Evaluating the Performance of Artificial Intelligence Models in Answering Facial Plastic Surgery Board Questions

Dev Shah, BS[1]; William Long, BS[1] ; Shervin Eskandari, BS[1]; Kush Patel, MD[2]

1. Medical College of Georgia, Augusta University, Augusta, Georgia
2. Department of Otolaryngology-Head and Neck Surgery, University of Mississippi Medical Center, Jackson, MS

INTRODUCTION

- Rapid advances in the development of artificial intelligence technology has led to the emergence of large language models (LLMs).
- These models are capable of understanding and generating human-like interactions. In recent years, these models have demonstrated steadily rising performance on standardized medical exams, posing the question of whether they could serve as a learning aid in medical education.
- Most of the existing literature has focused on broader licensing examinations such as the USMLE.
- Facial plastic and reconstructive surgery is a highly specialized field with distinct procedural techniques and postoperative management principles.
- This study addresses that gap by utilizing three of the most widely used LLMs: ChatGPT, Gemini, and Copilot.
- By comparing each model's accuracy and confidence against human performance metrics, there will be stronger metrics to determine the role of LLM in medical education.

METHODS

- Five hundred and nineteen text-based questions from the StatPearls facial plastic and reconstructive surgery question bank were answered by ChatGPT 3.5, Gemini 2.5 Flash, and Copilot.
- Question characteristics and human performance were recorded, and LLMs were tasked with providing confidence percentages alongside their answers.
- Each LLM was prompted one question at a time, and a unique conversation window was generated for every query as to not skew results. Relative performances were evaluated using the Z-test for proportions.

RESULTS

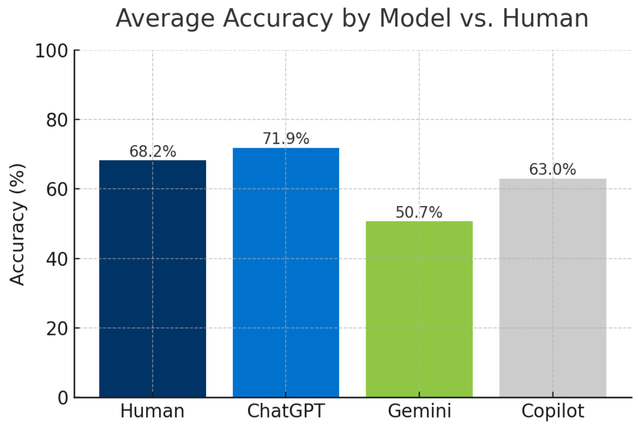

Figure 1. Accuracy (%) of LLM and humans

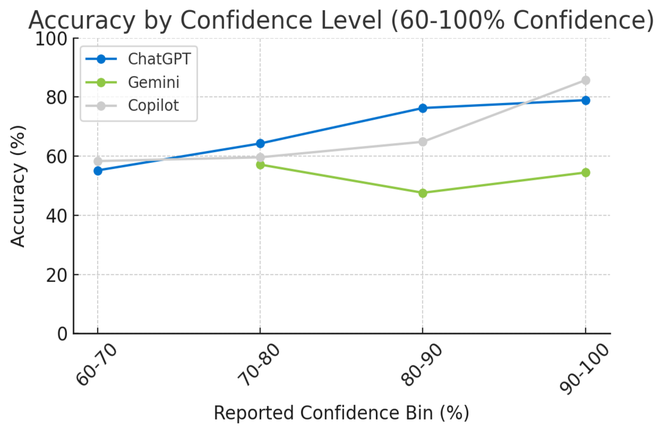

Figure 2. Confidence vs Accuracy of LLM's

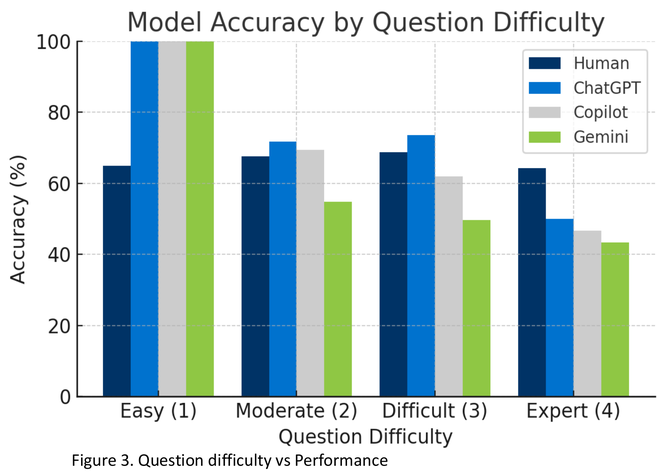

Figure 3. Question difficulty vs Performance

- 519 text based questions were analyzed with these LLMs. ChatGPT displayed a 71.9% accuracy rate, while humans scored 68.2%, followed by Copilot at 63.0% and Gemini at 50.7%.
- LLMs significantly outscored humans for questions that were classified as "easy", however they all lagged behind human performance on the hardest "expert" level questions.
- LLM performance was strongly correlated with one another, as when all three models agreed on the answer they showed a 86.1% accuracy, however when 2/3 agreed that number dropped to 60%.
- Overconfidence was prevalent: All three models indicated very high confidence with ChatGPT claiming 81.6%, Copilot 82.6%, and Gemini 91.3% on average, despite their lower accuracies, indicating a significant overconfidence bias.
- ChatGPT's confidence was the only LLM that showed confidence to be a predictor of accuracy ($R^2 = 0.025$, $p < 0.001$), whereas Gemini nor Copilot showed any meaningful confidence–accuracy relationship ($p > 0.1$).

CONCLUSIONS

- ChatGPT had the highest accuracy amongst the tested LLMs on the facial plastic surgery board style questions, far exceeding the performance of Copilot and Gemini.
- When all three AIs provided the same answer, their collective accuracy far exceeded human performance. However, the average accuracy of these models was recorded to be lower than the average known human performance,
- This signifies the ongoing need to be cautious when considering the use of artificial intelligence in medical educational settings.

IMPLICATIONS

- Although early performance is promising and the use of LLMs may provide a quick resource for checking answers and getting explanations, those studying for board examinations such as the facial plastics exam must be dilligent in not relying on their answers solely to guide their educational plans.
- These LLM's showed strong performance, and when the confidence was high alongside a question that was familiar, it not only gave the correct answer more often than not but it also provided a strong explanation.
- The role of LLM's in education is rapidly expanding, and as models progress, more importance may be placed upon utilizing them as a method of learning and practicing educational material.

REFERENCES

1) Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, Löffler CML, Schwarzkopf SC, Unger M, Veldhuizen GP, Wagner SJ, Kather JN. The future landscape of large language models in medicine. Commun Med (Lond). 2023 Oct 10;3(1):141. doi: 10.1038/s43856-023-00370-1. PMID: 37816837; PMCID: PMC10564921.
2) Cascella M, Semeraro F, Montomoli J, Bellini V, Piazza O, Bignami E. The Breakthrough of Large Language Models Release for Medical Applications: 1-Year Timeline and Perspectives. J Med Syst. 2024 Feb 17;48(1):22. doi: 10.1007/s10916-024-02045-3. PMID: 38366043; PMCID: PMC10873461.
3) Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023 Aug;29(8):1930-1940. doi: 10.1038/s41591-023-02448-8. Epub 2023 Jul 17. PMID: 37460753.