PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
UNIVERSITY OF AMAR TELIDJI LAGHOUAT
FACULTY OF MEDECINE – DEPARTEMENT OF MEDECINE



Graduation Thesis

Submitted in partial fulfillment of the requirements of the degree of the doctor of General Medicine

From Academic Theory to Clinical Reality: A Dual Evaluation of Artificial Intelligence in Medical Education and Emergency Case Management.

Presented by:

BOUFATAH Yacine Taha

SADEKI Fodhil

Jury members :

Pr BENAZZOUZ REDOUENE SID AHMED	Superviser
Pr BENYAGOUB MASSINISSA	President
Dr ZAABTA ABDELKRIM	Examiner

Academic Year: 2025 / 2026

Acknowledgements

At the completion of this doctoral dissertation, we would like to express our sincere gratitude to all those who contributed to its accomplishment.

First and foremost, we thank Almighty God for granting us the health, patience, and perseverance necessary to successfully complete this work.

We would like to express our deepest appreciation to our supervisor, **Professor Benazzouz Redouene Sid Ahmed**, for his guidance, availability, invaluable advice, and continuous support throughout this research.

We also extend our sincere thanks to the jury members for the honor of accepting to evaluate this dissertation.

Our gratitude also goes to all the professors and staff of the department for their contributions to our academic and professional development.

We would like to express our special appreciation to our families for their moral support, constant encouragement, patience, and sacrifices throughout this journey.

Finally, we warmly thank all our friends and colleagues who supported and encouraged us during the completion of this work.

To each and every one of them, we express our profound gratitude and appreciation.

Boufatah Yacine Taha

Sadeki Fodhil

Table of contents

General Introduction	**4**
Phase I – Artificial Intelligence in Medical Education	**8**
Published article / Abstract	9
Introduction	10
Materials and Methods	11
Results	12
Discussion	14
Conclusion	15
Ethical considerations and article declarations	15
References	16
Phase II – Artificial Intelligence in Emergency Case Management	**17**
Abstract	18
Introduction	20
Methods	22
Results	27
Discussion	29
Conclusion	33
References	34
Tables	36
Supplementary Material	42

General introduction

Recent advances in Large Language Models (LLMs) have opened new possibilities for supporting medical training and clinical decision support systems (CDSS), 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.

Concurrently, in high-acuity, time-sensitive environments like the emergency department (ED), where rapid and accurate clinical decision-making directly dictates patient outcomes, the rapid evolution of AI has introduced powerful new tools. Frontline care in these settings is predominantly managed by general practitioners (GPs), who must synthesize complex, undifferentiated patient presentations under significant cognitive load, time pressure, and resource constraints (1). Alongside general-purpose LLMs, specialized, medically-tailored AI platforms like Glass Health have emerged with the potential to serve as CDSS in acute care(2–4).

Among these digital developments, OpenAI's GPT-5 Pro has been introduced as a next-generation model with improved reasoning and stronger health-related benchmark performance than earlier OpenAI models. Similarly, Google DeepMind's Gemini 2.5 Pro is presented as an advanced reasoning model with strong long-context and multimodal capabilities relevant to medical question-answering tasks. Meanwhile, DeepSeek V3 offers a competitive open-weight alternative designed for efficient reasoning and multilingual comprehension, which is particularly relevant for resource-sensitive educational environments.

Despite these promising developments, independent evaluations show notable limitations, and current literature evaluating AI across both medical education and emergency medicine remains heavily fragmented. A critical bottleneck in current validation literature is that most benchmark datasets used to evaluate LLMs consist of publicly accessible or widely circulated questions, raising concerns about possible overlap with model training corpora.

Furthermore, much of the medical licensing literature has been dominated by earlier LLM versions and remains anchored in English-centric benchmark ecosystems, with relatively few evaluations of non-English examinations(5). 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(6–9). 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, leaving a significant research gap regarding the reliability of current-generation LLMs in underrepresented linguistic environments.

Beyond these regional and linguistic limitations, a critical gap exists because the majority of existing clinical studies evaluate AI models only within isolated, narrow dimensions of the clinical workflow. Several investigations restrict their focus entirely to emergency triage based purely on brief text inputs, or look exclusively at diagnostic accuracy and the generation of differential diagnoses (10–12).

However, in actual clinical practice, establishing a diagnosis represents only the initial phase of care. The medical literature routinely overlooks downstream therapeutic management, which involves selecting specific drug regimens, calculating exact dosages, screening for contraindications, and determining immediate life-saving interventions. Evaluating diagnostic reasoning in isolation introduces a profound safety risk: an AI model may generate a flawless differential diagnosis list while simultaneously proposing a catastrophic or unsafe treatment plan. In emergency medicine, recognizing clinical red flags and executing a safe, definitive management strategy are as critical as identifying the underlying pathology. Furthermore, existing medical AI benchmarks rely heavily on highly structured, pre-curated Western academic datasets that fail to reflect the raw, unstructured clinical documentation and unique epidemiological realities encountered in regional hospital settings.

To address these combined limitations, this research presents a comprehensive, multi-pronged comparative evaluation framework designed to assess current-generation AI models across both authentic regional educational standards and the practical continuum of acute care. This investigation is operationalized through two distinct yet complementary evaluation streams.

The first stream addresses the linguistic and regional benchmark gap by using 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, allowing a contextually grounded evaluation in a francophone North African setting where educational digital infrastructure 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.

Through this first arm, we conducted a cross-model comparative evaluation 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. We examined module and study cycle 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, whereas response format was analyzed because single-best-answer and combination-type items impose different levels of decision complexity. In contrast, temporal factors (weekday and time of day) were pre-specified as exploratory operational covariates and hypothesis-generating robustness checks to assess operational stability under varying platform usage loads. The primary aim of this first stream 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, while the secondary aim is to assess the influence of these contextual and structural factors.

The second phase shifts the evaluation directly to the clinical frontline to address the explicit lack of comparative data against practicing human physicians across a full acute care encounter. This component presents a comparative performance evaluation of general-purpose LLMs, a specialized clinical AI platform, and practicing general practitioners in solving real-world, uncurated emergency cases. By utilizing an independent, specialist-blinded evaluation framework that spans clinical reasoning, patient safety, and therapeutic management, this arm aims to rigorously assess the clinical readiness and safety profiles of these computational models in acute care medicine, establishing their capacity to navigate both initial diagnostics and safe downstream clinical execution.

1. Croskerry P. Clinical decision making. Ann Emerg Med. 2000;36(3):232–6. doi:10.1067/mem.2000.109403

2. Singhal K, Azizi S, Tu T, others. Large language models encode clinical knowledge. Nature. 2023;620(7973):172–80. doi:10.1038/s41586-023-06291-2

3. Rao A, Pang M, Kim J, others. Assessing the utility of ChatGPT throughout the entire clinical workflow: Development and usability study. J Med Internet Res. 2023;25:e48659. doi:10.2196/48659

4. Glass Health. AI clinical decision support: features. 2026.

5. Liu M, Okuhara T, Chang X, others. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res. 2024;26:e60807. doi:10.2196/60807

6. Fang C, Wu Y, Fu W, others. How does ChatGPT-4 perform on non-English national medical licensing examination? An evaluation in Chinese language. PLOS Digit Health. 2023;2(12):e0000397. doi:10.1371/journal.pdig.0000397

7. Jang D, Yun TR, Lee CY, others. GPT-4 can pass the Korean National Licensing Examination for Korean Medicine Doctors. PLOS Digit Health. 2023;2(12):e0000416. doi:10.1371/journal.pdig.0000416

8. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, others. 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):1460–87. doi:10.3390/clinpract13060130

9. Attal L, Shvartz E, Nakhoul N, Bahir D. Chat GPT 4o vs residents: French language evaluation in ophthalmology. AJO Int. 2025;2:100104. doi:10.1016/j.ajoint.2025.100104

10. Gottlieb M, Cydulka RK. Accuracy of a large language model on complex emergency medicine vignettes. Ann Emerg Med. 2023;82(6):708–10. doi:10.1016/j.annemergmed.2023.05.029

11. Kanjee Z, Crowe B, Harding M. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA. 2023;330(1):78–80. doi:10.1001/jama.2023.8288

12. Ramaswamy A, Tyagi A, Hugo H, others. ChatGPT Health performance in a structured test of triage recommendations. Nat Med. 2026;32(5):1671–5. doi:10.1038/s41591-026-04297-7

PHASE I

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

Volume 19, Issue 2
December 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.

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

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

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

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 $\alpha = 0.05$). Holm adjustment applies to pairwise comparisons only.
Abbreviations: MCQ, multiple-choice question; CI, confidence interval; OR, odds ratio; df, degrees of freedom.

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.

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,

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

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

PHASE II

Title

Comparative Clinical Performance of GPT-5.2 Pro, Gemini 3 Pro, Glass Health, and General Practitioners in Real-World Emergency Medicine: A Specialist-Blinded Case-Based Evaluation

Abstract

Background: In many resource-constrained emergency departments, general practitioners provide the first clinical assessment and initial management before specialist input is available. Whether advanced AI systems can produce clinically useful and safe written management suggestions from concise emergency case summaries remains uncertain. This study compared GPT-5.2 Pro, Gemini 3 Pro, Glass Health, and general practitioners in the initial management of emergency medicine cases from a regional hospital.

Methods: We conducted an observational, case-based comparative study in the emergency department of a 240-bed hospital in Algeria. A total of 147 adult emergency cases were retrospectively selected from routine records and reconstructed as standardized French vignettes limited to 100 words. Each vignette was submitted once to GPT-5.2 Pro, Gemini 3 Pro, Glass Health, and one of twelve practicing general practitioners. Responses were limited to 250 words, anonymized, and evaluated by blinded senior specialists using three predefined 0–5 clinical adequacy domains: clinical reasoning, therapeutic management, and clinical safety. The overall score was calculated as the mean of the three domain scores. Paired comparisons used Friedman tests, Kendall's W, and Holm-adjusted post-hoc Wilcoxon signed-rank tests.

Results: Overall clinical performance differed significantly across groups. Median overall scores were 4.33 [3.50–4.67] for GPT-5.2 Pro, 4.00 [3.67–4.67] for Gemini 3 Pro, 4.00 [3.67–4.67] for Glass Health, and 2.67 [2.00–3.33] for general practitioners ($p < 0.001$; Kendall's $W = 0.38$). Each AI system scored higher than general practitioners for the overall score and for all three domains, with Holm-adjusted p values <0.001. No significant pairwise differences were observed between the three AI systems. The AI advantage persisted across predefined difficulty levels. GPT-5.2 Pro showed higher overall scores in medical than surgical cases, whereas no significant medical–surgical difference was observed for Gemini 3 Pro, Glass Health, or general practitioners. Prompt length in characters was not significantly associated with overall performance.

Conclusions: In this specialist-blinded case-based evaluation using emergency cases reconstructed from routine hospital records, GPT-5.2 Pro, Gemini 3 Pro, and Glass Health achieved higher specialist-rated clinical performance than general practitioners for written initial management plans. These findings support further prospective evaluation of supervised AI-assisted decision support in emergency medicine, with attention to local validation, clinical governance, and data-protection requirements.

Keywords: artificial intelligence; emergency medicine; large language models; clinical decision support; general practitioners; patient safety; resource-constrained settings.

Introduction

The emergency department (ED) is a high-acuity, time-sensitive environment in which clinical decisions must often be made rapidly from incomplete, evolving, and sometimes poorly structured information. In many healthcare systems, especially outside highly resourced academic centers, the initial assessment and management of emergency patients are frequently performed by general practitioners (GPs), who must identify potentially life-threatening conditions, initiate treatment, and determine the need for specialist referral under substantial cognitive load and resource constraints (1).

In parallel, recent advances in artificial intelligence (AI), particularly large language models (LLMs), have introduced new possibilities for clinical decision support. General-purpose LLM-based tools such as GPT-5.2 Pro accessed through ChatGPT (OpenAI, San Francisco, CA, USA) and Gemini 3 Pro accessed through Google Gemini (Google LLC, Mountain View, CA, USA) can synthesize clinical information, generate differential diagnoses, and propose management plans, whereas specialized medical AI platforms such as Glass Health (Glass Health Inc., San Francisco, CA, USA) are specifically designed to support diagnostic and therapeutic reasoning in clinical practice (2,3). In this study, GPT-5.2 Pro and Gemini 3 Pro were selected to represent widely used general-purpose LLM-based tools available through major consumer AI applications. This choice was consistent with market intelligence data identifying ChatGPT and Google Gemini among the leading AI applications by global downloads during the study period (4). Glass Health was included as a specialized medical AI tool designed to support differential diagnosis and clinical plan generation (5).

Early evaluations of LLMs in medicine have demonstrated high performance on licensing examinations, medical knowledge tests, and structured clinical reasoning tasks (6,7). In emergency medicine, several studies have also suggested that LLMs may perform well in selected diagnostic or triage-related tasks (8–12). However, much of the existing evidence remains methodologically heterogeneous. Many studies use academic vignettes, examination-style questions, optimized prompts, or highly curated datasets that do not fully reproduce the clinical documentation, uncertainty, time pressure, and case-mix encountered in routine emergency care.

A further limitation is that many evaluations focus primarily on diagnostic accuracy or differential diagnosis generation. Although diagnosis is central to emergency medicine, it is not sufficient to define safe clinical performance. Initial emergency management also requires recognition of red flags, prioritization of urgent interventions, selection of appropriate therapeutic strategies, avoidance of contraindicated treatments, and anticipation of clinical deterioration. An AI response may appear diagnostically plausible while remaining incomplete or unsafe from a

management perspective. For this reason, evaluations of AI in emergency medicine should assess not only clinical reasoning, but also therapeutic management and clinical safety.

Prompt design is another important methodological issue in AI evaluation. Long, highly structured, or iteratively optimized prompts may improve model performance but reduce comparability with real clinical conditions, where physicians often work from concise initial notes. Similarly, unconstrained AI responses may appear more complete simply because they are longer. Prompt engineering has been shown to influence measured LLM performance in medical assessment settings, supporting the need to treat prompt design as a methodological variable rather than a neutral technical detail (13). Standardizing both prompt length and response length is therefore important when comparing AI systems with human clinicians. In this study, the clinical prompt was deliberately limited to 100 words, and responses were limited to 250 words for both AI systems and general practitioners.

Evidence from hospitals in resource-constrained settings remains particularly limited, despite the fact that these environments may be among those where supervised AI decision support could have the greatest practical relevance. Most AI benchmarks are generated in high-income or academic settings and may not reflect local documentation practices, epidemiology, specialist availability, or frontline workforce organization in regional hospitals. Locally generated evaluations are therefore needed to determine whether advanced AI systems can provide clinically meaningful support under realistic conditions.

Accordingly, the primary objective of this study was to compare the clinical performance of GPT-5.2 Pro, Gemini 3 Pro, Glass Health, and general practitioners in the initial management of real-world emergency medicine cases from a regional hospital, using specialist-blinded ratings across three predefined domains: clinical reasoning, therapeutic management, and clinical safety. The secondary objectives were to explore whether performance varied according to predefined case difficulty and medical versus surgical specialty category. An additional exploratory methodological analysis assessed the association between prompt length in characters and overall clinical performance.

Methods

Study design and setting

This observational, specialist-blinded, case-based comparative study evaluated the clinical performance of three AI systems and practicing general practitioners in the initial management of real-world emergency medicine cases. The study was conducted in the Emergency Department of the 240-bed Hospital of Laghouat, Algeria. The manuscript was prepared in accordance with the STROBE recommendations for observational studies where applicable (14).

Cases were retrospectively selected from routine emergency department records generated between November 2025 and March 2026. Sampling was performed using predefined randomly selected emergency department dates during the study period. For each selected date, available adult emergency department files were screened for completeness and eligibility. Clinical management had already been completed before case abstraction, and no aspect of patient care was influenced by the study. Case abstraction was performed from routinely generated clinical files after completion of the initial emergency department documentation.

Sample size

A sample size target was defined before the final analysis for the primary paired comparison across the four evaluated groups. Because no robust prior estimate was available for this specific setting, the calculation was based on the Friedman test, using Kendall's W as the target effect size. Assuming four paired groups, a two-sided alpha level of 0.05, 80% power, and a small target effect size of Kendall's W = 0.03, the minimum required number of complete cases was approximately 122. Allowing for incomplete or unusable records, the target sample size was set at approximately 140 to 145 cases.

Case selection

During each selected sampling date, adult emergency department files available in the records were screened for eligibility. Files were first assessed for completeness. Eligible complete files constituted the sampling frame for that date. When several eligible complete records were available, cases were selected at random, without stratification by diagnosis, specialty, sex, age, severity, or expected difficulty.

Inclusion criteria were adult patients aged 18 years or older presenting for acute care and availability of a sufficiently complete initial emergency department note, including the presenting complaint, relevant history, baseline vital signs, physical examination findings, and available initial investigations.

Exclusion criteria were pediatric encounters, obstetrical or gynecological emergencies, incomplete or illegible files, missing key baseline clinical information, and records subject to administrative or regulatory restrictions.

The final analytic sample comprised 147 adult medical and surgical emergency cases. The distribution of cases reflected real-world emergency department activity rather than an artificially balanced specialty distribution.

Case preparation and prompting

Investigators reconstructed each initial emergency encounter into a standardized text vignette. Subjective impressions, preliminary diagnostic labels, and therapeutic decisions made by the original treating physician were removed. The vignette included age and sex, relevant medical history, chronic treatments, presenting complaint, physical examination findings, available investigation results, and a final request for clinical interpretation and recommended initial management.

The clinical prompt was limited to 100 words and was written in French to reflect local clinical documentation and professional communication. The response instruction was standardized across AI systems and general practitioners, with a maximum response length of 250 words. The prompt followed a fixed structure including age and sex, relevant medical history, chronic treatments, presenting complaint, physical examination findings, available investigation results, and a final request for clinical interpretation and recommended initial management. The original French prompt template and an English translation are provided in Supplementary Material 1.

Prompt length was recorded in words and characters; character count was used for the exploratory correlation analysis.

AI systems and human comparator

The three AI systems evaluated were GPT-5.2 Pro accessed through ChatGPT (OpenAI, San Francisco, CA, USA), Gemini 3 Pro accessed through Google Gemini (Google LLC, Mountain View, CA, USA), and Glass Health (Glass Health Inc., San Francisco, CA, USA). Each case was submitted once to each AI system through a standard web browser interface on a secure local workstation. A zero-shot approach was used. No follow-up questions, regenerated answers, recursive corrections, or iterative prompt engineering were allowed. Default interface settings were used to reflect typical end-user conditions. For Glass Health, the standardized case prompt was pasted as a single free-text block into the chat interface. Optional contextual fields, specialty settings, or additional configuration options were not used, in order to preserve comparability with the prompts submitted to GPT-5.2 Pro and Gemini 3 Pro.

The human comparison group consisted of twelve practicing general practitioners from the Laghouat region who agreed to participate in the study, with more than one year of professional experience. The standardized clinical vignettes distributed to general practitioners were identical in content and format to those submitted to the AI systems. Each case contributed one general practitioner response. Cases were randomly allocated among participating general practitioners, while accounting for physician availability during the study period. General practitioners were instructed to provide their clinical interpretation and recommended initial management in 250 words or fewer, without access to AI-generated responses.

Blinded specialist assessment

All responses generated by GPT-5.2 Pro, Gemini 3 Pro, Glass Health, and general practitioners were anonymized before evaluation. Platform-specific identifiers, formatting, branding, and metadata were removed to produce a uniform response format. Responses were labelled using neutral codes, and the order of response sources was counterbalanced across cases according to a predefined permutation scheme to reduce order-related bias.

Senior specialists from the relevant clinical specialty evaluated cases within their own domain of expertise. Specialists were selected according to specialty relevance, seniority, and availability. Each case was assigned to the specialist corresponding to its predefined specialty domain. For each case, the anonymized responses from the four evaluated groups were assessed by the same specialist in a blinded format. Evaluators were unaware of whether a given response was produced by an AI system or by a general practitioner.

Case difficulty was classified before AI interrogation and before specialist evaluation using a 3-level clinical difficulty scale. Level 1 cases corresponded to straightforward presentations of a single, well-defined acute condition. Level 2 cases included ambiguous clinical signs or mild-to-moderate comorbidities. Level 3 cases included complex, undifferentiated, multisystem, unstable, or atypical presentations with potential red flags. Difficulty classification was performed by one of two coordinating investigators. Before case classification, the two investigators conducted a synchronization exercise using example cases to harmonize interpretation of the three-level scale.

Each case was assigned to a specialty domain by the coordinating investigators according to the main acute clinical problem and the specialty most appropriate for expert evaluation. For secondary analyses, specialty domains were collapsed into medical and surgical categories. Medical specialties initially eligible included pulmonology, cardiology, gastroenterology, hematology, infectious diseases, internal medicine, nephrology, neurology, critical care, endocrinology, and psychiatry. Surgical specialties included thoracic surgery, visceral surgery, neurosurgery, otorhinolaryngology, ophthalmology, orthopedics, and urology.

Outcomes

The primary outcome was the overall clinical performance score, calculated as the arithmetic mean of the three predefined specialist-blinded domains: clinical reasoning, therapeutic management, and clinical safety. Clinical reasoning assessed the relevance and coherence of the diagnostic interpretation, prioritization of differential diagnoses, and recognition of key clinical features. Therapeutic management assessed the appropriateness of the proposed initial management plan, including urgent interventions, treatment priorities, referral decisions, and avoidance of inappropriate or contraindicated therapies. Clinical safety assessed whether the response avoided dangerous omissions, unsafe recommendations, or failure to recognize potentially life-threatening features.

Each domain was scored on a 6-point clinical adequacy scale ranging from 0 to 5. The scoring rubric was anchored to predefined clinical criteria rather than general satisfaction. A score of 0 indicated an absent, irrelevant, or clinically dangerous response; 1 indicated major errors or omissions with substantial potential for harm; 2 indicated important limitations requiring major correction; 3 indicated a partially acceptable response with relevant omissions or imprecision; 4 indicated an appropriate response with only minor limitations; and 5 indicated a complete, appropriate, and safe response for the initial emergency management context.

Clinical reasoning, therapeutic management, and clinical safety were also analysed separately as predefined domain-specific outcomes. Secondary analyses assessed whether overall performance varied according to predefined case difficulty and medical versus surgical specialty category. The association between prompt length in characters and overall performance was analysed as an exploratory methodological analysis.

Statistical analysis

Statistical analyses were performed using R version 4.5.1 (R Foundation for Statistical Computing, Vienna, Austria). The case was considered the unit of analysis. Because each emergency medicine case was evaluated by GPT-5.2 Pro, Gemini 3 Pro, Glass Health, and general practitioners, comparisons between groups were treated as paired repeated-measures analyses.

Case characteristics were described using median [interquartile range] for continuous variables and counts with percentages for categorical variables. Prompt length was analysed using both word count and character count for descriptive purposes, and using character count for the exploratory correlation analysis. Scores were summarized as median [interquartile range], given the bounded 0–5 rating scale.

Global comparisons across the four groups were performed using the Friedman test. Kendall's W was reported as the effect size. When the global test was significant, pairwise post-hoc comparisons were performed using paired Wilcoxon signed-rank tests with Holm correction for

multiple comparisons. Comparisons between each AI system and general practitioners were reported as the main pairwise contrasts.

For secondary analyses, between-group comparisons within each difficulty level were performed using the Friedman test. Variation in overall performance across difficulty levels within each evaluated group was explored using the Kruskal–Wallis test. Differences between medical and surgical cases within each group were assessed using the Wilcoxon rank-sum test. The association between prompt length in characters and overall performance was assessed using Spearman correlation coefficients. Secondary analyses were considered exploratory.

After data verification and completion, all clinical performance ratings required for the primary and domain-specific analyses were available for the 147 included cases. No imputation was performed. All tests were two-sided, and p values below 0.05 were considered statistically significant.

Ethics

The study protocol was reviewed and approved by the Ethics Committee of Ben Ali Deghin Hospital of Laghouat, Algeria (approval No. 03/25, 7 November 2025). The study was conducted in accordance with the principles of the Declaration of Helsinki. Because this was a retrospective study based on routinely generated clinical records, with no intervention and no impact on patient management, the requirement for individual informed consent was waived by the ethics committee. All data were de-identified before AI interrogation, GP response collection, and specialist evaluation. No directly identifying patient information was submitted to any AI system. Participation of general practitioners and specialist evaluators was voluntary.

Results

Case characteristics

The final analytic dataset included 147 real-world emergency medicine cases. The median patient age was 61.00 [45.50–74.00] years, and 100 cases involved male patients (68.03%). One hundred cases were classified as medical specialty cases and 47 as surgical specialty cases. Case difficulty was classified as level 1 in 73 cases (49.66%), level 2 in 62 cases (42.18%), and level 3 in 12 cases (8.16%). The median prompt length was 58.00 [55.00–63.00] words and 396.00 [373.00–423.00] characters.

Medical specialty domains included pulmonology, cardiology, gastroenterology, hematology, infectious diseases, internal medicine, nephrology, neurology, and critical care. Surgical specialty domains included thoracic surgery, visceral surgery, neurosurgery, otorhinolaryngology, ophthalmology, orthopedics, and urology.

Medical cases involved older patients than surgical cases and included a higher proportion of level 2 and level 3 cases. Sex distribution and prompt length did not differ significantly between medical and surgical cases (Table 1).

Clinical performance across groups

Overall clinical performance differed significantly across the four evaluated groups. GPT-5.2 Pro, Gemini 3 Pro, and Glass Health had close median overall scores of 4.33 [3.50–4.67], 4.00 [3.67–4.67], and 4.00 [3.67–4.67], respectively. General practitioners had a lower median overall score of 2.67 [2.00–3.33]. The global paired comparison was statistically significant, with a Kendall's W of 0.38 (Table 2).

This pattern was consistent across the three predefined domains. In clinical reasoning, median scores were 4.00 [4.00–5.00] for GPT-5.2 Pro, Gemini 3 Pro, and Glass Health, compared with 3.00 [2.00–4.00] for general practitioners. In therapeutic management, median scores were 4.00 [3.00–5.00], 4.00 [3.00–4.00], 4.00 [3.00–5.00], and 2.00 [2.00–3.00], respectively. In clinical safety, median scores were 4.00 [4.00–5.00], 4.00 [3.00–4.50], 4.00 [3.00–5.00], and 3.00 [2.00–3.50], respectively. The global paired comparison was significant in all three domains, with Kendall's W ranging from 0.35 to 0.40 (Table 2).

Pairwise comparisons versus general practitioners

Each AI system scored higher than general practitioners for the overall performance score and for each predefined domain (Table 3). For the overall score, the median paired difference was 1.33 [0.67–2.33] points for GPT-5.2 Pro, 1.67 [0.67–2.00] points for Gemini 3 Pro, and 1.33 [0.67–2.33] points for Glass Health. All Holm-adjusted p values were <0.001.

The largest domain-specific differences were observed in clinical reasoning. Median paired differences versus general practitioners were 2.00 [1.00–3.00] points for GPT-5.2 Pro, 1.00 [1.00–2.00] point for Gemini 3 Pro, and 1.00 [1.00–3.00] point for Glass Health. Pairwise comparisons between the three AI systems did not show statistically significant differences for the overall score or for any of the three domains (Table S1).

Secondary analyses

When the analysis was stratified by predefined case difficulty, between-group differences in the overall score remained significant in level 1, level 2, and level 3 cases. Kendall's W was 0.42 for level 1 cases, 0.36 for level 2 cases, and 0.45 for level 3 cases (Table S2). Separate within-group analyses did not show a significant variation in overall performance across difficulty levels for any evaluated group.

In analyses according to medical versus surgical specialty category, GPT-5.2 Pro had a higher median overall score in medical cases than in surgical cases. No significant medical–surgical difference in overall performance was observed for Gemini 3 Pro, Glass Health, or general practitioners after correction for multiple testing. Prompt length in characters was not significantly associated with overall performance for any evaluated group (Table S3).

Discussion

In this specialist-blinded, case-based evaluation of emergency medicine cases reconstructed from routine hospital records, GPT-5.2 Pro, Gemini 3 Pro, and Glass Health achieved significantly higher clinical performance scores than general practitioners for written initial management plans. This difference was observed for the overall score and across the three predefined domains of clinical reasoning, therapeutic management, and clinical safety. No significant difference was observed between the three AI systems themselves.

The main contribution of this study lies in the way the comparison was framed. Rather than evaluating AI systems on examination questions alone or on isolated diagnostic tasks, we assessed their ability to produce concise initial management suggestions from emergency case summaries derived from routine hospital documentation. The same case information, prompt constraints, and response limits were applied to AI systems and general practitioners. This design addresses a practical question: whether these systems can generate clinically coherent, safe, and usable written responses when the available information resembles early emergency documentation rather than a fully developed academic case.

This distinction matters because emergency medicine is not limited to diagnostic formulation. The initial management of acute patients requires recognition of red flags, prioritization of urgent actions, avoidance of harmful omissions, and anticipation of clinical deterioration. A response can be diagnostically plausible but still clinically incomplete or unsafe. By assessing clinical reasoning, therapeutic management, and clinical safety separately, the present study attempted to capture a broader and more clinically relevant definition of performance. The superiority of the AI systems across all three domains suggests that their advantage was not restricted to naming likely diagnoses, but also extended to the structure and safety of proposed initial management.

Our findings are consistent with previous studies showing strong performance of large language models in medical reasoning tasks, including medical examinations and complex diagnostic challenges (6,7,10). They also align with emergency medicine studies reporting good performance of GPT-based systems in diagnostic or triage-related tasks (9,11,12). However, many previous evaluations remain focused on diagnostic accuracy, triage classification, or differential diagnosis generation. The present study adds to this literature by emphasizing management and safety, two dimensions that are particularly important in emergency care and in first-line settings where specialist input may not be immediately available.

The results should nevertheless be interpreted within the limits of the task evaluated. The AI systems were not tested as autonomous clinicians. They did not examine patients, communicate with families, reassess evolving symptoms, perform procedures, coordinate care, or assume responsibility for outcomes. They produced written responses to standardized vignettes.

Therefore, the comparison with general practitioners should not be read as evidence that AI can replace physicians in emergency departments. It is better understood as evidence that, under controlled written conditions, advanced AI systems can generate responses that senior specialists judge to be more complete, structured, and safe than those produced by the local human comparator.

The comparison with general practitioners remains clinically meaningful. In many emergency departments, particularly outside highly resourced academic centers, general practitioners are the first clinicians to evaluate and manage acute patients before referral or specialist review. Their performance therefore represents a pragmatic frontline benchmark rather than an artificial comparator. The findings suggest that AI tools may help reduce gaps in early reasoning and management planning when used as supervised support. The most realistic clinical implication is not replacement, but augmentation: helping clinicians structure differentials, identify red flags, check for omissions, and consider initial management priorities.

The absence of significant differences between GPT-5.2 Pro, Gemini 3 Pro, and Glass Health is also informative. Glass Health was included as a specialized medical AI platform, whereas GPT-5.2 Pro and Gemini 3 Pro were general-purpose AI systems. Their similar performance suggests that, for short written emergency vignettes, advanced general-purpose models may already approach the performance of a specialized medical AI tool. This finding should not be interpreted as proof of equivalence across clinical tasks. It does, however, support the need for direct, independent evaluation rather than assuming that medical specialization alone guarantees superior performance.

Prompt and response constraints are central to the interpretation of these findings. LLM performance can be influenced by prompt structure, wording, information density, and iterative refinement (11,13). In this study, prompts were limited to 100 words and responses to 250 words, and no follow-up questions or regenerated responses were allowed. This was intended to reproduce a constrained first-line decision-support situation rather than an optimized prompt-engineering exercise. The absence of a significant association between prompt length in characters and overall performance should therefore be interpreted only within this standardized framework. It does not mean that prompt design is unimportant; rather, it suggests that small variations in length did not substantially affect performance once prompts were deliberately kept short and structured.

The setting of this study is not only a geographical detail. A regional hospital determines how emergency information is recorded, who provides first-line care, how quickly specialists are available, and what decisions must be made before referral or escalation. Results obtained in academic centers, English-language datasets, or examination settings cannot automatically be transferred to such environments. Conversely, the present results should not be generalized

uncritically to all emergency departments. They show that, under the constraints of this study, AI systems performed well in a specific hospital context where general practitioners provide first-line emergency assessment. This supports the broader argument that AI performance should be evaluated locally, because it depends not only on the model, but also on the clinical task, language, documentation style, available resources, and expected standard of care (15,16).

In this context, the potential value of AI is practical rather than abstract. In resource-constrained emergency departments, where specialist access may be delayed and workload can be high, a supervised AI tool could serve as a structured second reading of the initial case summary. It may help clinicians verify whether important diagnoses, urgent actions, contraindications, or safety issues have been missed. However, this potential benefit cannot be inferred from global benchmarks alone. It requires local validation, clear clinical boundaries, and prospective evaluation in real workflows.

The deployment of AI in emergency care also raises issues beyond performance. Clinical data governance, institutional responsibility, and legal compliance are essential, especially when external cloud-based tools are accessible through public web interfaces. In Algeria, any future use of AI systems involving patient-level clinical information should remain aligned with national personal data protection requirements and institutional authorization procedures (17). The present study used de-identified prompts and did not submit directly identifying patient information to any AI system. This approach was appropriate for research benchmarking, but routine clinical deployment would require more formal safeguards, including data minimization, traceability, user responsibility, storage conditions, vendor access rules, and institutional oversight. In the long term, locally governed or institutionally approved deployment models may be preferable to informal individual use of public interfaces, particularly when patient data, medicolegal responsibility, and national data sovereignty are involved (17,18).

Several methodological choices strengthen the internal validity of this study. Cases were reconstructed from routine emergency records rather than written as examination questions, and they were selected from complete eligible files without stratification by diagnosis, specialty, sex, age, severity, or expected difficulty. Each case was evaluated across all four groups, allowing paired within-case comparison. AI and general practitioner responses were constrained to the same word limit. Responses were anonymized, stripped of platform identifiers, and assessed by senior specialists who were blinded to response source. In addition, the same specialist assessed the four responses for a given case, preserving internal comparability.

These strengths should be balanced against important limitations. The study was conducted in a single hospital, and the use of reconstructed vignettes also means that the study evaluated written decision-making from standardized summaries, not full bedside clinical practice. Real emergency care includes examination, repeated assessment, communication, procedural

judgment, teamwork, and responsibility for patient outcomes, none of which can be fully captured in a text-based comparison.

The evaluation process also has inherent limitations. Each response was rated by one senior specialist rather than several independent raters, so inter-rater variability could not be estimated. The use of a numerical 0–5 clinical adequacy scale allowed standardized scoring across many responses, facilitated paired statistical comparisons, and enabled separate analysis of reasoning, management, and safety. The use of predefined anchors and evaluator synchronization was intended to improve consistency. However, numerical scoring remains partly subjective and may not capture all dimensions of clinical quality, including the logic of prioritization, contextual appropriateness, strength of justification, or clinical consequences of specific omissions (19). Future evaluations should consider combining numerical scoring with qualitative expert adjudication, structured error classification, consensus panels, multiple independent raters, and formal assessment of inter-rater reliability (20).

Blinding also has practical limits. Although platform identifiers, formatting, and metadata were removed, some evaluators may have inferred the likely source of a response from its language, structure, or level of detail. This potential partial unblinding could have influenced ratings in either direction. Similarly, AI performance was judged against specialist opinion rather than against final patient outcomes. This was appropriate for a controlled comparative benchmark, but it cannot determine whether AI assistance would improve real clinical outcomes, reduce adverse events, or change referral decisions in practice.

The secondary analyses should also be interpreted cautiously. The AI advantage over general practitioners was preserved across predefined difficulty levels, suggesting that the main result was not driven only by straightforward cases. However, level 3 cases represented a small subgroup. The medical versus surgical analysis showed higher overall scores for GPT-5.2 Pro in medical than surgical cases, while no significant medical–surgical difference was observed for Gemini 3 Pro, Glass Health, or general practitioners after correction for multiple testing. This isolated finding is clinically plausible, because surgical emergency decision-making often depends heavily on physical examination nuance, imaging interpretation, procedural timing, operative pathways, and local resource availability. These elements may be less fully represented in a 100-word text prompt than medical diagnostic and therapeutic reasoning. Nevertheless, this result remains exploratory and should be tested in larger studies with richer case inputs and larger surgical subgroups.

The next step should not be limited to ranking AI models against clinicians in isolation. Future studies should test whether clinicians supported by AI perform better, faster, or more safely than clinicians working without AI. Multicenter prospective studies should evaluate decision quality, time to treatment, missed emergencies, medication safety, referral appropriateness, clinician

workload, and patient outcomes. They should also examine how AI suggestions are used, ignored, challenged, or over trusted in real clinical workflows. Such studies will be essential to determine whether the promising performance observed in written case-based evaluation can translate into safe and useful clinical support.

Conclusion

In this specialist-blinded evaluation of emergency medicine cases reconstructed from routine hospital records in a resource-constrained North African hospital, GPT-5.2 Pro, Gemini 3 Pro, and Glass Health outperformed general practitioners in overall clinical performance and across clinical reasoning, therapeutic management, and clinical safety. The three AI systems performed similarly to each other, and their advantage over general practitioners was preserved across predefined difficulty levels.

Beyond the comparative ranking of models, this findings support the potential of advanced AI systems as clinician-mediated decision-support tools for structured initial interpretation and management planning in emergency medicine. Before routine deployment, however, such tools require prospective clinical evaluation, local validation, robust assessment methods, clear governance, and compliance with institutional and national data-protection requirements.

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-5.5 model) was used only to improve grammar, spelling, and stylistic clarity of the manuscript. 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.

References

1. Croskerry P. Clinical decision making. Ann Emerg Med. 2000;36(3):232–6. doi:10.1067/mem.2000.109403

2. Singhal K, Azizi S, Tu T, others. Large language models encode clinical knowledge. Nature. 2023;620(7973):172–80. doi:10.1038/s41586-023-06291-2

3. Rao A, Pang M, Kim J, others. Assessing the utility of ChatGPT throughout the entire clinical workflow: Development and usability study. J Med Internet Res. 2023;25:e48659. doi:10.2196/48659

4. AppTweak. The most downloaded AI apps in 2025. 2026.

5. Glass Health. AI clinical decision support: features. 2026.

6. Kung TH, Cheatham M, Medenilla A, others. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198

7. Benazzouz RSA, Benyagoub M, Boufatah Y, Sadeki F, Benazzouz MS, Ould Setti M. 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 Educ Dev. 2026;19(2):4–11.

8. Berikol G, Kanbakan A, Ilhan B, Doğanay F. Mapping artificial intelligence models in emergency medicine: A scoping review on artificial intelligence performance in emergency care and education. Turk J Emerg Med. 2025;25(2):67–91. doi:10.4103/tjem.tjem_45_25

9. Gottlieb M, Cydulka RK. Accuracy of a large language model on complex emergency medicine vignettes. Ann Emerg Med. 2023;82(6):708–10. doi:10.1016/j.annemergmed.2023.05.029

10. Kanjee Z, Crowe B, Harding M. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA. 2023;330(1):78–80. doi:10.1001/jama.2023.8288

11. Hoppe JM, Auer MK, Strüven A, Massberg S, Stremmel C. ChatGPT with GPT-4 outperforms emergency department physicians in diagnostic accuracy: Retrospective analysis. J Med Internet Res. 2024;26:e56110. doi:10.2196/56110

12. Ramaswamy A, Tyagi A, Hugo H, others. ChatGPT Health performance in a structured test of triage recommendations. Nat Med. 2026;32(5):1671–5. doi:10.1038/s41591-026-04297-7

13. Hsieh MY, Wang TL, Su PH, Chou MC. Impact of prompt engineering on the performance of ChatGPT variants across different question types in medical student examinations: Cross-sectional study. JMIR Med Educ. 2025;11:e78320. doi:10.2196/78320

14. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology statement: Guidelines for reporting observational studies. Lancet. 2007;370(9596):1453–7. doi:10.1016/S0140-6736(07)61602-X

15. Ciecierski-Holmes T, Singh R, Axt M, Brenner S, Barteit S. Artificial intelligence for strengthening healthcare systems in low- and middle-income countries: A systematic scoping review. NPJ Digit Med. 2022;5:162. doi:10.1038/s41746-022-00700-y

16. Rutunda S, others. Large language models for frontline healthcare support in low-resource settings. Nat Health. 2026;1(2):191–7. doi:10.1038/s44360-025-00038-1

17. Algerian People's Democratic Republic. Law No. 18-07 of 10 June 2018 on the protection of individuals with regard to the processing of personal data. 2018.

18. World Health Organization. Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. Geneva: World Health Organization; 2025.

19. Govaerts MJB, Van de Wiel MWJ, Schuwirth LWT, Van der Vleuten CPM, Muijtjens AMM. Workplace-based assessment: Raters' performance theories and constructs. Adv Health Sci Educ. 2013;18(3):375–96. doi:10.1007/s10459-012-9376-x

20. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155–63. doi:10.1016/j.jcm.2016.02.012

Table 1. Case characteristics overall and according to medical versus surgical specialty category

Characteristic	**Overall, n = 147**	**Medical, n = 100**	**Surgical, n = 47**	**p value**
Age, years, median [IQR]	61.00 [45.50–74.00]	66.00 [54.75–76.00]	49.00 [34.00–58.50]	<0.001
Male sex, n (%)	100 (68.03)	69 (69.00)	31 (65.96)	0.858
Difficulty level, n (%)				0.002
Level 1	73 (49.66)	40 (40.00)	33 (70.21)	
Level 2	62 (42.18)	49 (49.00)	13 (27.66)	
Level 3	12 (8.16)	11 (11.00)	1 (2.13)	
Prompt length, words, median [IQR]	58.00 [55.00–63.00]	59.00 [55.00–63.00]	57.00 [54.00–63.50]	0.316
Prompt length, characters, median [IQR]	396.00 [373.00–423.00]	402.00 [380.75–423.00]	386.00 [372.00–429.00]	0.223

Footnote. Values are presented as median [IQR] or n (%). p values compare medical and surgical cases and are descriptive. Specialty categories were defined according to the main acute clinical problem and the specialty most appropriate for expert evaluation.

Table 2. Clinical performance by evaluated group

Domain	GPT-5.2 Pro	Gemini 3 Pro	Glass Health	General practitioners	p value	Kendall's W
Overall score	4.33 [3.50–4.67]	4.00 [3.67–4.67]	4.00 [3.67–4.67]	2.67 [2.00–3.33]	<0.001	0.38
Clinical reasoning	4.00 [4.00–5.00]	4.00 [4.00–5.00]	4.00 [4.00–5.00]	3.00 [2.00–4.00]	<0.001	0.40
Therapeutic management	4.00 [3.00–5.00]	4.00 [3.00–4.00]	4.00 [3.00–5.00]	2.00 [2.00–3.00]	<0.001	0.37
Clinical safety	4.00 [4.00–5.00]	4.00 [3.00–4.50]	4.00 [3.00–5.00]	3.00 [2.00–3.50]	<0.001	0.35

Footnote. Values are presented as median [IQR]. The overall score was calculated as the mean of the clinical reasoning, therapeutic management, and clinical safety domain scores. Global comparisons across the four groups were performed using the Friedman test. Kendall's W is reported as the effect size. Complete paired data were available for 147 cases for all outcomes.

Table 3. Pairwise comparisons between each AI system and general practitioners

Domain	Comparison	n pairs	Median paired difference [IQR]	Holm-adjusted p value
Overall score	GPT-5.2 Pro vs general practitioners	147	1.33 [0.67–2.33]	<0.001
Overall score	Gemini 3 Pro vs general practitioners	147	1.67 [0.67–2.00]	<0.001
Overall score	Glass Health vs general practitioners	147	1.33 [0.67–2.33]	<0.001
Clinical reasoning	GPT-5.2 Pro vs general practitioners	147	2.00 [1.00–3.00]	<0.001
Clinical reasoning	Gemini 3 Pro vs general practitioners	147	1.00 [1.00–2.00]	<0.001
Clinical reasoning	Glass Health vs general practitioners	147	1.00 [1.00–3.00]	<0.001
Therapeutic management	GPT-5.2 Pro vs general practitioners	147	1.00 [1.00–2.00]	<0.001
Therapeutic management	Gemini 3 Pro vs general practitioners	147	1.00 [0.00–2.00]	<0.001
Therapeutic management	Glass Health vs general practitioners	147	1.00 [1.00–2.00]	<0.001
Clinical safety	GPT-5.2 Pro vs general practitioners	147	1.00 [1.00–2.50]	<0.001
Clinical safety	Gemini 3 Pro vs general practitioners	147	1.00 [1.00–2.00]	<0.001
Clinical safety	Glass Health vs general practitioners	147	1.00 [0.00–2.00]	<0.001

Footnote. Paired differences are calculated as AI score minus general practitioner score. Pairwise comparisons were performed using paired Wilcoxon signed-rank tests. Holm correction was applied within each domain across all pairwise group comparisons.

Table S1. Pairwise comparisons between AI systems

Domain	Comparison	n pairs	Median paired difference [IQR]	Holm-adjusted p value
Overall score	GPT-5.2 Pro vs Gemini 3 Pro	147	0.00 [-0.33–0.67]	0.453
Overall score	GPT-5.2 Pro vs Glass Health	147	0.00 [-0.67–0.67]	0.955
Overall score	Gemini 3 Pro vs Glass Health	147	0.00 [-0.67–0.33]	0.955
Clinical reasoning	GPT-5.2 Pro vs Gemini 3 Pro	147	0.00 [0.00–1.00]	1.000
Clinical reasoning	GPT-5.2 Pro vs Glass Health	147	0.00 [-1.00–1.00]	1.000
Clinical reasoning	Gemini 3 Pro vs Glass Health	147	0.00 [-1.00–1.00]	1.000
Therapeutic management	GPT-5.2 Pro vs Gemini 3 Pro	147	0.00 [-0.50–1.00]	0.768
Therapeutic management	GPT-5.2 Pro vs Glass Health	147	0.00 [-1.00–1.00]	0.798
Therapeutic management	Gemini 3 Pro vs Glass Health	147	0.00 [-1.00–1.00]	0.687
Clinical safety	GPT-5.2 Pro vs Gemini 3 Pro	147	0.00 [-1.00–1.00]	0.438
Clinical safety	GPT-5.2 Pro vs Glass Health	147	0.00 [-1.00–1.00]	0.788
Clinical safety	Gemini 3 Pro vs Glass Health	147	0.00 [-1.00–0.00]	0.788

Footnote. Paired differences are calculated as the first AI system score minus the second AI system score. Pairwise comparisons were performed using paired Wilcoxon signed-rank tests. Holm correction was applied within each domain.

Table S2. Overall performance according to predefined case difficulty

Difficulty level	n cases	GPT-5.2 Pro	Gemini 3 Pro	Glass Health	General practitioners	p value	Kendall's W
Level 1	73	4.00 [3.33–4.67]	4.00 [3.67–4.67]	4.00 [4.00–5.00]	2.67 [2.00–3.33]	<0.001	0.42
Level 2	62	4.17 [3.75–4.67]	4.00 [3.33–4.33]	4.00 [3.67–4.33]	2.33 [1.67–3.33]	<0.001	0.36
Level 3	12	4.33 [3.92–4.75]	4.17 [4.00–4.50]	4.00 [3.25–4.42]	3.00 [2.58–3.75]	<0.001	0.45

Footnote. Values are presented as median [IQR]. Between-group comparisons within each difficulty level were performed using the Friedman test. Kendall's W is reported as the effect size. Separate within-group Kruskal–Wallis tests did not show significant variation in overall performance across difficulty levels for any evaluated group.

Table S3. Secondary analyses of overall performance according to medical versus surgical specialty category and prompt length

Group	Medical cases, median [IQR]	Surgical cases, median [IQR]	Medical minus surgical median difference	Adjusted p value	Spearman's rho with prompt length	Adjusted p value
GPT-5.2 Pro	4.33 [4.00–5.00]	4.00 [3.00–4.67]	0.33	0.037	0.10	0.928
Gemini 3 Pro	4.00 [3.67–4.67]	4.00 [3.50–4.33]	0.00	0.771	-0.11	0.536
Glass Health	4.00 [3.67–4.67]	4.00 [3.67–5.00]	0.00	1.000	-0.13	0.334
General practitioners	2.33 [1.67–3.33]	2.67 [2.00–3.58]	-0.33	0.368	-0.10	0.956

Footnote. Values are presented as median [IQR] unless otherwise specified. Medical versus surgical comparisons were performed using the Wilcoxon rank-sum test. Prompt length was analysed using the number of characters and assessed using Spearman correlation coefficients. Adjusted p values are Holm-adjusted p values from the exploratory secondary analyses. Only the overall score is shown.

Supplementary Material 1. Standardized prompt template and English translation

The standardized prompt was written and administered in French for both AI systems and general practitioners. The English version below is provided only for readability and was not used for data generation.

Original French prompt template

Âge et sexe : [Données]

Antécédents médicaux pertinents : [Données]

Traitements chroniques : [Données]

Motif de consultation : [Données]

Examen clinique : [Données]

Résultats d’examens disponibles : [Données]

Quelle est votre interprétation clinique et quelle conduite initiale recommandez-vous ? Veuillez répondre en moins de 250 mots.

English translation

Age and sex: [Data]

Relevant medical history: [Data]

Current long-term treatments: [Data]

Presenting complaint: [Data]

Physical examination findings: [Data]

Available investigation results: [Data]

What is your clinical interpretation and what initial management do you recommend? Please answer in fewer than 250 words.