Mongolian Journal of Economic Review-MJER **Volume 29/18 (2025)**
ARTICLE
DOI – https://doi.org/10.69588/mjer.v29i18.4262

Artificial Intelligence Models and Test Performance

(*Case Study in Accounting Program*)

Gantsetseg Sanjmyatav*[1], Namuun Tsendbayar[1], Enkhsuren Munkhjargal[2]
Enkh-Amgalan Erdenesuren[2], Zolboo Gantulga[2]

*[,1]*Department of Foreign Languages, Mandakh University, Ulaanbaatar, Mongolia*
[2]*Accounting and Analysis Department, Mandakh University, Ulaanbaatar, Mongolia*

**Corresponding author: gantsetseg@mandakh.edu.mn*

Received: 28 February 2025 / Accepted: 21 March 2025 / published online: 01 May 2025

Keywords: *Accounting program, final examination test, students' performance, Chat Generative Pre-trained Transformer, Copilot, Deepseek*

Abstract: *Integrating artificial intelligence (AI) in education has required considering the conservative approaches of teaching and assessment systems, as AI impacts how educators operate, how students learn, and how learning outcomes are evaluated. The primary goal of this research paper is to compare AI tools' performance with students' performance through experiments on the final examination tests in the accounting program. In the conducted experiment, ChatGBT achieved a performance rate of 77 percent in a series of tests that encompassed the final examination, which included 11 courses and a total of 100 questions. The performance of the top-scoring students exceeded the results of the Copilot and Deepseek models; however, it did not surpass that of ChatGBT. The study findings have shown that ChatGPT performed less effectively on multiple-choice questions that require extensive calculations and accounting estimations; however, it was good at simpler multiple-choice and matching tasks. Thus, it was concluded that permitting students to undertake online examinations in unsupervised settings or utilize AI tools without limitations can enhance student performance in unreal ways while simultaneously minimizing reliability in assessment outcomes.*

I. INTRODUCTION

The COVID-19 pandemic has demanded an inevitable shift to a global remote learning system. The situation has provided continuity of learning through the use of different high-tech resources and tools, which have led to advances in the development of e-learning. In the new learning environment, students have more autonomy, and the learners' ability to take advantage of technological advances is good enough in terms of using artificial intelligence technologies. Artificial intelligence (AI) has made a huge difference in the education system and almost every industry. This technology is beginning to create many new opportunities for increasing productivity and contributing to the development of different industries such as manufacturing, services, healthcare, arts, and education. The introduction of artificial intelligence, which is rapidly evolving in the field of education, has been a force to redefine and transform traditional teaching methods and assessment strategies. The increasing use of AI technology, such as ChatGPT, has had a major role in automating many tasks in the knowledge management, education, and research industries in terms of processing data efficiently, categorizing unstructured data, and generating accurate responses without human intervention. At the same time, there was a concern that teachers

DOI – https://doi.org/10.69588/mjer.v29i18.4262

had less direct control and direction over students' learning. It attracted attention that the students' digital platform-based assignments performed relatively better than classroom-based exams.

For this reason, this experiment was conducted to investigate learning outcomes, focusing on the final examination results, comparing different AI performances. This research is an experimental study conducted using final examination tests from the accounting program, with the aim of comparing student and AI tools' performances and the performance of commonly used AI tools. The experiment aimed to determine whether there were differences in exam performance in the following:

1. The performance differences among various AI tools
2. The differences in AI tools' capacity to complete simple versus analytical (calculation-based) test questions
3. The differences in performance between the highest-performing and average students compared to the AI tools

Literature review. Many studies have been conducted by researchers and educators focusing on the use of technological advances in online learning and exams, as well as the advantages and challenges of accepting AI in the classroom in accounting majors. A Researcher, Abeysekera, has investigated how ChatGPT and ChatGPT4 performed the tasks in an accounting major under the topic "ChatGPT and academia on accounting assessments." (Abeysekera, 2023). The study showed that ChatGPT performed better on the assignments in accounting basics than on tasks of the advanced accounting course. In the comparison of ChatGPT and ChatGPT4 models, ChatGPT4's performance was higher (Abeysekera, 2023).

Clare Baek et al. (2023), who examined how U.S. students understand and access ChatGPT, have highlighted the dual role of technology in educational contexts, illustrating how it can both empower and assist in their learning. They suggested there is a need to advocate for the equitable integration of artificial intelligence in academic settings to benefit a diverse student population in terms of the learners' age, gender, and occupation (Clare Baek, 2023). A study titled "ChatGPT: The End of Online Exam Integrity?" by Teo Susnjak (2022) has noted that educators and educational institutions need to recognize the potential for ChatGPT to be utilized for dishonest practices. They should explore strategies to mitigate this issue to ensure the integrity and fairness of online examinations for learners (Susnjak, 2022).

According to the study of Wood et al., ChatGPT performs better than the student average by 15.8 percent of assessments, by answering 56.5 percent of the questions. The findings highlighted that ChatGPT performs various questions, including open and closed assessments, and test bank questions on different accounting subjects at different course levels (Wood, 2023). The latest versions of ChatGPT have been investigated by Marc Eulerich, Aida Sanatizadeh, Hamid Vakilzadeh & David A. Wood (2024) and provide valuable guidance for accounting professionals, investors, and stakeholders on strategies to adapt to and mitigate the potential risks associated with AI technology in accounting and auditing firms. Their research findings determined that ChatGPT 3.5 was ineligible for exams, however, the ChatGPT 4 model could increase the performance score (Marc Eulerich, 2024). The effectiveness of ChatGPT on accounting and audit performance was tested at the University of North Macedonia. Using 11 course quizzes with a total of 401 questions, this study evaluated ChatGPT's capabilities to solve complex and contextual questions within these domains, with ChatGPT 3.5 successfully passing 8 out of 11 subjects, with a 73% rate (Atanasko Atanasovski, 2023). ChatGPT scored a C+ on 95 multiple-choice questions and 12 essay questions prepared by professors at the University of Minnesota, The School of Law (Jonathan H. Choi, 2022). Similarly, it received a B and a B - rating on the Business Administration course test at the Wharton School of the University of Pennsylvania (Terwiesch, 2023). According to a survey conducted by AICHEM, a research organization in the US, 70 percent of students said they use ChatGPT in their assignments and classes (Stokel-Walker, 2024). Some educators have been researching the effects of ChatGPT in the education field. Researchers have warned that increasing AI systems such as ChatGPT to help students complete assignments and take exams is causing specific problems in the learners' outcomes. Some electronic exams, for instance, are not directly supervised, so students can download from different sources to improve their exam scores. In such cases, artificial intelligence is a new challenge in digital exam evaluation. The study by Jonathan H. Choi has investigated whether ChatGPT is eligible for the Law, Medicine, and Business University exams. (Jonathan H. Choi, 2022). David Wood (2023) has noted that by using ChatGPT

ARTICLE
DOI – https://doi.org/10.69588/mjer.v29i18.4262

to take exams in an unsupervised environment, students increased their exam scores and eliminated the gap in their performance scores. Atanasko Atanasovski and others have summarized that the appropriate use of AI technology is a matter of carefully considering its relevance and adaptability to improving the educational process and creating value (Atanasko Atanasovski, 2023). The previous research reviews have studied how ChatGPT and similar artificial intelligence tools can be used effectively and ethically in education.

II. RESEARCH METHODOLOGY

Research questions

AI can provide accurate answers to complex questions that require advanced data analysis, integration, and use, and it can also process essential questions used to assess students' professional skills. Therefore, we made the first assumption that the development of AI tools could pass the online exam with 100% performance.

Question 1: Is the performance percentage of AI tools on the exam higher than the performance of students who completed the exam using traditional methods?

Based on previous studies, we summarized that ChatGPT has struggled to solve numerical tasks that require a few-step estimation; however, ChatGPT has performed well at answering simple multiple-choice and true/false questions. ChatGPT has performed true and false questions better than tasks that require more calculations. It also performed assignments in information systems and audits relatively well. This is because the information systems and audit questions were less mathematical. However, the result was relatively poor in tax, finance, and management accounting.

Question 2: Can the capability of AI tools work differently to perform simple multiple-choice tasks and computational tasks?

Research design

Compulsory subjects from the Accounting program, which include content for the final examination, were allocated into 6 groups, as shown in Figure 1.

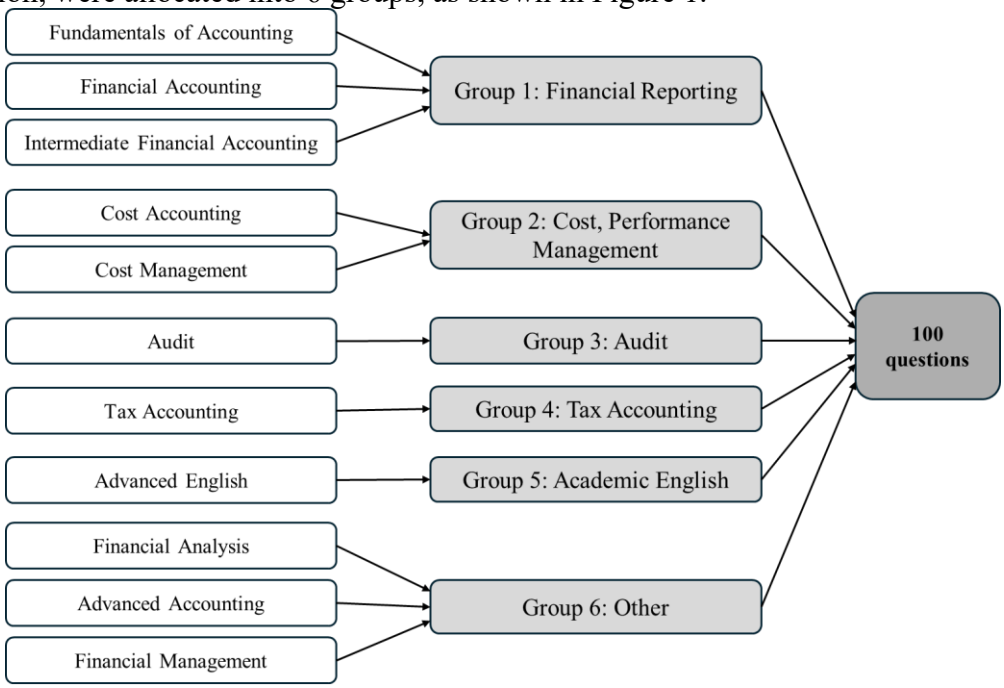

Figure 1. Grouping subject contents

ARTICLE
DOI – https://doi.org/10.69588/mjer.v29i18.4262

Source: Researcher's estimation

Research sampling and tools

The final examination test from the Accounting and English undergraduate program at Mandakh University, covering the 2022-2023 academic year, was used in the study. The graduate examination test comprises 100 questions with different types of assignments based on the content of 11 compulsory courses. Three artificial intelligence tools, ChatGPT, Copilot, and Deepseek, were experimented with in the study as research tools to achieve the research outcome.

III. RESEARCH RESULTS

Performance was evaluated by testing 100 scores in terms of the above 6 groups of courses using ChatGPT 3.5, Copilot, and Deepseek tools, which are available for free use in the study. Graduation exam questions comprise a simple question, a True/False statement, a basic or multiple-choice test, a problem-solving question, and matching tasks. When evaluating test results, a point is given for the correct answer and zero points for an incorrect answer.

The structure of the 100 questions used in the survey was expressed by question types in Figure 2, as demonstrated that 58% of the total questions were simple tests, 21% were problem-solving tests, 15% were True / False statements, and 6% were matching tests. The simple tests that have a lower technical complexity occupied 58% of the total questions. This has shown that the examination level was not high enough. The matching questions were included in the minority.

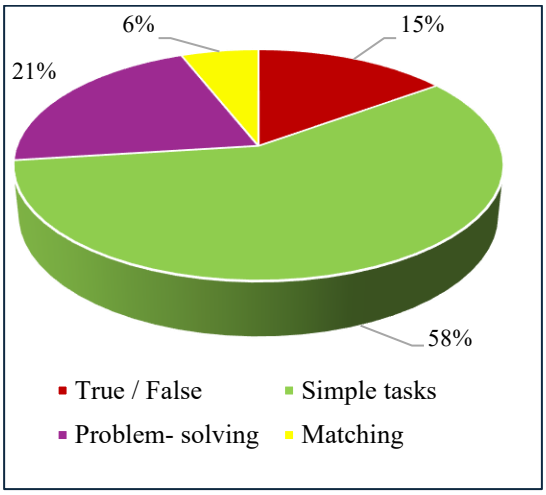

Figure 2. Percentage of questions

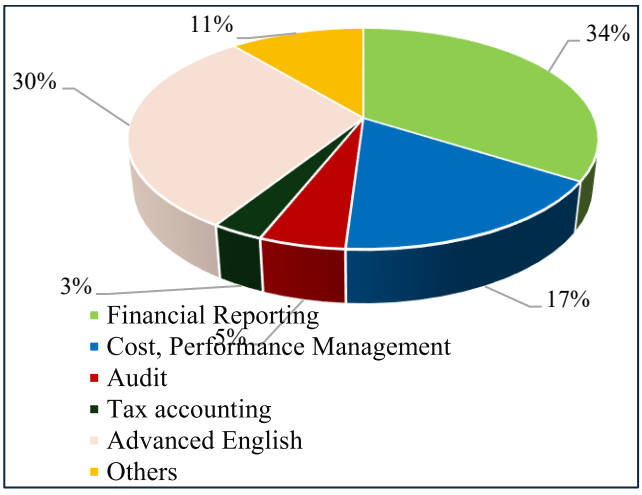

Figure 1 Percentage of subject groups

As illustrated in Figure 3, the highest percentages of the course groups were from the financial reporting group and English, at 34% and 30%, respectively. This has shown that financial reporting and English language knowledge were assessed relatively higher than other subjects.

A test performance of the artificial intelligence tools, such as ChatGPT, Copilot, and Deepseek, used in the survey is illustrated in Table 1 by question type compared to students' performance.

Table 1. Assessment of graduation tests of accounting majors/by question types/

Types of questions	Task range	ChatGPT	Copilot	Deepseek	Students' assessment	
					Max	Average
True / False statement question	15	66.67%	46.67%	60.00%	73.33%	59.32%
Simple multiple-choice question	58	82.76%	82.76%	74.14%	72.41%	46.93%
Problem-solving question	21	61.90%	52.38%	61.90%	80.95%	46.65%

ARTICLE
DOI – https://doi.org/10.69588/mjer.v29i18.4262

Matching question	6	100.00%	100.00%	100.00%	50.00%	41.13%
Total performance	**100**	**77%**	**72%**	**71%**	**73%**	**50.00%**

For the questions assigned in the exam, the artificial intelligence tools performed simple multiple-choice and matching tasks, over 74.14% and 100%, respectively. On the other hand, the capacity to answer true/false and multiple-choice tests has resulted in poor as ranged between 47% and 67%. Compared to average student scores, the performance of artificial intelligence tools is significantly higher across all types of questions, so students with average and below-average scores can improve their test scores using AI tools. The AI tools could work better on simple tasks, such as matching, simple multiple-choice questions, while students could perform complex tasks better than the AI. In terms of overall test results, ChatGPT performed the best, answering 77% of the total questions correctly, while the students' performance was shown in 50% in total. This result showed that students may use AI technologies in some ways when they take exams.

Table 1 General assessment of the graduation test of the accounting program / by the subject groups/

Subject groups	Questions	ChatGPT Correct answer percentage	Copilot Correct answer percentage	Deepseek Correct answer percentage	The highest percentage of individual srudent
Financial Reporting	34	79.41%	67.65%	64.71%	91.18%
Cost, Performance management	17	58.82%	52.94%	52.94%	76.47%
Audit	5	60.00%	80.00%	100.00%	80.00%
Tax accounting	3	100.00%	100.00%	33.33%	33.33%
Advanced English	30	86.67%	86.67%	83.33%	50.00%
Others	11	72.73%	63.64%	81.82%	81.82%
Total performance	**100**	**77%**	**72%**	**71%**	**73%**

Based on the results of the experiment in Table 2, ChatGPT and the Copilot had 100% performance in the tax accounting group. For the highest-performing classes of the AI tools, most of the English tests consisted of a matching task, and all assignments for the tax accounting test consisted of simple multiple-choice tasks. On the other hand, subjects that need high performance, such as financial reporting and cost management courses, were composed of true or false statements, simple multiple-choice tasks, and problem-solving tasks. The result showed that there were some capacity differences of AI tools in performing lower and higher complexity tasks, such as simple multiple-choice and problem-solving questions.

IV. CONCLUSION

The study concludes that the final exam performance of the accounting program was evaluated using three types of artificial intelligence tools that were used free of charge.

Initially, it was summarized that ChatGPT had the best performance, while Deepseek was the worst. At the same time, the AI models performed significantly higher than the average student score. It has been observed that AI tools, particularly ChatGPT, have a significant impact on enhancing exam performance and mitigating the volatility of assessments.

Then, artificial intelligence tools have been observed to respond more effectively to simple multiple-choice tasks and matching tasks. However, the outcomes of artificial intelligence were relatively low compared to problem-solving tasks that required some complex accounting calculations.

As a result of this study, the AI tools have performed the given tasks between 58% and 100%. This consequence can be changed as AI models are progressing to an advanced level.

AI models, that are paid, tend to have more advanced algorithms and access to broader datasets, allowing them to deliver higher performance compared to free models. Therefore, it is necessary to continue evaluating the performance differences between paid and free AI tools, as well as examining the impact of ongoing updates.

DOI – https://doi.org/10.69588/mjer.v29i18.4262

In terms of the research limitation, it is suggested that it may be beneficial to use the paid version of AI tools for the better performance of AI platforms in further studies. Additionally, it would be evident if any studies were conducted using various types of tasks in different contextual disciplines.

V. REFERENCES

Abeysekera, I. (2023). ChatGPT and academia on accounting assessments. Technology, Market, and Complexity, www.sciencedirect.com/journal/journal-of-open-innovation-technology- market-and-complexity. https://doi.org/10.1016/j.joitmc.2024.100213

Atanasko Atanasovski, T. T. (2023). EVALUATING THE PERFORMANCE OF CHATGPT IN ACCOUNTING AND. Economic and Business Trends Shaping the Future. Research Gate. https://doi.org/10.47063/EBTSF.2023.0003

Atanasovski, A. (2023). Evaluating the Performance of ChatGPT in Accounting and Auditing Exams: An Experimental Study in North Macedonia. Proceedings of the 4th International Conference Economic and Business Trends Shaping the Future. https://doi.org/10.47063/EBTSF.2023.0003

ChatGPT and academia on accounting assessments. (огноо байхгүй).

Clare Baek, T. T. (2023). ChatGPT seems too good to be true": College students' use and perceptions of generative AI. Artificial Intelligence , www.sciencedirect.com/journal/computers-and-education-artificial-intelligence. https://doi.org/10.31219/osf.io/6tjpk

Jonathan H. Choi, 1. K. (2022). CHATGPT GOES TO LAW SCHOOL . Journal of Legal Education, 387-400.

Marc Eulerich, A. S. (2024). Is it all hype? ChatGPT's performance and disruptive. Review of Accounting Studies. https://doi.org/10.1007/s11142-024-09833-9

Sarangoo.I. (2024 оны 10 27). https://eagle.mn/: https://eagle.mn/r/131945-ээс Гаргасан

Stokel-Walker, C. (2024 оны 6 24). newscientist.com: https://www.newscientist.com/article/2436888-university-examiners-fail-to-spot-chatgpt-answers-in-real-world-test/?utm_source=chatgpt.com-ээс Гаргасан

Stokel-Walker, C. (2024 оны 12 13). (https://www.newscientist.com/. Over 70 per cent of students in US survey use AI for school work. https://doi.org/10.1016/S0262-4079(24)01217-X

Susnjak, T. (2022). ChatGPT: The End of Online Exam Integrity? https://doi.org/10.48550/arXiv.2212.09292 . ChatGPT: The End of Online Exam Integrity?-ээс Гаргасан

Terwiesch, C. (2023). Would Chat GPT3 Get a Wharton MBA? A Prediction Based on Its Performance in the Operations Management Course . Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania.

Wood, D. A. (2023). The ChatGPT artificial intelligence chatbot: how well does it answer accounting assessment questions? Issues Account. Educ. 38 (3), 1-28. .., Educ. 38 (3), 1-28. https://doi.org/10.2308/ISSUES-2023-013