frontiers | Frontiers in Artificial Intelligence

TYPE Brief Research Report
PUBLISHED 15 September 2025
DOI 10.3389/frai.2025.1662203

OPEN ACCESS

EDITED BY
Tim Hulsen,
Rotterdam University of Applied Sciences, Netherlands

REVIEWED BY
Siquan Wang,
Columbia University, United States
Alberto Bustillos,
Technical University of Ambato, Ecuador

*CORRESPONDENCE
Cristian N. Rivera-Rosas
✉ md.cristian.rivera@gmail.com

RECEIVED 08 July 2025
ACCEPTED 28 August 2025
PUBLISHED 15 September 2025

CITATION
Rivera-Rosas CN, Calleja-López JRT, Larios-Camacho SJ and Trujillo-López S (2025) Using ChatGPT as an assessment tool for medical residents in Mexico: a descriptive experience. *Front. Artif. Intell.* 8:1662203. doi: 10.3389/frai.2025.1662203

Using ChatGPT as an assessment tool for medical residents in Mexico: a descriptive experience

Cristian N. Rivera-Rosas[1]*, J. R. Tadeo Calleja-López[2], Sandra J. Larios-Camacho[3] and Sergio Trujillo-López[4]

[1]General Hospital Zone 89, Mexican Social Security Institute, Guadalajara, Mexico, [2]General Hospital Zone 14, Mexican Social Security Institute, Hermosillo, Mexico, [3]General Hospital Zone 14, Mexican Social Security Institute, Guadalajara, Mexico, [4]Department of Medicine and Health Sciences, University of Sonora, Hermosillo, Mexico

Introduction: Artificial intelligence (AI) in medical education has progressed gradually, with numerous authors debating whether to prohibit, restrict, or adopt its use in academic contexts. Growing evidence exists regarding the capabilities and applications of AI in this field, particularly in supporting educational tasks such as student assessment. In this article we described our experience using ChatGPT to evaluate medical residents.

Materials and methods: A descriptive cross-sectional study was conducted involving 35 medical residents from different specialty's at a secondary-level hospital. Two different exams were generated using ChatGPT in topics of Rocky Mountain Spotted Fever (RMSF) and *Pertussis*. Additionally, an opinion survey—previously validated was administered to assess participants' perceptions of ChatGPT ability to generate multiple-choice questions.

Results: Overall average score for the *Pertussis* examination was 8.46, while the average for the RMSF examination was 8.29. All participants reported that the examination was well written and that the language used was coherent; 34 residents (97.14%) stated that the language was clear, concise, and easy to understand; 9 residents (25.71%) agreed that the language used was confusing; 33 residents (94.28%) rated the exams questions as difficult; 32 residents (91.42%) felt that they had adequately prepared for both examinations.

Discussion: ChatGPT exhibits a promising faculty as a tool to support teaching activities in the training of medical specialists, mainly in reducing the human workload of healthcare personnel, and becoming integral to the next phase of medical education through AI-assisted content creation supervised by educators.

KEYWORDS

ChatGPT, artificial intelligence, medical education, resident physicians, multiple choice question exams

1 Introduction

Large language models (LLMs) such as ChatGPT continue to revolutionize human activities across various clinical and professional domains. The integration of artificial intelligence (AI) into medical education has progressed gradually, with numerous authors debating whether to prohibit, restrict, or adopt its use in academic contexts (Sánchez Mendiola, 2023). Nonetheless, there is growing evidence regarding the capabilities and applications of

AI in this field, particularly in passing medical licensing examinations and supporting educational tasks such as student assessment, clinical scenario development, and the creation of formative feedback (Rivera-Rosas et al., 2024; Aster et al., 2024). These developments have sparked renewed reflection on the future of medical education.

Despite the increasing anecdotal recognition of AI's plausible utility in the training of medical residents and the educational responsibilities of faculty members, there remains a scarcity of empirical evidence documenting medical educators' experiences with these tools to address learning needs inherent to medical residence training. LLMs like ChatGPT represent potentially transformative tools that could support a new paradigm in medical education by alleviating the workload of both faculty and trainees, particularly in routine academic tasks such as exam generation and assessment.

The objective of this article is to describe our experience utilizing ChatGPT for the creation of multiple-choice question (MCQ) exams administered to medical residents from various specialties at a secondary-level hospital in Mexico, as well as to report resident perceptions regarding the AI-generated questions.

2 Materials and methods

A descriptive cross-sectional study was conducted involving 35 medical residents from the specialties of anesthesiology, emergency medicine, and internal medicine at a secondary-level hospital in Mexico. Initially, a general session on the topic *Pertussis-like Syndrome and Pertussis* was delivered over the course of one week. One week later, a session was held on *Rocky Mountain Spotted Fever caused by Rickettsia rickettsii (RMSF)*.

To generate the exams, ChatGPT-3.5, in its "Scholar GPT" mode, was prompted using two different inputs (Table 1) to create two questionnaires consisting of 15 multiple-choice questions each, based on the material presented in the corresponding class (Supplementary Table 1). Subsequently, using the Delphi method, three physicians reviewed the AI-generated questions and selected 10 questions from each questionnaire. Items were excluded due to inappropriate focus, inaccuracies in the AI-generated answers, or misalignment with the content covered during the instructional sessions.

Additionally, an opinion survey—previously validated in a Mexican student population was administered to assess participants' perceptions of ChatGPT-3.5's ability to generate multiple-choice questions (Rivera-Rosas et al., 2024). The survey responses were later dichotomized for analysis. Importantly, none of the participants were aware that the exam questions had been generated using ChatGPT-3.5.

After the final selection of questions for each exam and one week after the RMSF session, the two separate exams created with ChatGPT-3.5 were administered to the residents covering the respective topics previously mentioned. The Google Forms platform was used to administer both the assessments and the opinion survey to the residents. When answering the google forms survey, residents were first asked to give their consent for using the results of their responses for academic and research purposes. The evaluation results were recorded in a Microsoft Office Excel 360® spreadsheet. Descriptive statistical analysis was conducted using frequency measures, and the results were presented through bar charts (Figure 1). No additional statistical tests were performed.

TABLE 1 Prompts used for generating MCQ exams.

Prompts used for generating MCQ exams	
Whooping Cough Exam	You are a physician and professor at a hospital. You are in charge of the teaching area and conducted a general session at the hospital where you work, training staff on the topic of Pertussis-like Syndrome and Whooping Cough. Create 15 multiple-choice questions covering the etiology, clinical presentation, diagnosis, and treatment of this topic. Each question should have 4 options and only one correct answer. Show me the correct answer.
RMSF Exam	You are a physician and professor at a hospital. You are in charge of the teaching area and conducted a general session at the hospital where you work, training staff on the topic of Rickettsiosis and Rocky Mountain Spotted Fever (*Rickettsia rickettsii*). Create 15 multiple-choice questions covering the etiology, clinical presentation, diagnosis, and treatment of this topic. Each question should have 4 options and only one correct answer. Show me the correct answer.

RMSF, Rocky Mountain Spotted Fever; MCQ, multiple-choice questions.

3 Results

A total of two assessments were administered to 35 medical residents at various stages of specialty training. Of these, 18 were male (51.43%) and 17 were female (48.57%). Regarding specialty distribution, 8 residents were from anesthesiology (22.86%), 20 from emergency medicine (57.14%), and 7 from internal medicine (20.0%). On a scale of 1 to 10, the overall average score for the *Pertussis* examination was 8.46, while the average for the RMSF examination was 8.29. The score range for the *Pertussis* exam was 5 to 10, whereas scores for the RMSF exam ranged from 2 to 10.

Regarding the opinion survey administered to the residents, a summary of the results is presented in Figure 1. All participants (100%) reported that the examination was well written and that the language used was coherent. Additionally, 34 residents (97.14%) stated that the language was clear, concise, and easy to understand. However, when asked if the language used was confusing 9 residents (25.71%) agreed, while 26 (74.29%) disagreed. In terms of the difficulty of the exam questions, 33 residents (94.28%) rated them as difficult. Finally, 32 residents (91.42%) felt that they had adequately prepared for both examinations.

4 Discussion

To our knowledge, this is one of the first documented studies involving medical residents in Mexico that describes the experience of faculty using ChatGPT as a tool to support teaching activities in the training of medical specialists. Previous studies have reported the use of LLMs for generating examination questions in specialties such as otolaryngology and emergency medicine (Lotto et al., 2024; Law et al., 2025), demonstrating the model's strong ability to formulate multiple-choice questions. Moreover, a systematic review has evaluated the

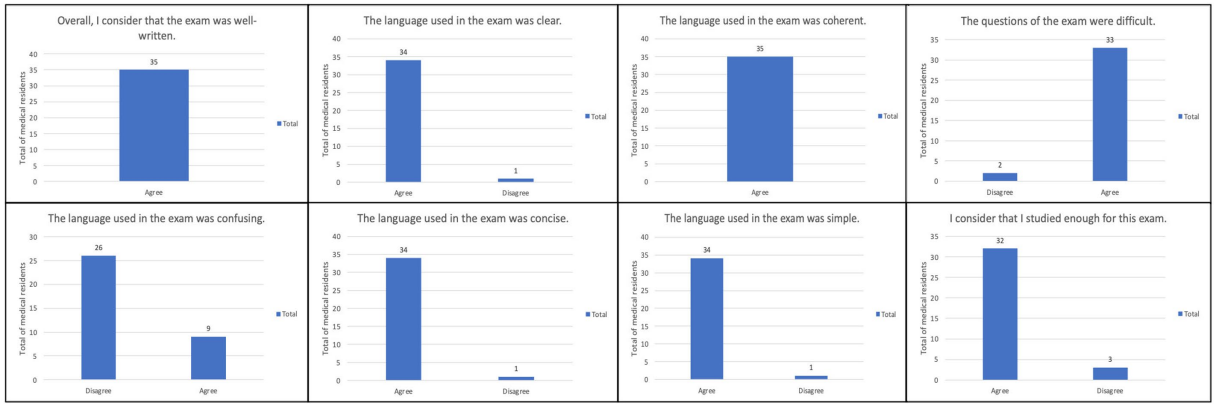

FIGURE 1
Overall results from the satisfaction survey used for medical residents' perception about the exams' questions. *n* = 35.

efficacy of LLMs in creating high-quality multiple-choice questions (Kıyak and Emekli, 2024).

Although no statistical analysis was performed in our study, the results provide preliminary evidence of the model's ability to generate complex questions suitable for postgraduate-level health education. Notably, most residents considered the questions generated by ChatGPT to be difficult, indicating that they could represent a cognitive challenge comparable to questions traditionally developed by experienced medical professors. This underscores the importance of prompt quality (Kıyak, 2024), as the AI's output quality strongly depends on the input provided. Nevertheless, the integration of AI tools like ChatGPT in hospital-based educational activities either for teaching or resident learning remains understudied, and it is still unclear whether statistically significant educational benefits exist compared to conventional teaching strategies.

Our findings align with prior reports of student acceptance and satisfaction regarding the quality of AI-generated questions (Rivera-Rosas et al., 2024). Similar studies have reported high levels of user satisfaction among medical students using AI for various academic tasks (Boris Miclin et al., 2024). Otherwise, medical professors' opinions should also be evaluated in residence training contexts to overview their acceptance or rejection of its usefulness and their perceived knowledge about this tool and how to use them in their teaching activities in the hospital.

In this study, we highlighted some empirical uses of AI within hospital settings for specialist training, ranging from question generation for assessments to the integration of smart platforms such as Google Forms. These tools in addition to other AI models could offer two key advantages: (a) reducing the human workload of healthcare personnel that participates in residence training, and (b) becoming integral to the next phase of medical education through AI-assisted content creation supervised by professors.

While our work focuses specifically on ChatGPT's utility in generating MCQ, further exploration of other applications is warranted. These may include drafting high-quality clinical notes, assisting with emergency department triage, creating presentations, enhancing scientific literature searches, or translating scientific articles applications that have been discussed in other studies (Hallquist et al., 2025). These use cases could indirectly reduce physician workload, improve learning outcomes, and enhance patient care. Otherwise, professors using LLM as ChatGPT should be aware that the knowledge (or "AI training), accuracy, context recognition and the ability of handling more complex prompts can vary depending on the LLMs version used for the MCQ creation, as their capabilities for answering or generating MQC can be vary (Mistry et al., 2024; Liu et al., 2025). This supports the needed supervision and domain that professors should have of AI tools for its adequate uses.

Despite the promising outlook, it is essential to remain aware of the limitations of AI, including ethical concerns such as plagiarism and authorship attribution in scientific writing, as well as philosophical questions regarding AI's impact on critical thinking and its susceptibility to generating factual inaccuracies or "hallucinations" (Jeyaraman et al., 2023). Among the limitations of our study is its exploratory nature. We did not perform statistical validation of the AI-generated questions. Also, our results display promising utility of this AI tool in creating difficult MCQ, which also could diminish workload for teachers, but a quality comparison against human questions in medical residence scenarios should be assess in prospective studies to clarify its validated utility. Therefore, further research with greater scientific rigor are required to substantiate our findings.

5 Conclusion

Artificial intelligence and large language models like ChatGPT have become part of our new reality. Increasing evidence supports their potential benefits in the medical field, from education to clinical care. For this reason, their integration into healthcare processes should not be resisted but rather approached pragmatically as tools to be used thoughtfully. Developing AI literacy and competencies among medical students and faculty is essential and should progressively be incorporated into medical curricula to better prepare professionals for AI-assisted medical education and clinical practice.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The participants or the participants legal guardian/next of kin, provided their written informed consent to participate in this study.

Author contributions

CR-R: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. JC-L: Writing – original draft, Writing – review & editing, Methodology. SL-C: Data curation, Investigation, Project administration, Supervision, Validation, Writing – review & editing. ST-L: Conceptualization, Funding acquisition, Supervision, Validation, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was funded by the Department of Medicine and Health Sciences from the University of Sonora, Hermosillo.

Acknowledgments

To the Mexican Social Security Institute and the University of Sonora.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that Gen AI was used in the creation of this manuscript. We used ChatGPT to check the grammar and translation of some words from Spanish to English in the manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2025.1662203/full#supplementary-material

References

Aster, A., Laupichler, M. C., Rockwell-Kollmann, T., Masala, G., Bala, E., and Raupach, T. (2024). ChatGPT and other large language models in medical education - scoping literature review. *Medical Science Educator* 35:555. doi: 10.1007/s40670-024-02206-6

Boris Miclin, C. D., Estrada Rodríguez, Y., and Leyva Argibay, S. R. (2024). Uso de ChatGPT por estudiantes de medicina en su proceso de enseñanza - aprendizaje. [Internet], Cuba. 30 de diciembre de 2024 [citado 3 de septiembre de 2025], 6. Available online at: https://revunimed.sld.cu/index.php/revestud/article/view/398

Hallquist, E., Gupta, I., Montalbano, M., and Loukas, M. (2025). Applications of artificial intelligence in medical education: a systematic review. *Cureus* 17:e79878. doi: 10.7759/cureus.79878

Jeyaraman, M. K. S. P., Jeyaraman, N., Nallakumarasamy, A., Yadav, S., and Bondili, S. K. (2023). ChatGPT in medical education and research: a boon or a bane? *Cureus* 15:e44316. doi: 10.7759/cureus.44316

Kıyak, Y. S. (2024). Chatgpt's ability or prompt quality: what determines the success of generating multiple-choice questions. *Academic Pathology* 11:100119. doi: 10.1016/j.acpath.2024.100119

Kıyak, Y. S., and Emekli, E. (2024). ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: a literature review. *Postgrad. Med. J.* 100, 858–865. doi: 10.1093/postmj/qgae065

Law, A. K., So, J., Lui, C. T., Choi, Y. F., Cheung, K. H., Kei-ching Hung, K., et al. (2025). AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. *BMC Med. Educ.* 25:208. doi: 10.1186/s12909-025-06796-6

Liu, M., Okuhara, T., Dai, Z., Huang, W., Gu, L., Okada, H., et al. (2025). Evaluating the effectiveness of advanced large language models in medical knowledge: a comparative study using Japanese national medical examination. *Int. J. Med. Inform.* 193:105673. doi: 10.1016/j.ijmedinf.2024.105673

Lotto, C., Sheppard, S. C., Anschuetz, W., Stricker, D., Molinari, G., Huwendiek, S., et al. (2024). ChatGPT generated otorhinolaryngology multiple-choice questions: quality, psychometric properties, and suitability for assessments. *OTO Open* 8:e70018. doi: 10.1002/oto2.70018

Mistry, N. P., Saeed, H., Rafique, S., Le, T., Obaid, H., and Adams, S. J. (2024). Large language models as tools to generate radiology board-style multiple-choice questions. *Acad. Radiol.* 31, 3872–3878. doi: 10.1016/j.acra.2024.06.046

Rivera-Rosas, C. N., Calleja-López, J. T., Ruibal-Tavares, E., Villanueva-Neri, A., Flores-Felix, C. M., and Trujillo-López, S. (2024). Exploring the potential of ChatGPT to create multiple-choice question exams. *Educ. Med.* 25:100930. doi: 10.1016/j.edumed.2024.100930

Sánchez Mendiola, M. (2023). La inteligencia artificial generativa y la evaluación:¿ qué pasará con los exámenes? *Investig. Educ. Med.* 12, 5–8. doi: 10.22201/fm.20075057e.2023.48.23550