

Comparative benchmarking of three artificial intelligence chatbots (ChatGPT-5.1, Qwen-3 Max, and Perplexity AI) on faculty validated undergraduate respiratory physiology multiple-choice examinations

Évaluation comparative de trois chatbots d'intelligence artificielle (ChatGPT-5.1, Qwen-3 Max et Perplexity AI) sur des questions à choix multiples validées par la faculté en physiologie respiratoire de premier cycle

Khouloud Kchaou, Salma Mokaddem, Soumaya Khaldi, Yacine Ouahchi, Khadija Ayed, Rym Baati

University of Tunis El Manar, Faculty of Medicine of Tunis, Laboratory of Physiology

ABSTRACT

Background: Artificial intelligence (AI) chatbots are increasingly used by medical students for learning and examination preparation. However, their reliability in mechanistically demanding disciplines such as respiratory physiology remains insufficiently evaluated using authentic examination material. This study aimed to assess and compare the performance of three AI chatbots on validated undergraduate respiratory physiology examination questions.

Methods: We conducted a cross-sectional analytical study including all respiratory physiology multiple-choice questions (MCQs), each comprising five propositions, used in official undergraduate examinations at the Faculty of Medicine of Tunis during the academic years 2023–2024 and 2024–2025 (101 questions). Three AI chatbots (ChatGPT-5.1, Qwen-3 Max, and Perplexity AI) were evaluated using standardized French-language prompts. Exact question-level concordance with the faculty-validated answer key, established by the teaching staff responsible for respiratory physiology examinations, was compared using Cochran's Q test and pairwise McNemar tests. At the proposition level, responses were analyzed as binary outcomes (true/false) to compute accuracy, sensitivity, and specificity.

Results: Question-level exact concordance differed significantly across models (Cochran's $Q = 10.18$, $p = 0.006$). ChatGPT showed the highest concordance rate (79.2%), followed by Qwen (74.3%) and Perplexity (62.4%). At the proposition level (505 propositions), ChatGPT achieved the highest overall accuracy (94.1%) and sensitivity for true propositions (96.5%), whereas Qwen demonstrated the highest specificity for false propositions (94.2%).

Conclusion: Although all evaluated AI chatbots performed well on respiratory physiology MCQs, substantial variability was observed across models. These findings indicate that AI chatbots are not interchangeable and underscore the need for institution-level evaluation using local examination material before their integration into medical education.

Key words: Machine Learning; Medical Education; Multiple-Choice Questions; Physiology

RÉSUMÉ

Introduction : Les chatbots d'intelligence artificielle (IA) sont de plus en plus utilisés par les étudiants en médecine. Cependant, leur fiabilité dans la physiologie respiratoire reste insuffisamment évaluée. Notre objectif a été d'évaluer et de comparer les performances de trois chatbots d'IA sur des questions d'examen en physiologie respiratoire.

Méthodes : Nous avons mené une étude transversale incluant les questions à choix multiples (QCM) de physiologie respiratoire, comportant chacune cinq propositions, utilisées lors des examens de premier cycle à la Faculté de Médecine de Tunis au cours des années universitaires 2023–2024 et 2024–2025 (101 questions). Trois chatbots d'IA (ChatGPT-5.1, Qwen-3 Max et Perplexity AI) ont été évalués. La concordance exacte, au niveau de chaque question, avec la correction validée par l'équipe enseignante responsable des examens de physiologie respiratoire a été comparée à l'aide du test Q de Cochran et de tests de McNemar. Au niveau des propositions, les réponses ont été analysées comme des variables binaires (vrai/faux) afin de calculer l'exactitude, la sensibilité et la spécificité.

Résultats : La concordance au niveau des questions a différé significativement entre les modèles ($Q=10,18$; $p=0,006$). ChatGPT a présenté le taux de concordance le plus élevé (79,2%), suivi de Qwen (74,3%) et de Perplexity (62,4%). Au niveau des propositions (505 propositions), ChatGPT a obtenu la meilleure exactitude globale (94,1%) ainsi que la sensibilité la plus élevée pour les propositions vraies (96,5%), tandis que Qwen a présenté la meilleure spécificité pour les propositions fausses (94,2%).

Conclusion : Une variabilité significative entre les modèles a été observée. Ces résultats indiquent que les chatbots d'IA ne sont pas interchangeables et soulignent la nécessité d'évaluations réalisées au niveau de chaque institution, à partir de ses propres questions d'examen, avant leur intégration dans l'enseignement médical.

Mots-clés : Apprentissage automatique; Enseignement médical; Physiologie; Questions à choix multiples

Correspondance

Khouloud Kchaou

University of Tunis El Manar, Faculty of Medicine of Tunis, Laboratory of Physiology

Email: kchaoukhouloud14@gmail.com

INTRODUCTION

Artificial intelligence (AI) chatbots are increasingly integrated into medical education and clinical reasoning workflows (1,2). Their capacity to process complex biomedical questions, generate structured explanations, and assist decision-making has driven their rapid adoption by students, residents, and educators (3,4). However, despite this widespread use, the reliability of AI chatbots in high-stakes domains requiring precise mechanistic reasoning remains insufficiently characterized, particularly in physiology.

Respiratory physiology represents a uniquely demanding discipline within preclinical medical education (5). Examinations require analytical understanding of respiratory mechanics, gas exchange, alveolar–capillary diffusion, gas transport, neural control of breathing, and exercise-related adaptations (6). These concepts are commonly assessed through multiple-choice questions (MCQs) that demand discrimination between subtly correct and incorrect propositions (7,8). Performance on such items reflects core competencies expected from medical students and is tightly standardized by physiology teaching boards.

Although several studies have evaluated AI chatbots performance on broad medical examinations or open-ended clinical tasks (9–12), a few (13) have conducted direct comparisons of multiple AI systems using the same faculty-validated physiology curriculum. As of November 2025, no study has specifically benchmarked AI chatbots on respiratory physiology examinations, a domain in which conceptually plausible but mechanistically incorrect answers may carry important educational and clinical implications. Moreover, the relative performance of different state-of-the-art AI chatbots, such as ChatGPT, Qwen, and Perplexity AI, remains unknown in this context.

To address this gap, the present study aimed to benchmark the performance of three AI chatbots on validated respiratory physiology MCQs used in undergraduate medical education. By comparing AI-generated responses with the official answer key, we sought to quantify their accuracy, analyze their error profiles, and assess their potential role and current limitations in supporting learning and assessment in respiratory physiology.

METHODS

Study design and data sources

This was a cross-sectional analytical study conducted in November 2025. All MCQs related to respiratory physiology used in the official examinations of the Faculty of Medicine of Tunis during the academic years 2023–2024 and 2024–2025 were included.

Each academic year comprised two examination sessions (main and resit), each containing 24–28 items covering the full respiratory physiology curriculum, including respiratory mechanics, gas exchange, alveolar–capillary diffusion, gas transport, arterial blood formation,

bronchomotor tone, pulmonary function tests, neural control of breathing, and exercise-related respiratory adaptations.

All correct answers had been previously validated by the Respiratory Physiology Teaching Unit (faculty board responsible for the course). Validation was performed through collective review and consensus during exam preparation meetings, ensuring consistency with the official curriculum and established physiological principles. These faculty-validated answer keys were used as the gold standard for all analyses. Importantly, none of the examination questions had ever been published online or disseminated through any digital platform, ensuring that all items were inaccessible to the AI models prior to testing.

Artificial intelligence models evaluated

Three AI chatbots were evaluated: ChatGPT-5.1 (OpenAI), Qwen-3 Max (Alibaba Cloud), and Perplexity AI. Each question was submitted to all three models during the same time period in November 2025 to minimize potential temporal update bias.

Question prompting and evaluation procedure

All questions were originally written in French, the official language of medical education and examinations at the Faculty of Medicine of Tunis. Accordingly, all prompts were administered in French to preserve authentic examination conditions.

Each question was entered independently into each model using the standardized prompt:

“Among the following statements, which one(s) is/are true?”

Models were instructed to answer by explicitly specifying the correct option(s) (A–E). Questions were submitted to each model in separate sessions to prevent memory contamination.

Each model's response was compared with the gold standard. A response was considered concordant only when the exact set of correct options was correctly identified.

Statistical analysis

All statistical analyses were performed using SPSS Statistics version 25.0.

To assess global differences in question-level exact concordance across the three AI chatbots, Cochran's Q test was applied, taking into account the paired nature of the data (each examination question evaluated by all three models). When a significant overall effect was observed, pairwise comparisons were performed using McNemar tests (ChatGPT vs. Qwen, ChatGPT vs. Perplexity, and Qwen vs. Perplexity), with Bonferroni correction applied to adjust for multiple testing.

At the proposition level, each answer option (A–E) was treated as a binary outcome (true vs. false) relative to the faculty-validated answer key. Aggregated confusion matrices were constructed for each model. Sensitivity was

defined as the proportion of true propositions correctly identified by the model, and specificity as the proportion of false propositions correctly rejected. Overall accuracy was calculated as the proportion of correctly classified propositions among all evaluated propositions. A p value < 0.05 was considered statistically significant.

Ethical considerations

This study did not involve human participants, biological samples, or identifiable personal data. It was exclusively based on the analysis of previously administered, anonymized undergraduate examination questions. No student-level data were collected or analyzed. In accordance with institutional and international guidelines, formal review and approval by an ethics committee were therefore not required, and the study was considered exempt from ethical approval.

RESULTS

A total of 101 items from four summative examinations in respiratory physiology were included.

Question-level exact concordance with gold standard answers

Exact concordance with the faculty-validated answer key differed significantly across models (Cochran's $Q = 10.18$, $p = 0.006$). ChatGPT showed the highest level of agreement, followed by Qwen and Perplexity (Figure 1).

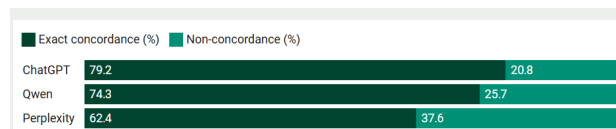


Figure 1. Question-level exact concordance and non-concordance rates with the faculty-validated answer key across the three evaluated artificial intelligence chatbots (figure created using Datawrapper)

Pairwise McNemar tests indicated that ChatGPT achieved a significantly higher exact concordance rate than Perplexity ($\chi^2 = 7.76$, $p = 0.005$), whereas no significant difference was observed between ChatGPT and Qwen ($\chi^2 = 0.55$, $p = 0.458$). Qwen also demonstrated a higher exact concordance rate than Perplexity ($\chi^2 = 4.32$, $p = 0.038$).

Diagnostic performance at the proposition level

When each proposition was considered as a binary decision ("true" vs "false" relative to the faculty-validated answer key), aggregated confusion matrices across the 505 evaluated propositions revealed differences in the numbers of true positives, false negatives, true negatives, and false positives between models. ChatGPT yielded the highest number of true positive classifications, Qwen the highest number of true negatives, and Perplexity the highest number of false positives. Detailed counts are presented in Table 1.

Table 1. Aggregated proposition-level confusion matrices for the evaluated chatbots

Outcome	ChatGPT	Qwen	Perplexity
True positives	220	208	216
False negatives	8	20	12
True negatives	255	261	240
False positives	22	16	37

At the proposition level, ChatGPT achieved the highest overall accuracy, correctly classifying 475 out of 505 propositions, followed by Qwen and Perplexity. Sensitivity for identifying true propositions was highest for ChatGPT and Perplexity, whereas Qwen showed lower sensitivity. Specificity for rejecting false propositions was highest for Qwen and lowest for Perplexity. Corresponding performance metrics are summarized in Table 2.

Table 2. Proposition-level accuracy, sensitivity, and specificity of the evaluated chatbots

Metric	ChatGPT	Qwen	Perplexity
Overall accuracy, %	94.1	92.9	90.3
Sensitivity for true propositions, %	96.5	91.2	94.7
Specificity for false propositions, %	92.1	94.2	86.6

Performance according to answer option

Performance was further examined according to the position of propositions within multiple-choice questions (options A–E). Across all models, classification accuracy was consistently high across answer options, with modest variation between positions. Proposition-level accuracy by answer option is detailed in Table 3.

Table 3. Proposition-level classification accuracy by answer option (A–E)

Proposition	ChatGPT (%)	Qwen (%)	Perplexity (%)
A	91.1	91.1	95.0
B	95.0	92.1	91.1
C	94.1	92.1	88.1
D	96.0	94.1	86.1
E	94.1	95.0	91.1

Note: Percentages represent classification accuracy for each answer option (A–E), calculated as the proportion of correctly classified propositions divided by the total number of propositions for that option across all questions ($n = 101$ per option).

DISCUSSION

This study evaluated the performance of three AI chatbots on respiratory physiology MCQs. ChatGPT demonstrated the highest concordance with the reference answer key at both the question and proposition levels, followed by Qwen and then Perplexity. At the proposition level, the models showed distinct diagnostic profiles, with ChatGPT achieving the highest sensitivity for true propositions, Qwen the highest specificity for false propositions, and Perplexity generating the largest number of false positives. By evaluating 101 MCQs, this study provides a rigorous assessment of AI performance in a highly

mechanistic domain of preclinical medical education.

Model-specific performance

At the question level, ChatGPT achieved the highest concordance with the reference answer key, followed by Qwen and Perplexity. A similar hierarchy has been reported in previous evaluations conducted in related educational domains. For example, Agarwal et al. (13) demonstrated notable differences between ChatGPT-3.5 and Claude-2 on conceptual physiology MCQs, while Alshammari et al. (14) observed marked variability across ChatGPT-4, Gemini, Qwen 2.5, and DeepSeek-V3 when answering dental examination questions. These findings support the notion that AI performance should not be assumed to be uniform across models. In this context, our study contributes additional evidence through a direct three-way comparison using a large set of official examination questions rather than curated or author-constructed questions.

Training data, language, architecture, and question structure

Several hypotheses may account for the performance differences observed between models. One potential factor is cross-model variation in training strategies and data composition. Differences in alignment procedures, exposure to educational content, and the breadth of biomedical corpora may influence how effectively models internalize and combine physiological concepts.

Model architecture and deployment objectives may also play a role. Perplexity is designed to leverage retrieval-augmented generation, a feature that may confer advantages in open-book or information-seeking contexts (15,16). In the present closed-book setting, where no external retrieval was available, this design choice may have limited its relative performance. Given that all examination items were administered in French, differences in language representation and optimization across models may have influenced their responses.

Language-related factors may further contribute to performance variability. Previous studies have shown that multilingual performance in Large Language Models (LLMs) correlates with the proportion of each language represented in the pre-training corpus, resulting in uneven performance across languages (17–19).

Another plausible explanation relates to the intrinsic structure of MCQs. Many items integrate several physiological domains and require multistep conceptual integration. This goes beyond isolated factual recall. Such questions place demands on internal consistency across sequential reasoning steps, particularly when several propositions within the same item must be evaluated jointly. Previous studies have suggested that some LLMs powering AI chatbots may perform more reliably on tasks requiring structured multistep reasoning (9,20). Conversely, reliance on local plausibility or surface-level associations may increase the selection of distractors. These distractors can appear physiologically plausible in isolation but remain internally inconsistent within a

broader mechanistic framework. This interpretation is consistent with findings reported by Chytas et al. (21), who showed that AI models may hallucinate or over-generalize when required to synthesize information across multiple conceptual layers.

More broadly, LLMs generate responses through probabilistic token prediction rather than explicit mechanistic computation (22). In physiology, where reasoning relies on well-defined physical and biological principles, model performance depends on how consistently such relationships are reflected in learned patterns (23). When internal representations are incomplete or inconsistently structured, models may generate responses that appear plausible. However, these responses can remain mechanistically incorrect, as reported in prior evaluations of AI reasoning in complex scientific and medical domains (24–26).

Beyond accuracy, the observed differences between models also raise questions regarding stability and reliability. Previous studies have emphasized that consistency across responses is essential for high-stakes educational use (14).

Implications for medical education

Taken together, these findings indicate that AI tools, although increasingly capable, remain heterogeneous in their reasoning behavior. Given their widespread use by medical students for examination preparation, this variability has direct implications for medical education. It also supports the need for clearer guidance on AI use and for integrating basic AI literacy into physiology teaching. From a methodological perspective, this study highlights the value of evaluating AI systems using authentic examination questions and of complementing global performance measures with option-level metrics, while avoiding the treatment of AI as a single, homogeneous tool. At the proposition level, models also differed in their error profiles: ChatGPT showed the highest sensitivity for true propositions, whereas Qwen achieved the highest specificity, indicating a greater ability to correctly reject false options.

From a practical perspective, AI chatbots should be positioned as adjunct study aids rather than standalone learning authorities in respiratory physiology. Students should be explicitly instructed to systematically verify AI-generated answers against faculty-approved resources. Prior to any institutional adoption, faculties are encouraged to conduct local performance testing using their own examination items, as chatbot accuracy varies across models. Clear institutional policies should define appropriate educational use of AI tools, including transparency about their limitations.

Limitations and future directions

This study has several limitations. First, it was conducted in a single institution, focused on a single discipline (respiratory physiology), and based on French-language examinations, which may limit the generalizability of the findings to other institutions, languages, and curricula.

Second, the evaluation was performed in a closed-book setting, without access to external retrieval tools. Each question was administered only once to each model, which prevented assessment of response reproducibility and intra-model variability. These aspects are relevant given the probabilistic nature of LLMs (27,28). The standardized prompting strategy, while necessary for comparability, did not allow exploration of prompt optimization or retrieval-augmented scenarios. Third, the evaluation was limited to MCQs and did not assess open-ended responses, case-based reasoning, or explanation quality, which represent increasingly important educational applications of AI chatbots. Moreover, the study evaluated a limited number of AI chatbots and was not designed as an exhaustive benchmark of all available systems; other widely used tools such as Gemini or DeepSeek were not included. Nevertheless, the selected chatbots reflect commonly used tools and distinct interaction paradigms in educational practice (29–33). In addition, a direct comparison between AI performance and student performance on the same examinations was not performed. Finally, the standardized prompting strategy, while necessary for comparability, did not allow exploration of the effects of prompt variation. Future studies should therefore assess AI performance across institutions, languages, question formats, and in both closed-book and retrieval-augmented settings, and further compare AI outputs with human student performance.

CONCLUSION

All evaluated AI chatbots demonstrated a generally good level of performance when answering mechanistically oriented respiratory physiology MCQs. However, performance varied across models, indicating that they do not behave identically when confronted with conceptually demanding items. These findings suggest that, although AI tools may support learning and examination preparation, they should not be considered interchangeable in medical education settings. Careful evaluation and contextualized guidance are therefore required to support their responsible integration into physiology teaching and assessment. As AI systems continue to evolve, ongoing methodologically rigorous evaluation will remain essential to ensure their effective and equitable use in medical training.

REFERENCES

- Vrdoljak J, Boban Z, Vilović M, Kumrić M, Božić J. A Review of Large Language Models in Medical Education, Clinical Decision Support, and Healthcare Administration. *Healthcare*. 2025;13(6):603. doi:10.3390/healthcare13060603
- Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Med Inform Decis Mak*. 2025;25(1):117. doi:10.1186/s12911-025-02954-4
- Lu Z, Peng Y, Cohen T, Ghassemi M, Weng C, Tian S. Large language models in biomedicine and health: current research landscape and future directions. *J Am Med Inform Assoc*. 2024;31(9):1801-1811. doi:10.1093/jamia/ocae202
- Mondal H, Karri JKK, Ramasubramanian S, Mondal S, Juhi A, Gupta P. A qualitative survey on perception of medical students on the use of large language models for educational purposes. *Adv Physiol Educ*. 2025;49(1):27-36. doi:10.1152/advan.00088.2024
- Elmoselhi A. Clinically oriented laboratory illuminates challenging respiratory concepts to the pre-clinical medical students. *Medical Teacher*. 2012;34(3):258-258. doi:10.3109/01421591003719012
- Crozier T. Respiratory Physiology and Terminology. In: Lapinsky SE, Plante LA, eds. *Respiratory Disease in Pregnancy*. 1st ed. Cambridge University Press; 2020:1-8. doi:10.1017/9781108163705.001
- Cheng HM, Hoe SZ. Students' responses under "negative pressure" to respiratory questions at the 15th physiology quiz international event: 100 medical school teams from 22 countries. *BLDE Univ J Health Sci*. 2017;2(2):105. doi:10.4103/bjhs.bjhs_27_17
- Badenhorst E, Mamede S, Abrahams A, et al. First-year medical students' naive beliefs about respiratory physiology. *Adv Physiol Educ*. 2016;40(3):342-348. doi:10.1152/advan.00193.2015
- Bharatha A, Ojeh N, Rabbi AMF, et al. Comparing the Performance of ChatGPT-4 and Medical Students on MCQs at Varied Levels of Bloom's Taxonomy. *AMEP*. 2024;15:393-400. doi:10.2147/AMEP.S457408
- Sadeq MA, Ghorab RMF, Ashry MH, et al. AI chatbots show promise but limitations on UK medical exam questions: a comparative performance study. *Sci Rep*. 2024;14(1):18859. doi:10.1038/s41598-024-68996-2
- Zhang R, Cai Q, Sartori A, Gayed N, Collette H. Comparing Artificial Intelligence Large Language Models in Medical Training: A Performance Analysis of ChatGPT and DeepSeek on United States Medical Licensing Examination (USMLE) Style Questions. *Cureus*. 2025;17(8):e90212. doi:10.7759/cureus.90212
- Uppalapati VK, Nag DS. A Comparative Analysis of AI Models in Complex Medical Decision-Making Scenarios: Evaluating ChatGPT, Claude AI, Bard, and Perplexity. *Cureus*. 2024;16(1):e52485. doi:10.7759/cureus.52485
- Agarwal M, Goswami A, Sharma P. Evaluating ChatGPT-3.5 and Claude-2 in Answering and Explaining Conceptual Medical Physiology Multiple-Choice Questions. *Cureus*. 2023;15(9):e46222. doi:10.7759/cureus.46222
- Alshammari AF, Madfa AA, Anazi BA, Alenezi YE, Alkurdi KA. Comparison of accuracy and consistency of AI Language models when answering standardised dental MCQs. *BMC Med Educ*. 2025;25(1):1507. doi:10.1186/s12909-025-07624-7
- Bhayana R, Fawzy A, Deng Y, Bleakney RR, Krishna S. Retrieval-Augmented Generation for Large Language Models in Radiology: Another Leap Forward in Board Examination Performance. *Radiology*. 2024;313(1):e241489. doi:10.1148/radiol.241489
- Masanneck L, Epping PZ, Meuth SG, Pawlitzki M. Evaluating Web Retrieval-Assisted Large Language Models With and Without Whitelisting for Evidence-Based Neurology: Comparative Study. *J Med Internet Res*. 2025;27:e79379. doi:10.2196/79379
- Kim Y, Shin S, Yoo SH. Performance of large language models in non-English medical ethics-related multiple choice questions: comparison of ChatGPT performance across versions and languages. *BMC Med Ethics*. 2025;26:168. doi:10.1186/s12910-025-01316-z
- Perez-Toro PA, Dineley J, Iniesta R, et al. Exploring biases related to the use of large language models in a multilingual depression corpus. *Sci Rep*. 2025;15(1):36197. doi:10.1038/s41598-025-19980-x
- Qin L, Chen Q, Zhou Y, et al. A survey of multilingual large language models. *Patterns (N Y)*. 2025;6(1):101118. doi:10.1016/j.patter.2024.101118
- Soulage CO, Van Coppenolle F, Guebre-Egziabher F. The conversational AI "ChatGPT" outperforms medical students on a physiology university examination. *Adv Physiol Educ*. 2024;48(4):677-684. doi:10.1152/advan.00181.2023
- Chytas D, Noussios G, Salmas M, Vasiladis AV, Troupis T. Investigation of Studies on ChatGPT's Ability to Answer Anatomy Questions: A Self-Evaluation by ChatGPT and Comparison with an Evaluation by Gemini. *Cureus*. 2025;17(8):e90572. doi:10.7759/

cureus.90572

22. Blank IA. What are large language models supposed to model? *Trends Cogn Sci.* 2023;27(11):987-989. doi:10.1016/j.tics.2023.08.006
23. Banerjee A, Ahmad A, Bhalla P, Goyal K. Assessing the Efficacy of ChatGPT in Solving Questions Based on the Core Concepts in Physiology. *Cureus.* 2023;15(8):e43314. doi:10.7759/cureus.43314
24. Ullah E, Parwani A, Baig MM, Singh R. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology – a recent scoping review. *Diagn Pathol.* 2024;19:43. doi:10.1186/s13000-024-01464-7
25. Omar M, Sorin V, Collins JD, et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. *Commun Med (Lond).* 2025;5:330. doi:10.1038/s43856-025-01021-3
26. Roustan D, Bastardot F. The Clinicians' Guide to Large Language Models: A General Perspective With a Focus on Hallucinations. *Interact J Med Res.* 2025;14:e59823. doi:10.2196/59823
27. Herrera-Poyatos D, Peláez-González C, Zuheros C, et al. An overview of model uncertainty and variability in LLM-based sentiment analysis: challenges, mitigation strategies, and the role of explainability. *Front Artif Intell.* 2025;8:1609097. doi:10.3389/frai.2025.1609097
28. Waldo J, Boussard S. GPTs and Hallucination: Why do large language models hallucinate? *Queue.* 2024;22(4):19-33. doi:10.1145/3688007
29. Sallam M, Alasfoor IM, Khalid SW, et al. Chinese generative AI models (DeepSeek and Qwen) rival ChatGPT-4 in ophthalmology queries with excellent performance in Arabic and English. *Narra J.* 2025;5(1):e2371. doi:10.52225/narra.v5i1.2371
30. Kalam KA, Masoud FD, Muntaser A, Ranga R, Geng X, Goyal M. ChatGPT as a Learning Tool for Medical Students: Results From a Randomized Controlled Trial. *Cureus.* 17(6):e85767. doi:10.7759/cureus.85767
31. Gravina AG, Pellegrino R, Palladino G, Imperio G, Ventura A, Federico A. Charting new AI education in gastroenterology: Cross-sectional evaluation of ChatGPT and perplexity AI in medical residency exam. *Dig Liver Dis.* 2024;56(8):1304-1311. doi:10.1016/j.dld.2024.02.019
32. Chimenea A, García-Díaz L, Antiñolo G. Exploring the Potential of Artificial Intelligence Language Models in Obstetrics with a Focus on Fetal Medicine: An Evaluation of the Perplexity AI Model. *Fetal Diagn Ther.* 2024;51(2):125-132. doi:10.1159/000535345
33. Zhu S, Hu W, Yang Z, Yan J, Zhang F. Qwen-2.5 Outperforms Other Large Language Models in the Chinese National Nursing Licensing Examination: Retrospective Cross-Sectional Comparative Study. *JMIR Med Inform.* 2025;13:e63731. doi:10.2196/63731