Journal of
Information System and Technology Auditing

Research Paper

Evaluating the Performance of Large Language Models on Doctoral Accounting Exams: A Comparative Study of Six Generative AI Chatbots [1]

Sasan Khademi[2]

Journal of Information System and Technology Auditing
Iranian Information Technology Audit Scientific Association
Vol. 1, No. 2, Autumn & Winter 2025 - 2026
pp. 17-25

Received: 2025.11.26
Revised: 2026.01.12
Accepted: 2026.02.18

1. Introduction

Recent advances in artificial intelligence (AI), particularly the rapid development of large language models (LLMs), have profoundly transformed educational and professional landscapes worldwide. LLMs, including ChatGPT, Gemini, Perplexity, Grok, DeepSeek, and Qwen, leverage sophisticated deep learning architectures and extensive natural language processing (NLP) frameworks to generate coherent text, comprehend complex instructions, perform multi-step reasoning, and simulate cognitive processes that resemble human analytical capabilities. Their versatility and apparent intelligence have sparked growing scholarly interest, particularly in higher education and professional training, where the evaluation of cognitive and conceptual understanding is crucial.

Accounting, as a discipline, presents unique challenges for automated reasoning systems due to its reliance on analytical

[1] https://doi.org/10.22034/JISTA. 2025.545440.1062
[2] Ph.D in Accounting, Department, Faculty of Economics, Management and Social Sciences, Shiraz University, Shiraz, Iran. (Corresponding Author). Email: s.khademi@shirazu.ac.ir

thinking, interpretation of theoretical frameworks, application of accounting standards, integration of qualitative and quantitative information, and judgment in complex scenarios. Standardized examinations, particularly at the PhD level, assess not only factual recall but also conceptual understanding, analytical reasoning, and the ability to discriminate between subtly different alternatives. These exams therefore provide a rigorous benchmark for evaluating the baseline capabilities of LLMs in professional and academic contexts.

Prior international studies have examined LLM performance in fields such as medicine, law, computer science, and language education. These studies demonstrate that advanced models can achieve accuracy levels comparable to, and sometimes exceeding, those of human examinees. In accounting, LLMs such as ChatGPT and GPT-4 have shown promising results, though performance varies depending on model version, question format, language, and contextual complexity. Despite this growing body of research, a notable gap exists regarding LLM performance in non-English contexts and nationally specific assessments. In Iran, the PhD entrance examination in accounting represents a high-stakes, standardized assessment that emphasizes auditing, accounting theory, and management accounting. No prior study has systematically compared multiple LLMs on this examination. This study aims to fill this gap by evaluating six prominent LLMs on official Iranian PhD-level accounting questions, exploring their potential as educational support tools, and examining implications for teaching, assessment, and policy.

2. MATERIALS AND METHODS

This research adopts a descriptive–analytical, empirical, comparative design. The study population consists of all officially released multiple-choice questions from the Iranian PhD entrance examination in accounting between the years 1400 and 1404. The dataset includes 300 questions across three core domains: auditing (75 questions), management accounting (100 questions), and

accounting theory (125 questions). As the complete set of available questions was analyzed, no sampling procedure was required.

Each of the six LLMs (ChatGPT, Gemini, Perplexity, Grok, DeepSeek, and Qwen) was tested under standardized conditions using their publicly available interfaces. Questions were presented in their original text format, without advanced prompt engineering, few-shot learning, external tool integration, or any additional guidance. Responses were coded in binary form (correct = 1, incorrect = 0), and each model's performance was assessed across all 300 questions. To ensure reliability, models were run independently three times with three-day intervals, mitigating short-term memory effects. Mean accuracy, standard deviation, and Fleiss' Kappa were calculated to quantify reproducibility, with all models exceeding Kappa values of 0.85. Additionally, 95% confidence intervals for overall and subject-level accuracy were computed to provide statistical precision and assess expected variability.

Data analysis was conducted using SPSS version 27. Descriptive statistics summarized overall and subject-specific performance, while inferential analyses addressed two categories of research hypotheses. One-sample proportion tests evaluated whether each model performed significantly above the chance level (25%) and a minimum acceptable baseline (50%). Nonparametric Cochran's Q test was used to assess whether differences among models were statistically significant. This approach allowed robust comparison of performance across models while accommodating non-normality in accuracy distributions.

Notably, the evaluation occurred in an open-book scenario, reflecting practical usage conditions. Models operated under default configurations, potentially leveraging stored knowledge from prior training, which introduces the possibility of data leakage. Consequently, performance reflects operational effectiveness rather than independent reasoning, emphasizing that observed accuracy should not be interpreted as evidence of conceptual understanding or high-level cognitive reasoning. This distinction between open-book

and closed-book contexts is explicitly acknowledged throughout the analysis and interpretation.

3. RESULTS AND DISCUSSION

Descriptive analysis revealed that all six LLMs performed substantially above both reference thresholds. Gemini achieved the highest overall accuracy at 67.3%, followed by Perplexity (65.7%), ChatGPT (65.0%), Grok (64.0%), DeepSeek (64.3%), and Qwen (63.3%). While these differences suggest relative variation, Cochran's Q test indicated no statistically significant differences, suggesting convergence in baseline performance across models when applied to standardized accounting questions.

Subject-level analysis highlighted domain-specific performance differences. In management accounting, involving applied problem-solving and numerical reasoning, Perplexity and Qwen achieved the highest accuracy (~75%), while Gemini performed comparatively lower. In accounting theory, emphasizing abstract reasoning, Gemini outperformed other models with 68% accuracy. In auditing, Gemini again led with 64% accuracy, whereas DeepSeek recorded the lowest performance. These findings illustrate the importance of content specificity: excelling in one domain does not guarantee superior overall performance, reflecting the interaction between question type, cognitive demand, and model architecture.

Inferential analyses confirmed that all models performed significantly above the chance level (25%) and minimum acceptable baseline (50%) across both one-sample proportion test categories. Large Z-statistics and p-values below 0.001 indicate robust statistical evidence for baseline competence in handling PhD-level accounting questions. Despite observable descriptive differences, nonparametric analysis shows that these are not statistically robust, emphasizing that nominal performance variations among contemporary LLMs are small relative to their overall capabilities.

Comparison with prior literature indicates alignment with international findings. Wood et al. (2023) reported ~56.5% accuracy for ChatGPT on accounting MCQs, while more advanced models

reached higher rates. Amoah et al. (2024) demonstrated that ChatGPT-4 and Claude could achieve human-comparable accuracy without specialized fine-tuning. Differences among studies likely reflect model version, language, exam structure, and interaction strategy. The present study adds value by evaluating multiple LLMs in a non-English, nationally specific context, providing insights on model generalizability, domain sensitivity, and practical utility.

4. CONCLUSION

This study provides the first comprehensive comparative evaluation of six LLMs on the Iranian PhD accounting entrance examination. Findings demonstrate that all models perform significantly above chance and baseline thresholds, reflecting practical competence in producing correct responses to complex accounting questions. While descriptive differences exist, Gemini achieved the highest accuracy and Qwen the lowest, nonparametric analysis indicates no statistically significant differences, suggesting broadly comparable baseline capabilities across LLM platforms.

The study highlights the importance of distinguishing open-book from closed-book performance: observed accuracy may partly reflect retrieval of previously encoded knowledge rather than independent reasoning. As such, results should not be interpreted as evidence of deep conceptual understanding. Educators can leverage LLMs as complementary tools for learning support, practice, and self-assessment, while policymakers and assessment designers should consider AI capabilities in exam design and pedagogical strategies.

Limitations include reliance on multiple-choice questions, potential data leakage, exclusive focus on quantitative accuracy, and context-specificity to the Iranian PhD exam. Future research should explore closed-book assessments, qualitative analysis of reasoning, prompt engineering effects, cross-linguistic comparisons, and adaptive learning integrations. Overall, this study provides robust empirical evidence on LLM performance in advanced accounting education, supporting informed, responsible integration of AI technologies into teaching, assessment, and curriculum design.

Keywords: Large Language Models, Accounting Education, Accounting PhD Entrance Examination, Performance Evaluation, Artificial Intelligence in Education

JEL classification: *A22, C12, M41*

References

Adnan Hammood, M., Piri, P., & Ashtab, A. (2025). Feasibility of utilizing advanced artificial intelligence technologies to improve auditing processes in the country. *Accounting and Auditing Review*, 32(3), 535-559. (in Persian) https://doi.org/10.22059/acctgrev.2025.391837.1009085

Agarwal, P., & Gaur, F. (2020). A historical perspective of artificial intelligence in accounting: Evolution, current developments, and future opportunities. *Journal of Accounting and Organizational Change*, 16(1), 1–12. https://doi.org/10.1108/JAOC-04-2017-0035

AI Index Steering Committee. (2025). *The AI Index 2025 annual report*. Institute for Human-Centered AI, Stanford University. https://doi.org/10.48550/arXiv.2504.07139

Alibaba Group. (2024, September 19). *Alibaba Cloud unveils Qwen2.5, full-stack AI infrastructure enhancements at 2024 Apsara Conference*. Alibaba Group. https://www.alibabagroup.com/en-US/document-1773855135127044096

Albuquerque, F., & Gomes dos Santos, P. (2024). Can ChatGPT Be a Certified Accountant? Assessing the Responses of ChatGPT for the Professional Access Exam in Portugal. *Administrative Sciences*, 14(7), 152. https://doi.org/10.3390/admsci14070152

Amoah, N., Fianko, S. K., Dake, S., Agyemang, K., Nyame, I., Adjaye-Gyamfi, O., ... & Lartey, R. (2024). The Impact of Ai Chatbots on the Landscape of Professional Accountancy Examination: An Experimental Study. *Available at SSRN 4991304*. http://dx.doi.org/10.2139/ssrn.4991304

Bordt, S., & von Luxburg, U. (2023). Chatgpt participates in a computer science exam. *arXiv preprint arXiv:2303.09461*. https://doi.org/10.48550/arXiv.2303.09461

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

Chippagiri, S. (2025, March 4). *DeepSeek: Revolutionizing AI with Open-Source Large Language Models. DEV Community.* https://dev.to/srinivas_chippagiri_e01c8/deepseek-revolutionizing-ai-with-open-source-large-language-models-127i

Dell, S., & Akpan, M. (2024). You are the auditor: A ChatGPT-based multiple choice exam. *Advances in Online Education: A Peer-Reviewed Journal*, 3(2), 111–120. https://doi.org/10.69554/EINF1743

de Freitas, M. M., Sallaberry, J. D., & de Jesus Silva, T. B. (2024). Application of Chat GPT 4.0 for solving accounting problems. *GCG: revista de globalización, competitividad y gobernabilidad*, 18(2), 49-64. https://dialnet.unirioja.es/servlet/articulo?codigo=9498637

de Winter, J. C. (2024). Can ChatGPT pass high school exams on English language comprehension?. *International Journal of Artificial Intelligence in Education*, 34(3), 915-930. https://doi.org/10.1007/s40593-023-00372-z

Eulerich, M., Sanatizadeh, A., Vakilzadeh, H., & Wood, D. A. (2023). Can artificial intelligence pass accounting certification exams? ChatGPT: CPA, CMA, CIA, and EA. ChatGPT: CPA, CMA, CIA, and EA. *Available at SSRN*. http://www.ais.nptu.edu.tw/bsacc/1121%20materials/SSRN-id4452175_ChatGPT%E8%80%83%E6%9C%83%E8%A8%88%E8%AD%89%E7%85%A7.pdf

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. https://doi.org/10.1007/s11142-024-09833-9

Foote, K. D. (2023, December 28). *A brief history of large language models*. DATAVERSITY. https://www.dataversity.net/a-brief-history-of-large-language-models/

Glover, E. (2025, July 16). *Grok: What we know about Elon Musk's AI chatbot. Built In*. https://builtin.com/articles/grok

Greenman, C., Esplin, D., Johnston, R., & Richards, J. (2024). An Analysis of the Impact of Artificial Intelligence on the Accounting Profession. *Journal of Accounting, Ethics & Public Policy*, *JAEPP*, 25(2), 188-188. https://doi.org/10.60154/jaepp.2024.v25n2p188

Guinness, H. (2024, April 3). *What is Perplexity AI? How to use it + how it works. Zapier Blog*. https://zapier.com/blog/perplexity-ai

Hashemi-Pour, C., Kerner, S. M., & Patrizio, A. (2025, January 8). What is the Google Gemini AI model (formerly Bard)? TechTarget. https://www.techtarget.com/searchenterpriseai/definition/Google-Gemini

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2024). Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270), 20230254. https://doi.org/10.1098/rsta.2023.0254

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., ... & Tseng, V. (2023). Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS digital health*, 2(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198

Martínez, E. (2024). Re-evaluating GPT-4's bar exam performance. *Artificial intelligence and law*, 1-24. https://doi.org/10.1007/s10506-024-09307-6

Mashayekhi, B., & Amrollahi, M. R. (2025). The effect of internal auditors' knowledge and professional skepticism on the artificial intelligence utilization. *Journal of Empirical Research in Accounting*, 15(2), 1-28. (in Persian) https://doi.org/10.22051/jera.2025.50268.3523

Mendonça, N. C. (2024). Evaluating chatgpt-4 vision on brazil's national undergraduate computer science exam. *ACM Transactions on Computing Education*, 24(3), 1-56. https://dl.acm.org/doi/abs/10.1145/3674149

Mikalef, P., & Gupta, M. (2021). Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Information & Management*, 58(3), 103434. https://doi.org/10.1016/j.im.2020.103434

National Aeronautics and Space Administration. (2024). What is artificial intelligence? NASA. https://www.nasa.gov/what-is-artificial-intelligence/

Nourahmadi, M., & Parsi, F. (2025). The role of artificial intelligence in enhancing green accounting and sustainable development: a bibliometrix methode. *Journal of Empirical Research in Accounting*, 15(2), 211-238. (in Persian) https://doi.org/10.22051/jera.2025.50235.3512

Pierotti, M., Monreale, A., & De Santis, F. (2024). *Artificial Intelligence in Accounting and Auditing: Accessing the Corporate Implications*. Palgrave Macmillan, Switzerland. ISBN. https://doi.org/10.1007/978-3-031-31299-1

Rahmaini, A., Maanavi, S., & Haddadi, N. (2025). Integration of Artificial Intelligence in Auditing: Challenges and Benefits. *Journal of Information System and Technology Audit (JISTA)*, 1(1). 1-27. (in Persian) https://doi.org/10.22034/jista.2025.528769.1051

Rahnama, M., & Rafati, H. (2025). The Ethical Implications of Adopting Artificial Intelligence (AI) in Financial Decision-Making. *Journal of Information System and Technology Audit (JISTA)*, 1(1). 284-301. (in Persian) https://doi.org/10.22034/jista.2025.509536.1032

Saghafi, A., & Parsapoor, M. (2025). Examining impact of accounting data analysis with generative ai on the quality of digital sustainability reporting with the mediating role of green sustainability internal control system. *Financial Accounting Knowledge*, 12(1), 1-31. (in Persian) https://doi.org/10.30479/jfak.2025.21533.3270

SecureNinja. (2025, March 18). Comparison of Top AI Models: DeepSeek AI, ChatGPT, Gemini, and Perplexity AI. SecureNinja Blog. https://secureninja.com/news/comparison-of-top-ai-models-deepseek-ai-chatgpt-gemini-and.html

Sharida, A., & Hashlamon, I. (2021). Real-time vision-based controller for delta robots. International *Journal of Intelligent Systems Technologies and Applications*, 20 (4), 271–295. https://doi.org/10.1504/IJISTA.2021.10045532

Sharida, A., Hamdan, A., & Al-Hashimi, M. (2020). Smart cities: The next urban evolution in delivering a better quality of life. Toward Social Internet of Things (SIoT): Enabling Technologies, Architectures and Applications: Emerging Technologies for Connected and Smart Social Objects, 287–298. https://doi.org/10.1007/978-3-030-24513-9_16

Stengel, F. C., Stienen, M. N., Ivanov, M., Gandía-González, M. L., Raffa, G., Ganau, M., ... & Motov, S. (2024). Can AI pass the written European Board Examination in Neurological Surgery?-Ethical and practical issues. *Brain and Spine*, 4, 102765. https://doi.org/10.1016/j.bas.2024.102765

SY Partners. (2025, February 10). *The history of GPT: A journey through generative pre-trained transformers*. https://syp.vn/en/article/the-history-of-GPT

TechCrunch. (2025, May 20). *DeepThink boosts the performance of Google's flagship Google Gemini AI model*. https://techcrunch.com/2025/05/20/deep-think-boosts-the-performance-of-googles-flagship-google-gemini-ai-model

Vaˇrzaru, A. A. (2022). Assessing artificial intelligence technology acceptance in managerial accounting. *Electronics, 11*, 1–13. https://doi.org/10.3390/electronics11142256

Vasarhelyi, M. A., Moffitt, K. C., Stewart, T., & Sunderland, D. (2023). Large language models: An emerging technology in accounting. *Journal of Emerging Technologies in Accounting*, 20(2), 1–10. https://doi.org/10.2308/JETA-2023-047. https://doi.org/10.2308/JETA-2023-047

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wood, D. A., Achhpilia, M. P., Adams, M. T., Aghazadeh, S., Akinyele, K., Akpan, M., ... & Kuruppu, C. (2023). The ChatGPT artificial intelligence chatbot: How well does it answer accounting assessment questions?. *Issues in Accounting Education*, 38(4), 81-108. https://doi.org/10.2308/ISSUES-2023-013

World Economic Forum. (2020). Future of Jobs Report 2020. https://www.weforum.org/publications/the-future-of-jobs-report-2020/

Wutzler, J. (2024). Outsmarting Artificial Intelligence in the Classroom—Incorporating Large Language Model-Based Chatbots into Teaching. *Issues in Accounting Education*, 39(4), 183-206. https://doi.org/10.5555/ISSUES-2023-064tn

Zacher, W., & Kuppannagari, S. (2024). Can LLMs Pass the CPA Exam? Evaluating Large Language Model Performance on the Certified Public Accountant Test. *Available at SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4788096*

Zhangyang, Q., Fang, Y., Zhang, M., Sun, Z., Wu, T., Liu, Z., Lin, D., Wang, J., & Zhao, H. (2023, December 22). Gemini vs GPT-4V: A preliminary comparison and combination of vision-language models through qualitative cases. arXiv. https://doi.org/10.48550/arXiv.2312.15011

نشریه علمی
حسابرسی سیستم‌ها و فناوری اطلاعات
انجمن علمی حسابرسی فناوری اطلاعات ایران

مقالهٔ پژوهشی

ارزیابی عملکرد مدل‌های زبانی بزرگ در آزمون دکتری حسابداری: مطالعه‌ای مقایسه‌ای از شش چت‌بات هوش مصنوعی مولد[1]

ساسان خادمی[2]

نشریه علمی حسابرسی سیستم‌ها و فناوری اطلاعات
انجمن حسابرسی فناوری اطلاعات ایران
سال اول، پیاپی۲، پاییز و زمستان ۱۴۰۴
صص ۵۷ - ۹۱

تاریخ دریافت: ۱۴۰۴/۰۹/۰۵
تاریخ بازنگری: ۱۴۰۴/۱۰/۲۲
تاریخ پذیرش: ۱۴۰۴/۱۱/۲۹

چکیده

پیشرفت شتابان مدل‌های زبانی بزرگ، توجه پژوهشگران را به عملکرد این ابزارها در پاسخ‌گویی به پرسش‌های تخصصی و پیامدهای بالقوه آن‌ها برای یادگیری و ارزشیابی معطوف کرده است. هدف پژوهش حاضر، ارزیابی و مقایسه عملکرد شش مدل زبانی بزرگ شامل ChatGPT، Gemini، Perplexity، Grok، DeepSeek و Qwen در پاسخ‌گویی به سؤالات آزمون دکتری حسابداری ایران است. داده‌های پژوهش شامل ۳۰۰ سؤال چهارگزینه‌ای رسمی آزمون دکتری حسابداری طی سال‌های ۱۴۰۰ تا ۱۴۰۴ در سه درس حسابرسی، حسابداری مدیریت و تئوری حسابداری است. پاسخ‌های هر مدل به‌صورت دودویی (صحیح/غلط) کدگذاری شد و با استفاده از آزمون نسبت تک‌نمونه‌ای، عملکرد آن‌ها نسبت به دو سطح مرجع ۰٫۲۵ (عملکرد تصادفی) و ۰٫۵۰ (سطح پایه قابل قبول) ارزیابی گردید. همچنین، برای مقایسه عملکرد نسبی مدل‌ها از آزمون Cochran's Q استفاده شد. نتایج نشان داد که عملکرد تمامی مدل‌ها به‌طور معناداری فراتر از هر دو سطح مرجع است. اگرچه مدل Gemini بالاترین و مدل Qwen پایین‌ترین درصد پاسخ صحیح را ثبت کردند، آزمون Cochran's Q تفاوت معناداری میان عملکرد کلی مدل‌ها نشان نداد. با این حال، نتایج در چارچوب یک سناریوی عملیاتی open-book تفسیر می‌شوند و با توجه به احتمال نشت داده و ماهیت چندگزینه‌ای سؤالات، نباید به‌عنوان شواهدی از درک مفهومی عمیق یا استدلال مستقل مدل‌ها تلقی شوند. به‌طور کلی، یافته‌ها نشان می‌دهد که مدل‌های زبانی بزرگ، حتی بدون تنظیمات پیشرفته یا آموزش اختصاصی، از توان قابل توجهی در عملکرد صحیح در آزمون‌های استاندارد حسابداری برخوردارند و می‌توانند به‌عنوان ابزارهای مکمل در آموزش و طراحی فعالیت‌های ارزشیابی در آموزش عالی حسابداری مورد توجه قرار گیرند.

واژه‌های کلیدی: *مدل‌های زبانی بزرگ، آموزش حسابداری، آزمون دکتری حسابداری، ارزیابی عملکرد، هوش مصنوعی در آموزش.*

طبقه‌بندی موضوعی: *A22, C12, M41*

---

[1] https://doi.org/10.22034/JISTA.2026.568921.1077

[2] دانش‌آموخته دکتری حسابداری، بخش حسابداری، دانشکده اقتصاد، مدیریت و علوم اجتماعی، دانشگاه شیراز، شیراز، ایران. Email: s.khademi@shirazu.ac.ir

مقدمه

تحولات شتابان در حوزه هوش مصنوعی[1]، به‌ویژه با ظهور مدل‌های زبانی بزرگ[2] نظیر ChatGPT، Gemini، Perplexity، Grok، DeepSeek و Qwen، به‌طور بنیادین ماهیت تعامل انسان با دانش تخصصی، آموزش دانشگاهی و نظام‌های ارزیابی را دستخوش تغییر کرده است. این مدل‌ها که بر معماری‌های پیشرفته یادگیری عمیق و پردازش زبان طبیعی استوارند، توانایی‌هایی فراتر از تولید متن ساده از خود نشان داده‌اند و قادرند در سطوحی از استدلال، تحلیل مفهومی و پاسخ‌گویی به مسائل پیچیده شبه‌ساختاریافته عمل کنند. ازاین‌رو، توجه پژوهشگران و نهادهای حرفه‌ای در رشته‌هایی نظیر پزشکی، حقوق، مهندسی و حسابداری به‌طور فزاینده‌ای به ارزیابی قابلیت‌ها، محدودیت‌ها و پیامدهای آموزشی این فناوری‌ها معطوف شده است (وود و همکاران[3]، ۲۰۲۳؛ گرینمن و همکاران[4]، ۲۰۲۴؛ دِ فریتاس و همکاران[5]، ۲۰۲۴؛ سکیور نینجا[6]، ۲۰۲۵).

در این میان، یکی از رویکردهای رایج و قابل اتکا برای سنجش عملکرد عملی مدل‌های زبانی بزرگ، ارزیابی آن‌ها در پاسخ‌گویی به سؤالات استاندارد آزمون‌های دانشگاهی و حرفه‌ای است. چنین آزمون‌هایی، به‌ویژه در سطوح تحصیلات تکمیلی، صرفاً حافظه‌محور یا محاسباتی نیستند، بلکه مستلزم درک مفهومی، تحلیل موقعیت‌محور، تشخیص گزینه‌های گمراه‌کننده و اعمال قضاوت حرفه‌ای هستند. از همین رو، مطالعات متعددی عملکرد این مدل‌ها را در آزمون‌های حرفه‌ای و دانشگاهی در حوزه‌هایی نظیر پزشکی (مانند: کونگو همکاران[7]، ۲۰۲۳؛ استنگلو همکاران[8]، ۲۰۲۴)؛ حقوق (مانند: کاتز و همکاران[9]، ۲۰۲۴؛ مارتینِز[10]، ۲۰۲۴)؛ علوم رایانه (مانند: بُردت و وون لوکسبرگ[11]، ۲۰۲۳؛ مِندونکا[12]، ۲۰۲۴) و آموزش زبان انگلیسی

1. Artificial Intelligence (AI)
2. Large Language Models (LLMs)
3 .Wood et al.
4. Greenman et al.
5. de Freitas et al.
6. SecureNinja
7. Kung et al.
8. Stengel et al.
9. Katz et al.
10. Martínez
11. Bordt & von Luxburg
12. Mendonça

(مانند: دی وینتر[1]، ۲۰۲۴) بررسی کرده‌اند و نشان داده‌اند که این مدل‌ها در برخی سطوح شناختی می‌توانند به عملکرد انسانی نزدیک شوند یا حتی از آن پیشی بگیرند.

در حوزه حسابداری، اهمیت این موضوع دوچندان است. حسابداری دانشی است که در سطوح پیشرفته خود، بیش از محاسبات صرف، متکی بر قضاوت حرفه‌ای، تفسیر استانداردها، تحلیل مفهومی و ارزیابی موقعیت‌های پیچیده است. به همین دلیل، سنجش توان مدل‌های زبانی بزرگ در درک و تحلیل سؤالات مفهومی و نظری حسابداری، برای ارزیابی نقش بالقوه آن‌ها در آموزش، طراحی آزمون و پشتیبانی از یادگیری عمیق اهمیت اساسی دارد. پژوهش‌های پیشین نشان داده‌اند که عملکرد مدل‌هایی نظیر ChatGPT در آزمون‌های حسابداری ناهمگن است و به عواملی مانند نسخه مدل، نوع سؤال و سطح شناختی آزمون وابسته است (وود و همکاران، ۲۰۲۳؛ آموآه و همکاران[2]، ۲۰۲۴؛ دِل و اکپان[3]، ۲۰۲۴). برای مثال، ChatGPT در مطالعه‌ای گسترده تنها به‌طور متوسط به ۵۶.۵ درصد سؤالات استاندارد حسابداری پاسخ صحیح داده است، در حالی که عملکرد دانشجویان انسانی به‌مراتب بالاتر بوده است (وود و همکاران، ۲۰۲۳). با این حال، شواهد جدید نشان می‌دهد که نسخه‌های پیشرفته‌تر و آموزش‌دیده‌تر این مدل‌ها، به‌ویژه در سؤالات مفهومی و تحلیلی، به سطح عملکرد انسانی نزدیک شده‌اند یا حتی از آن فراتر رفته‌اند (آموآه و همکاران، ۲۰۲۴).

با این حال، تأکید بر این نکته ضروری است که هدف از چنین ارزیابی‌هایی، جایگزینی مدل‌های زبانی با داوطلبان انسانی یا تجویز استفاده از این ابزارها در آزمون‌های رسمی نیست. بلکه هدف اصلی، سنجش عملکرد عملی این مدل‌ها در شرایطی مشابه محیط‌های واقعی ارزیابی آموزشی و بررسی این پرسش است که آن‌ها در چه سطحی و با چه محدودیت‌هایی می‌توانند به‌عنوان ابزارهای پشتیبان آموزش و ارزشیابی مورد استفاده قرار گیرند. از این منظر، آزمون دکتری حسابداری ــ با تمرکز غالب بر سؤالات مفهومی، نظری و مبتنی بر قضاوت حرفه‌ای ــ نه به‌عنوان نماینده کامل آموزش حسابداری، بلکه به‌عنوان یک ابزار معتبر برای ارزیابی عملکرد مدل‌ها در سطحی پیشرفته و استاندارد مورد توجه قرار می‌گیرد این آزمون به‌ویژه برای بررسی توان پاسخ‌گویی مدل‌ها در بستر زبان فارسی و متون تخصصی حسابداری مناسب است.

1. de Winter
2. Amoah et al.
3. Dell & Akpan

با وجود رشد سریع ادبیات بین‌المللی در این زمینه، در ایران تاکنون پژوهشی که به‌طور نظام‌مند و مقایسه‌ای عملکرد مدل‌های زبانی بزرگ را در پاسخ‌گویی به سؤالات استاندارد و سطح‌بالای حسابداری در زبان فارسی بررسی کند، انجام نشده است. مطالعات داخلی موجود عمدتاً بر کاربردهای کلی هوش مصنوعی در حسابداری و حسابرسی تمرکز داشته‌اند و شواهد تجربی مبتنی بر آزمون‌های استاندارد و مقایسه چندمدلی ارائه نکرده‌اند. این خلأ پژوهشی از آن جهت حائز اهمیت است که بدون شواهد تجربی بومی، امکان سیاست‌گذاری آگاهانه در حوزه آموزش حسابداری، طراحی آزمون‌ها و بهره‌گیری مسئولانه از ابزارهای هوش مصنوعی فراهم نخواهد شد.

بر این اساس، هدف پژوهش حاضر آن است که با اتخاذ رویکردی تجربی و مقایسه‌ای، عملکرد شش مدل زبانی پرکاربرد شامل ChatGPT، Gemini، Perplexity، Grok، DeepSeek و Qwen را در پاسخ‌گویی به سؤالات چهارگزینه‌ای آزمون دکتری حسابداری ایران مورد ارزیابی قرار دهد. تمرکز مطالعه بر سنجش عملکرد نتیجه‌محور مدل‌ها در قالب آزمون‌های چندگزینه‌ای و مقایسه دقت پاسخ‌گویی آن‌ها در یک چارچوب اجرایی یکنواخت است، بدون آنکه ادعایی درباره استدلال مفهومی مستقل یا تحلیل شناختی عمیق مدل‌ها مطرح شود. یافته‌های این پژوهش می‌تواند برای مدرسان دانشگاهی در طراحی فعالیت‌های آموزشی مفهومی، برای طراحان آزمون در بازاندیشی ساختار سؤالات، و برای سیاست‌گذاران آموزشی در تدوین چارچوب‌های استفاده مسئولانه از هوش مصنوعی در آموزش عالی حسابداری، پیامدهای عملی معناداری به همراه داشته باشد.

در ادامه، ابتدا پیشینه نظری و تجربی مرتبط مرور و فرضیه‌های پژوهش تدوین می‌شود. سپس روش‌شناسی و یافته‌های پژوهش ارائه شده و در نهایت، بحث و نتیجه‌گیری به‌همراه محدودیت‌ها و مسیرهای پژوهش آتی مطرح خواهد شد.

مبانی نظری و توسعه فرضیه‌ها

هوش مصنوعی و مدل‌های زبانی بزرگ

هوش مصنوعی به‌عنوان یکی از مهم‌ترین فناوری‌های تحول‌آفرین قرن بیست‌ویکم، به‌تدریج جایگاهی راهبردی در طیف گسترده‌ای از حوزه‌ها از جمله خدمات مالی، آموزش

عالی، سلامت و نظام حقوقی به دست آورده است. شواهد تجربی ارائه‌شده در گزارش شاخص هوش مصنوعی ۲۰۲۵ دانشگاه استنفورد نشان می‌دهد که حجم سرمایه‌گذاری شرکتی در این حوزه در سال ۲۰۲۴ به ۲۵۲.۳ میلیارد دلار رسیده است. بر اساس این گزارش، سرمایه‌گذاری در هوش مصنوعی طی دهه گذشته رشدی چشمگیر را تجربه کرده و مجموع آن از سال ۲۰۱۴ تاکنون بیش از سیزده برابر افزایش یافته است. افزون بر این، سرمایه‌گذاری خصوصی ایالات متحده در سال ۲۰۲۴ به ۱۰۹.۱ میلیارد دلار بالغ شده که به‌مراتب فراتر از سرمایه‌گذاری چین (۹.۳ میلیارد دلار) و بریتانیا (۴.۵ میلیارد دلار) بوده و به‌ترتیب حدود ۱۲ و ۲۴ برابر آن‌هاست. این نابرابری در حوزه هوش مصنوعی مولد حتی برجسته‌تر است؛ به‌گونه‌ای که سرمایه‌گذاری ایالات متحده با فاصله‌ای معادل ۲۵.۴ میلیارد دلار از مجموع سرمایه‌گذاری چین و اتحادیه اروپا به‌علاوه بریتانیا فراتر رفته و شکاف موجود در سال ۲۰۲۳ (۲۱.۸ میلیارد دلار) را بیش از پیش تعمیق کرده است (کمیته راهبری شاخص هوش مصنوعی[1]، ۲۰۲۵).

مطابق با گزارش مجمع جهانی اقتصاد[2] (۲۰۲۰)، هوش مصنوعی در کنار فناوری‌هایی نظیر رایانش ابری[3]، کلان‌داده[4]، بلاک‌چین[5]، رمزنگاری[6]، اینترنت اشیاء[7] و تجارت الکترونیک، از مهم‌ترین پیشران‌های تحول دیجیتال در جهان محسوب می‌شود. این فناوری به سامانه‌هایی اطلاق می‌شود که توانایی انجام وظایفی همچون استدلال، تصمیم‌گیری، یادگیری و پردازش زبان طبیعی را دارند؛ وظایفی که پیش‌تر به‌طور سنتی در قلمرو توانمندی‌های انسانی قرار می‌گرفتند (ناسا، ۲۰۲۴). در ادبیات علمی، هوش مصنوعی غالباً به‌عنوان تلاشی نظام‌مند برای شبیه‌سازی سازوکارهای شناختی و تصمیم‌گیری انسان در بسترهای محاسباتی تعریف می‌شود؛ تلاشی که از طریق پردازش اطلاعات، استنتاج داده‌محور و یادگیری تدریجی، دستیابی به اهداف مشخص را دنبال می‌کند (میکالف و گوپتا[8]، ۲۰۲۱). از این منظر، هوش مصنوعی ماهیتی میان‌رشته‌ای دارد و تلفیقی از علوم کامپیوتر، ریاضیات، آمار، فلسفه، فیزیولوژی و زبان‌شناسی است که

1. AI Index Steering Committee
2. World Economic Forum
3. Cloud Computing
4. Big Data
5. Blockchain
6. Cryptography
7. Internet of Things (IoT)
8. Mikalef & Gupta

می‌کوشد ویژگی‌های شناختی انسان را در قالب سیستم‌های رایانه‌ای بازنمایی کند (ثقفی و پارساپور، ۱۴۰۴).

کاربست‌های اصلی هوش مصنوعی شامل سیستم‌های خبره، یادگیری ماشین[1]، یادگیری عمیق[2]، پردازش زبان طبیعی[3]، بینایی ماشین[4] و تشخیص صوت است (شریدا و هشلامون[5]، ۲۰۲۱؛ شریدا و همکاران، ۲۰۲۰؛ پیروتی و همکاران[6]، ۲۰۲۴). تحولات این حوزه از دهه ۱۹۸۰ با توسعه سیستم‌های خبره آغاز شد؛ سامانه‌هایی که با اتکا بر قواعد از پیش تعریف‌شده، در حل مسائل تخصصی از جمله در حوزه حسابداری و حسابرسی به کار گرفته می‌شدند (وارزارو[7]، ۲۰۲۲). با پیشرفت یادگیری ماشین در دهه ۲۰۰۰، الگوریتم‌هایی توسعه یافتند که بدون برنامه‌ریزی صریح، قادر به یادگیری از داده‌ها و بهبود مستمر عملکرد خود بودند (پیروتی و همکاران، ۲۰۲۴). هم‌زمان، پردازش زبان طبیعی به‌عنوان یکی از زیرشاخه‌های کلیدی هوش مصنوعی مطرح شد که هدف آن تسهیل تعامل مؤثر میان انسان و ماشین در سطح زبان است (آگروال و گور[8]، ۲۰۲۰). در ادامه، ظهور یادگیری عمیق و شبکه‌های عصبی مصنوعی چندلایه، امکان تحلیل الگوهای پیچیده در داده‌های حجیم را فراهم ساخت و زیرساخت فنی لازم برای توسعه مدل‌های زبانی بزرگ را مهیا کرد (پیروتی و همکاران، ۲۰۲۴).

مدل‌های زبانی بزرگ حاصل همگرایی پیشرفت‌ها در حوزه یادگیری عمیق، پردازش زبان طبیعی و توان محاسباتی هستند و هسته اصلی آن‌ها بر معماری ترنسفورمر[9] استوار است. معماری ترنسفورمر با هدف رفع محدودیت‌های شبکه‌های عصبی بازگشتی[10] طراحی شده و به‌جای پردازش ترتیبی داده‌ها، امکان پردازش هم‌زمان کل توالی زبانی را فراهم می‌سازد. این ویژگی سبب می‌شود مدل بتواند وابستگی‌های کوتاه‌مدت و بلندمدت میان واژگان را با دقت بالاتری شناسایی و مدل‌سازی کند. عنصر کلیدی در این معماری، سازوکار خودتوجهی[11] است که به

1. Machine Learning (ML)
2. Deep Learning
3. Natural Language Processing (NLP)
4. Machine Vision
5. Sharida & Hashlamon
6. Pierotti et al.
7. Va˘rzaru
8. Agarwal & Gaur
9. Transformer
10. Recurrent Neural Netw
11. Self-Attention

مدل اجازه می‌دهد میزان اهمیت هر واژه را نسبت به سایر واژگان در همان جمله یا متن محاسبه کند. در این فرایند، برای هر واژه بردارهایی موسوم به پرسش (Query)، کلید (Key) و مقدار (Value) ایجاد می‌شود و با محاسبه وزن‌های توجه، نقش هر واژه در تفسیر معنای کلی متن تعیین می‌گردد. بدین ترتیب، مدل قادر است روابط معنایی پیچیده، حتی میان واژگانی که در فاصله‌های دور از یکدیگر قرار دارند، را به‌طور مؤثر درک کند. پس از مرحله خودتوجهی، اطلاعات حاصل از طریق لایه‌های پیش‌خور[1] پردازش شده و با بهره‌گیری از لایه‌های نرمال‌سازی و اتصال‌های میان‌بر[2]، پایداری یادگیری و کارایی مدل افزایش می‌یابد. عملکرد نهایی مدل‌های زبانی بزرگ مبتنی بر پیش‌بینی واژه بعدی در یک توالی زبانی است؛ فرآیندی که در جریان آموزش بر روی حجم عظیمی از داده‌های متنی، به تدریج منجر به یادگیری ساختارهای نحوی، الگوهای معنایی و دانش ضمنی زبان می‌شود (واسوانیو همکاران[3]، ۲۰۱۷).

در نتیجه این معماری، مدل‌های زبانی بزرگ توانایی تولید متونی منسجم، معنادار و نزدیک به زبان انسانی را کسب می‌کنند و دامنه کاربرد آن‌ها به حوزه‌هایی نظیر ترجمه ماشینی، خلاصه‌سازی متون، پاسخ‌گویی به پرسش‌ها، تولید محتوا و توسعه چت‌بات‌ها گسترش یافته است. در حوزه حسابداری و حسابرسی نیز، این مدل‌ها به دلیل توانایی در تحلیل متون تخصصی، تفسیر مفاهیم پیچیده و پردازش زبان طبیعی، به ابزارهایی کارآمد برای پشتیبانی از آموزش دانشگاهی، تحلیل استانداردها و تصمیم‌گیری حرفه‌ای تبدیل شده‌اند.

در سطح کاربردی، هوش مصنوعی تأثیرات قابل‌توجهی بر فرایندهای حسابداری و حسابرسی برجای گذاشته است. پیشرفت‌های سریع در الگوریتم‌های یادگیری ماشین، پردازش زبان طبیعی و تحلیل کلان‌داده‌ها، زمینه‌ساز بهبود فرایندهای سنتی و ارائه راهکارهای نوین برای ارتقای کارایی و دقت در این حوزه‌ها شده است (رحمانی و همکاران، ۱۴۰۴). به‌طور مشخص، هوش مصنوعی با بهینه‌سازی فرایندهای حسابرسی، شناسایی ناهنجاری‌های مالی، پردازش حجم عظیمی از داده‌ها و کاهش خطاهای انسانی، نقش مؤثری در افزایش اثربخشی فعالیت‌های حرفه‌ای ایفا می‌کند. از این‌رو، سازمان‌ها به‌طور فزاینده‌ای از این فناوری برای جمع‌آوری، یکپارچه‌سازی و تبدیل داده‌ها از منابع متنوع و استخراج اطلاعات معنادار به‌منظور بهبود

1. Feedforward
2. Residual Connections
3. Vaswani et al.

تصمیم‌گیری در محیط‌های پیچیده و دستیابی به منافع اقتصادی استفاده می‌کنند (مشایخی و امراللهی، ۱۴۰۴).

علاوه بر این، هوش مصنوعی نقشی اساسی در ارتقای شفافیت و پاسخگویی در گزارش‌های مالی و گزارشگری پایداری ایفا می‌کند؛ به‌گونه‌ای که کاهش مداخله انسانی در فرایند تهیه گزارش‌ها می‌تواند میزان خطا را محدود کرده و عینیت داده‌های افشاشده را افزایش دهد (نوراحمدی و پارسی، ۱۴۰۴). با وجود این مزایا، به‌کارگیری هوش مصنوعی با چالش‌ها و مخاطرات اخلاقی نیز همراه است. مهم‌ترین چالش‌های اخلاقی در حوزه مالی شامل بروز سوگیری و تبعیض الگوریتمی، فقدان شفافیت در سازوکارهای تصمیم‌گیری، ابهام در مسئولیت‌پذیری و مخاطرات بالقوه برای ثبات بازارهای مالی است؛ مسائلی که مستلزم توجه نظام‌مند پژوهشگران، نهادهای حرفه‌ای و سیاست‌گذاران می‌باشد (رهنما و رفعتی، ۱۴۰۴).

سیر تحول مدل‌های زبانی بزرگ

سیر تکاملی مدل‌های زبانی بزرگ، بازتابی از رشد نمایی توان پردازشی، بهبود معماری‌های الگوریتمی و توسعه داده‌های آموزشی در سطح جهانی است. نقطه آغاز این تحول را می‌توان به معرفی مدل GPT-1 توسط شرکت OpenAI در سال ۲۰۱۸ نسبت داد. این مدل، با اتکا به داده‌های متنی و بدون نیاز به برچسب‌گذاری، توانایی تولید متون معنادار را برای نخستین بار اثبات کرد. با توسعه نسخه‌های پیشرفته‌تر مانند GPT-2، GPT-3 و در نهایت GPT-4، قدرت پیش‌بینی، انسجام زبانی، و دامنه عملکرد این مدل‌ها به‌طرز چشمگیری ارتقاء یافت (اس وای پارتنرز[1]، ۲۰۲۵). همچنین در سال ۲۰۲۴، شرکت OpenAI مدل‌هایی مانند o1 و o3 را معرفی کرد که برای استدلال تکرارشونده بر روی خروجی‌های خود طراحی شده‌اند. این رویکرد «محاسبه در زمان آزمون» به‌طور چشمگیری عملکرد را بهبود داد؛ به‌طوری‌که مدل o1 در آزمون ورودی المپیاد بین‌المللی ریاضی امتیاز ۷۴/۴٪ کسب کرد، در حالی که GPT-4o تنها ۹/۳٪ به دست آورد. با این حال، این استدلال پیشرفته هزینه‌بر است: مدل o1 تقریباً شش برابر گران‌تر و ۳۰ برابر کندتر از GPT-4o است (کمیته راهبری شاخص هوش مصنوعی، ۲۰۲۵).

تحول دوم در این حوزه با معرفی مدل‌های چندوجهی همچون Gemini از سوی شرکت Google DeepMindرقم خورد. این مدل‌ها با قابلیت پردازش هم‌زمان داده‌های متنی،

1. SY Partners

تصویری، صوتی و ویدئویی، گامی فراتر از مدل‌های صرفاً زبانی برداشتند و امکان تعامل چندحسی و استدلال چندلایه را فراهم آوردند (هاشمی‌پور و همکاران، ۲۰۲۵؛ ژانگیانگ و همکاران[1]، ۲۰۲۳). نسخه‌های پیشرفته‌تر نظیر Gemini 2.5 Pro، با توانایی تولید صوت چندزبانه، تفکر سطح بالا و پایداری امنیتی بیشتر، مرزهای جدیدی در تعامل انسان–ماشین گشودند (تک‌کراچ[2]، ۲۰۲۵).

در کنار این پیشرفت‌ها، سایر مدل‌های نوآورانه نیز مطرح شدند. Perplexity AI با بهره‌گیری همزمان از مدل زبانی و جست‌وجوی بلادرنگ، پاسخ‌هایی مبتنی بر داده‌های به‌روز ارائه می‌دهد و با ارائه منابع درون‌متنی، رویکردی شفاف و مستند در پاسخ‌گویی را دنبال می‌کند (گینس[3]، ۲۰۲۴). Grok مدل توسعه‌یافته توسط شرکت xAI، با تمرکز بر تعامل اجتماعی و داده‌های شبکه‌ای، سعی دارد تولید محتوای شخصی‌سازی‌شده را در بسترهای اجتماعی تسهیل کند (گلوور[4]، ۲۰۲۵). مدل Qwen متعلق به شرکت Alibaba نیز با پشتیبانی از ۲۷ زبان و بهره‌گیری از ساختارهای نوین، یکی از پیشرفته‌ترین مدل‌های متن‌باز آسیایی به‌شمار می‌رود (گروه علی‌بابا، ۲۰۲۴).

مدل DeepSeek نیز، با اتخاذ رویکرد منبع‌باز و استفاده از معماری Mixture-of-Experts (MoE)، امکان تخصیص هوشمند منابع محاسباتی را فراهم کرده و با فعال‌سازی تنها بخشی از ۶۷۱ میلیارد پارامتر خود، هزینه پردازش را تا ۹۵ درصد کاهش داده است (چیپاگیری[5]، ۲۰۲۵). این مدل از لحاظ شفافیت علمی و کارایی محاسباتی، گامی مهم در مسیر توسعه مسئولانه هوش مصنوعی محسوب می‌شود.

در مجموع، شواهد موجود حاکی از آن است که مدل‌های زبانی بزرگ از سامانه‌هایی با کارکرد محدود زبانی به فناوری‌هایی چندمنظوره، انسان‌محور و مبتنی بر استدلال تحول یافته‌اند؛ به‌گونه‌ای که توانایی تحلیل، تبیین و حتی تولید دانش را از خود نشان می‌دهند. یافته‌های پژوهشی اخیر نشان می‌دهد که توان محاسباتی موردنیاز برای آموزش مدل‌های شاخص هوش مصنوعی تقریباً هر پنج ماه دو برابر می‌شود، اندازه مجموعه‌داده‌های آموزشی مدل‌های زبانی

1. Zhangyang et al.
2. TechCrunch
3. Guinness
4. Glover
5. Chippagiri

بزرگ در بازه‌های حدوداً هشت‌ماهه افزایش دوبرابری را تجربه می‌کند و میزان انرژی مصرفی برای فرایند آموزش نیز به‌صورت سالانه روندی فزاینده دارد. در این میان، سرمایه‌گذاری گسترده بخش صنعت همچنان نقش محرک اصلی در مقیاس‌پذیری مدل‌ها و بهبود عملکرد آن‌ها ایفا می‌کند (کمیته راهبری شاخص هوش مصنوعی، ۲۰۲۵).

در نتیجه این تحولات، مدل‌های زبانی بزرگ به‌طور فزاینده‌ای در حوزه‌هایی نظیر حسابداری، حقوق، آموزش و سلامت به‌عنوان دستیاران هوشمند به‌کار گرفته می‌شوند و سهم معناداری در ارتقای دقت، سرعت و کیفیت تصمیم‌گیری حرفه‌ای دارند. افزون بر این، شواهد تجربی ارائه‌شده در مطالعات اخیر نشان می‌دهد که نسخه‌های پیشرفته این مدل‌ها در آزمون‌های تخصصی حسابداری از جمله CPA و CMA عملکردی فراتر از سطح انتظار داشته‌اند و از این منظر می‌توانند به‌عنوان ابزارهای مکمل مؤثر در فرایندهای آموزشی و ارزیابی حرفه‌ای مورد استفاده قرار گیرند (واسارهلی و همکاران[1]، ۲۰۲۳).

پیشینه تجربی

پیشرفت‌های چشمگیر در حوزه هوش مصنوعی، به‌ویژه توسعه مدل‌های زبانی بزرگ نظیر ChatGPT، Gemini، Grok، Perplexity، DeepSeek، Qwen، LlaMA و Claude از زمینه‌ساز تحولی نوین در آموزش و ارزیابی دروس تخصصی همچون حسابداری شده‌اند. در سال‌های اخیر، تمرکز مطالعات بین‌المللی به بررسی عملکرد این مدل‌ها در آزمون‌های دانشگاهی و حرفه‌ای حسابداری معطوف شده است. یافته‌های پژوهش‌های خارجی، تصویری پیچیده از توانمندی‌های این مدل‌ها ارائه می‌دهد که بسته به نسخه، نوع آزمون و شرایط ارزیابی متفاوت است.

در پژوهشی گسترده، وود و همکاران (۲۰۲۳) عملکرد ChatGPT را با داده‌هایی از ۱۸۶ مؤسسه از ۱۴ کشور مورد بررسی قرار دادند. نتایج نشان داد که ChatGPT توانسته است به‌طور میانگین ۵۶.۵٪ از ۲۸۰۸۵ پرسش آزمون‌های حسابداری را به‌درستی پاسخ دهد، و در ۹.۴٪ دیگر نیز پاسخ‌هایی نسبتاً درست ارائه کرده است. در مقایسه، میانگین نمرات دانشجویان در همین

1. Vasarhelyi et al.

آزمون‌ها ۷۶.۷٪ بوده است. با این حال، ChatGPT در ۱۵.۸٪ آزمون‌ها عملکردی بهتر از میانگین دانشجویان داشت، به‌ویژه در سؤالات مفهومی و توصیفی.

الریش و همکاران[1] (۲۰۲۳، ۲۰۲۴) در دو مطالعه مجزا با استفاده از آزمون‌های حرفه‌ای بین‌المللی از جمله CPA، CMA، CIA و EA، نشان دادند که نسخه ChatGPT-3.5 تنها موفق به کسب میانگین نمره ۵۳.۱٪ شد و در هیچ‌یک از این آزمون‌ها قبول نشد. اما پس از ارتقاء به مدل ChatGPT-4، میانگین نمرات به ۶۹.۶٪ افزایش یافت. سپس با آموزش به روش "۱۰-شات" (افزایش ۶.۶٪) و فعال‌سازی توانایی استدلال و استفاده از ابزارهای کمکی مانند ماشین‌حساب (افزایش ۸.۹٪)، نمرات به سطح ۸۵.۱٪ رسید و مدل در تمامی بخش‌های آزمون‌ها موفق شد.

مطالعه آموآه و همکاران (۲۰۲۴) نیز به‌صورت تجربی عملکرد چهار مدل (Claude، ChatGPT-4، ChatGPT-3.5 و Gemini) را با استفاده از آزمون ICAG (یک نوع آزمون حرفه‌ای حسابداری بین‌المللی) بررسی کرده است. مدل‌های آموزش‌ندیده به‌ترتیب نمرات ۷۹.۷۵٪، ۷۷٪، ۵۴.۳۸٪ و ۵۰.۲۵٪ کسب کردند. پس از آموزش، نمرات به‌ترتیب به ۷۹.۸۸٪، ۸۰.۲۵٪، ۵۹.۳۸٪ و ۵۹٪ افزایش یافت. تفاوت عملکرد بین مدل‌ها از نظر آماری معنادار بود (p=۰.۰۲۹) و مشخص شد که Claude و ChatGPT-4 حتی در حالت آموزش‌ندیده نیز از داوطلبان انسانی عملکرد بهتری داشتند (p=۰.۰۱۹).

در پژوهشی دیگر، زاکر و کوپاناگاری[2] (۲۰۲۴) پنج مدل زبانی از جمله ChatGPT-4، Gemini، Claude، Mixtral و Llama-2b را در پاسخ به سؤالات واقعی آزمون CPA با نرخ قبولی داوطلبان انسانی مقایسه کردند. نتایج نشان داد ChatGPT-4 و Claude Opus توانستند در برخی از بخش‌ها نظیر حسابداری و گزارشگری مالی (FAR) و مقررات (REG) عملکرد بهتری از داوطلبان انسانی ارائه دهند، در حالی‌که Gemini و Mixtral عملکردی نوسانی و ضعیف‌تر داشتند. مدل Llama-2b نیز بدون آموزش و راهنمایی اضافی در هیچ‌یک از بخش‌ها عملکرد قابل‌قبولی از خود نشان نداد.

1. Eulerich et al.
2. Zacher & Kuppannagari

بوماریتو و همکاران[1] (۲۰۲۳) نیز با تمرکز بر مدل TEXT-davINcI-003 از OpenAI دریافتند که این مدل در بخش مقررات (REG) از آزمون CPA تنها توانسته است در ۱۴.۴٪ موارد پاسخ صحیح دهد. هرچند در برخی سطوح شناختی همچون «درک و حافظه» و در صورت حذف مؤلفه‌های محاسباتی، عملکرد آن به سطح انسانی نزدیک شده است. با استفاده از شیوه پرامپت نویسی مؤثر، دقت پاسخ‌ها تا ۵۷.۶٪ ارتقا یافته و در ۸۲.۱٪ موارد، یکی از دو پاسخ پیشنهادی برتر، پاسخ صحیح بوده است.

مطالعه دِل و اکپان (۲۰۲۴) نشان می‌دهد که ChatGPT 3.5 در یک محیط آموزشی مبتنی بر آزمون چندگزینه‌ای، حدود ۵۰٪ مواقع پاسخ‌های درستی به سؤالات حسابداری ارائه می‌دهد. این پژوهش تأکید دارد که کاربرد مدل‌های زبانی در محیط‌های آموزشی، علاوه بر مزایا، نیازمند آگاهی نسبت به خطاهای احتمالی و پدیده «توهم پاسخ» است و باید دانشجویان را به تفکر انتقادی تشویق کرد.

از سوی دیگر، دِ فریتاس و همکاران (۲۰۲۴) ChatGPT-4 را در آزمون‌های کفایت حسابداری ارزیابی کرده‌اند. مدل در هر چهار دوره آزمون موفق به قبولی شد و میانگین پاسخ‌های صحیح آن برابر با ۷۱٪ بود؛ در حالی که تنها ۲۰ تا ۲۳ درصد از داوطلبان انسانی توانسته بودند این آزمون‌ها را با موفقیت بگذرانند.

در پرتغال، آلبوکرکی و گومس دوس سانتوس[2] (۲۰۲۴) عملکرد ChatGPT را در آزمون حسابدار رسمی این کشور بررسی کردند. نتایج نشان داد با وجود اینکه این ابزار در درک زمینه سؤالات موفق بوده، اما در بخش‌هایی که نیاز به قضاوت حرفه‌ای داشته، دقت پایین‌تری داشته است. ChatGPT در نهایت نتوانست نمره قبولی کسب کند، هرچند به آن نزدیک شده بود.

با مرور این مطالعات، مشخص می‌شود که عملکرد مدل‌های زبانی بزرگ در پاسخ به سؤالات حسابداری بسته به نسخه مدل، نوع آزمون، سطح تخصصی سؤال و وجود یا عدم وجود راهنمایی اولیه (پرامپت نویسی) متغیر است. درحالی‌که نسخه‌های جدیدتر مانند ChatGPT-4 و Claude در بسیاری از ارزیابی‌ها به عملکرد انسانی نزدیک شده‌اند یا حتی از آن پیشی گرفته‌اند، مدل‌هایی چون Gemini و Llama-2b هنوز در سطحی پایین‌تر قرار دارند.

1. Bommarito et al.
2. Albuquerque & Gomes dos Santos

در ایران با وجود انجام برخی پژوهش‌ها پیرامون کاربرد هوش مصنوعی در حوزه‌های حسابداری و حسابرسی (مانند پژوهش‌های ثقفی و پارساپور، ۱۴۰۴؛ نوراحمدی و پارسی، ۱۴۰۴؛ عدنان حمود و همکاران، ۱۴۰۴؛ مشایخی و امراللهی، ۱۴۰۴)، تاکنون هیچ پژوهشی به‌طور مشخص و مقایسه‌ای عملکرد مدل‌های زبانی بزرگ را در پاسخ‌گویی به پرسش‌های چندگزینه‌ای آزمون‌های تخصصی در حوزه حسابداری یا حسابرسی بررسی نکرده است. از این‌رو، پژوهش حاضر با رویکردی نوآورانه و مقایسه‌ای، به بررسی عملکرد شش مدل زبانی پرکاربرد در پاسخ به سؤالات آزمون دکترای حسابداری پرداخته و تلاش می‌کند تصویری روشن از میزان قابلیت‌ها این مدل‌ها در پاسخ دهی به سؤالات چهارگزینه‌ای زمینه‌های مختلف این آزمون در ایران ارائه دهد. این مطالعه نه‌تنها می‌تواند به درک بهتر مخاطرات و ظرفیت‌های بالقوه هوش مصنوعی در آموزش حسابداری کمک کند، بلکه بستری برای توسعه چارچوب‌های ارزیابی آینده‌نگرانه در محیط آموزشی و حرفه‌ای ایران فراهم می‌آورد.

فرضیه‌های پژوهش

تحولات سریع در حوزه هوش مصنوعی و به‌ویژه توسعه مدل‌های زبانی بزرگ، زمینه‌ساز گسترش استفاده از این مدل‌ها در حوزه‌های تخصصی علمی، از جمله حسابداری شده است. پژوهش‌های متعددی در حوزه‌هایی نظیر پزشکی، حقوق، و علوم رایانه نشان داده‌اند که این مدل‌ها قادرند به‌گونه‌ای مؤثر به پرسش‌های تخصصی پاسخ دهند (کونگ و همکاران، ۲۰۲۳؛ کاتز و همکاران، ۲۰۲۴؛ مندونکا، ۲۰۲۴). در حوزهٔ حسابداری نیز، مطالعاتی نظیر وود و همکاران (۲۰۲۳)، آموآه و همکاران (۲۰۲۴)، و دل و اکپان (۲۰۲۴) به ارزیابی توانایی مدل‌هایی چون ChatGPT در آزمون‌های حسابداری پرداخته‌اند و حاکی از آن هستند که این مدل‌ها می‌توانند به درصدی قابل قبول از پاسخ‌های صحیح دست یابند. برای نمونه، وود و همکاران (۲۰۲۳) گزارش کردند که ChatGPT در پاسخ به ۲۸۰۰۰ سؤال آزمون حسابداری موفق به کسب میانگین ۵۶.۵٪ پاسخ صحیح شده است، در حالی‌که عملکرد انسانی در همان آزمون‌ها حدود ۷۶.۷٪ بوده است. همچنین، نتایج برخی مطالعات حاکی از عملکرد بالای ۶۰٪ مدل‌هایی چون Gemini و DeepSeek در پرسش‌های تخصصی هستند (آموآه و همکاران، ۲۰۲۴؛ گینس، ۲۰۲۴).

با تکیه بر مبانی نظری و تجربی مذکور، پژوهش حاضر دو دسته فرضیه را برای هر یک از شش مدل زبانی بزرگ شامل ChatGPT، Gemini، Perplexity، Grok، DeepSeek و Qwen مطرح می‌کند.

فرضیه‌های دسته اول: آزمون توانایی مفهومی فراتر از حد تصادفی

هدف این دسته از فرضیه‌ها بررسی این است که آیا عملکرد مدل‌ها به‌طور معنادار فراتر از سطح شانس در آزمون‌های چهارگزینه‌ای است.

فرضیه ۱-۱: مدل ChatGPT در پاسخگویی به پرسش‌های آزمون دکترای حسابداری، عملکردی معنادار بالاتر از سطح تصادفی (۲۵ درصد) دارد.

فرضیه ۱-۲: مدل Gemini در پاسخگویی به پرسش‌های آزمون دکترای حسابداری، عملکردی معنادار بالاتر از سطح تصادفی (۲۵ درصد) دارد.

فرضیه ۱-۳: مدل Perplexity در پاسخگویی به پرسش‌های آزمون دکترای حسابداری، عملکردی معنادار بالاتر از سطح تصادفی (۲۵ درصد) دارد.

فرضیه ۱-٤: مدل DeepSeek در پاسخگویی به پرسش‌های آزمون دکترای حسابداری، عملکردی معنادار بالاتر از سطح تصادفی (۲۵ درصد) دارد.

فرضیه ۱-۵: مدل Grok در پاسخگویی به پرسش‌های آزمون دکترای حسابداری، عملکردی معنادار بالاتر از سطح تصادفی (۲۵ درصد) دارد.

فرضیه ۱-٦: مدل Qwen در پاسخگویی به پرسش‌های آزمون دکترای حسابداری، عملکردی معنادار بالاتر از سطح تصادفی (۲۵ درصد) دارد.

فرضیه‌های دسته دوم: بررسی دستیابی به حداقل عملکرد قابل قبول (۵۰٪ پاسخ صحیح)

این دسته از فرضیه‌ها بررسی می‌کند که آیا مدل‌ها قادر به دستیابی به حداقل ۵۰٪ پاسخ صحیح در آزمون‌های چهارگزینه‌ای هستند، که به‌عنوان آستانه‌ای عملی برای عملکرد قابل قبول در سطح تحصیلات تکمیلی در نظر گرفته شده است.

فرضیه ۲-۱: مدل ChatGPT در پاسخ به پرسش‌های آزمون دکترای حسابداری بیش از ۵۰ ٪ پاسخ صحیح ارائه می‌دهد.

فرضیه ۲-۲: مدل Gemini در پاسخ به پرسش‌های آزمون دکترای حسابداری بیش از ۵۰ ٪ پاسخ صحیح ارائه می‌دهد.

فرضیه ۲ -۳: مدل Perplexity در پاسخ به پرسش‌های آزمون دکترای حسابداری بیش از ۵۰٪ پاسخ صحیح ارائه می‌دهد.

فرضیه ۲ -٤: مدل DeepSeek در پاسخ به پرسش‌های آزمون دکترای حسابداری بیش از ۵۰٪ پاسخ صحیح ارائه می‌دهد.

فرضیه ۲ -۵: مدل Grok در پاسخ به پرسش‌های آزمون دکترای حسابداری بیش از ۵۰٪ پاسخ صحیح ارائه می‌دهد.

فرضیه ۲ -٦: مدل Qwen در پاسخ به پرسش‌های آزمون دکترای حسابداری بیش از ۵۰ ٪ پاسخ صحیح ارائه می‌دهد.

روش‌شناسی پژوهش

پژوهش حاضر از نظر هدف، کاربردی و از حیث روش، توصیفی–تحلیلی با رویکرد تجربی و مقایسه‌ای است. هدف اصلی مطالعه، ارزیابی و مقایسه عملکرد شش مدل زبانی بزرگ شامل ChatGPT، Gemini، Perplexity، Grok، DeepSeek و Qwen در پاسخ‌گویی به سؤالات آزمون دکتری حسابداری ایران است. تمرکز پژوهش بر سنجش عملکرد عملی این مدل‌ها در یک چارچوب رسمی و استاندارد است که شباهت بالایی با محیط‌های واقعی ارزشیابی آموزشی و رقابتی دارد.

جامعه و داده‌های پژوهش

جامعه آماری پژوهش شامل کلیه سؤالات تخصصی آزمون دکتری رشته حسابداری ایران طی دوره پنج‌ساله ۱۴۰۰ تا ۱۴۰۴ است. انتخاب این بازه زمانی با هدف کاهش اثر نوسانات مقطعی در سطح دشواری و ساختار سؤالات و افزایش نمایندگی محتوایی داده‌ها انجام شد. در مجموع، داده‌های پژوهش شامل ۳۰۰ پرسش چهارگزینه‌ای است که در سه حوزه اصلی طبقه‌بندی شدند: ۷۵ سؤال حسابرسی، ۱۰۰ سؤال حسابداری مدیریت و ۱۲۵ سؤال تئوری حسابداری.

از آنجا که تمامی پرسش‌های موجود در بازه زمانی تعریف‌شده مورد استفاده قرار گرفت، نمونه‌گیری انجام نشد و داده‌ها معادل کل جامعه آماری تلقی شدند. این حجم داده، با توجه به ماهیت مطالعات تجربی مشابه در حوزه ارزیابی مدل‌های زبانی بزرگ، برای دستیابی به نتایج

پایدار و قابل اتکا کفایت دارد. لازم به ذکر است که غالب سؤالات آزمون دکتری حسابداری ماهیتی مفهومی و نظری دارند؛ ازاین‌رو، تعمیم نتایج به سؤالات صرفاً محاسباتی نیازمند احتیاط است.

معیار انتخاب مدل‌های زبانی بزرگ

انتخاب مدل‌های زبانی بزرگ در این پژوهش بر اساس سه معیار انجام شد:

۱- سطح بلوغ فنی و اعتبار علمی–کاربردی در ادبیات بین‌المللی

۲- تنوع نهادی و فنی به‌منظور کاهش سوگیری ناشی از وابستگی به یک توسعه‌دهنده خاص

۳- دسترس‌پذیری عملی و پایدار برای کاربران داخل ایران در زمان اجرای پژوهش.

بر این اساس، شش مدل ChatGPT، Gemini، Perplexity، Grok، DeepSeek و Qwen به‌عنوان نمایندگان پرکاربرد نسل جدید مدل‌های زبانی بزرگ انتخاب شدند. اگرچه مدل‌هایی مانند LLaMA و Claude از نظر فنی در زمره مدل‌های پیشرفته قرار می‌گیرند، اما به دلیل محدودیت‌های دسترسی رسمی و نبود امکان استفاده پایدار در ایران در زمان اجرای پژوهش، از نمونه نهایی حذف شدند. این تصمیم با هدف حفظ یکپارچگی روش‌شناسی، افزایش قابلیت تکرارپذیری نتایج و جلوگیری از ایجاد شرایط نابرابر در ارزیابی مدل‌ها اتخاذ شد. در نتیجه، مدل‌های منتخب، بازتاب‌دهنده عملکرد عملی سامانه‌های هوش مصنوعی در دسترس در بستر آموزش عالی ایران هستند.

پروتکل اجرا و مشخصات فنی

به‌منظور افزایش تکرارپذیری و امکان ممیزی نتایج، تمامی مدل‌ها با تنظیمات تولید حداقلی (Temperature = 0) و بدون استفاده از مثال‌های راهنما (Zero-shot) مورد ارزیابی قرار گرفتند؛ رویکردی که در مطالعات پیشین (مانند: الریش و همکاران، ۲۰۲۳ و ۲۰۲۴؛ بوماریتو و همکاران، ۲۰۲۳) نیز به‌طور گسترده به کار رفته است. هر مدل در سه نوبت مستقل و با فاصله زمانی اجرا شد و تنها پاسخ نهایی تولیدشده به‌عنوان خروجی معتبر ثبت گردید.

تمامی پرسش‌ها به‌صورت متن ساده و بدون ارائه هرگونه راهنمای اضافی، مثال حل‌شده یا نشانه‌های زمینه‌ای، در محیط رسمی هر ابزار وارد شدند تا ثبات شرایط اجرا در میان شش مدل تضمین شود. جدول (۱) مشخصات دقیق اجرای هر یک از مدل‌های زبانی بزرگ، شامل نسخه، تاریخ اجرا، تنظیمات تولید، تعداد ران و قالب پرامپت را نشان می‌دهد.

جدول ۱. مشخصات اجرای مدل‌های زبانی بزرگ

Table 1. Execution and Coding Details of Large Language Models

Prompt Format	Runs	Top_p	Temperature	Execution Date	Version	Model
Zero-shot, text-only question input	3	Default (platform-defined)	0.0	2025/05/05	GPT-4	ChatGPT
Zero-shot, text-only question input	3	Default (platform-defined)	0.0	2025/05/05	Gemini-1.5	Gemini
Zero-shot, text-only question input	3	Default (platform-defined)	0.0	2025/05/06	Perplexity AI v2	Perplexity
Zero-shot, text-only question input	3	Default (platform-defined)	0.0	2025/05/06	Grok v1.2	Grok
Zero-shot, text-only question input	3	Default (platform-defined)	0.0	2025/05/07	DeepSeek v3.0	DeepSeek
Zero-shot, text-only question input	3	Default (platform-defined)	0.0	2025/05/07	Qwen-7B	Qwen

گردآوری داده‌ها و رویکرد ارزیابی

پاسخ‌های تولیدشده توسط هر یک از مدل‌های زبانی ثبت و با استفاده از کلیدهای رسمی پاسخ منتشرشده توسط سازمان سنجش آموزش کشور صحت‌سنجی شدند. برای هر مدل، نسبت پاسخ‌های صحیح به کل پرسش‌ها به‌عنوان شاخص اصلی عملکرد در نظر گرفته شد. رویکرد پژوهش حاضر، یک ارزیابی نتیجه‌محور است؛ بدین معنا که معیار قضاوت صرفاً درستی یا نادرستی پاسخ نهایی بوده و فرآیند استدلال یا مسیر تولید پاسخ مورد ارزیابی قرار نگرفته است. این رویکرد با منطق سنجش در آزمون‌های رسمی آموزشی هم‌راستا است.

شایان تأکید است که ارزیابی مدل‌های زبانی در این پژوهش در شرایط استفاده عمومی و پیش‌فرض آن‌ها انجام شده است؛ به‌گونه‌ای که هیچ‌گونه مداخله‌ای در فعال یا غیرفعال‌سازی قابلیت‌های بالقوه جست‌وجوی وب، بازیابی اطلاعات یا دسترسی به منابع بیرونی صورت نگرفته است. ازاین‌رو، سناریوی ارزیابی حاضر را باید نزدیک به وضعیت open-book تلقی کرد، نه به‌عنوان یک سنجش از توان درونی مدل‌ها در شرایط کاملاً بسته یا closed-book.

بر این اساس، نتایج به‌دست‌آمده منعکس‌کننده عملکرد عملی مدل‌ها در شرایط استفاده واقعی کاربران است و نباید به‌منزله سنجش مستقیم ظرفیت شناختی مستقل یا حافظه درونی آن‌ها تفسیر شود. افزون بر این، با توجه به ماهیت رسمی و استاندارد سؤالات آزمون دکتری حسابداری، احتمال نشت داده[1] یا بازیابی دانش پیش‌آموخته در مرحله آموزش مدل‌ها به‌طور صریح به‌عنوان یک محدودیت روش‌شناختی در نظر گرفته شده است. ازاین‌رو، تفسیر یافته‌ها با احتیاط انجام شده و هرگونه برداشت فراتر از سطح عملکرد در پاسخ‌گویی به سؤالات چهارگزینه‌ای اجتناب شده است.

راهنمای کدگذاری پاسخ‌ها

با توجه به ماهیت مولد پاسخ در مدل‌های زبانی بزرگ، تمامی پاسخ‌ها لزوماً به‌صورت انتخاب صریح یکی از گزینه‌های چهارگانه ارائه نمی‌شوند. ازاین‌رو، برای افزایش شفافیت، تکرارپذیری و جلوگیری از تفسیر ذهنی، یک راهنمای کدگذاری[2] دقیق تدوین و به‌صورت یکنواخت اعمال شد. پاسخ‌ها در پنج دسته عملیاتی طبقه‌بندی شدند.

(۱) پاسخ صریح و منطبق با گزینه صحیح؛

(۲) پاسخ صریح اما نادرست؛

(۳) پاسخ مبهم یا چندگزینه‌ای؛

(۴) امتناع از پاسخ یا اعلام عدم قطعیت؛

(۵) پاسخ خارج از قالب سؤال.

1. data leakage
1. Codebook

در تحلیل‌های آماری، تنها پاسخ‌های دسته (۱) با کد ۱ به‌عنوان پاسخ صحیح ثبت شدند و سایر موارد-صرف‌نظر از میزان نزدیکی محتوایی- با کد ۰ به‌عنوان پاسخ نادرست در نظر گرفته شدند. این رویکرد محافظه‌کارانه از بیش‌برآورد عملکرد مدل‌ها جلوگیری می‌کند.

بررسی روایی و پایایی ابزار

روایی صوری و محتوایی ابزار پژوهش، به دلیل استفاده از سؤالات رسمی و استاندارد آزمون دکتری حسابداری، قابل قبول تلقی می‌شود. افزون بر این، انطباق محتوای سؤالات با اهداف آموزشی مقطع دکتری توسط دو عضو هیئت علمی حسابداری تأیید شد.

برای ارزیابی پایایی و تکرارپذیری، هر مدل در سه اجرای مستقل با فاصله زمانی ۳ روز اجرا شد تا اثر حافظه کوتاه‌مدت کاهش یابد. میانگین دقت، انحراف معیار و بازه اطمینان ٪۹۵ برای هر مدل محاسبه شد و شاخص Fleiss' Kappa میزان توافق بین اجراها را نشان داد. نتایج، سطح بالای ثبات و تکرارپذیری عملکرد مدل‌ها را تأیید می‌کنند (Fleiss' Kappa ≥ 0.85). جدول (۲) خلاصه شاخص‌های پایایی، تکرارپذیری و بازه‌های اطمینان مدل‌های زبانی را ارائه می‌کند.

جدول ۲. پایایی، تکرارپذیری و بازه اطمینان ٪۹۵ دقت عملکرد مدل‌های زبانی در سه اجرای مستقل

Table 2. Reliability, Reproducibility, and 95% Confidence Intervals of Language Model Accuracy Across Three Independent Runs

Time Interval Between Runs	Fleiss' Kappa	± Standard Deviation (%)	CI %95 (%)	Mean Accuracy (%)	Model
3 days	0.87	±1.2	67.98-62.02	65.0	ChatGPT
3 days	0.88	±1.0	69.78-64.82	67.3	Gemini
3 days	0.86	±1.1	68.43-62.97	65.7	Perplexity
3 days	0.85	±1.3	66.98-61.02	64.0	Grok
3 days	0.85	±1.2	67.53-61.07	64.3	DeepSeek
3 days	0.86	±1.4	66.78-59.82	63.3	Qwen

تحلیل داده‌ها

تحلیل داده‌ها با استفاده از نرم‌افزار SPSS نسخه ۲۷ انجام شد. با توجه به ساختار چهارگزینه‌ای سؤالات و وجود تنها یک پاسخ صحیح برای هر پرسش، متغیر عملکرد به‌صورت دودویی تعریف شد؛ به‌طوری‌که به هر پاسخ صحیح مقدار یک و به هر پاسخ نادرست مقدار

صفر اختصاص یافت. در این چارچوب، هر پرسش به‌عنوان یک واحد تحلیل مستقل در نظر گرفته شد (n= ۳۰۰ برای هر مدل).

پیش از انجام آزمون‌های استنباطی، نرمال بودن توزیع داده‌ها با استفاده از آزمون‌های Kolmogorov–Smirnov و Shapiro–Wilk بررسی شد. نتایج نشان داد که توزیع داده‌ها برای تمامی مدل‌ها از نرمالیت تبعیت نمی‌کند ($p < 0.001$). ازاین‌رو، استفاده از آزمون‌های ناپارامتریک در تحلیل‌های مقایسه‌ای ضروری تشخیص داده شد.

تحلیل‌های آماری پژوهش بر مبنای رویکرد ارزیابی پیامد-محور انجام گرفت؛ رویکردی که در مطالعات تجربی ارزیابی مدل‌های زبانی بزرگ، به‌ویژه در آزمون‌های استاندارد آموزشی و حرفه‌ای، به‌طور گسترده به کار گرفته می‌شود. در این رویکرد، تمرکز بر درستی یا نادرستی پاسخ نهایی است و نه تحلیل فرایند تولید پاسخ یا بررسی کیفی استدلال‌ها. این رویکرد روش‌شناختی امکان مقایسه عملکرد عملی مدل‌ها در شرایط مشابه محیط‌های واقعی ارزیابی آموزشی را فراهم می‌کند.

برای آزمون فرضیه‌های پژوهش، از آزمون نسبت تک‌نمونه‌ای (One-sample proportion test) استفاده شد. در این آزمون، نسبت پاسخ‌های صحیح هر مدل زبانی با دو مقدار مرجع مقایسه گردید: نخست، مقدار مرجع ۰.۲۵ به‌منظور بررسی معناداری عملکرد فراتر از سطح تصادفی در آزمون‌های چهارگزینه‌ای؛ و دوم، مقدار مرجع ۰.۵۰ به‌عنوان حداقل آستانه قابل قبول برای عملکرد مفهومی در سطح تحصیلات تکمیلی. برای هر آزمون، مقدار آماره Z و سطح معناداری (p-value) محاسبه و گزارش شد و سطح خطای ۰.۰۵ به‌عنوان معیار تصمیم‌گیری آماری در نظر گرفته شد.

به‌منظور بررسی تفاوت عملکرد مدل‌ها نسبت به یکدیگر، با توجه به دودویی بودن داده‌ها، غیرنرمال بودن توزیع متغیرها و پاسخ‌گویی همه مدل‌ها به مجموعه‌ای یکسان از پرسش‌ها، از آزمون ناپارامتریک Cochran’s Q استفاده شد. این آزمون امکان مقایسه هم‌زمان چند گروه وابسته را فراهم می‌کند و برای داده‌های پژوهش حاضر مناسب تشخیص داده شد. نتایج نشان داد که تفاوت عملکرد کلی شش مدل از نظر آماری معنادار نیست ($p > 0.05$). در نتیجه، انجام آزمون‌های پسین زوجی یا تحلیل‌های پیشرفته‌تر ضروری تشخیص داده نشد.

لازم به تأکید است که این پژوهش تحلیل کیفی مسیرهای استدلال، سبک پاسخ‌دهی یا کیفیت توضیحات مدل‌ها را شامل نمی‌شود. چنین تحلیل‌هایی نیازمند طراحی پژوهشی جداگانه و چارچوب‌های تفسیری ذهنی هستند که خارج از دامنه و اهداف مطالعه حاضر قرار دارند. تمرکز این تحقیق بر تحلیل کمی استاندارد و قابل تکرار است و تصویری شفاف و قابل مقایسه از عملکرد عملی مدل‌های زبانی بزرگ در آزمون رسمی حسابداری ارائه می‌دهد. بررسی کیفی پاسخ‌ها می‌تواند به‌عنوان مسیر پژوهشی تکمیلی در مطالعات آینده مورد توجه قرار گیرد.

یافته‌های پژوهش

آمار توصیفی

به‌منظور بررسی عملکرد شش مدل زبانی بزرگ در پاسخ‌گویی به پرسش‌های تخصصی آزمون دکتری حسابداری، داده‌های حاصل از ۳۰۰ پرسش چهارگزینه‌ای طی دوره پنج‌ساله ۱۴۰۰ تا ۱۴۰۴ در سه درس «حسابرسی»، «حسابداری مدیریت» و «تئوری حسابداری» مورد تجزیه‌وتحلیل قرار گرفت. جدول (۳) تعداد پاسخ‌های صحیح هر مدل و جدول (۴) درصد پاسخ‌های صحیح مدل‌ها را به تفکیک درس و عملکرد کل آزمون گزارش می‌کند.

بر اساس جدول (۴)، مدل Gemini با دقت کلی ۶۷٫۳۳ درصد بالاترین عملکرد را به خود اختصاص داد، در حالی که مدل Qwen با ۶۳٫۳۳ درصد پایین‌ترین عملکرد کلی را داشت. سایر مدل‌ها نیز عملکردی نزدیک به یکدیگر ارائه کردند؛ به‌گونه‌ای که دقت کلی ChatGPT، Perplexity، Grok و DeepSeek به‌ترتیب ۶۵ درصد، ۶۵.۶۷ درصد، ۶۴ درصد و ۶۴.۳۳ درصد گزارش شد. این نزدیکی نسبی عملکردها نشان‌دهنده همگرایی سطح توان پاسخ‌گویی مدل‌ها در چارچوب آزمون استاندارد دکتری حسابداری است.

جدول ۳. تعداد پاسخ‌های صحیح مدل‌های زبانی بزرگ به سؤالات آزمون دکتری حسابداری به تفکیک سال و درس

Table 3. Number of Correct Responses by Large Language Models to PhD Accounting Examination Questions by Year and Subject

Year	Subject	Number of Questions	Total Correct Answers					
			ChatGPT	Gemini	Perplexity	Grok	DeepSeek	Qwen
2021	Audit	15	11	9	9	9	7	6
	Management Accounting	20	16	17	17	18	16	18
	Accounting Theory	25	14	16	15	15	15	15
2022	Audit	15	7	8	7	8	7	9
	Management Accounting	20	14	15	15	15	17	16
	Accounting Theory	25	14	16	15	14	18	14
2023	Audit	15	10	9	10	10	7	7
	Management Accounting	20	12	12	15	11	12	11
	Accounting Theory	25	15	20	16	16	20	16
2024	Audit	15	8	11	11	10	9	9
	Management Accounting	20	17	13	16	15	14	14
	Accounting Theory	25	17	14	16	15	12	13
2025	Audit	15	11	11	9	10	9	11
	Management Accounting	20	12	12	12	13	13	16
	Accounting Theory	25	17	17	14	13	17	15
All Years	Audit	75	47	48	46	47	39	42
	Management Accounting	100	71	69	75	72	72	75
	Accounting Theory	125	77	85	76	73	82	73
Total		300	195	202	197	192	193	190

جدول ٤. درصد پاسخ‌های صحیح مدل‌های زبانی بزرگ در آزمون دکتری حسابداری به تفکیک درس و عملکرد کل آزمون

Table 4. Percentage of Correct Responses of Large Language Models by Subject and Overall Performance

Subject	ChatGPT	Gemini	Perplexity	Grok	DeepSeek	Qwen
Audit	62.67%	64%	61.33%	62.67%	52%	56%
Management Accounting	71%	69%	75%	72%	72%	75%
Accounting Theory	61.60%	68%	60.80%	58.40%	65.60%	58.40%
Total	65%	67.33%	65.67%	64%	64.33%	63.33%

تحلیل درس‌محور نتایج نشان داد که عملکرد مدل‌ها به نوع محتوای علمی سؤالات حساس است. در درس «حسابداری مدیریت»، که عمدتاً شامل پرسش‌های مفهومی و مسئله‌محور است، Qwen و Perplexity بالاترین دقت (۷۵ درصد) را ثبت کردند، در حالی که Gemini کمترین عملکرد را داشت. در درس «تئوری حسابداری»، که ماهیتی انتزاعی و نظری دارد، Gemini با دقت ۶۸ درصد برتر بود و Qwen پایین‌ترین عملکرد را نشان داد. در درس «حسابرسی»، مدل Gemini با ۶۴ درصد عملکرد بالاتر و DeepSeek کمترین درصد پاسخ‌های صحیح را ارائه داد.

به‌طور کلی، نتایج نشان می‌دهد که برتری یک مدل در یک درس خاص الزاماً به معنای برتری آن در کل آزمون نیست و نوع درس، سطح انتزاع مفاهیم و ساختار شناختی پرسش‌ها نقش تعیین‌کننده‌ای در عملکرد مدل‌های زبانی بزرگ ایفا می‌کند.

بررسی نرمال بودن داده‌ها

پیش از انجام تحلیل‌های استنباطی، توزیع داده‌های هر مدل از نظر نرمال بودن با آزمون‌های Kolmogorov–Smirnov و Shapiro–Wilk بررسی شد. نتایج نشان داد که داده‌ها برای تمامی مدل‌ها از نرمالیت تبعیت نمی‌کنند (۰/۰۵ > سطح معناداری). البته این نتیجه، با توجه به ماهیت دودویی متغیر عملکرد (صحیح/غلط)، قابل انتظار بوده و استفاده از آزمون‌های ناپارامتریک را برای تحلیل‌های مقایسه‌ای توجیه می‌کند.

آمار استنباطی

آزمون نسبت تک‌نمونه‌ای

برای بررسی معناداری عملکرد هر مدل به‌صورت مستقل، آزمون نسبت تک‌نمونه‌ای بر اساس دو مقدار مرجع ۰.۲۵ (سطح تصادفی در آزمون چهارگزینه‌ای) و ۰.۵۰ (حداقل آستانه عملکرد قابل قبول در سطح تحصیلات تکمیلی) انجام شد. نتایج این آزمون‌ها در جداول (۵) و (۶) گزارش شده‌اند.

جدول ۵. نتایج آزمون نسبت تک‌نمونه‌ای عملکرد مدل‌ها نسبت به سطح تصادفی (۰.۲۵)

Table 5. Results of the One-Sample Proportion Test of Model Performance Relative to the Random Baseline (0.25)

Test Result	Significance Level (One-tailed)	Z-Statistic	Difference from Test Value (0.25)	Success Rate	Model
Hypothesis H1-1 Supported	< 0.001	16.000	+0.400	65.0%	ChatGPT
Hypothesis H1-2 Supported	< 0.001	16.933	+0.423	67.3%	Gemini
Hypothesis H1-3 Supported	< 0.001	16.267	+0.407	65.7%	Perplexity
Hypothesis H1-4 Supported	< 0.001	15.733	+0.393	64.3%	Grok
Hypothesis H1-5 Supported	< 0.001	15.600	+0.390	64.0%	DeepSeek
Hypothesis H1-6 Supported	< 0.001	15.333	+0.383	63.3%	Qwen

همان‌طور که در جدول (۵) مشاهده می‌شود، تمامی شش فرضیه دسته اول پژوهش تأیید شده و مدل‌های زبانی مورد بررسی عملکردی معنادار فراتر از سطح تصادفی (۲۵٪) از خود نشان دادند. نسبت موفقیت برای مدل‌ها بین ۶۳٪ تا ۶۸٪ متغیر بود که همگی به‌طور معناداری بالاتر از مقدار آزمون (۰/۲۵) قرار داشتند (تمامی سطح‌های معناداری کمتر از ۰/۰۱). بزرگ‌ترین مقدار

آماره Z مربوط به مدل Gemini با میزان ۱۶/۹۳۳ و کمترین مقدار Z مربوط به مدل Qwen با ۱۵/۳۳۳ است. این نتایج نشان می‌دهد که همه مدل‌ها توانسته‌اند به‌طور معناداری از حد تصادفی عملکرد بهتری داشته باشند.

همچنین طبق نتایج مندرج در جدول (۶)، تمامی شش فرضیه دسته دوم پژوهش تأیید شده و عملکرد همه مدل‌ها به‌طور معناداری فراتر از سطح پایه ۵۰٪ نیز بوده است که نشان‌دهنده توان قابل قبول آن‌ها در پاسخ‌گویی به پرسش‌های تخصصی آزمون دکتری حسابداری است. آماره Z در این مرحله نیز در همه مدل‌ها به‌طور معناداری مثبت و سطح معناداری آزمون‌ها کمتر از ۰.۰۱ است. بالاترین اختلاف نسبت موفقیت با سطح پایه مربوط به مدل Gemini با تفاوت ۱۵/۷ درصد و آماره ۶/۰۰۴ و کمترین آن مربوط به Qwen با اختلاف ۱۶/۳ درصد و ۴/۶۱۹ است.

جدول ٦. نتایج آزمون نسبت تک‌نمونه‌ای عملکرد مدل‌ها نسبت به سطح پایه قابل قبول (۰/۵۰)

Table 6. Results of the One-Sample Proportion Test of Model Performance Relative to the Acceptable Baseline (0.50)

Test Result	Significance Level (One-tailed)	Z-Statistic	Difference from Test Value (0.50)	Success Rate	Model
Hypothesis H1-1 Supported	< 0.001	5.196	+0.150	65.0%	ChatGPT
Hypothesis H1-2 Supported	< 0.001	6.004	+0.173	67.3%	Gemini
Hypothesis H1-3 Supported	< 0.001	5.427	+0.157	65.7%	Perplexity
Hypothesis H1-4 Supported	< 0.001	4.965	+0.143	64.3%	Grok
Hypothesis H1-5 Supported	< 0.001	4.850	+0.140	64.0%	DeepSeek
Hypothesis H1-6 Supported	< 0.001	4.619	+0.133	63.3%	Qwen

به گونه کلی این نتایج نشان می‌دهد که مدل‌ها عملکردی معنادار فراتر از سطح شانس و سطح پایه ارائه می‌کنند، هرچند این نتایج به معنای استدلال مفهومی مستقل مدل‌ها نیست و صرفاً نشان‌دهنده توان عملی آن‌ها در پاسخ‌گویی به پرسش‌های استاندارد است.

مقایسه عملکرد مدل‌ها: آزمون Cochran's Q

برای بررسی تفاوت عملکرد مدل‌ها نسبت به یکدیگر، از آزمون ناپارامتریک Cochran's Q استفاده شد که برای داده‌های دودویی و وابسته مناسب است. نتایج آزمون Cochran's Q نشان داد که تفاوت عملکرد کلی شش مدل معنادار نیست (Q= ۲/۹۰۸؛ df= ۵ و p=۰/۷۱۴). با این حال، مقادیر دقت توصیفی نشان‌دهنده تفاوت‌های جزئی و عملی بین مدل‌هاست که از نظر کاربردی قابل توجه هستند.

بنابراین نمی‌توان ادعا کرد که یک مدل مشخص، مانند Gemini، به‌طور معنادار بر سایر مدل‌ها برتری دارد. نزدیکی عملکرد مدل‌ها و عدم معناداری آزمون بیانگر سطح مشابه توان پاسخ‌گویی آن‌ها در چارچوب آزمون استاندارد دکتری حسابداری است. از این رو، انجام آزمون‌های مقایسه‌ای زوجی یا تحلیل‌های پیشرفته‌تر به‌عنوان آزمون‌های پسین ضروری تشخیص داده نشد.

بحث و نتیجه‌گیری

هدف این پژوهش، ارزیابی و مقایسه عملکرد شش مدل زبانی بزرگ شامل ChatGPT، Gemini، Perplexity، Grok، DeepSeek و Qwen در پاسخ‌گویی به سؤالات آزمون دکتری حسابداری ایران و تبیین ظرفیت بالقوه آن‌ها برای کاربردهای آموزشی و ارزشیابی در آموزش عالی حسابداری بود. این ارزیابی با استفاده از ورودی‌های متنی یکسان و بدون بهره‌گیری از راهبردهای پیشرفته پرامپت‌نویسی یا مثال‌های راهنما انجام شد؛ با این حال، همان‌گونه که در بخش روش‌شناسی تصریح شد، مدل‌ها در شرایط استفاده عمومی و پیش‌فرض خود مورد آزمون قرار گرفتند که از منظر تفسیری به سناریویی نزدیک به open-book شباهت دارد، نه یک وضعیت کاملاً بسته یا closed-book.

نتایج نشان داد که تمامی مدل‌های مورد بررسی، عملکردی معنادار بالاتر از سطح تصادفی (۲۵ درصد) و سطح پایه قابل قبول (۵۰ درصد) در پاسخ‌گویی به سؤالات چندگزینه‌ای داشته‌اند. این یافته نشان می‌دهد که مدل‌های زبانی بزرگ قادرند الگوهای دانشی و ساختار مفهومی نهفته در سؤالات استاندارد حسابداری در سطح تحصیلات تکمیلی را به‌نحو مؤثری پردازش کرده و پاسخ‌های صحیح تولید کنند. با این حال، این نتایج صرفاً ناظر بر عملکرد خروجی‌محور در قالب سؤالات چندگزینه‌ای است و نباید به‌عنوان شواهدی قطعی از درک مفهومی عمیق، استدلال مستقل یا توان حل مسئله در شرایط closed-book تفسیر شود.

از منظر توصیفی، تفاوت‌هایی در میانگین دقت مدل‌ها مشاهده شد؛ به‌گونه‌ای که مدل Gemini بالاترین و مدل Qwen پایین‌ترین نرخ پاسخ صحیح را ثبت کردند. با این وجود، نتایج آزمون ناپارامتریک Cochran’s Q نشان داد که این تفاوت‌ها از نظر آماری معنادار نیستند. این عدم معناداری به‌معنای برابری کامل عملکرد مدل‌ها نیست، بلکه بیانگر آن است که در چارچوب طراحی پژوهش، حجم نمونه و واریانس بین اجراها، شواهد آماری کافی برای ادعای برتری قطعی یک مدل خاص وجود ندارد.

تفسیر یافته‌ها مستلزم توجه جدی به ماهیت سناریوی ارزیابی است. موفقیت نسبی مدل‌ها در پاسخ‌گویی به سؤالات رسمی و پرتکرار آزمون دکتری حسابداری می‌تواند، دست‌کم تا حدی، تحت تأثیر نشت داده‌های آموزشی، هم‌پوشانی محتوایی با داده‌های پیش‌آموخته یا بازیابی الگوهای ذخیره‌شده قرار گرفته باشد. ازاین‌رو، یافته‌های این پژوهش بیش از آنکه بازتاب‌دهنده «درک مفهومی مستقل» مدل‌ها در یک وضعیت closed-book باشد، نشان‌دهنده آمادگی عملی آن‌ها برای ارائه پاسخ صحیح در شرایط واقعی وopen-book گونه استفاده کاربران است. این تمایز تفسیری برای کاربردهای آموزشی، طراحی فعالیت‌های ارزشیابی و سیاست‌گذاری آموزشی اهمیت اساسی دارد.

یافته‌های پژوهش حاضر با بخش قابل توجهی از مطالعات پیشین در حوزه ارزیابی مدل‌های زبانی در آموزش حسابداری هم‌راستا است. برای مثال، وود و همکاران (۲۰۲۳) دقت ChatGPT را در آزمون‌های حسابداری حدود ۵۶/۵ درصد گزارش کرده‌اند که با نتایج بالاتر این مدل در پژوهش حاضر قابل مقایسه است. این تفاوت می‌تواند ناشی از به‌روزرسانی نسخه‌های مدل، تفاوت زبان آزمون و زمینه آموزشی باشد. همچنین، آموآه و همکاران (۲۰۲۴)

نشان داده‌اند که مدل‌هایی نظیر ChatGPT-4 و Claude قادرند بدون آموزش اختصاصی، در شرایط خاص به دقت‌هایی نزدیک یا حتی بالاتر از کاربران انسانی دست یابند. اختلاف میان نتایج آن‌ها و یافته‌های پژوهش حاضر را می‌توان به تفاوت نسخه مدل‌ها، نوع آزمون، سطح استانداردسازی پرسش‌ها و نحوه تعامل با مدل‌ها نسبت داد.

مطالعات الزریش و همکاران (۲۰۲۳، ۲۰۲۴) نیز نشان می‌دهد که بهره‌گیری از راهبردهایی مانند یادگیری چندمثالی، پرامپت‌نویسی هدفمند و ابزارهای کمکی می‌تواند عملکرد مدل‌های زبانی را به‌طور چشمگیری افزایش دهد. در پژوهش حاضر، مدل‌ها عمداً بدون چنین مداخلاتی و در یک چارچوب پایه مورد ارزیابی قرار گرفتند؛ ازاین‌رو، نزدیکی عملکرد مدل‌ها و عدم معناداری تفاوت‌ها را می‌توان به‌عنوان بازتابی از سطح پایه مشترک عملکرد آن‌ها در یک سناریوی open-book عملیاتی تلقی کرد که در صورت بهینه‌سازی محیط استفاده، قابلیت ارتقا دارد. در مقابل، نتایج بوماریتو و همکاران (۲۰۲۳) که عملکرد ضعیف‌تری را برای برخی مدل‌ها در آزمون CPA گزارش کرده‌اند، بار دیگر بر نقش تعیین‌کننده نوع آزمون، زبان، نسخه مدل و شرایط ارزیابی در تبیین نتایج تأکید می‌کند.

پیامدهای آموزشی و حرفه‌ای برای حسابداری

از منظر آموزشی، نتایج این پژوهش نشان می‌دهد که مدل‌های زبانی بزرگ می‌توانند به‌عنوان ابزارهای کمکی مؤثر در آموزش حسابداری مورد استفاده قرار گیرند؛ از جمله در تولید سؤالات تمرینی، شبیه‌سازی آزمون‌ها، ارائه بازخورد اولیه و پشتیبانی از یادگیری خودراهبر دانشجویان. نزدیکی عملکرد مدل‌ها حاکی از آن است که اثربخشی آموزشی این ابزارها بیش از آنکه به انتخاب یک مدل خاص وابسته باشد، به طراحی سناریوهای یادگیری، نوع پرسش‌ها و چارچوب استفاده از آن‌ها وابسته است.

برای طراحان آزمون‌های آموزشی و حرفه‌ای، یافته‌ها نشان می‌دهد که سؤالات صرفاً مفهومی و نظری بیش از پیش توسط مدل‌های زبانی قابل پاسخ‌گویی هستند. در نتیجه، حرکت به سمت طراحی سؤالات سناریومحور، ترکیبی و تحلیلی - محاسباتی می‌تواند تمایزپذیری ارزیابی‌ها را افزایش داده و ریسک تقلب مبتنی بر هوش مصنوعی را کاهش دهد.

محدودیت‌ها و مسیرهای پژوهش آتی

با وجود یافته‌های معنادار، نتایج پژوهش حاضر باید در پرتو مجموعه‌ای از محدودیت‌های روش‌شناختی تفسیر شوند. این محدودیت‌ها نه‌تنها دامنه تعمیم نتایج را مشخص می‌کنند، بلکه مسیرهای روشنی برای پژوهش‌های آتی در زمینه ارزیابی مدل‌های زبانی بزرگ در آموزش حسابداری فراهم می‌آورند:

۱) **محدودیت قالب سؤالات و پرهیز از برداشت شناختی**: داده‌های پژوهش صرفاً شامل سؤالات چندگزینه‌ای بودند که ارزیابی را به درستی یا نادرستی پاسخ نهایی محدود می‌کند. چنین قالبی امکان استنتاج درباره درک مفهومی عمیق، استدلال گام‌به‌گام یا توان حل مسئله مستقل مدل‌ها را فراهم نمی‌سازد. پژوهش‌های آتی می‌توانند با استفاده از سؤالات تشریحی، محاسباتی و مسئله‌محور، تمایز دقیق‌تری میان عملکرد خروجی‌محور و توان استدلالی ایجاد کنند.

۲) **نشت داده‌ها به‌عنوان محدودیت مرکزی**: با توجه به رسمی، پرتکرار و در دسترس بودن سؤالات آزمون دکتری حسابداری، احتمال حضور مستقیم یا غیرمستقیم این پرسش‌ها یا پاسخ‌های مشابه در داده‌های آموزشی مدل‌ها وجود دارد. ازاین‌رو، بخشی از عملکرد مشاهده‌شده ممکن است ناشی از بازیابی الگوهای ذخیره‌شده باشد، نه استنتاج مستقل. تفکیک سؤالات بر اساس احتمال نشت داده بالا/پایین و تحلیل حساسیت عملکرد نسبت به این تفکیک، مسیر مهمی برای مطالعات آتی محسوب می‌شود.

۳) **نبود آزمون‌های استحکام برای تفکیک بازیابی از استدلال**: در این پژوهش، پارافریز نظام‌مند سؤالات، تغییر مقادیر عددی، یا بازطراحی سناریوهای مفهومی انجام نشد. انجام چنین آزمون‌هایی در پژوهش‌های آینده می‌تواند به شناسایی دقیق‌تر نقش حفظ‌محوری در برابر استدلال مفهومی کمک کرده و تفسیر شناختی نتایج را معتبرتر سازد.

۴) **غلبه سناریوی open-book بر ارزیابی عملیاتی:** مدل‌ها در شرایط پیش‌فرض و عمومی در دسترس کاربران ارزیابی شدند و کنترل صریحی بر فعال یا غیرفعال بودن قابلیت‌های بازیابی اطلاعات یا جست‌وجوی وب اعمال نشد. بنابراین، نتایج پژوهش بازتاب‌دهنده عملکرد مدل‌ها در یک سناریوی عملیاتی نزدیک به open-book است و نه

توان آن‌ها در یک وضعیت کاملاً بسته (closed-book). اجرای آزمایش‌های کنترل‌شده و مقایسه مستقیم این دو سناریو، مسیر ضروری پژوهش‌های آتی است.

۵) **محدودیت دامنه دروس تخصصی**: سؤالات مورد بررسی به سه درس حسابداری مدیریت، حسابرسی و تئوری حسابداری محدود بود. تعمیم نتایج به کل حوزه حسابداری مستلزم گسترش دامنه به سایر زمینه‌ها، از جمله حسابداری مالی، حسابداری بخش عمومی و حسابداری بین‌الملل، در پژوهش‌های آینده است.

۶) **تمرکز بر شاخص دقت و فقدان تحلیل کیفی پاسخ‌ها**: ارزیابی عملکرد صرفاً بر اساس شاخص دقت کمی انجام شد و کیفیت استدلال، انسجام توضیحات و الگوهای خطای مفهومی مدل‌ها تحلیل نشد. پژوهش‌های آتی می‌توانند با افزودن تحلیل‌های کیفی، نرخ توهم[1] و کیفیت تبیین پاسخ‌ها، تصویر جامع‌تری از محدودیت‌ها و ظرفیت‌های آموزشی مدل‌ها ارائه دهند.

۷) **تغییرپذیری نسخه‌ها و پایداری نتایج:** مدل‌های زبانی بزرگ به‌طور مستمر به‌روزرسانی می‌شوند و عملکرد آن‌ها به نسخه، تاریخ اجرا و تنظیمات سامانه وابسته است. بررسی پایداری نتایج در نسخه‌های مختلف و در بازه‌های زمانی متفاوت، از مسیرهای مهم پژوهش‌های آتی محسوب می‌شود.

۸) **محدودیت تعمیم‌پذیری بافتی و زبانی**: نتایج این پژوهش به بافت آموزشی، زبانی و آزمونی ایران محدود است. تکرار مطالعه در سایر کشورها، زبان‌ها و نظام‌های آزمونی می‌تواند به افزایش اعتبار بیرونی و تعمیم‌پذیری یافته‌ها کمک کند.

ملاحظات اخلاقی

حامی مالی: مقاله حامی مالی ندارد.

مشارکت نویسندگان: تمام نویسندگان در آماده‌سازی مقاله مشارکت داشته‌اند.

تعارض منافع: بنا بر اظهار نویسندگان در این مقاله هیچ‌گونه تعارض منافعی وجود ندارد.

1. hallucination

منابع

ثقفی، علی؛ پارساپور، محمدرضا. (۱۴۰۴). بررسی تأثیر تحلیل داده‌های حسابداری با هوش مصنوعی مولد بر کیفیت گزارش دهی دیجیتال پایداری با توجه به نقش میانجی سیستم کنترل داخلی سبز پایداری. *دانش حسابداری مالی*، ۱۲(۱)، ۱-۳۱. https://doi.org/10.30479/jfak.2025.21533.3270

رحمانی، علی؛ معنوی، سمیرا؛ حدادی، نفیسه. (۱۴۰۴). ادغام هوش مصنوعی در حسابرسی؛ چالش‌ها و مزایا. *حسابرسی سیستم‌ها و فناوری اطلاعات*، ۱(۱)، ۱-۲۷. https://doi.org/10.22034/jista.2025.528769.1051

رهنما، مریم؛ رفعتی، حمیدرضا. (۱۴۰۴). پیامدهای اخلاقی پذرش هوش مصنوعی در تصمیم‌گیری‌های مالی. *حسابرسی سیستم‌ها و فناوری اطلاعات*، ۱(۱)، ۲۸۴-۳۰۱. https://doi.org/10.22034/jista.2025.509536.1032

عدنان حمود، محمد؛ پیری، پرویز؛ آشتاب، علی. (۱۴۰۴). امکان‌سنجی بهره‌گیری از فناوری‌های نوین هوش مصنوعی در بهبود فرایندهای حسابرسی در کشور. *بررسی‌های حسابداری و حسابرسی*، ۳۲(۳)، ۵۳۵-۵۵۹. https://doi.org/10.22059/acctgrev.2025.391837.1009085

مشایخی، بیتا؛ امراللهی، محمدرضا. (۱۴۰۴). تأثیر دانش و تردید حرفه‌ای حسابرسان داخلی بر به کارگیری هوش مصنوعی. *پژوهش‌های تجربی حسابداری*، ۱۵(۲)، ۱-۲۸. https://doi.org/10.22051/jera.2025.50268.3523

نوراحمدی، مرضیه؛ پارسی، فاطمه. (۱۴۰۴). نقش هوش مصنوعی در ارتقای حسابداری سبز و توسعه پایدار: رویکرد نگاشت دانش. *پژوهش‌های تجربی حسابداری*، ۱۵(۲)، ۲۱۱-۲۳۸. https://doi.org/10.22051/jera.2025.50235.3512

References

Adnan Hammood, M., Piri, P., & Ashtab, A. (2025). Feasibility of utilizing advanced artificial intelligence technologies to improve auditing processes in the country. *Accounting and Auditing Review*, 32(3), 535-559. (in Persian) https://doi.org/10.22059/acctgrev.2025.391837.1009085

Agarwal, P., & Gaur, F. (2020). A historical perspective of artificial intelligence in accounting: Evolution, current developments, and future opportunities. *Journal of Accounting and Organizational Change*, 16(1), 1–12. https://doi.org/10.1108/JAOC-04-2017-0035

AI Index Steering Committee. (2025). *The AI Index 2025 annual report*. Institute for Human-Centered AI, Stanford University. https://doi.org/10.48550/arXiv.2504.07139

Alibaba Group. (2024, September 19). *Alibaba Cloud unveils Qwen2.5, full-stack AI infrastructure enhancements at 2024 Apsara Conference*. Alibaba Group. https://www.alibabagroup.com/en-US/document-1773855135127044096

Albuquerque, F., & Gomes dos Santos, P. (2024). Can ChatGPT Be a Certified Accountant? Assessing the Responses of ChatGPT for the Professional Access

Exam in Portugal. *Administrative Sciences*, 14(7), 152. https://doi.org/10.3390/admsci14070152

Amoah, N., Fianko, S. K., Dake, S., Agyemang, K., Nyame, I., Adjaye-Gyamfi, O., ... & Lartey, R. (2024). The Impact of Ai Chatbots on the Landscape of Professional Accountancy Examination: An Experimental Study. *Available at SSRN 4991304*. http://dx.doi.org/10.2139/ssrn.4991304

Bordt, S., & von Luxburg, U. (2023). Chatgpt participates in a computer science exam. *arXiv preprint arXiv:2303.09461*. https://doi.org/10.48550/arXiv.2303.09461

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

Chippagiri, S. (2025, March 4). *DeepSeek: Revolutionizing AI with Open-Source Large Language Models. DEV Community*. https://dev.to/srinivas_chippagiri_e01c8/deepseek-revolutionizing-ai-with-open-source-large-language-models-127i

Dell, S., & Akpan, M. (2024). You are the auditor: A ChatGPT-based multiple choice exam. *Advances in Online Education: A Peer-Reviewed Journal*, 3(2), 111–120. https://doi.org/10.69554/EINF1743

de Freitas, M. M., Sallaberry, J. D., & de Jesus Silva, T. B. (2024). Application of Chat GPT 4.0 for solving accounting problems. *GCG: revista de globalización, competitividad y gobernabilidad*, 18(2), 49-64. https://dialnet.unirioja.es/servlet/articulo?codigo=9498637

de Winter, J. C. (2024). Can ChatGPT pass high school exams on English language comprehension?. *International Journal of Artificial Intelligence in Education*, 34(3), 915-930. https://doi.org/10.1007/s40593-023-00372-z

Eulerich, M., Sanatizadeh, A., Vakilzadeh, H., & Wood, D. A. (2023). Can artificial intelligence pass accounting certification exams? ChatGPT: CPA, CMA, CIA, and EA. ChatGPT: CPA, CMA, CIA, and EA. *Available at SSRN*. http://www.ais.nptu.edu.tw/bsacc/1121%20materials/SSRN-id4452175_ChatGPT%E8%80%83%E6%9C%83%E8%A8%88%E8%AD%89%E7%85%A7.pdf

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. https://doi.org/10.1007/s11142-024-09833-9

Foote, K. D. (2023, December 28). *A brief history of large language models*. DATAVERSITY. https://www.dataversity.net/a-brief-history-of-large-language-models/

Glover, E. (2025, July 16). *Grok: What we know about Elon Musk's AI chatbot. Built In*. https://builtin.com/articles/grok

Greenman, C., Esplin, D., Johnston, R., & Richards, J. (2024). An Analysis of the Impact of Artificial Intelligence on the Accounting Profession. *Journal of Accounting, Ethics & Public Policy, JAEPP*, 25(2), 188-188. https://doi.org/10.60154/jaepp.2024.v25n2p188

Guinness, H. (2024, April 3). *What is Perplexity AI? How to use it + how it works. Zapier Blog*. https://zapier.com/blog/perplexity-ai

Hashemi-Pour, C., Kerner, S. M., & Patrizio, A. (2025, January 8). What is the Google Gemini AI model (formerly Bard)? TechTarget. https://www.techtarget.com/searchenterpriseai/definition/Google-Gemini

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2024). Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270), 20230254. https://doi.org/10.1098/rsta.2023.0254

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., ... & Tseng, V. (2023). Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS digital health*, 2(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198

Martínez, E. (2024). Re-evaluating GPT-4's bar exam performance. *Artificial intelligence and law*, 1-24. https://doi.org/10.1007/s10506-024-09307-6

Mashayekhi, B., & Amrollahi, M. R. (2025). The effect of internal auditors' knowledge and professional skepticism on the artificial intelligence utilization. *Journal of Empirical Research in Accounting*, 15(2), 1-28. (in Persian) https://doi.org/10.22051/jera.2025.50268.3523

Mendonça, N. C. (2024). Evaluating chatgpt-4 vision on brazil's national undergraduate computer science exam. *ACM Transactions on Computing Education*, 24(3), 1-56. https://dl.acm.org/doi/abs/10.1145/3674149

Mikalef, P., & Gupta, M. (2021). Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Information & Management*, 58(3), 103434. https://doi.org/10.1016/j.im.2020.103434

National Aeronautics and Space Administration. (2024). What is artificial intelligence? NASA. https://www.nasa.gov/what-is-artificial-intelligence/

Nourahmadi, M., & Parsi, F. (2025). The role of artificial intelligence in enhancing green accounting and sustainable development: a bibliometrix methode. *Journal of Empirical Research in Accounting*, 15(2), 211-238. (in Persian) https://doi.org/10.22051/jera.2025.50235.3512

Pierotti, M., Monreale, A., & De Santis, F. (2024). *Artificial Intelligence in Accounting and Auditing: Accessing the Corporate Implications*. Palgrave Macmillan, Switzerland. ISBN. https://doi.org/10.1007/978-3-031-31299-1

Rahmaini, A., Maanavi, S., & Haddadi, N. (2025). Integration of Artificial Intelligence in Auditing: Challenges and Benefits. *Journal of Information System and Technology Audit (JISTA)*, 1(1). 1-27. (in Persian) https://doi.org/10.22034/jista.2025.528769.1051

Rahnama, M., & Rafati, H. (2025). The Ethical Implications of Adopting Artificial Intelligence (AI) in Financial Decision-Making. *Journal of Information System and Technology Audit (JISTA)*, 1(1). 284-301. (in Persian) https://doi.org/10.22034/jista.2025.509536.1032

Saghafi, A., & Parsapoor, M. (2025). Examining impact of accounting data analysis with generative ai on the quality of digital sustainability reporting with the mediating role of green sustainability internal control system. *Financial*

Accounting Knowledge, 12(1), 1-31. (in Persian) https://doi.org/10.30479/jfak.2025.21533.3270

SecureNinja. (2025, March 18). Comparison of Top AI Models: DeepSeek AI, ChatGPT, Gemini, and Perplexity AI. SecureNinja Blog. https://secureninja.com/news/comparison-of-top-ai-models-deepseek-ai-chatgpt-gemini-and.html

Sharida, A., & Hashlamon, I. (2021). Real-time vision-based controller for delta robots. International *Journal of Intelligent Systems Technologies and Applications*, 20 (4), 271–295. https://doi.org/10.1504/IJISTA.2021.10045532

Sharida, A., Hamdan, A., & Al-Hashimi, M. (2020). Smart cities: The next urban evolution in delivering a better quality of life. Toward Social Internet of Things (SIoT): Enabling Technologies, Architectures and Applications: Emerging Technologies for Connected and Smart Social Objects, 287–298. https://doi.org/10.1007/978-3-030-24513-9_16

Stengel, F. C., Stienen, M. N., Ivanov, M., Gandía-González, M. L., Raffa, G., Ganau, M., ... & Motov, S. (2024). Can AI pass the written European Board Examination in Neurological Surgery?-Ethical and practical issues. *Brain and Spine*, 4, 102765. https://doi.org/10.1016/j.bas.2024.102765

SY Partners. (2025, February 10). *The history of GPT: A journey through generative pre-trained transformers*. https://syp.vn/en/article/the-history-of-GPT

TechCrunch. (2025, May 20). *DeepThink boosts the performance of Google's flagship Google Gemini AI model*. https://techcrunch.com/2025/05/20/deep-think-boosts-the-performance-of-googles-flagship-google-gemini-ai-model

Vaˇrzaru, A. A. (2022). Assessing artificial intelligence technology acceptance in managerial accounting. *Electronics*, *11*, 1–13. https://doi.org/10.3390/electronics11142256

Vasarhelyi, M. A., Moffitt, K. C., Stewart, T., & Sunderland, D. (2023). Large language models: An emerging technology in accounting. *Journal of Emerging Technologies in Accounting*, 20(2), 1–10. https://doi.org/10.2308/JETA-2023-047. https://doi.org/10.2308/JETA-2023-047

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wood, D. A., Achhpilia, M. P., Adams, M. T., Aghazadeh, S., Akinyele, K., Akpan, M., ... & Kuruppu, C. (2023). The ChatGPT artificial intelligence chatbot: How well does it answer accounting assessment questions?. *Issues in Accounting Education*, 38(4), 81-108. https://doi.org/10.2308/ISSUES-2023-013

World Economic Forum. (2020). Future of Jobs Report 2020. https://www.weforum.org/publications/the-future-of-jobs-report-2020/

Wutzler, J. (2024). Outsmarting Artificial Intelligence in the Classroom—Incorporating Large Language Model-Based Chatbots into Teaching. *Issues in*

Accounting Education, 39(4), 183-206. https://doi.org/10.5555/ISSUES-2023-064tn

Zacher, W., & Kuppannagari, S. (2024). Can LLMs Pass the CPA Exam? Evaluating Large Language Model Performance on the Certified Public Accountant Test. *Available at SSRN*. *https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4788096*

Zhangyang, Q., Fang, Y., Zhang, M., Sun, Z., Wu, T., Liu, Z., Lin, D., Wang, J., & Zhao, H. (2023, December 22). Gemini vs GPT-4V: A preliminary comparison and combination of vision-language models through qualitative cases. arXiv. https://doi.org/10.48550/arXiv.2312.15011