Comparing the Performance of Popular Large Language Models on the National Board of Medical Examiners Sample Questions

Ali Abbas, B.S., Mahad Rehman, B.S.

UTSouthwestern Medical Center

Introduction:

- Large language models (LLMs) have transformed various domains in medicine, aiding in complex tasks and clinical decision-making
- OpenAI's GPT-4, GPT-3.5, Google's Bard, and Anthropic's Claude are among the most widely used LLMs
- LLMs have the potential to aid students and serve as a supplemental educational resource
- This study aims to compare the accuracy of popular LLMs on National Board of Medical Examiners (NBME) clinical subject exam sample questions

Methods:

- The questions used in this study were multiple-choice questions obtained from the official NBME website and are publicly available
- Questions from the NBME subject exams in medicine, pediatrics, obstetrics and gynecology, clinical neurology, ambulatory care, family medicine, psychiatry, and surgery were used to query each LLM
- The response by each LLM was compared to the answer provided by the NBME and checked for accuracy. Statistical analysis was performed using one-way analysis of variance (ANOVA)

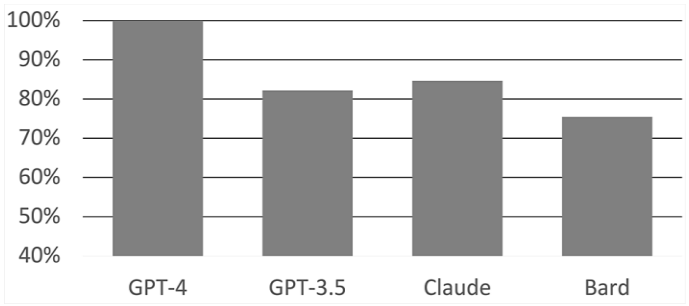

Figure 1: Overall performance of the LLMs on all NBME sample questions
LLM: large language model, NBME: National Board of Medical Examiners

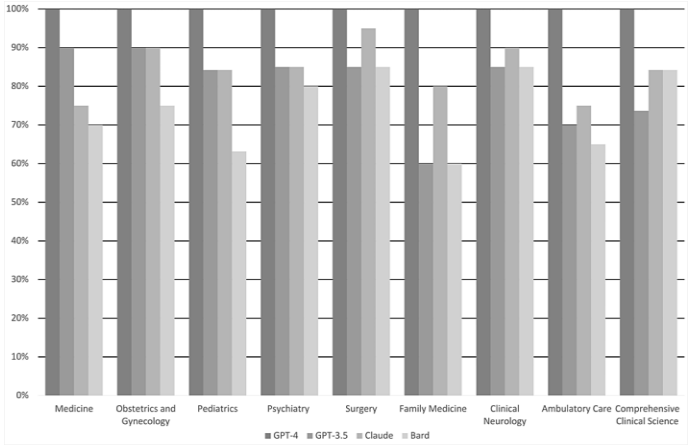

Figure 2: Performance of the LLMs on each subject exam
LLM: large language model

Results:

- A total of 163 questions were queried by each LLM
- GPT-4 scored 163/163 (100%), GPT-3.5 scored 134/163 (82.2%), Bard scored 123/163 (75.5%), and Claude scored 138/163 (84.7%)
- The total performance of GPT-4 was statistically superior to that of GPT-3.5, Claude, and Bard by 17.8%, 15.3%, and 24.5%, respectively
- Across all LLMs, the surgery exam had the highest average score (18.25/20), while the family medicine exam had the lowest average score (3.75/5)

Conclusions:

- GPT-4's superior performance on NBME clinical subject exam sample questions underscores its potential in medical education and practice
- While LLMs exhibit promise, discernment in their application is crucial, considering occasional inaccuracies
- As technological advancements continue, regular reassessments and refinements are imperative to maintain their reliability and relevance in medicine

For more information: DOI: 10.7759/cureus.55991