Original Research Article

Measuring AI Accuracy on Standardized Tests: A Comparative Study of ChatGPT, Copilot and Gemini

Raudeen Roodgarmi

John Paul II High School, 900 Coit Rd, Plano, TX 75075, United States

ABSTRACT

This study evaluates the performance of three widely used artificial intelligence systems, ChatGPT, Microsoft Copilot, and Google Gemini, on standardized test questions in Math, Reading, and English derived from SAT and ACT examinations. A total of 90 questions (30 per subject) were selected from multiple test forms across different years to reduce potential bias and ensure broad content coverage. All questions, including those with visual components, were presented to each AI system in a standardized format, and responses were scored for accuracy. A chi-square test for homogeneity was conducted to assess differences in performance among the models. Results indicate that all three AI systems performed strongly in language-based tasks, Reading and English. In contrast, performance in Math was notably lower across all models, with common errors involving advanced mathematical concepts and misinterpretation of visual and graphical information. Despite observable differences in error patterns, statistical analysis revealed no significant differences in overall performance among the three systems. These findings suggest that current AI models are highly proficient in processing and interpreting textual information but remain less reliable in mathematical reasoning and multimodal tasks. The study highlights both the capabilities and limitations of AI in standardized testing contexts and underscores the importance of prompt design and continued model development.

Keywords: Artificial intelligence; ChatGPT; Microsoft Copilot; Google Gemini; SAT; ACT

INTRODUCTION

Artificial intelligence has become a central focus in the technological world (1). Computers have traditionally been viewed as non-sentient tools designed to follow programmed instructions rather than think or make decisions independently (2). Computers are present in nearly every aspect of modern life, serving vital roles in everything from medicine to finance to education to business and beyond.

Computers have pushed humans toward the idea of an efficient, progressive world, enabling programs to accomplish tasks that were once considered impossible. But one characteristic that all computers have had since their creation has been that they can only use what they are given. A program can only achieve the level of quality defined by the standards established during its creation (3).

Artificial Intelligence refers to computer systems that can learn, adapt, and perform tasks that typically require human intelligence, making them a major advancement beyond traditional programmed machines (4). The concept of Artificial Intelligence has transformed the capabilities of computers, introducing possibilities

Corresponding author: Raudeen Roodgarmi, E-mail: raudeenroodgarmi1@gmail.com.

Accepted May 5, 2026
https://doi.org/10.70251/HYJR2348.434451

so expansive that they raise concerns among skeptics regarding its potential (5). A historical parallel to AI's development might be the Gutenberg printing press. The Gutenberg press was criticized in its time for taking jobs from scribes. However, the Gutenberg press also increased literacy rates in Europe and revolutionized the academic landscape by making textbooks more accessible to students (6). This raises the question: why is AI regarded as so consequential, and in what ways might its potential impact mirror that of the Gutenberg press?

AI has been labeled the biggest invention since the Internet and compared to the atomic bomb in terms of its potential societal impact. However, AI is still relatively new and is just taking its first steps (7). Like humans, AI can make mistakes, and it is crucial to understand AI's abilities and limitations. Without understanding these boundaries, the integration of AI into certain fields or subjects could have harmful consequences (8).

One area in which AI has rapidly advanced is its use as a study tool for students. From elementary school through graduate studies, AI software serves both as a tutor and study companion for students in need. Today, AI platforms are being increasingly used by students to enhance their skills in core academic areas, providing support that extends beyond the classroom (9). In American schools, the three traditional core subjects that are taught to all children in schools are Reading, Writing, and Math. College entrance standardized tests, such as the SAT and the ACT, both give questions in these three core subjects, with the SAT having a mix of reading and English questions within one section and the ACT having separate sections for each (8). In education, AI could greatly enhance test preparation by offering intelligent, adaptive support that strengthens students' understanding and performance on these standardized exams (10).

This study evaluated the performance of various AI software, ChatGPT, Microsoft Copilot, and Google Gemini, on real standardized test questions from the ACT and SAT. Each system was presented with a set of 30 questions representative of those encountered by students across the United States. Their strengths and weaknesses were then analyzed across three core subject areas, Math, Reading, and Writing, with a focus on accuracy and patterns of error. Accordingly, the study addresses several key questions regarding how accurately ChatGPT, Microsoft Copilot, and Google Gemini can answer Math, English, and Reading questions from the SAT and ACT, how their accuracy compares across these subject areas, and what common or unique errors these systems produce when answering incorrectly.

LITERATURE REVIEW

AI is changing education quickly. It's reshaping classrooms, tutoring, and even how people study for tests. Many students now rely on AI for personalized feedback, detailed study guides, and assistance with subjects that previously required a teacher (9). These programs are used for test prep, class reviews, and career guidance. Many education experts agree that AI can make studying easier by giving quick answers, creating practice questions, and showing students how to solve them. It can also change topic difficulty level based on how well a student is doing, making learning more personalized than ever before (11). However, researchers caution that AI is not flawless. It can make reasoning errors, misinterpret complex topics, and struggle to explain the rationale behind its answers (12). For this reason, most believe that AI should serve as a tool to support learning rather than replace genuine understanding (13) As AI's role in education expands, it is crucial to understand how effectively it performs in real classroom settings. This study examines AI's effectiveness in key school subjects, analyzing its accuracy and error patterns to identify how it can best support students and where improvements are needed.

ChatGPT

ChatGPT has gained popularity in schools as an AI tool, offering quick and clear responses to students' questions (14). Students use ChatGPT to enhance their reading, grammar, and writing skills. It excels at generating questions that test content recall, correcting grammar, expanding vocabulary, condensing lengthy texts, summarizing key ideas, and simplifying complex language (15). Research has shown that ChatGPT tends to face challenges with problems that require multiple steps of reasoning or higher levels of critical thinking (16). In more complex subjects such as math and science, studies report that it can skip intermediate steps or produce answers that sound correct but lack full logical support (17). Other researchers have observed that ChatGPT often presents its responses with high confidence even when they are inaccurate, which can mislead users who rely on its output without verification (12). Overall, the literature suggests that while ChatGPT is accessible and efficient, its main limitation lies in providing reasoning and explanation for complex or abstract questions (18).

Google Gemini

Gemini is an AI model highly capable in language comprehension and processing large volumes of structured text (19). Studies suggest that it performs well on reading and fact-based questions, particularly when the questions are straightforward. (20). Researchers have found that Gemini can retain details from lengthy texts, enabling it to effectively organize and summarize information (21). Teachers have observed that it provides clear, step-by-step answers when questions are well-structured (12). However, Gemini can struggle with tone, hidden meanings, and author subtext, which are key for reading and writing (20). Some studies indicate that its answers can vary depending on how a question is phrased, suggesting it may rely more on specific wording than on understanding the underlying meaning (19).

Microsoft Copilot

Copilot is an AI system designed to assist with productivity, problem-solving, and structured tasks across various fields, including education (22). Within classroom settings, Copilot has been tested on technical questions, rule-based reasoning, and assignments that require step-by-step logic (23). Studies show that it performs well in areas such as Math, grammar, and data analysis because it follows consistent rules and recognizable patterns (19). Research also notes that Copilot maintains steady accuracy when handling familiar question types and performs reliably when given clear instructions (22). However, studies indicate that one of Copilot's main limitations is its brief explanations, as it tends to provide quick answers without elaborating on its reasoning process (23). This makes it efficient but less effective for helping students understand the steps or logic behind an answer (24). Additional findings suggest that Copilot can struggle with open-ended or creative problems that require flexible or original thinking (23). Despite these challenges, the research concludes that Copilot remains a reliable tool for structured academic tasks and technical work, making it useful for organized study and classroom applications (19).

Comparing AI Software

Artificial intelligence has become a tool that meaningfully impacts education. However, the researchers are still in agreement that no system can be entirely accurate or self-sufficient (19). ChatGPT, Gemini, and Copilot differ in their functionality, with each excelling at different tasks. ChatGPT is particularly strong in writing and reading, Gemini performs well at organizing and interpreting structured information, and Copilot is most effective for rule-based or technical questions (19). Studies have often found that these systems can answer simple and factual questions satisfactorily; however, they are not capable of performing creative, reasoning, or abstract analysis tasks (23). Several experts have highlighted that AI tools can enhance student learning by providing personalized support, generating practice problems, and adjusting difficulty levels to match individual abilities (22). Conversely, the studies also note that these systems can make logical errors or offer incomplete explanations, meaning human guidance remains essential to ensure their effectiveness in the classroom (13). Overall, AI is transforming the ways students learn; however, it is most effective when used in conjunction with teachers and human judgment rather than as a replacement (24).

METHODS AND MATERIALS

This study aimed to evaluate the performance of Microsoft Copilot, ChatGPT, and Google Gemini on standardized test questions in Math, Reading, and Writing. To accomplish this, a set of 30 questions was compiled for each subject area, drawn from a combination of SAT and ACT examinations. These nationally standardized assessments were selected because they are administered uniformly across the United States and are not subject to regional variation.

To minimize the likelihood that the AI systems could identify the original test sources and retrieve known answer keys, questions were drawn from multiple test forms administered across different years. ACT materials were drawn from Form 59F (2005–06), Form 0661C (2008–09 retired test), and Form 1874FPRE (2018–19), all developed by ACT, Inc., the organization responsible for administering the ACT in the United States. SAT materials were taken from official full-length SAT practice tests #1, #2, #6, #9, and #10, published by the College Board, which administers the SAT. From these sources, 30 questions were randomly selected for each of the three subjects, Math, Reading, and English, for a total of 90 questions across the study.

Questions were selected to ensure coverage of key concepts within each subject area. For Math, topics included Algebra, Advanced Math, Problem-Solving and Data Analysis, Geometry and Trigonometry. For English and reading, the categories covered evidence-based reading, main idea comprehension, vocabulary and text structure, as well as grammar and punctuation.

Within each subject, questions were randomly ordered prior to administration to prevent any ordering effects. All questions including those accompanied by graphs, diagrams, or data displays were compiled into a single PDF file per subject. These PDF files were uploaded directly to each AI system so that visual materials were rendered and presented in an identical format across all three platforms, ensuring consistency in how image-based content was accessed and interpreted.

The three AI systems tested in this study were ChatGPT-5.3, Google Gemini 3, and Microsoft Copilot Wave 3. All systems were accessed via their standard consumer interfaces during the same testing period. To control for variability in response generation, each system was set to its default temperature setting, and no custom configurations, plugins, or retrieval-augmented tools were enabled. No prior contextual information or example questions were provided before testing began. All three systems received the same standardized prompt, which instructed each AI to answer every question to the best of its ability and to provide a brief 2–3 sentence explanation of its reasoning for each response. These explanations were retained for error analysis to help identify patterns in incorrect responses. After each AI completed the tests, responses were scored for accuracy based on the number of correct answers out of 30 per subject. To determine whether observed differences in accuracy rates across AI systems were statistically significant, a chi-square test for homogeneity was conducted for each subject. This test evaluates whether the proportion of correct and incorrect responses is distributed equally across the three models. An alpha level of .05 was used as the threshold for significance, with a critical value of 5.991 at degree of freedom d_f = 2. Results were subsequently compared across subjects and AI systems to identify overall performance trends, subject-specific strengths, and areas of weakness.

RESULTS

Each AI was evaluated on 30 questions per subject, with performance measured by the number of correct responses. A chi-square test for homogeneity was used to assess whether correct-answer rates differed meaningfully across the three systems. Findings were then analyzed by subject and model to identify overall trends and subject-specific strengths and weaknesses. Tables 1, 2, and 3 present each AI's correct and incorrect responses, accuracy rates, and chi-square outcomes. Note that the reading and writing questions were initially missing the standard instructions typically included at the start of SAT and ACT sections. Copilot and Gemini handled this gap reasonably well, but ChatGPT had difficulty determining how to approach the questions without them. Once the official exam instructions were added and all three AIs were retested, Copilot and Gemini's scores remained the same, while ChatGPT showed marked improvement, as reflected in Tables 1 and 2.

English

Table 1 displays accuracy rates and chi-square test results for the English section. The chi-square test results, $\chi^2(2, N = 90) = 0.829$, $p = .661$, revealed no statistically significant differences among the three models, indicating comparable performance overall.

Table 1. *Performance comparison of ChatGPT, Copilot, and Gemini on English questions (accuracy and Chi-Square test results).*

	ChatGPT	Copilot	Gemini
Correct answers	27	25	27
Incorrect answers	3	5	3
Total answers	30	30	30
Accuracy rate (%)	90.0%	83.3%	90.0%
Chi-square test results	$\chi^2(2, N = 90) = 0.829, p = .661$		

ChatGPT answered 27 of 30 questions correctly for a 90% accuracy rate. Two errors stemmed from incorrect punctuation choices and one from improper use of a possessive form. These mistakes were isolated and did not point to any broader comprehension difficulties.

Copilot correctly answered 25 of 30 questions, earning an 83.3% accuracy rate. Four errors involved text organization, where Copilot tended to revise or remove content that was already acceptable and consistently favored wordier constructions over concise ones in style-based editing tasks. A fifth error arose from misidentifying the grammatical subject, leading to a subject–verb agreement mistake.

Gemini answered 27 of 30 questions correctly, achieving a 90% accuracy rate. Its three errors share a common thread: the overapplication of stylistic rules without adequate attention to broader grammatical context. In one case, a drive for conciseness led to the removal of necessary information; in another, a business-

writing convention was applied without accounting for the punctuation required in a non-restrictive appositive; and in a third, a preference for parallel structure produced a wordier answer than the correct one.

Reading

Table 2 presents accuracy rates and chi-square test results for the reading section. The chi-square test, $\chi^2(2, N = 90) = 3.567$, $p = 0.168$, found no statistically significant differences across the three models.

Table 2. *Performance comparison of ChatGPT, Copilot, and Gemini on reading questions (accuracy and Chi-Square test results).*

	ChatGPT	Copilot	Gemini
Correct answers	25	29	28
Incorrect answers	5	1	2
Total answers	30	30	30
Accuracy rate (%)	83.3%	96.7%	93.3%
Chi-square test results	$\chi^2(2, N = 90) = 3.567, p = .168$		

ChatGPT answered 25 of 30 questions correctly for an 83.3% accuracy rate. Most errors occurred in literary passages, particularly when identifying main ideas. In some instances, ChatGPT drew conclusions beyond what the text supported; in others, it overlooked key qualifiers in the question stem.

Copilot answered 29 of 30 questions correctly, achieving a 96.7% accuracy rate. Its single error involved a misreading of narrative perspective and character relationships, pointing to occasional difficulty with temporally layered narration and subtle character dynamics in complex literary texts.

Gemini answered 28 of 30 questions correctly for a 93.3% accuracy rate. Both errors reflected a tendency to over-interpret passages at the cost of textual accuracy, in one case selecting a thematic inference over a straightforward summary, and in the other failing to recognize the resolution of a building tension.

Math

Table 3 presents accuracy rates and chi-square test results for the Math section. The chi-square test, $\chi^2(2, N = 90) = 0.757$, $p = 0.685$, indicated no statistically significant difference in performance across the three models.

Table 3. *Performance comparison of ChatGPT, Copilot, and Gemini on math questions (accuracy and Chi-Square test results).*

	ChatGPT	Copilot	Gemini
Correct answers	23	21	20
Incorrect answers	7	9	10
Total answers	30	30	30
Accuracy rate (%)	76.7%	70.0%	66.7%
Chi-square test results	$\chi^2(2, N = 90) = 0.757, p = 0.685$		

ChatGPT answered 23 of 30 questions correctly for a 76.7% accuracy rate. Two errors occurred in advanced math, one involving a misinterpretation of fractional exponent notation and another involving a sign error when rearranging a linear equation. Two errors fell under geometry and trigonometry: one from undercounting angles in a parallel lines figure, and one from misreading a trigonometric graph and assigning the wrong sign to an amplitude constant. The remaining errors involved misreading visual and graphical information data analysis.

Copilot answered 21 of 30 questions correctly, earning a 70% accuracy rate. Five errors reflected substantive conceptual gaps in advanced math and algebra, including incorrect identification of the median, miscounting a data set, and selecting an algebraically non-equivalent expression. Three more errors involved misinterpreting visual information, such as scatter plot relationships, wave graph amplitudes, and geometric angle diagrams. One additional error resulted from misapplying the order of operations and incorrectly solving a system of equations.

Gemini answered 20 of 30 questions correctly for a 66.7% accuracy rate, with most errors concentrated in advanced math and data analysis. Two notable patterns emerged. First, Gemini frequently arrived at the correct numerical or algebraic result but then selected the wrong answer choice, pointing to a systematic disconnect between computation and answer selection rather than a fundamental failure in mathematical reasoning. Second, it consistently struggled to accurately interpret visual and graphical information, including scatter plots, axis scale labels, and trigonometric graphs, even when its underlying reasoning was otherwise sound.

DISCUSSION

The present study compared the performance of three contemporary AI systems, ChatGPT, Microsoft

Copilot, and Google Gemini, on standardized SAT and ACT style questions across Math, Reading, and English. Overall, results indicate that while all three models demonstrate strong capabilities in language-based tasks, their performance is less consistent in mathematical reasoning, particularly when visual interpretation is required. Notably, despite observable differences in accuracy across subjects and systems, none of the chi-square test analyses reached statistical significance, suggesting that performance variations were not large enough to establish clear superiority among the models under the study conditions.

A key finding is the consistently high performance across English and Reading sections relative to Math. All three AI systems achieved accuracy rates above 80% in language-based tasks, with particularly strong results in Reading comprehension and vocabulary-in-context questions. This suggests that current large language models are highly effective at processing and interpreting textual information, identifying main ideas, and applying grammatical conventions. However, the error analyses reveal subtle but meaningful differences in how each model approaches language tasks. ChatGPT's errors were largely mechanical (e.g., punctuation and possessives), indicating strong comprehension but occasional lapses in rule application. In contrast, Copilot and Gemini showed a tendency to overapply stylistic rules, such as conciseness and text revision, sometimes leading to unnecessary or incorrect edits. These patterns suggest that while all models are proficient in language understanding, they rely on slightly different heuristics when making decisions about style and structure.

Math proved to be the most challenging domain for all three systems, with accuracy rates dropping significantly compared to language sections. Errors in Math were more varied and often reflected deeper conceptual or procedural misunderstandings. ChatGPT performed the strongest in this category but still exhibited issues with symbolic manipulation, sign errors, and interpreting graphical information. Copilot's errors suggest more fundamental difficulties with core mathematical concepts, including data analysis and algebraic equivalence, as well as occasional computational mistakes. Gemini's performance revealed a unique pattern in which correct solutions were sometimes paired with incorrect answer selections, indicating a disconnect between reasoning and final response execution. Across all models, misinterpretation of visual and graphical information such as scatter plots, geometric diagrams, and trigonometric graphs was a consistent source of error, highlighting a shared limitation in processing non-textual data.

Another important observation is the impact of prompt completeness on performance. In both the English and Reading sections, the absence of standard test instructions initially affected ChatGPT. Once the instructions were added, ChatGPT's performance improved significantly, while Copilot and Gemini remained unchanged. This suggests that ChatGPT may be more sensitive to prompt structure and contextual framing, whereas the other models may rely more on implicit assumptions about task format. This finding underscores the importance of prompt design when evaluating AI performance and suggests that results may vary depending on how tasks are presented.

Despite these findings, several limitations should be considered. The performance of each AI model may vary depending on factors such as model version, prompt design, interface, and subsequent updates. Given the rapid pace of AI development, this variability should be taken into account. First, the sample size of 30 questions per subject may not fully capture the range of skills evaluated in standardized testing. Second, AI performance is highly dependent on model version, interface, and ongoing updates, meaning the results represent only a snapshot in time. Third, although efforts were made to standardize testing conditions, differences in how each platform processes uploaded PDFs and visual content may have influenced the outcomes. Finally, the study focused exclusively on accuracy and did not assess other relevant factors such as response time, confidence, or consistency across repeated trials.

CONCLUSION

This study evaluated the performance of ChatGPT, Microsoft Copilot, and Google Gemini on standardized SAT and ACT style questions across Math, Reading, and English. The findings show that all three AI systems perform strongly in language-based tasks, achieving high accuracy in Reading and English, while demonstrating comparatively weaker and less consistent performance in Math.

Although no statistically significant differences were found among the models, each exhibited distinct error patterns. Language-related errors were typically due to overapplication of stylistic or interpretive rules, while mathematical errors were more often linked to procedural mistakes, conceptual gaps, and difficulties interpreting visual or graphical information. Additionally, the results

highlight the importance of clear and complete prompts, as performance improved, particularly for ChatGPT, when standard instructions were provided.

Overall, these results suggest that current AI systems are well-suited for language comprehension and analysis but remain limited in mathematical reasoning and multimodal interpretation. As AI continues to evolve, improvements in handling visual data, strengthening reasoning consistency, and optimizing response accuracy will be critical for broader application in academic and assessment settings. Future research should expand the sample size, include additional AI models, and explore the impact of prompt design and multimodal enhancements. Evaluating additional factors such as response consistency, reasoning transparency, and real-world applicability would also provide a more comprehensive understanding of AI performance in educational contexts.

CONFLICT OF INTEREST

The author declares that there are no conflicts of interest related to this work.

REFERENCES

1. Hirsch-Kreinsen H. Artificial intelligence: a "promising technology". *AI Soc*. 2024; 39: 1641–1652. doi:10.1007/s00146-023-01629-w.
2. Signorelli CM. Can computers become conscious and overcome humans? *Front Robot AI.* 2018; 5: 121. doi:10.3389/frobt.2018.00121.
3. Sunny A, Joseph S. *Computers' Essential Role in Society and Health*. 2023. Available from: https://www.researchgate.net/publication/376645247_Computers (accessed on 2025-10-15)
4. Russell S, Norvig P. Artificial Intelligence: A Modern Approach. *Pearson*; 2021. ISBN-13: 9780137505135.
5. Alalaq AS. The history of the Artificial Intelligence Revolution and the Nature of Generative AI Work. *DS Journal of Artificial Intelligence and Robotics*. 2024; 18; 2 (4): 1–15. doi:10.59232/AIR-V2I4P101.
6. Raj S, Bansal R. Impact of the Printing Press: Revolutionizing Communication in the Renaissance. *IJIEMR*. 2024; 13 (4): 382-387.
7. Littman ML, Ajunwa I, Berger G, Boutilier C, Currie M, Doshi-Velez F, *et al*. *Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report*. 2021. Available from: https://doi.org/10.48550/arXiv.2210.15767 (accessed on 2025-10-15)
8. Balalle H, Pannilage S. Reassessing academic integrity in the age of AI: a systematic literature review on AI and academic integrity. *Soc Sci Humanit Open*. 2025; 11 (1): 101299. doi:10.1016/j.ssaho.2025.101299.
9. Cardona M, Rodríguez R, Ishmael K. Artificial Intelligence and the Future of Teaching and Learning Insights and Recommendations. *U.S. Department of Education, Office of Educational Technology*. 2023; Available from: https://www.ed.gov/sites/ed/files/documents/ai-report/ai-report.pdf. (accessed on 2025-10-15)
10. Hemmer P, Westphal M, Schemmer M, Vetter S, Vössing M, Satzger G. Human-AI collaboration: the effect of AI delegation on human task performance and task satisfaction. *In: Proceedings of the 28th International Conference on Intelligent User Interfaces*. New York, NY, USA: ACM. 2023; p. 453–63. doi:10.48550/arXiv.2303.09224.
11. McGehee N**.** *AI in Education: Student Usage in Online Learning. Michigan Virtual*. 2024. Available from: https://michiganvirtual.org/research/publications/ai-in-education-student-usage-in-online-learning/?utm_source=chatgpt.com. (accessed on 2025-10-15)
12. Abbas M, Jam FA, Khan TI. Is It Harmful or helpful? Examining the Causes and Consequences of Generative AI Usage among University students. *International Journal of Educational Technology in Higher Education*. 2024; 21 (1). doi:10.1186/s41239-024-00444-7.
13. Nguyen A, Ngo HN, Hong Y, Dang B, Nguyen BT. Ethical principles for artificial intelligence in education. *Educ Inf Technol (Dordr).* 2023; 28 (4): 4221-4241. doi: 10.1007/s10639-022-11316-w.
14. Mai DTT, Van Da C, Van Hanh N. The Use of ChatGPT in Teaching and learning: a Systematic Review through SWOT Analysis Approach. *Frontiers in Education*. 2024; 9: 1328769. doi: 10.3389/feduc.2024.1328769.
15. Fütterer T, Fischer C, Alekseeva A, Chen X, Tate T, Warschauer M, *et al*. ChatGPT in education: global reactions to AI innovations. *Sci Rep*. 2023; 13 (1): 15310. doi:10.1038/s41598-023-42227-6.
16. Lee D, Arnold M, Srivastava A, Plastow K, Strelan P, Ploeckl F, *et al*. The impact of generative AI on higher education learning and teaching: A study of educators' perspectives. *Computers and Education: Artificial Intelligence.* 2024; 6 (100221). doi:10.1016/j.caeai.2024.100221.
17. Goorts L, Hollevoet R, Xia V, Cammaerts F, Güngör A. How do LLMs perform in the context of MCQs across different levels of thinking skills in a business education course at higher education? A comparison of ChatGPT, Gemini, and Copilot. *Computers and*

Education: Artificial Intelligence. 2025; 9 (100475). doi:10.1016/j.caeai.2025.100475.

18. Dimeli M, Kostas A. The role of ChatGPT in education: Applications, challenges: Insights from a systematic review. *J Inf Technol Educ Res*. 2025; 24. doi:10.28945/5422.
19. Giacomo Rossettini, Rodeghiero L, Corradi F, Cook C, Paolo Pillastrini, Turolla A, *et al*. Comparative accuracy of ChatGPT-4, Microsoft Copilot and Google Gemini in the Italian entrance test for healthcare sciences degrees: a cross-sectional study. *BMC medical education*. 2024; 24 (1): 694. doi:10.1186/s12909-024-05630-9.
20. Imran M, Almusharraf N. Google Gemini as a next generation AI educational tool: a review of emerging educational technology. *Smart learning environments*. 2024; 11 (1). doi:10.1186/s40561-024-00310-z.
21. Mustafa MY, Tlili A, Lampropoulos G, Huang R, Petar Jandrić, Zhao J, *et al*. A systematic review of literature reviews on artificial intelligence in education (AIED): a roadmap to a future research agenda. *Smart Learning Environments.* 2024; 11 (1): 1-33. doi:10.1186/s40561-024-00350-5.
22. Nga NTH, Nhung NTP, Anh NTQ. Exploring teacher adoption of AI: A structural analysis of Microsoft Copilot in education. *Journal of Education and e-Learning Research*. 2025; 12 (3): 479–87. doi: 10.20448/jeelr.v12i3.7395.
23. Bano M, Zowghi D, Whittle J, Zhu L, Reeson A, Martin R, *et al*. *A Qualitative Study of User Perception of M365 AI Copilot.* 2025. Available from: https://doi.org/10.48550/arXiv.2503.17661. (accessed on 2025-10-15)
24. Kateryna Osadcha, Justyna Szynkiewicz, Chishti MS. Using Microsoft Copilot Chat in the Work of IT Educators: Pilot Study. *Norsk IKT-konferanse for forskning og utd*anning. 2024 Nov 24; (4). doi:10.5324/nikt.6202.