Generative AI's performance on emergency medicine boards questions: observational study

STEN KAJITANI[1], MILA PASTRAK[1], ANTHONY GOODINGS[1], AUDREY NGUYEN[1], AUSTIN DREWEK[2], ANDREW LAFREE[3], ADRIAN MURPHY[1]

[1] School of Medicine, University College Cork, Cork, Ireland

[2] Department of Emergency Medicine, John's Hopkins University, Baltimore, Maryland, USA

[3] Department of Emergency Medicine, University of California San Diego, San Diego, California, USA

https://doi.org/10.33178/SMJ.2025.1.39

Abstract

BACKGROUND: The evolving field of medicine has introduced ChatGPT as a potential assistive platform, though its use in medical board exam preparation remains debated [1-2]. This study aimed to evaluate the performance of a custom-modified version of ChatGPT-4, tailored with emergency medicine board exam preparatory materials (Anki deck), compared to its default version and previous iteration (3.5) [3]. The goal was to assess the accuracy of ChatGPT-4 answering board- style questions and its suitability as a tool for medical education.

MATERIALS & METHODS: A comparative analysis was conducted using a random selection of 598 questions from the Rosh In--Training Exam Question Bank [4]. The subjects of the study included three versions of ChatGPT: the Default, a Custom, and ChatGPT-3.5. Accuracy, response length, medical discipline subgroups, and underlying causes of error were analyzed.

RESULTS: Custom ChatGPT-4 did not significantly improve accuracy over Default ($p>0.05$), but both significantly outperformed ChatGPT-3.5 ($p<0.05$) (Table 1). Default produced longer responses than Custom ($p<0.05$). Subgroup analysis showed no significant difference across medical sub-disciplines ($p>0.05$). ChatGPT-4 had a 99% probability of passing, while ChatGPT-3.5 had 85%.

	Custom ChatGPT-4 (n=598)	Default ChatGPT-4 (n=598)	Default ChatGPT-3.5 (n=269)
Number of Correct Questions	481	480	169
Correct (%)	80.4	80.3	62.8

Table 1. The performance of 3 language models on American Emergency Board Exam using Rosh Review.

CONCLUSIONS: The findings suggest that while newer versions of ChatGPT exhibit improved performance in emergency medicine board exam preparation, specific enhancements do not significantly impact accuracy. The study highlights the potential of ChatGPT-4 as a tool for medical education, capable of providing accurate support across a wide range of topics in emergency medicine.