BMC Nursing

RESEARCH

Open Access

Comparative performance of artificial intelligence models in intensive care nursing questions: an evaluation of ChatGPT, DeepSeek, and Google Gemini

Seçil Gülhan Güner[1*], Ziya Tan[2] and Serhat Gülpınar[3]

Abstract

Background The complexity and rapidly evolving nature of critical patient care in Intensive Care Units underscore the importance of the accuracy and timeliness of nursing decisions, further highlighting the significance of nursing education. This study aims to examine the accuracy of four generative artificial intelligence tools (ChatGPT 5.0 Plus, ChatGPT 5.0, DeepSeek, and Google Gemini) in answering multiple-choice questions related to the intensive care nursing exam, a fundamental area in nursing education.

Methods In the study, the ChatGPT 5.0 Plus, ChatGPT 5.0, DeepSeek, and Google Gemini models were evaluated using a test data set consisting of 55 questions. The questions were classified according to their difficulty levels as easy ($n=16$), medium ($n=17$), and difficult ($n=22$). The models' correct response rates and standard or unique correct/incorrect response distributions were examined. Computer-assisted statistical analysis used the Chi-square, one-way ANOVA, and Post-hoc Tukey tests. The study was reported according to STROBE.

Results According to the study results, the success rates of all models were similar for easy and medium-level questions (70–82%), and the difference between them was not statistically significant ($p>0.05$). Under difficult questions, however, the performance of the models diverged significantly, with Google Gemini achieving the highest success rate at 77.27% and DeepSeek showing the lowest performance at 45.45%. The chi-square analysis revealed no statistically significant difference in the correct/incorrect distribution among the models ($\chi^2=3.69$; $p=0.296$), but at the observational level, Google Gemini had a higher number of unique correct answers ($n=6$) compared to the other models. ChatGPT 5.0 was found to have no unique errors.

Conclusion In conclusion, while AI models generally showed similar levels of success in intensive care nursing exam questions, Google Gemini demonstrated superior performance in difficult questions, and DeepSeek showed the lowest level of success among the models. The study provides an essential comparative framework regarding the usability of AI-based learning and assessment tools in nursing education. It offers guidance for the future development of AI-based educational technologies.

*Correspondence:
Seçil Gülhan Güner
secilgulhan@ktu.edu.tr; secilgulhan@gmail.com

Full list of author information is available at the end of the article

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Clinical trial number Not applicable.

Keywords Nursing, Large language model, Artificial intelligence, ChatGPT, DeepSeek, Google Gemini

Background

Critical care nursing is one of the most complex areas of patient care, and nurses working in this field are expected to possess theoretical knowledge and the ability to make quick decisions, think analytically and critically, and solve problems. The literature emphasizes that minor errors in intensive care units directly impact patient morbidity and mortality. Therefore, it is necessary to test nurses' knowledge using reliable measurement and evaluation methods [1]. However, traditional examination methods are mostly limited to measuring knowledge levels and are insufficient for assessing the critical, analytical, and problem-solving skills required in clinical scenarios. The proliferation of artificial intelligence-based technologies in education in recent years has offered new opportunities to fill this gap [2]. In particular, Chatbot applications based on generative language models are among the innovative digital tools attracting attention in nursing education. AI-based chatbots can simulate clinical scenarios, provide instant feedback, and support self-directed learning, potentially enhancing clinical reasoning skills.

In their study, Zhang (2024) emphasized that chatbots in nursing education increase student engagement and knowledge retention [3]. A study by Han, Park, and Lee (2025) reported that these tools significantly improve self-directed learning and motivation [4]. Similarly, Srinivasan (2024) noted that artificial intelligence (AI) and chatbots contribute to critical thinking and clinical reasoning in nursing, while Cucci (2025) stated that AI offers personalized learning [2, 5]. Labrague (2025) emphasized that AI offers innovative opportunities for foundational pedagogical approaches, but ethical concerns should not be overlooked [1]. Existing research has primarily examined the supportive role of AI in learning engagement, student participation, and knowledge reinforcement (Table 1) [6–11].

The study provides an objective comparative analysis of AI model performance in intensive care nursing examination questions.

Therefore, this study aims to fill a gap in the literature. Based on this rationale, the study aimed to examine the performance of four generative AI tools (ChatGPT 5.0 Plus, ChatGPT 5.0, DeepSeek, and Google Gemini) in answering multiple-choice questions related to the intensive care nursing exam, which is a fundamental area in nursing education. In this respect, the study provides a general pedagogical contribution analysis and objectively compares exam performance. Furthermore, not only success rates but also common correct and incorrect answers, unique correct answers, and statistical significance levels were examined, thus providing a quantitative and comprehensive evaluation unlike previous studies. This approach is original and innovative in providing the first comparative evidence regarding the exam performance of AI models in intensive care nursing education. In this regard, the research questions are as follows:

1. What are the correct answer rates of ChatGPT 5.0 Plus, ChatGPT 5.0, DeepSeek, and Google Gemini models on easy, medium, and difficult intensive care nursing questions?
2. Are there statistically significant differences in the correct response rates of ChatGPT 5.0 Plus, ChatGPT 5.0, DeepSeek, and Google Gemini models on easy, medium, and difficult intensive care nursing questions?

Materials and methods

Study design

This descriptive, observational, cross-sectional study evaluated four generative AI chatbots, including one paid-access model and three publicly available models. The evaluated models were ChatGPT 5.0 Plus, ChatGPT 5.0, DeepSeek, and Google Gemini. The selection of these models was based on several methodological considerations.

First, the ChatGPT series (5.0 and 5.0 Plus) is one of the most widely used and referenced generative language models in education and research today [12]. Evaluating the performance of these models, which are most frequently preferred in academic and clinical settings, on intensive care nursing questions provides an essential basis for comparison, considering their widespread use. The rationale for considering ChatGPT 5.0 and ChatGPT 5.0 Plus together is that, despite belonging to the same model family, they differ in processing capacity and contextual understanding levels. This allows for an objective demonstration of how two models built on the same architecture but with different versions perform on intensive care questions.

Another model included in the study, DeepSeek, is an alternative large language model developed primarily in China that has recently gained popularity [13]. Since it has been trained on different cultural and linguistic datasets, it was deemed meaningful to examine the performance of this model as well, as it could lead to differences in knowledge transfer and accuracy levels in the context of nursing education.

Table 1 Some studies using different chatbots in the field of nursing and key findings

Article	Models	Question Type / Field	Main Findings
[6]	ChatGPT 4.0, ChatGPT 3.5, Google Bard	NCLEX-RN & Chinese National Nursing Licensing Examination (MCQ)	ChatGPT 4.0 demonstrated the best performance with an accuracy of 88.7% for original questions and 79.3% for translation questions.
[7]	GPT-3.5, Gemini	NCLEX-PN & NCLEX-RN (nursing licensing questions)	In the NCLEX-RN, GPT-3.5 (69.2%) performed better than Gemini (65.8%).
[8]	ChatGPT-3.5, Gemini, Microsoft Copilot, Llama 2	100 Nursing Exam Questions (MCQ)	Copilot achieved 84% accuracy, ChatGPT-3.5 77%, Gemini 75%, and Llama 2 68%; each model had some unique correct answers.
[9]	ChatGPT-4, Microsoft Copilot, Google Gemini	Italy Health Sciences-Based Examination (CINECA)	ChatGPT-4 and Copilot achieved significantly higher accuracy than Gemini ($p < 0.001$).
[10]	ChatGPT (various versions)	China Medical Licensing Examination (CNMLE)	Performance differences were found to be related to ChatGPT's version development and system role definitions.
[11]	ChatGPT (depending on the model version)	Taiwan RNLE (Nursing Licensing Examination)	ChatGPT's advantages and limitations were identified; strategic use in education was recommended.

Finally, Google Gemini stands out with its multimodal structure and problem-solving-oriented design and is considered one of the models that will shape the future of AI research [14]. It has been reported that Google Gemini has higher success rates, particularly on questions requiring complex reasoning [15]. Therefore, testing whether Google Gemini outperforms other models on difficult intensive care questions will contribute new insights to the literature in the context of nursing education.

In conclusion, the rationale for selecting these four models is to include the most widely used models globally (the ChatGPT family), an alternative regional model (DeepSeek), and a new generation, powerful multimodal model (Google Gemini), thereby enabling a broad comparative analysis of AI-based assessment tools. All of these tools were accessed in July 2025. The study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) high-quality reporting standards [16].

Application process

Each model was exposed to the same set of questions in separate sessions and under standard conditions. Questions were posed to all models using the same guidelines (prompt standardization) to reduce the randomness factor, and responses were generated. To ensure rigorous and unbiased evaluation, "standard conditions" were operationally defined as follows:

Session Independence; each of the 55 questions was presented to the LLMs in a new, independent chat session. This ensured that the models' internal conversation memory was reset, preventing any contextual bias from previous interactions. Zero-Shot Prompting; no prior examples or "training" prompts were provided before the questions. System Settings; all models were accessed via their primary web interfaces using default system parameters" (e.g., temperature and top-p settings) to ensure the results represent the models' standard clinical reasoning performance without manual tuning.

To ensure a standardized and unbiased evaluation, a zero-shot prompting strategy was utilized. Each question was presented to the models only once to prevent potential response bias or drifting from the models' inherent knowledge state. To ensure a consistent academic framing across all evaluated models, a standardized system prompt was used to guide the responses toward an undergraduate-level nursing perspective. This prompt was not intended to control model parameters, but rather to reduce variability in response style and contextual interpretation across different AI systems. To calibrate the models' academic persona, the following system prompt was consistently utilized: *"I want you to think at the level of an undergraduate nursing student within the scope of this project and answer the questions given to you in order."* This minimal prompting approach was intentionally chosen to evaluate the models' inherent knowledge and reasoning capabilities in intensive care nursing without the influence of complex prompt engineering, contextual cues, or role-playing instructions. To ensure data integrity and minimize human error, all questions were directly copied and pasted from a standardized master Word document into the respective AI chatbot interfaces. This procedure ensured that the syntax, punctuation, and formatting remained identical across all evaluations. Two researchers (SG and ZT) independently managed the data entry process and cross-verified the recorded inputs to maintain strict consistency. Any potential inconsistencies or discrepancies during this process were resolved by consulting a third senior researcher (SGG).

Implementation of the study

The dataset comprises a total of 55 questions, consisting of 25 questions from midterm exams and 30 questions from final exams at a university in northeastern Türkiye during the 2024–2025 academic year. The intensive care nursing course has been offered as an elective course in the nursing department's undergraduate education since 2021 by a faculty member (SGG) with a doctoral degree.

The nursing department where the course is taught is accredited by the Association for Evaluation and Accreditation of Nursing Education Programs (HEPDAK), authorized by the Higher Education Quality Council in Türkiye, and provides undergraduate and graduate-level education.

The course lasts 14 weeks, with two hours per week devoted to theoretical topics taught interactively through demonstrations, questions and answers, and case studies. During the period when the questions were applied, 100 students selected the course. The course has three exams: a midterm exam, a final exam, and a make-up exam. In the department where the study was conducted, the measurement and evaluation committee calculates the discriminative power (easy, medium, and difficult) of the questions before each exam. The classification of question difficulty and the assurance of item validity were based on objective psychometric metrics. The Item Difficulty Index (p) was employed to categorize questions into three tiers: easy ($p > 0.70$), medium ($0.30 \leq p \leq 0.70$), and difficult ($p < 0.30$). Additionally, the Discrimination Index (r) was used as a threshold for item inclusion; only questions demonstrating an r-value of ≥ 0.30 were utilized. This rigorous selection process, overseen by a specialized measurement and evaluation committee, ensured that the test items were statistically sound and capable of effectively distinguishing clinical proficiency levels. The dataset comprises a total of 55 multiple-choice questions obtained from midterm ($n = 25$) and final ($n = 30$) exams conducted between 2021 and 2025. Based on their difficulty levels, the questions were categorized as easy ($n = 16$), medium ($n = 17$), and difficult ($n = 22$). The questions covered fundamental topics in intensive care nursing, such as clinical knowledge, monitoring, mechanical ventilation, sedation and analgesia, patient safety, complication management, and ethical practices. Easy-level questions were simple, including definitions and basic applications. Medium-level questions focused on a more profound understanding, explanations, and comparisons. Difficult-level questions involved critical thinking, analysis, synthesis, and complex case analysis. Questions consisted of five options, with only one correct answer expected for each question. All questions in this study were original, and no online questions or question bank books were used. All test questions and model answers were prepared by the author of this article (SGG) and a nursing associate professor with an intensive care nursing certificate. The other author (SG), who holds a doctorate in nursing, reviewed the questions to ensure clarity and appropriate difficulty levels. To minimize selection and confirmation bias, a structured validation protocol was followed. The independent reviewer (SG) conducted a blind review of the questions, focusing on their alignment with evidence-based intensive care guidelines and international nursing standards. The reviewer was blinded to the study's preliminary hypotheses regarding specific LLM performances. Furthermore, a consensus-based reconciliation process was used to resolve any ambiguities in the questions or model answers, ensuring that the test items provided a neutral and rigorous challenge to all evaluated AI models.

Operational definitions for model evaluation

To standardize the qualitative analysis of the LLMs' performances, two interpretative metrics were defined. 'Consensus Alignment' was defined as a situation where at least three out of the four models ($\geq 75\%$) provided the identical correct answer, indicating high inter-model reliability. 'Idiosyncratic Risk' was defined as the tendency of a single model to generate a 'unique' response (either a unique correct answer or a unique error) that diverges from the collective output of the other three models. This metric was used to evaluate the independent reasoning patterns of the models.

When scoring the answers given in the exam, the average question score was calculated, and this average score was added equally to the easy and difficult questions. To evaluate and compare the performances of different AI models on a consistent scale, a normalization process was applied. The normalization to a 0–100 scale was achieved through a weighted scoring system, where each difficulty level (easy, medium, and difficult) was assigned a specific point value. The weighted scores for each difficulty level were derived using a linear scaling formula to ensure the total score of the 55-question assessment equals exactly 100 points. The weighted scores for each difficulty level were derived using a linear scaling formula to ensure the total score of the 55-question assessment equals exactly 100 points. To reflect the increasing cognitive demand of the questions, an incremental weight of 0.2 points was assigned between each difficulty level. Let x represent the point value of an easy question. The scoring distribution was calculated as follows:

$$16\,(\mathrm{X}) + 17\,(\mathrm{X} + 0.2) + 22\,(\mathrm{X} + 0.4) = 100$$

Solving for x yields a base score of approximately 1.597 for easy questions. Consequently, the points were determined as 1.597 for easy, 1.797 for medium, and 1.997 for difficult questions. This systematic weighting ensures that higher-level clinical reasoning is proportionally rewarded while maintaining a standardized 100-point performance metric. Representing the total correct answers through this standardized weighted metric facilitates easier interpretation and statistical comparison. This approach ensures that raw scores are converted into percentage success levels, making model performances more intuitive and aligned with similar studies in the literature,

while also contributing to a more robust interpretation of statistical analyses such as ANOVA and post-hoc tests.

Statistical analysis

A three-step approach was used to analyze the data. First, a pilot study was conducted to ensure the comprehensibility of the prompts, and a single-prompt strategy was adopted for each question to maintain the integrity of the chatbots' performance and minimize systemic errors. Second, the dataset consisting of 55 questions was entered into four different Large Language Models (LLMs) (ChatGPT 5.0 Plus, ChatGPT 5.0, DeepSeek, and Google Gemini). To ensure data entry accuracy and minimize individual bias, two researchers (SG and ZT) independently performed the data entry process and cross-checked the recorded responses. Each model was tested once with the complete set of questions. The accuracy rates of the LLMs were calculated, and the differences between the models were determined. As a final step, the performance of the LLMs was analyzed based on the difficulty levels and discriminative power of the questions. Subsequently, total scores were calculated each time to assess whether each large language model passed the test (Fig. 1). Data were analyzed using the SPSS 25.0 software package. First, descriptive statistics (frequency, percentage, mean, and standard deviation) were calculated. For the success levels between models, the following tests were applied: Chi-square test (χ^2): To evaluate the difference between models in correct-incorrect distributions; One-way analysis of variance (ANOVA): To compare the mean success scores; Post-hoc Tukey test: To determine which models differed when a significant difference was detected. The 95% Confidence Intervals (CI) for the accuracy rates were calculated using the Wilson Score Interval method, which is particularly robust for smaller sample sizes ($n = 55$) and provides more accurate coverage than the normal approximation method. The statistical significance level was set at $p < 0.05$.

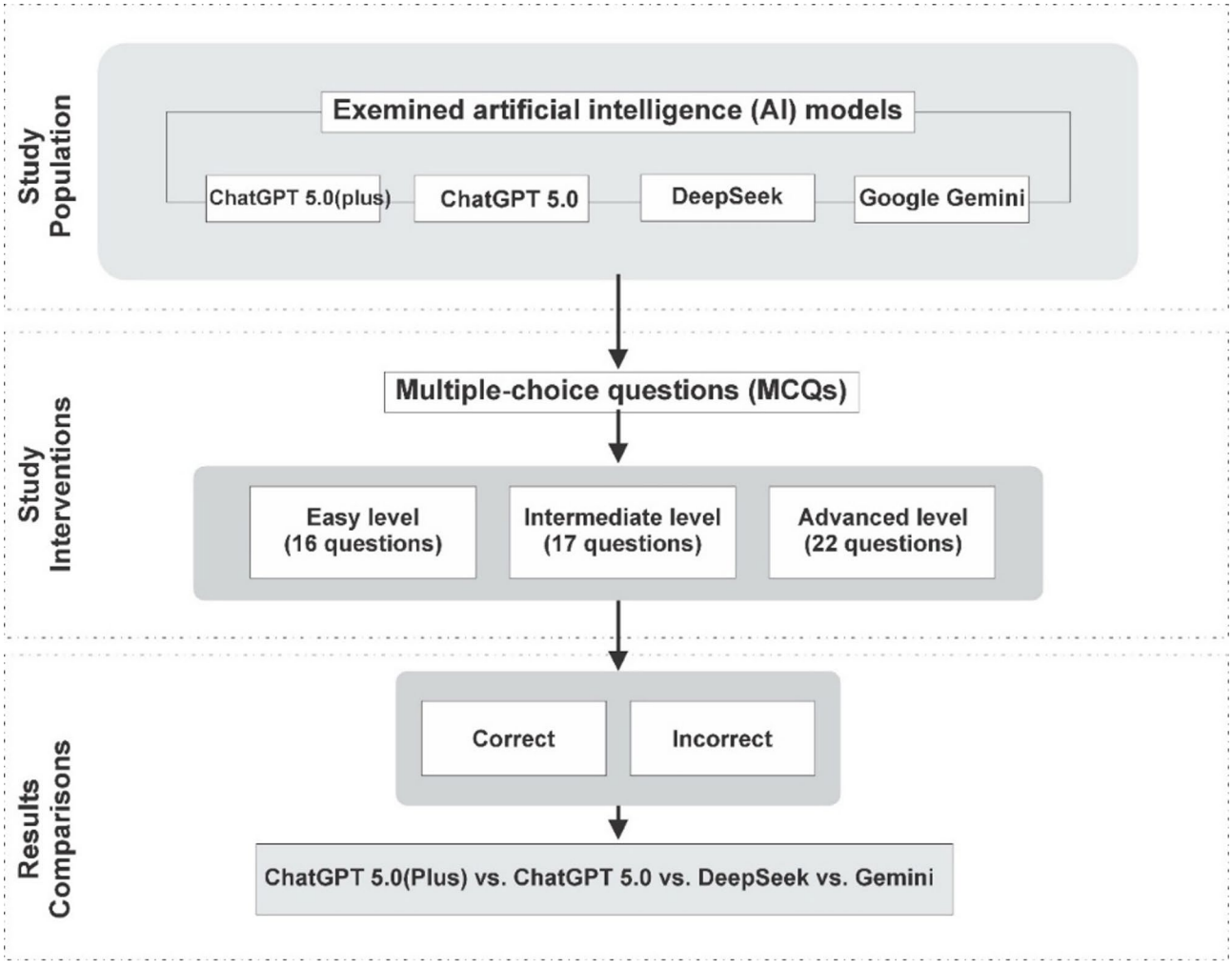

Fig. 1 Research diagram

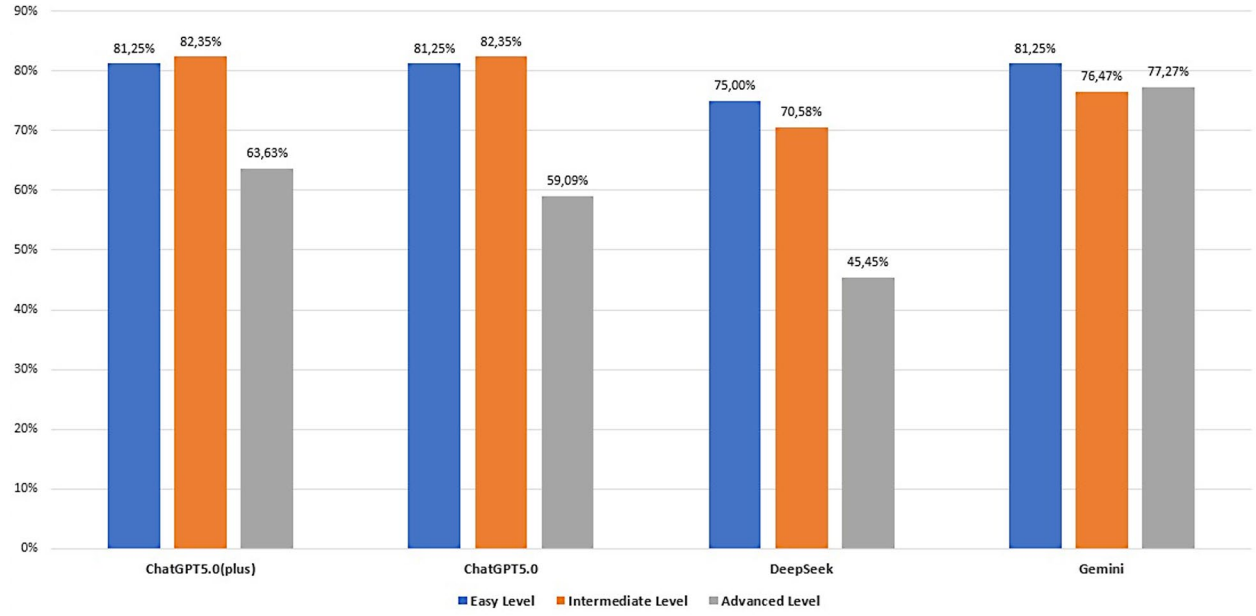

Fig. 2 Accuracy of chatbot responses according to question levels

Table 2 Comparative accuracy of the models ($n=55$)

Chatbots	Correct answer		95% CI	Incorrect answer		Test statistic	
	n	%		n	%	X^2	p
Google Gemini	43	78.18	78.18% (65.0–88.2%)	12	21.82	3.694	0.296
ChatGPT5_Plus	41	74.55	74.55% (61.0–85.3%)	14	25.45		
ChatGPT5	40	72.73	72.73% (59.0–83.9%)	15	27.27		
DeepSeek	34	61.82	61.82% (47.7–74.6%)	21	38.18		

Ethical considerations

This study did not involve human participants or animals; therefore, institutional ethics committee approval was not required. The assessment questions employed in this study were originally developed by the authors, who also serve as the instructors for the relevant course. As such, the intellectual property rights and sole authorization for the use of these materials reside with the authors. The utilization of these items as a research dataset does not contravene any institutional policies regarding examination confidentiality. Furthermore, since the questions were authored specifically by the researchers and are not subject to third-party copyright, no additional institutional or ethical approval was required for their use.

Results

Accuracy of chatbot responses to critical care nursing questions by question level

The number and percentage of correct answers for critical care nursing questions from each chatbot model at each level and overall are shown in Fig. 2. The analysis revealed that the chatbots' accuracy rates for answering questions were as follows: Google Gemini 78.18%, ChatGPT 5.0 (plus) 74.55%, ChatGPT 5.0 72.73%, and Deepseek 61.82%. In answering easy-level questions, ChatGPT 5.0 Plus, ChatGPT 5.0, and Google Gemini showed the highest performance with an equal success rate of 81.25%, while DeepSeek lagged behind other models with 75%. In answering medium-level questions, ChatGPT 5.0 Plus and ChatGPT 5.0 achieved the highest success rate at 82.35%; Google Gemini lagged slightly behind these models at 76.47%, while DeepSeek showed the lowest success rate at 70.58%. For difficult-level questions, the performance ranking was as follows: Google Gemini 77.27%, ChatGPT 5.0 Plus 63.63%, ChatGPT 5.0 59.09%, and DeepSeek 45.45%.

Table 2 shows the comparative accuracy of the models. According to this, there is no statistically significant difference between the total true-false distributions of the four models (X^2=3.694, $p=0.296$). However, it is observed that Google Gemini has more correct answers than expected, while DeepSeek has fewer than expected. In the comparison between ChatGPT 5.0 Plus and ChatGPT 5.0, the two models provided identical answers for 85.45% ($n=47$) of the questions. Specifically, they both failed to identify the correct answer in 21.82% ($n=12$) of the cases, while they reached a consensus on the correct choice in 63.64% ($n=35$) of the questions. Google Gemini

Table 3 Performance analysis of models

Comparison	Average Difference	*p*-value	Lower Confidence Interval	Upper Confidence Interval
ChatGPT5_Plus – ChatGPT5	0.67	0.905	-3.77	5.10
ChatGPT5_Plus – DeepSeek	2.67	0.203	-1.77	7.10
ChatGPT5_Plus – Google Gemini	-1.00	0.812	-5.43	3.43
ChatGPT5 – DeepSeek	2.00	0.379	-2.43	6.43
ChatGPT5 – Google Gemini	-1.67	0.585	-6.10	2.77
DeepSeek – Google Gemini	-3.67	0.081	-8.10	0.77

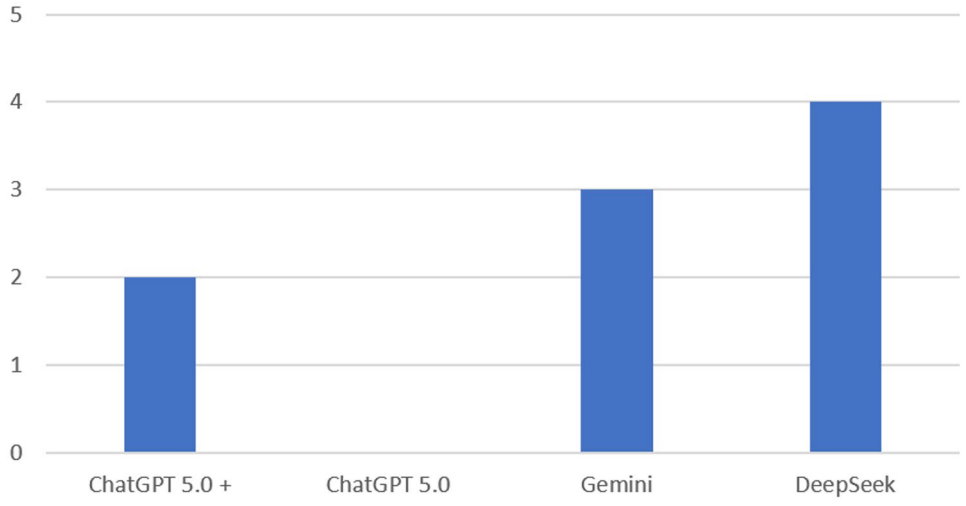

Fig. 3 Unique incorrect answers

demonstrated the highest accuracy rate at 78.18% (95% CI: 65.0%–88.2%), followed by ChatGPT5_Plus at 74.55% (95% CI: 61.0%–85.3%) and ChatGPT5 at 72.73% (95% CI: 59.0%–83.9%). DeepSeek showed a lower accuracy rate of 61.82% (95% CI: 47.7%–74.6%). The overlapping confidence intervals indicate that while performance trends are visible, the differences between the top-performing models are not statistically disparate at the 95% confidence level.

According to the results of the one-way analysis of variance (ANOVA) conducted on the number of correct answers of AI models at different difficulty levels in Table 3, no statistically significant difference was found between the average success scores of the models ($F(3,8) = 2.73$, $p = 0.114$). This finding indicates that the four models performed similarly regarding overall accuracy rates.

According to the post-hoc Tukey test results conducted after the ANOVA, no pairwise comparisons were statistically significant ($p > 0.05$). The difference between Gemini and DeepSeek (mean difference = -3.67, $p = 0.081$) approached the threshold of statistical significance. While the differences between the other model pairs did not reach statistical significance ($p > 0.05$), this marginal p-value suggests a potential trend toward performance variation between Gemini and DeepSeek that might be further elucidated with a larger dataset. Therefore, the lack of statistical significance in these comparisons should not be interpreted as definitive functional equivalence between the models.

When evaluating the data on "the number of questions only the relevant model got wrong" along with the previous accuracy rates in Fig. 3, a clear distinction between consistency and difficulty robustness is observed among the models. As shown in Fig. 3, Google Gemini (77.27%), which demonstrated the highest success at the difficult level, also carries idiosyncratic risk with three unique errors. ChatGPT 5.0 Plus maintains the highest/most stable performance at the easy–medium level (≈81–82%) while making two unique errors. Although ChatGPT 5.0's difficulty level success is lower than Plus (59.09%), it offers the highest consensus alignment with zero unique errors. DeepSeek, on the other hand, ranks low in accuracy (especially at the difficult level, with 45.45%) and has the highest idiosyncratic error load with four unique errors.

Consistency of chatbot responses to intensive care nursing questions: Correct and incorrect answers

Figure 4 examines four AI models' "correct" and "incorrect" answer rates based on their responses to easy, medium, and difficult exam questions. At the easy level, it was determined that 62.5% ($n = 10$) of the total 16 questions were answered correctly by all models, while only

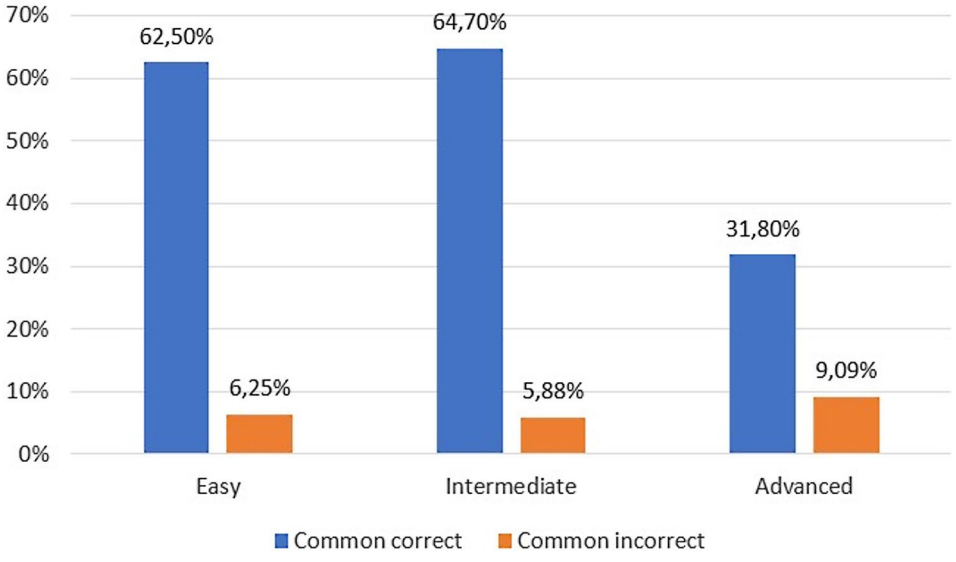

Fig. 4 Common correct and incorrect response rates for all models

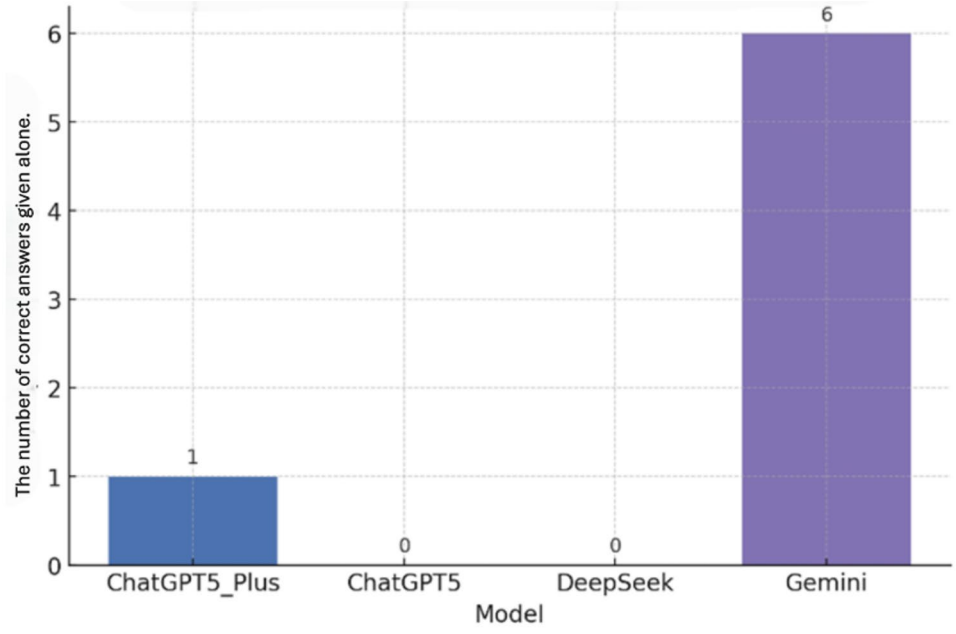

Fig. 5 Unique correct answer performance of models

6.25% ($n = 1$) were answered incorrectly by all models. At the medium level, 64.7% ($n = 11$) of the 17 questions were answered correctly by all models, while 5.88% ($n = 1$) were answered incorrectly. At the difficult level, only 31.8% ($n = 7$) of the 22 questions were answered correctly by all models, while 9.09% ($n = 2$) were answered incorrectly. The response rate was calculated as the ratio of questions answered with a definitive selection of an option to the total number of questions ($n = 55$). All evaluated LLMs achieved a 100% response rate, as they provided a specific answer for every question in the dataset without any refusals or inconclusive outputs.

Figure 5 shows the models' "unique correct" answer performance and reveals which model could provide the correct answer differently from the others, especially on difficult questions. According to the analysis results, ChatGPT 5.0 Plus provided the correct answer in only 1 question, unlike all other models. ChatGPT 5.0 and DeepSeek did not demonstrate this type of "single correct" performance on any question. Google Gemini, on the other hand, demonstrated correct answer performance on six questions.

Discussion

Nursing education is a broad field that equips individuals with the skills to protect, improve, and maintain health. This education covers basic sciences (anatomy, physiology, microbiology, etc.), nursing fundamentals, internal medicine, patient communication, and ethics. Critical care nursing is a specialized branch built upon this education, focusing on supporting the vital functions of critically ill patients. With the integration of artificial intelligence into nursing education, how students and educators will use these new technological approaches has also gained importance. Therefore, this study examined the accuracy of four generative artificial intelligence tools (ChatGPT 5.0 Plus, ChatGPT 5.0, DeepSeek, and Google Gemini) in answering multiple-choice questions related to intensive care nursing. Due to the limited number of studies examining the accuracy of multiple AI chatbots, there are some challenges in comparing the data obtained [9, 17–20]. Furthermore, a literature review revealed no studies examining generative AI tools in multiple-choice questions related to intensive care nursing. Therefore, it is believed that evaluating the correct response performance of four generative AI tools (ChatGPT 5.0 Plus, ChatGPT 5.0, DeepSeek, and Google Gemini) on multiple-choice questions related to intensive care nursing will contribute to the nursing discipline, including applied education.

When the chatbots used in the study were ranked from highest to lowest in terms of their accuracy rates in answering questions, Google Gemini, ChatGPT 5.0 (plus), ChatGPT 5.0, and Deepseek were identified. Furthermore, it was determined that these models showed distinct performance differences in accuracy rates in answering easy, medium, and difficult-level intensive care nursing questions. ChatGPT 5.0 Plus showed consistent and high success in easy and medium-level questions, while Google Gemini stood out significantly in difficult-level questions. DeepSeek lagged behind other models at all levels, offering more limited competence in overall performance. Therefore, while ChatGPT 5.0 Plus or ChatGPT 5.0 may be preferred for reliable performance on easy and medium-level questions, Google Gemini is arguably more advantageous at high difficulty levels. The analysis of 'unique correct answers' and 'unique errors' provides critical insights into the divergent knowledge bases of the evaluated LLMs. A 'unique correct answer' suggests that a specific model may possess a more refined understanding of a particular clinical nuance in internal medicine nursing that other models fail to capture. Conversely, 'unique errors' may indicate specific algorithmic biases or gaps in the training data of a single model. Discussing these outliers is essential for understanding the reliability of LLMs in specialized clinical domains, as they highlight that even models with similar overall accuracy rates may exhibit significantly different reasoning paths for the same clinical scenario. Although not statistically significant, the descriptive data suggests that Google Gemini may offer a potential advantage in specific scenarios, particularly in 'unique correct' answers where other models underperformed. This observed trend could imply a more assertive response strategy, though it also carries an inherent 'idiosyncratic risk' of error. This is because the tendency to provide unique correct answers, while reflecting greater independent reasoning and advanced data interpretation, may also carry an inherent idiosyncratic risk of generating incorrect responses. Indeed, our analysis indicates that Google Gemini produced three unique errors, demonstrating a moderate level of idiosyncratic risk. These findings align with previous studies; for instance, in an evaluation of AI performance in China's Nursing License Exams, ChatGPT 4.0 outperformed both ChatGPT 3.5 and Google Bard in terms of accuracy, further highlighting the disparate reasoning capabilities across different generations of LLMs [6]. In Taiwan, ChatGPT achieved a high accuracy (80.75%) in preparation for the National Nursing Examination but encountered obstacles with complex questions and logical consistency [21]. A study found that Google Gemini performed highly in accuracy on questions from an ear, nose, and throat question bank [22]. A systematic review emphasized that ChatGPT-3.5 took medical exams and performed satisfactorily [23]. Contrary to the results of this study, another study determined that ChatGPT failed the Chinese national medical licensing exams from 2017 to 2021 [24]. Again, in contrast to this study, when the comparative performance of ChatGPT and Google Gemini in answering multiple-choice questions by final-year emergency medicine students was examined, ChatGPT was found to be strong and sound in text-based questions. At the same time, Google Gemini had low accuracy [25]. When comparing different models for American Board of Surgery Education Exam questions, ChatGPT had the highest correct answer rate, while Google Gemini had the lowest [26]. These differing results indicate a need for future studies in this area to explore the potential of AI models in nursing education fully. The findings indicate that statistical significance may not have been achieved due to the limited number of questions in the study, but Gemini showed a clear tendency to outperform DeepSeek. Retesting this with larger sample sizes or different exam sets would help clarify the actual performance differences between the models. Another significant factor to consider is the potential reasons for the divergent results observed between our study and previous literature. First, the domain-specific complexity of intensive care nursing necessitates high-level clinical reasoning, whereas studies reporting higher accuracy rates often utilize more generalized medical or

licensing exam questions. Second, due to the rapid iterative updates of LLMs, performance metrics from earlier models may not align with current results, as seen in the evolution from Bard to Gemini. Finally, the linguistic nuances of Turkish-language prompts may influence model performance compared to studies conducted primarily in English, potentially leading to varied proficiency levels across different languages.

According to the study results, no statistically significant difference exists between the total correct-incorrect distributions of the four models used. In contrast to this result, a study evaluating the performance of large language models (ChatGPT-3.5, ChatGPT-4, and Microsoft Copilot) for the national nursing exam in Japan found a statistically significant difference between the models used [27]. However, the study found that Google Gemini had a higher-than-expected number of correct answers, while DeepSeek had a lower-than-expected number. In addition, the data obtained showed that Google Gemini, despite its high reasoning power, occasionally tended to "deviate on its own"; ChatGPT 5.0 Plus is a strong option for general-purpose use, but it can rarely produce model-specific errors; ChatGPT 5.0's success at the difficult level is lower than Plus, and it implies that errors are primarily clustered in "commonly difficult" questions and that the model does not produce deviations on its own. DeepSeek, on the other hand, demonstrates that it is the model requiring the most careful verification of overall accuracy and consistency. Furthermore, the linguistic context of the questions represents another significant factor. The use of Turkish-language questions is a significant factor that may influence the performance of the evaluated LLMs. While models like Google Gemini and ChatGPT are multilingual, they are predominantly trained on English-centric datasets. Turkish, being an agglutinative language with complex morphology, presents unique challenges for AI tokenization and semantic understanding. Furthermore, the nuances of Turkish nursing terminology and clinical jargon may not be as extensively represented in the training data as their English counterparts. This linguistic barrier might lead to subtle misinterpretations of clinical scenarios, potentially explaining why some models showed 'idiosyncratic' errors or varied performance trends in our study compared to English-language literature.

Limitations of the study

This study has some limitations, but it provides opportunities for future work. First, this study was conducted solely to test questions from an intensive care nursing course in a nursing department in Türkiye; therefore, the results cannot be generalized to all areas in this field. The second limitation is that the study examined four available AI models (ChatGPT 5.0 plus, ChatGPT 5.0, Deepseek, and Google Gemini). Third, both paid (ChatGPT 5.0 plus) and free versions (ChatGPT 5.0, Deepseek, and Google Gemini) of the models were used. Third limitation of this study is the stochastic variability inherent in LLMs. Each question was submitted to the models only once, which may not account for potential variations in responses across different iterations. While this approach reflects a single-point evaluation of the models' current knowledge base, future studies involving multiple testing cycles for the same dataset could provide more robust insights into the consistency and reliability of AI-generated clinical responses. The fourth limitation was that the questions were asked in Turkish. Fifth, only a small number of questions were asked of the chatbots. The sixth limitation was that the response time of AI chatbots was not calculated because the data obtained was affected by the speed of the internet connection and data traffic. The seventh limitation of the study was that intensive care nursing questions in nursing departments in Türkiye do not include common questions at the national level and are prepared by the instructor of the relevant course, so the responses of AI chatbots could not be compared with those of students. The eight limitation, furthermore, since all four LLMs evaluated the same set of 55 questions, the observations may not be entirely independent. The inherent difficulty or ease of certain questions could lead to correlated performance across different models. Nineth, while standard comparative statistical tests (ANOVA and Chi-square) were used to analyze general performance trends, this potential dependence between observations should be considered a methodological limitation. Future studies could employ more complex repeated-measures or paired-design analyses to further explore these inter-model correlations. Furthermore, the sample size of 55 questions is a tenth limitation that may affect the robustness and generalizability of the statistical outcomes. Although these questions represent a comprehensive range of difficulty levels and topics within intensive care nursing, the relatively small dataset means that the results should be interpreted as preliminary. Larger-scale studies incorporating a broader array of nursing specialties and a higher volume of questions are needed to confirm these findings. The last limitation of this study is the reliance on a multiple-choice question format. While this format allows for objective scoring and comparison, it does not provide a direct window into the models' internal reasoning mechanisms. Finally, a high accuracy rate should be interpreted as a measure of performance success rather than a definitive confirmation of complex clinical reasoning. Future research incorporating qualitative analysis of open-ended responses is needed to better understand how these models synthesize clinical data in intensive care nursing.

Conclusion

The results obtained from the comparative evaluation of four different chatbot performances in multiple-choice intensive care nursing exam questions were generally as follows: Google Gemini is preferred for high-difficulty questions, with a second check recommended for critical responses; ChatGPT 5.0 is a safer choice when prioritizing answers close to consensus in institutional/clinical settings; ChatGPT 5.0 Plus stands out for general-aim and balanced performance; DeepSeek should be limited to low-risk use cases with additional verification steps. In summary, Google Gemini represents an innovative and exploratory approach to relevant questions, while ChatGPT 5.0 Plus produced fewer but strategically accurate responses. ChatGPT 5.0 and DeepSeek, conversely, have a highly consistent but less exploratory response strategy due to their lack of unique correct answers. This finding indicates that the strategy should be determined based on whether the goal in model selection is "high common accuracy" or "making a difference on difficult questions." Difficulty level is a decisive factor in AI-based exam performance evaluations; therefore, it is essential to consider common correct and incorrect patterns in model comparisons.

Based on these results, AI chatbots showed promising accuracy in quickly predicting correct answers in the intensive care nursing course. The findings contribute to current discussions regarding the academic use of AI chatbots in nursing. Further research is needed to explore the potential of AI chatbots in nursing education and other healthcare fields and to address their limitations as an innovative tool for exam preparation.

Abbreviations

AI Artificial Intelligence

Author contributions

S.G.G., made work design, conception, and analysis of literature; drafting; final approval; and accountability for all aspects of the work. Z.T., who analyzed the data and drafted the manuscript, was responsible for collecting and analyzing the data, prepared figures and tables.S.G., who analyzed the data and drafted the manuscript, prepared figures, as well as reviewing the literature. All authors reviewed the manuscript.

Funding

None.

Data availability

The transcript dataset, AI prompt, and associated analysis are available from the corresponding author upon reasonable request.

Declarations

Human ethics and consent to participate

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Internal Medicine Nursing, Faculty of Health Sciences, Karadeniz Technical University, Üniversite Neighbourhood, Farabi Street, Number: 88, Ortahisar , Trabzon 61080, Türkiye
[2]Tokat Gaziosmanpaşa University Pazar Vocational School Computer Technology Department, Tokat, Türkiye
[3]Tokat Gaziosmanpaşa University Pazar Vocational School Health Care Services Department, Tokat, Türkiye

Received: 2 February 2026 / Accepted: 24 April 2026
Published online: 02 May 2026

References

1. LabragueLJ, Al Sabei S, Al Yahyaei A. Artificial intelligence in nursing education: A review of AI-based teaching pedagogies. Teach Learn Nurs. 2025;20(3):210–21. https://doi.org/10.1016/j.teln.2025.01.019.
2. CucciF, Marasciulo D, Romani M, Soldano G, Cascio D, De Nunzio G, et al. The Contribution of Artificial Intelligence in Nursing Education: A Scoping Review of the Literature. Nurs Rep (Pavia). 2025;15(8):283. https://doi.org/10.3390/nursrep15080283.
3. ZhangF, Liu X, Wu W, Zhu S. Evolution of Chatbots in Nursing Education: Narrative Review. JMIR Med Educ. 2024;10:e54987. https://doi.org/10.2196/54987.
4. HanJW, Park J, Lee H. Development and effects of a chatbot education program for self-directed learning in nursing students. BMC Med Educ. 2025;25(1):825. https://doi.org/10.1186/s12909-025-07316-2.
5. SrinivasanM, Venugopal A, Venkatesan L, Kumar R. Navigating the Pedagogical Landscape: Exploring the Implications of AI and Chatbots in Nursing Education. JMIR Nurs. 2024;7:e52105. https://doi.org/10.2196/52105.
6. WuZ, Gan W, Xue Z, Ni Z, Zheng X, Zhang Y. Performance of ChatGPT on Nursing Licensure Examinations in the United States and China: Cross-Sectional Study. JMIR Med Educ. 2024;10:e52746. https://doi.org/10.2196/52746.
7. García-RudolphA, Sanchez-Pinsach D, Fernandez-Mira C, Cunyat S, Opisso E, Hernandez-Pena E. Accuracy analysis of AI chatbots GPT-3.5 and GEMINI on English NCLEX-style and Spanish EU general nursing multiple choice questions: challenges and performance insights. Teach Learn Nurs. 2025;20(3):e730–5. https://doi.org/10.1016/j.teln.2025.02.013.
8. HiwaDS, Abdalla SS, Muhialdeen AS, Hamasalih HM, Karim SO. Assessment of Nursing Skill and Knowledge of ChatGPT, Gemini, Microsoft Copilot, and Llama: A Comparative Study. Barw Med J. 2024;2(3):3–6. https://doi.org/10.58742/bmj.v2i2.87.
9. RossettiniG, Rodeghiero L, Corradi F, Cook C, Pillastrini P, 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 Med Educ. 2024;24(1):694. https://doi.org/10.1186/s12909-024-05630-9.
10. MingS, Guo Q, Cheng W, Lei B. Influence of Model Evolution and System Roles on ChatGPT's Performance in Chinese Medical Licensing Exams: Comparative Study. JMIR Med Educ. 2024;10:e52784. https://doi.org/10.2196/52784.
11. HuangH. Performance of ChatGPT on Registered Nurse License Exam in Taiwan: A Descriptive Study. Healthc (Basel). 2023;11(21):2855. https://doi.org/10.3390/healthcare11212855.
12. OpenAI. Introducing GPT-5 OpenAI for developers 2024. Available: https://openai.com/tr-TR/index/introducing-gpt-5-for-developers/.
13. Wang C, Kantarcioglu M. A review of DeepSeek models' key innovative techniques. arXiv preprint arXiv:2503.11486. 2025.
14. Yang L, Xu S, Sellergren A, Kohlberger T, Zhou Y, Ktena I et al. Advancing multimodal medical capabilities of Gemini. arXiv preprint arXiv:2405.03162. 2024.
15. ChuH, Noelle PE, Stephanie Y, Gary C. Assessing the Ethical and Professional Capabilities of AI: A Study of ChatGPT and Google Gemini versus PREview (Situational Judgement Test) for Medical Student Applicant. Lancet Oncol. 2024;1:82–9. https://doi.org/10.69854/jcq.2024.0011.
16. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med. 2007;4(10):e296.
17. KumariA, Kumari A, Singh A, Singh SK, Juhi A, Dhanvijay AKD, et al. Large Language Models in Hematology Case Solving: A Comparative Study of

ChatGPT-3.5, Google Bard, and Microsoft Bing. Cureus. 2023;15(8):e43861. https://doi.org/10.7759/cureus.43861.
18. DhanvijayAKD, Pinjar MJ, Dhokane N, Sorte SR, Kumari A, Mondal H. Performance of Large Language Models (ChatGPT, Bing Search, and Google Bard) in Solving Case Vignettes in Physiology. Cureus. 2023;15(8):e42972. https://doi.org/10.7759/cureus.42972.
19. GiannakopoulosK, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the Performance of Generative AI Large Language Models ChatGPT, Google Bard, and Microsoft Bing Chat in Supporting Evidence-Based Dentistry: Comparative Mixed Methods Study. J Med Internet Res. 2023;25:e51580. https://doi.org/10.2196/51580.
20. Torres-ZegarraBC, Rios-Garcia W, Ñaña-Cordova AM, Arteaga-Cisneros KF, Chalco XCB, Ordoñez MAB, et al. Performance of ChatGPT, Bard, Claude, and Bing on the Peruvian National Licensing Medical Examination: a cross-sectional study. J Educ Eval Health Prof. 2023;20:30. https://doi.org/10.3352/jeehp.2023.20.30.
21. SuMC, Lin LE, Lin LH, Chen YC. Assessing question characteristic influences on ChatGPT's performance and response-explanation consistency: Insights from Taiwan's Nursing Licensing Exam. Int J Nurs Stud. 2024;153:104717. https://doi.org/10.1016/j.ijnurstu.2024.104717.
22. RamchandaniR, Guo E, Mostowy M, Kreutz J, Sahlollbey N, Carr MM, et al. Comparison of ChatGPT-4, Copilot, Bard and Gemini Ultra on an Otolaryngology Question Bank. Clin Otolaryngol. 2025;50(4):704–11. https://doi.org/10.1111/coa.14302.
23. SumbalA, Sumbal R, Amir A. Can ChatGPT-3.5 Pass a Medical Exam? A Systematic Review of ChatGPT's Performance in Academic Testing. J Med Educ Curric Dev. 2024;11:23821205241238641. https://doi.org/10.1177/23821205241238641.
24. ZongH, Li J, Wu E, Wu R, Lu J, Shen B. Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. BMC Med Educ. 2024;24(1):143. https://doi.org/10.1186/s12909-024-05125-7.
25. Al-ThaniSN, Anjum S, Bhutta ZA, Bashir S, Majeed MA, Khan AS, et al. Comparative performance of ChatGPT, Gemini, and final-year emergency medicine clerkship students in answering multiple-choice questions: implications for the use of AI in medical education. Int J Emerg Med. 2025;18(1):146. https://doi.org/10.1186/s12245-025-00949-6.
26. Sanli AN, Tekcan Sanli DE, Karabulut A. Can American Board of Surgery in Training Examinations be passed by Large Language Models? Comparative assessment of Gemini, Copilot, and ChatGPT. Am Surg. 2025;91(11),1923–29 https://doi.org/10.1177/00031348251341956.
27. KuribaraT, Hirayama K, Hirata K. Performance evaluation of large language models for the national nursing examination in Japan. Digit Health. 2025;11:20552076251346571. https://doi.org/10.1177/20552076251346571.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.