from https://www.taxcomplexity.org/.

Teubner, T., Flath, C.M., Weinhardt, C., Van der Aalst, W. & Hinz, O., 2023, 'Welcome to the era of ChatGPT et al.', *Business & Information Systems Engineering* 65, 95–101. https://doi.org/10.1007/s12599-023-00795-x

Van Wyk, E., 2011, 'A note: The SAICA part I qualifying examinations: Factors that may influence candidates' success', *South African Journal of Accounting Research* 25(1), 145–174. https://doi.org/10.1080/10291954.2011.11435157

Wood, D.A., Achhpilia, M.P., Adams, M.T., Aghazadeh, S., Akinyele, K., Akpan, M. et al, 2023, 'The ChatGPT artificial intelligence chatbot: How well does it answer accounting assessment questions?' *Issues in Accounting Education* 38(4), 1–28. https://doi.org/10.2308/ISSUES-2023-013