Journal of Medical Education Development
e-ISSN: 2980-7670
https://edujournal.zums.ac.ir

Volume 19, Issue 2
April 2026
Pages: 4–11

Original Article

Evaluation of AI support for medical training in resource-constrained settings: performance of GPT-5 Pro, Gemini 2.5 Pro, and DeepSeek V3 on real examination questions

Redouene Sid Ahmed Benazzouz [1*], Massinissa Benyagoub [1], Yacine Boufatah [1], Fodhil Sadeki [1], Mohamed Safouane Benazzouz [2,3], Mounir Ould Setti [4,5]

[1]Faculty of Medicine, Laghouat University, Laghouat, Algeria
[2]Pasteur Institute of Algeria, Algiers, Algeria
[3]Faculty of Pharmacy, University of Algiers, Algiers, Algeria
[4]Alma Mater Europaea University, Vienna, Austria
[5]Observational Studies Germany, Real World Solutions, IQVIA, Frankfurt am Main, Germany

Article info

Article history:
Received 11 Oct. 2025
Revised 17 Oct. 2025
Accepted 22 Feb. 2026
Published 1 Apr. 2026

[*]**Corresponding author:**
Redouene Sid Ahmed Benazzouz, Faculty of Medicine, Laghouat University, Laghouat, Algeria
Email: r.benazzouz@lagh-univ.dz

How to cite this article:
Benazzouz RSA, Benyagoub M, Sadeki F, Benazzouz MS, Setti MO. Evaluation of AI support for medical training in resource-constrained settings: performance of GPT-5 Pro, Gemini 2.5 Pro, and DeepSeek V3 on real examination questions. *J Med Edu Dev*. 2026;19(2):4-11. http://dx.doi.org/10.66224/edcj.19.2.4

Abstract

Background & Objective: Recent advances in Large Language Models (LLMs) have expanded their potential applications in medical education and assessment. This study compared the performance of GPT-5 Pro (OpenAI), Gemini 2.5 Pro (Google DeepMind), and DeepSeek V3 (DeepSeek AI) on authentic, faculty-validated Multiple-Choice Questions (MCQs) from an Algerian francophone Medical Faculty.

Materials & Methods: This parallel, cross-sectional comparative evaluation was carried out under standardized online conditions. A total of 480 faculty-validated, non-public MCQs from a private subscription repository, covering four pre-clinical modules and four clinical modules, were presented to each model in independent chat sessions. Accuracy was compared across models using Cochran's Q and pairwise McNemar tests with Holm correction. Intra-model subgroup analyses (module, study cycle, question type, response format, and temporal factors) used chi-square or Mann–Whitney tests, with $p < 0.05$ considered significant.

Results: Gemini 2.5 Pro achieved the highest accuracy (447/480, 93.1% [95% CI: 90.5 – 95.1]), followed by GPT-5 Pro (430/480, 89.6% [95% CI: 86.5 – 92.0]) and DeepSeek V3 (429/480, 89.4% [95% CI: 86.3 – 91.8]). The overall difference in accuracy was significant (Cochran's Q = 8.65, p = 0.013), with a small global effect size (Kendall's W = 0.009). Pairwise testing showed Gemini 2.5 Pro performed better than both competitors (p = 0.049), whereas GPT-5 Pro and DeepSeek V3 did not differ (p = 1.000). Within-model accuracy was stable across subgroups; non-responses were rare (< 2%) and did not change ranking.

Conclusion: All tested LLMs demonstrated strong competence on structured medical MCQs and may support supervised formative learning in resource-constrained settings. However, although between-model differences were statistically significant, their absolute educational impact was modest, and their effect on real learning outcomes remains uncertain. Key limitations include potential residual training overlap, single-source MCQ sampling, and absence of explanation-quality assessment; future multicenter longitudinal studies should evaluate open-ended clinical reasoning and learning outcomes.

Keywords: large language models; education, medical; artificial intelligence; formative assessment; resource-limited settings

[Downloaded from edujournal.zums.ac.ir on 2026-11-29]

[DOI: 10.66224/edcj.19.2.4]

Copyright © 2026 Zanjan University of Medical Sciences. Published by Zanjan University of Medical Sciences.
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International license (https://creativecommons.org/licenses/by-nc/4.0/). Noncommercial uses of the work are permitted, provided the original work is properly cited.

Introduction

Recent advances in Large Language Models (LLMs) have opened new possibilities for supporting medical training, particularly in settings where teaching resources, faculty availability, and standardized materials may be limited. These models have shown strong potential in structured assessment tasks, including Multiple-Choice Questions (MCQs), which remain a central component of medical education worldwide, including in low-resource and lower-middle-income settings.

OpenAI's GPT-5 Pro (OpenAI, San Francisco, USA) has been introduced as a next-generation model with improved reasoning and stronger health-related benchmark performance than earlier OpenAI models [1]. Similarly, Google DeepMind's Gemini 2.5 Pro (Google DeepMind, London, UK) is presented as an advanced reasoning model with strong long-context and multimodal capabilities relevant to medical question-answering tasks [2]. DeepSeek V3 (DeepSeek, Beijing, China) is a competitive open-weight alternative designed for efficient reasoning and multilingual comprehension, which is particularly relevant for resource-sensitive educational environments [3]. Despite these promising developments, independent evaluations show notable limitations.

For example, a recent study in emergency medicine found that GPT-4.0 and Gemini-1.5 performed worse than final-year medical students on both text-only and image-based MCQs [4]. In addition, most benchmark datasets used to evaluate LLMs consist of publicly accessible or widely circulated questions, raising concerns about possible overlap with model training corpora. Recent evidence syntheses indicate that much of the medical licensing literature before late-generation models was dominated by earlier LLM versions and was often anchored in English-centric benchmark ecosystems, with relatively few evaluations of non-English examinations [1].

Available studies of non-English medical exams (e.g., Chinese, Korean, Spanish, and French contexts) support feasibility but remain heterogeneous in scope, specialty focus, and model generation, which limits direct transferability to francophone, resource-constrained training settings [2–5]. Critically, most existing evaluations have focused on earlier-generation models or English-language datasets, leaving the real-world performance of next-generation LLMs uncertain in non-English and resource-constrained educational environments. Because linguistic structure, curriculum design, and educational infrastructure differ across regions, performance observed in English-dominant benchmark settings cannot be assumed to generalize to francophone North African medical training contexts. This creates an important and unresolved research gap regarding the reliability and educational value of current-generation LLMs in underrepresented linguistic and resource-constrained environments.

To reduce this risk, the present study used a private, paywalled repository of Algerian medical school examination MCQs (2022–2025), accessible only through subscription and not publicly available online. Only items validated through faculty-issued answer keys were included. This approach allows a contextually grounded evaluation in a francophone North African setting where educational digital infrastructure and access to standardized preparation resources may be uneven.

To our knowledge, we did not identify any prior published head-to-head evaluations of GPT-5 Pro, Gemini 2.5 Pro, and DeepSeek V3 using a private French-language undergraduate medical exam repository from a North African context.

In this context, we conducted a cross-model comparative evaluation of GPT-5 Pro, Gemini 2.5 Pro, and DeepSeek V3 using authentic French-language MCQs covering both pre-clinical (Biochemistry, Immunology, Genetics, Anatomic Pathology) and clinical (Cardiology, Otorhinolaryngology, Endocrinology, Gynecology) curricula. These models were selected because they were publicly accessible, supported multilingual prompting, and ranked among the most visited AI chatbot platforms globally at the time of writing [6]. We also examined time of day and weekday as exploratory operational covariates, along with module, study cycle, question structure (isolated vs vignette-based), and response format (single vs combination-type answers), to identify where performance variability may arise in real educational use. Module and study cycle were included to test whether accuracy varies with disciplinary content and expected cognitive level (pre-clinical recall-oriented knowledge vs more clinically integrated reasoning). Question structure was analyzed because isolated stems and vignette-based items differ in contextual load and clinical inference demands, whereas response format was analyzed because single-best-answer and combination-type items impose different levels of decision complexity and distractor management.

In contrast, temporal factors were pre-specified as hypothesis-generating robustness checks rather than

[Downloaded from edujournal.zums.ac.ir on 2026-11-29]

[DOI: 10.66224/edcj.19.2.4]

causal predictors, because prior educational evidence for true temporal effects on LLM accuracy is limited and real-world users may suspect output variability during periods of high platform demand [7]. Accordingly, weekday and time-of-day analyses were intended to assess operational stability under varying usage load, not to infer a causal temporal mechanism. The primary aim of this study is to compare the performance of GPT-5 Pro, Gemini 2.5 Pro, and DeepSeek V3 in solving MCQs derived from authentic medical school examinations. The secondary aim is to assess the influence of contextual factors—including time of day, weekday, academic module, study cycle, question structure (isolated vs vignette-based), and answer format (single-best-answer vs combination-type)—on the performance of these AI models.

Materials & Methods

Design and setting(s)

This cross-sectional comparative study was conducted in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement for cross-sectional studies [8]. The study evaluated the performance of three LLMs: GPT-5 Pro (OpenAI), Gemini 2.5 Pro (Google DeepMind), and DeepSeek V3 (DeepSeek AI). Each model was tested on the same standardized set of validated Algerian medical school MCQs to compare their accuracy under identical conditions. The study was conducted online from 22 August 2025 to 13 September 2025. Potential sources of bias included the use of a single educational context and a private question bank that might partially overlap with model training data. To reduce these risks, only faculty-validated questions with official answer keys were included, and all models were tested using an identical prompting strategy.

Participants and sampling

This study did not involve human participants. The analytical units consisted of 480 MCQs obtained from Quizy-DZ (www.quizy-dz.com), a private subscription-based educational platform that requires paid access and is not freely accessible or indexed as a public benchmark dataset. The platform reproduces official Algerian medical school examinations from 2022 to 2025. All items were faculty-validated using official answer keys issued by university teaching staff, ensuring that each question had a verified correct response. The dataset covered eight academic modules: Biochemistry, Immunology, Genetics, and Anatomic Pathology in the pre-clinical cycle, and Cardiology, Otorhinolaryngology, Endocrinology, and Gynecology in the clinical cycle. Sixty questions were included from each module, yielding a total of 480 items. Questions were sampled chronologically in descending order to ensure consistency across modules.

Tools/Instruments

Three LLM systems served as the analytical tools in this study. GPT-5 Pro was accessed through the ChatGPT Plus interface, Gemini 2.5 Pro through Google One AI Premium, and DeepSeek V3 through its public web interface. All testing accounts were newly created with no prior conversation history, and the default ‘auto’ mode was used for all models.

Data collection methods

The Multiple-Choice Questions (MCQs) were exclusively in French and utilized a single-best-answer format. The dataset comprised standalone items, progressive vignette-based questions, and combination answer formats like ‘1+2’ or ‘1+2+3’. Questions were submitted using a standardized French instruction: ‘Choisissez la bonne réponse’ (‘Choose the right answer’), followed by the question stem and answer options. No prompt engineering, fine-tuning, or additional contextual information was provided.

All questions were evaluated using a single-turn zero-shot approach, meaning that each item was presented with one instruction and no examples or follow-up prompts.

Only the first response generated by each model was recorded for analysis, and the models were required to provide a single final answer option. Explicit chain-of-thought prompting was not requested.

For GPT-5 Pro and Gemini 2.5 Pro, each question was entered in a new temporary chat session using study-specific accounts with empty memory. Temporary chat functionality was not available for DeepSeek V3 during the study period; therefore, conversation history was manually cleared after each question. Each MCQ was entered in a separate chat session, except for sequential vignette questions, which were grouped to preserve the continuity of clinical reasoning.

Responses were manually recorded in a standardized extraction file and compared with the official answer keys. The primary outcome was model accuracy, defined as the proportion of correctly answered MCQs. Non-responses or inconclusive outputs were classified as incorrect. Additional recorded variables included

[Downloaded from edujournal.zums.ac.ir on 2026-11-29]

[DOI: 10.66224/edcj.19.2.4]

academic module (pre-clinical or clinical), study cycle, question structure (isolated or case-based), and response format (simple or combination-type). Each query was also time-stamped and categorized into four predefined time intervals (00:00–09:59, 10:00–11:59, 12:00–15:59, and 16:00–23:59) as well as by weekday to explore possible contextual influences. A non-response was defined as a first-shot output that did not commit to a single final answer option.

Data analysis

Statistical analyses were performed using R Project for Statistical Computing, version 4.5.1 (R Foundation for Statistical Computing, Vienna, Austria). Model performance was expressed as the proportion of correct answers with corresponding 95% confidence intervals (95% CI) calculated using the Wilson score method.

Descriptive statistics summarized overall accuracy and subgroup distributions according to subject module, cycle of study (pre-clinical vs clinical), question structure (isolated vs vignette-based), response format (simple vs combination-type), weekday of testing, and time of day.

Comparisons between subgroups were performed using the chi-square test or the Mann–Whitney U test, as appropriate. Global differences in accuracy across the three models were evaluated using Cochran's Q test for paired binary outcomes.

When significant differences were observed, pairwise McNemar tests were conducted with Holm correction for multiple comparisons. Effect sizes were reported as Kendall's W for Cochran's Q and matched odds ratios with 95% confidence intervals for McNemar comparisons. A two-sided p-value < 0.05 was considered statistically significant.

Weekday and query-hour analyses were prespecified as exploratory secondary analyses and were interpreted as hypothesis-generating assessments of operational stability rather than confirmatory inferential tests.

After the quantitative analyses, the 14 items missed by all three LLMs were reviewed by a panel of four experts whose expertise collectively covered the included modules. This complementary review classified items by domain (pre-clinical or clinical), convergence of model errors (same vs different distractors), and key validity (single clearly correct answer vs multiple defensible answers). This analysis was descriptive and was not part of the primary inferential endpoint.

Results

The dataset included 480 validated medical MCQs across eight modules. No human participants were involved.

Gemini 2.5 Pro achieved the highest accuracy, correctly answering 447 items (93.1%, 95% CI 90.5–95.1), followed by GPT-5 Pro (430/480, 89.6%, 95% CI 86.5–92.0) and DeepSeek V3 (429/480, 89.4%, 95% CI 86.3–91.8).

Cochran's Q test indicated a significant overall difference among models (Q = 8.65, df = 2, p = 0.013). The overall effect size was small (Kendall's W = 0.009). Holm-adjusted pairwise McNemar comparisons showed that Gemini 2.5 Pro outperformed both GPT-5 Pro and DeepSeek V3 (both p = 0.049), whereas no significant difference was observed between GPT-5 Pro and DeepSeek V3 (p = 1.000; **Table 1**).

Table 1. Comparative performance of GPT-5 Pro, Gemini 2.5 Pro, and DeepSeek V3 on 480 medical MCQs

Model / Comparison	Correct answers, n/N (%)	95% CI	Adjusted p-value*	Effect size
Panel A. Accuracy by model				
GPT-5 Pro	430/480 (89.6)	86.5–92.0	—	—
Gemini 2.5 Pro	447/480 (93.1)	90.5–95.1	—	—
DeepSeek V3	429/480 (89.4)	86.3–91.8	—	—
Panel B. Global and pairwise comparisons				
Global (all three models)	—	—	0.013	Kendall's W = 0.009
GPT-5 Pro vs Gemini 2.5 Pro	—	—	0.049	Matched OR = 0.485 (95% CI 0.267–0.881)
GPT-5 Pro vs DeepSeek V3	—	—	1.000	Matched OR = 1.048 (95% CI 0.576–1.905)
Gemini 2.5 Pro vs DeepSeek V3	—	—	0.049	Matched OR = 2.125 (95% CI 1.173–3.850)

Note: Ninety-five percent confidence intervals (CIs) for accuracy were calculated using Wilson score intervals. The global p-value was obtained using Cochran's Q test (Q = 8.65, df = 2). Pairwise p-values were calculated using McNemar tests with Holm correction (two-sided α = 0.05). Holm adjustment applies to pairwise comparisons only.
Abbreviations: MCQ, multiple-choice question; CI, confidence interval; OR, odds ratio; df, degrees of freedom.

[Downloaded from edujournal.zums.ac.ir on 2026-11-29]

[DOI: 10.66224/edcj.19.2.4]

Non-responses were rare: GPT-5 Pro produced 7 (1.5%), DeepSeek V3 3 (0.6%), and Gemini 2.5 Pro 1 (0.2%). These outputs were treated as incorrect but are reported separately for transparency. In three cases (two with GPT-5 Pro and one with Gemini 2.5 Pro), the model mentioned the correct option while hesitating between two alternatives but did not commit to a single final answer; according to the prespecified rule, these outputs were scored as incorrect. Across all items, 14 questions (2.9%) were missed by every model; in 10 of these items, all models converged on the same incorrect option. Concordance analysis showed strong agreement between systems: Gemini 2.5 Pro and GPT-5 Pro produced identical correct answers on 429 items (89.4%), DeepSeek V3 and GPT-5 Pro on 426 items (88.8%), and DeepSeek V3 and Gemini 2.5 Pro on 427 items (89.0%). Model accuracy remained stable across contextual and academic factors. No significant variation was observed by weekday or query hour; median query times were similar across models (all $p > 0.15$; **Table 2**).

Table 2. Temporal analysis of model performance according to weekday and query hour

Variable	GPT-5 Pro	p	DeepSeek V3	p	Gemini 2.5 Pro	p
Day of week	—	0.186	—	0.165	—	0.210
Query hour, median [IQR]	12.0 [10–16]	0.945	12.0 [10–16]	0.673	16.0 [9–17]	0.150

Note: Values are presented as median [interquartile range]. P-values represent within-model comparisons between correct and incorrect responses. For Day of week, p-values were obtained using a chi-square test across weekday categories within each model. For Query hour, p-values were obtained using a Mann–Whitney U test comparing hour distributions between correct and incorrect responses within each model. No between-model comparisons are reported in this table.
Abbreviations: IQR, interquartile range; p, probability value.

Accuracy was also consistent across the eight medical modules and between pre-clinical and clinical questions, with no significant within-model differences (all $p > 0.05$; **Table 3**). Question structure (isolated vs. vignette-based) and answer format (single vs. combination type) showed no measurable effect on performance, with all within-model p-values exceeding 0.30 (range 0.34–0.75). Among the 14 items missed by all models, 8 were pre-clinical and 6 were clinical. All three systems converged on the same incorrect option in 10 of 14 items, whereas in 4 items the selected incorrect options differed across models.

Expert panel review indicated that 11 of the 14 items allowed more than one defensible answer and required selecting the 'most correct' option. In the remaining three items, experts concluded that the answer selected by the models was correct and that the reference key was likely incorrect.

Table 3. Model accuracy across modules, study cycle, question structure, and answer format (N = 480 MCQs)

Subgroup	GPT-5 Pro Correct n/N (%)	p†	DeepSeek V3 Correct n/N (%)	p†	Gemini 2.5 Pro Correct n/N (%)	p†
Modules (N = 60 each)						
Anatomic Pathology	56/60 (93.3)		57/60 (95.0)		59/60 (98.3)	
Biochemistry	49/60 (81.7)		53/60 (88.3)		56/60 (93.3)	
Genetics	49/60 (81.7)		49/60 (81.7)		54/60 (90.0)	
Immunology	55/60 (91.7)		53/60 (88.3)		56/60 (93.3)	
Cardiology	54/60 (90.0)		55/60 (91.7)		57/60 (95.0)	
Endocrinology	57/60 (95.0)		56/60 (93.3)		55/60 (91.7)	
Gynecology	54/60 (90.0)		51/60 (85.0)		52/60 (86.7)	
Otorhinolaryngology	56/60 (93.3)		55/60 (91.7)		58/60 (96.7)	
Module comparison (omnibus)		0.099		0.271		0.247
Study cycle (N = 240 each)						
Pre-clinical items	209/240 (87.1)		212/240 (88.3)		225/240 (93.8)	
Clinical items	221/240 (92.1)		217/240 (90.4)		222/240 (92.5)	
Cycle comparison		0.100		0.554		0.718
Question structure						
Isolated questions (N = 443)	398/443 (89.8)		397/443 (89.6)		414/443 (93.5)	
Case-based vignette items (N = 37)	32/37 (86.5)		32/37 (86.5)		33/37 (89.2)	
Structure comparison		0.718		0.752		0.518
Answer format						
Single-best-answer (N = 420)	378/420 (90.0)		378/420 (90.0)		393/420 (93.6)	
Combination-type (N = 60)	52/60 (86.7)		51/60 (85.0)		54/60 (90.0)	
Answer-format comparison		0.572		0.341		0.453

Note: Values represent accuracy within each subgroup (correct/total, %). †P-values represent within-model comparisons. 'Module comparison' corresponds to an omnibus test across the eight modules; other p-values compare the two subgroup levels (e.g., pre-clinical vs clinical, isolated vs vignette, single-best-answer vs combination-type). Chi-square or Fisher's exact tests were used as appropriate. No between-model comparisons are presented in this table.
Abbreviations: MCQ, multiple-choice question.

[Downloaded from edujournal.zums.ac.ir on 2026-11-29]

[DOI: 10.66224/edcj.19.2.4]

Discussion

This study evaluated whether three widely used general-purpose LLMs can support structured medical assessment in a francophone, resource-constrained educational context using authentic local examination materials. An analysis of 480 faculty-validated MCQs revealed that all three models achieved high accuracy rates exceeding 89%. Specifically, Gemini 2.5 Pro scored 93.1%, followed by GPT-5 Pro at 89.6% and DeepSeek V3 at 89.4%. While the overall performance differences were statistically significant (Cochran's Q $p = 0.013$), the effect size was minimal (Kendall's $W = 0.009$), suggesting negligible practical distinction. This is an important educational consideration, as statistical significance does not automatically translate into a substantial pedagogical benefit. Pairwise comparisons indicated practical equivalence between GPT-5 Pro and DeepSeek V3 (0.2% absolute difference; $p = 1.000$; OR 1.05, 95% CI 0.58–1.91). Conversely, Gemini's modest yet significant lead (3.5–3.7%) appears to represent an incremental improvement rather than a fundamental paradigm shift in learner outcomes. When viewed alongside previous syntheses [1, 9] and recent 2025 comparative studies [5, 10], these findings likely reflect a combination of methodological and generational factors. The models assessed here belong to newer releases than most systems included in earlier meta-analyses. In addition, our benchmark relied on a private, paywalled repository with faculty-validated answer keys, whereas many prior evaluations used public or widely circulated datasets.

The French-language and North-African curricular context further differentiates our data from the predominantly English-language literature. Finally, the balanced inclusion of eight modules spanning pre-clinical and clinical disciplines may have enhanced measurement stability compared with narrower, single-specialty samples. Nevertheless, part of the high observed accuracy likely stems from the task format. Single-best-answer MCQs reward option discrimination but do not capture open-ended reasoning, uncertainty management, or communication competence.

This suggests that LLMs should be integrated primarily as formative support tools rather than as substitutes for summative assessment. The observed accuracy thus reflects both model enhancements and the inherently constrained nature of MCQ testing. Additional interpretive insight came from expert review of the 14 universally missed items: 8 were pre-clinical and 6 clinical. Models converged on the same incorrect option in 10 of these questions and diverged in 4. Experts judged that 11 items contained more than one defensible answer ('most-correct' framing), while in three cases the reference key was likely incorrect. Hence, a portion of the apparent 'model error' likely represents question subjectivity rather than reasoning failure. In assessment design, such 'most-correct' items are frequent yet may diminish scoring fairness and benchmarking accuracy. Accordingly, these findings should be interpreted as accuracy relative to operational keys, with caution regarding disputed items.

The scoring protocol was intentionally conservative: models were required to provide a single final option on first attempt. Three outputs (two originating from GPT-5 Pro and one from Gemini 2.5 Pro) mentioned the correct option but did not finalize an answer and were classified as incorrect. This criterion may have slightly lowered observed accuracy yet improved consistency and minimized post-hoc bias.

Taken together, these results highlight three practical levels of implication.

For students, LLMs may serve as adjunctive tutors for structured revision, provided outputs are verified against course materials and used with explicit acknowledgment of uncertainty.

For educators, convergence and disagreement patterns among models can flag potentially flawed items before high-stakes deployment, though final judgment should remain faculty-led. For institutions and policymakers, integration should emphasize governance, academic integrity, and AI literacy—ensuring that certification decisions remain under human oversight. Future research should extend beyond response accuracy toward explanation quality and clinical reasoning. Next steps include multicenter validation across francophone contexts, use of newly generated post-training questions, and longitudinal study designs examining durable learning outcomes.

From an assessment standpoint, further work should also test whether AI-assisted review enhances psychometric indices such as discrimination, reliability, and challenge rates. Establishing predefined multi-expert adjudication for ambiguous items will strengthen interpretive validity in subsequent research.

Several limitations should temper interpretation. Although the source repository is private and paywalled, complete exclusion of prior model exposure cannot be guaranteed. All items were drawn from a single faculty source, which may limit generalizability across institutions. The study evaluated answer selection only,

[Downloaded from edujournal.zums.ac.ir on 2026-11-29]

[DOI: 10.66224/edcj.19.2.4]

not explanation quality or reasoning safety. Explicit chain-of-thought prompting was not applied, and first-shot scoring may underestimate potential performance under structured deliberation prompts. Finally, the presence of 'most-correct' items implies residual label subjectivity, affecting interpretability for both human examinees and LLM benchmarking.

Conclusion

This study evaluated the performance of GPT-5 Pro, Gemini 2.5 Pro, and DeepSeek V3 on a dataset of 480 authentic, faculty-validated Algerian medical MCQs sourced from a private repository. Results indicated that Gemini 2.5 Pro attained the highest accuracy at 93.1%, demonstrating a modest yet statistically significant advantage over GPT-5 Pro (89.6%) and DeepSeek V3 (89.4%).

Model performance remained stable across modules, academic cycles, question structures, response formats, weekdays, and time-of-day categories, supporting the reliability of current general-purpose LLMs for structured educational assessment in francophone, resource-limited settings. While statistically significant, the magnitude of superiority was small—fewer than four additional correct answers per 100 questions—and unlikely to produce a major difference in supervised learning outcomes.

These findings suggest that LLMs are ready to be adopted as supervised formative tools, not as autonomous decision systems. Responsible implementation, grounded in institutional oversight and AI literacy, could help reduce educational inequities between low-resource and high-resource environments while advancing global health education goals. Broader multicenter studies exploring open-ended reasoning and real-world learning outcomes are warranted before any high-stakes integration.

Ethical considerations

This study was conducted in accordance with the ethical principles of the Declaration of Helsinki and was approved by the Ethics Committee of the Faculty of Medicine, University of Laghouat, Algeria (Protocol No. 19/2025; Decision Date: June 8, 2025). The research involved no patient data or personally identifiable information.

Examination materials were obtained from an external faculty repository under a formal license, and permission for use and analysis was secured from the rights holder prior to data processing.

Artificial intelligence utilization for article writing

AI-assisted tools were used solely for minor grammar and language refinement in accordance with ethical guidelines. They were not used for content generation, data analysis, or interpretation. All outputs were carefully reviewed and verified by the authors prior to submission. Specifically, ChatGPT (OpenAI; GPT-5.2 model) was used only to improve grammar, spelling, and stylistic clarity of the manuscript text. It was not used to generate scientific claims, conduct statistical analyses, interpret findings, select references, or make publication decisions.

Acknowledgment

None.

Conflict of interest statement

The authors declare that there are no conflicts of interest relevant to the content of this publication.

Author contributions

RSB led the conceptualization of the study and was responsible for project administration. Conceptualization was performed by RSB, YB, FS, and MO. Methodology, formal analysis, and validation were conducted by RSB, MSB, and MO. Investigation was carried out by RSB, YB, FS, MB, and MSB, while data curation was undertaken by YB, FS, MB, and MSB. The original draft of the manuscript was prepared by RSB, YB, and FS. Critical review and editing of the manuscript were performed by RSB, YB, FS, MB, MSB, and MO. Supervision of the research process was provided by RSB and MO. Funding acquisition was not applicable. All authors reviewed and approved the final version of the manuscript.

Funding

None.

Data availability statement

For transparency and independent verification, the raw dataset used in this study has been deposited in Zenodo as Dataset 1 (DOI: 10.5281/zenodo.17666696). The deposited files include the MCQ dataset, the corresponding official reference answers, the raw outputs generated by GPT-5 Pro, Gemini 2.5 Pro, and DeepSeek V3, as well as all temporal and contextual variables used in the analyses. A detailed and readable codebook

[Downloaded from edujournal.zums.ac.ir on 2026-11-29]

[DOI: 10.66224/edcj.19.2.4]

describing variable definitions and coding procedures accompanies the dataset.

References

1. Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. *J Med Internet Res*. 2024;26:e60807. https://doi.org/10.2196/60807
2. Fang C, Wu Y, Fu W, Ling J, Wang Y, Liu X, et al. How does ChatGPT-4 perform on non-English national medical licensing examination? An evaluation in Chinese language. *PLOS Digit Health*. 2023;2(12):e0000397. https://doi.org/10.1371/journal.pdig.0000397
3. Jang D, Yun TR, Lee CY, Kwon YK, Kim CE. GPT-4 can pass the Korean National Licensing Examination for Korean Medicine Doctors. *PLOS Digit Health*. 2023;2(12):e0000416. https://doi.org/10.1371/journal.pdig.0000416
4. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, Alas-Brun R, Onambele L, Ortega W, et al. Evaluating the efficacy of ChatGPT in navigating the Spanish Medical Residency Entrance Examination (MIR): promising horizons for AI in clinical medicine. *Clin Pract*. 2023;13(6):1460–1487. https://doi.org/10.3390/clinpract13060130
5. Attal L, Shvartz E, Nakhoul N, Bahir D. Chat GPT 4o vs residents: French language evaluation in ophthalmology. *AJO Int*. 2025;2:100104. https://doi.org/10.1016/j.ajoint.2025.100104
6. Similarweb. *Top AI chatbots and tools websites ranking* [Internet]. Similarweb; n.d. [cited 2026 Mar 12]. Available from: https://www.similarweb.com/top-websites/ai-chatbots-and-tools/
7. Dean J, Barroso LA. The tail at scale. *Commun ACM*. 2013;56(2):74–80. https://doi.org/10.1145/2408776.2408794
8. 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. *Epidemiology*. 2007;18(6):800–804. https://doi.org/10.1097/EDE.0b013e3181577654
9. Kasagga A, Sapkota A, Changaramkumarath G, Abucha JM, Wollel MM, Somannagari N, et al. Performance of ChatGPT and large language models on medical licensing exams worldwide: a systematic review and network meta-analysis with meta-regression. *Cureus*. 2025;17(1):e94300. https://doi.org/10.7759/cureus.94300
10. Al-Thani SN, 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;18:146. https://doi.org/10.1186/s12245-025-00949-6

[Downloaded from edujournal.zums.ac.ir on 2026-11-29]

[DOI: 10.66224/edcj.19.2.4]