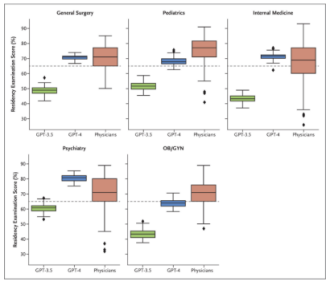
HealthManagement
Promoting Management and Leadership

LLMs' Comparable Performance to Physicians on Medical Board Examination



A recent article published in the New England Journal of Medicine AI discusses the role of large language models (LLMs), particularly ChatGPT, in artificial intelligence (AI) as virtual problem solvers with human-like text generation abilities. It highlights the potential of LLMs in supporting clinicians and suggests comparing their performance with trained physicians. Previous studies have mainly evaluated LLMs using simulated medical exams and open-source data. However, recent research focused on real exam settings, such as official medical board exams. The study examined 849 resident physicians taking Israeli board certification exams in five core medical specialities. The primary goal was to assess the performance of Generative Pretrained Transformer 3.5 (GPT-3.5) and GPT-4 compared to practising physicians, and the secondary objective was to compare the performance of GPT-3.5 and GPT-4. The article provides the exams as a benchmark dataset for the medical machine learning and natural language processing communities, potentially informing future LLM studies.

Evaluating Physician and LLM Performance on Medical Board Examinations

The study conducted a retrospective analysis of physicians' performance on the 2022 medical board certification examinations across five core medical specialities: internal medicine, general surgery, psychiatry, paediatrics, and OB/GYN. The dataset included scores achieved by 849 physicians on 655 multiple-choice questions translated from Hebrew to English. Questions requiring imaging analysis were excluded. The analysis estimated how GPT-3.5 and GPT-4 model test scores ranked among physicians in each speciality, repeating the process 120 times per model. Median percentiles and 95% confidence intervals were reported. A secondary comparison between GPT-3.5 and GPT-4 was conducted using a two-sided independent-sample t-test with Bonferroni correction. Box plots of GPT-4 performance with different parameter values were provided for internal medicine. Statistical analysis was performed using Python 3.11.3 and SciPy version 1.11.4.

GPT-4's Remarkable Performance

The study compared the performance of GPT-4 and GPT-3.5 with physicians across different medical specialities. GPT-4 outperformed a significant portion of physicians in all specialities, with its highest performance seen in psychiatry. However, it ranked lower in paediatrics and OB/GYN. In contrast, GPT-3.5 performed weaker than GPT-4 across all disciplines, ranking below all physicians in general surgery and OB/GYN. Additionally, GPT-3.5 had lower median percentiles compared to GPT-4 in paediatrics, internal medicine, and psychiatry. GPT-4 achieved median scores above the passing score in most disciplines, while GPT-3.5 fell below the passing score in all examinations. The study also noted reduced variance in test scores among GPT models compared to physicians, attributed to model stochasticity.

This study marks a significant advancement in AI technology by showcasing the capability of large language models (LLMs), particularly GPT-4, to achieve performance levels comparable to physicians on official medical board examinations. The research involved a thorough comparison between the performance of LLMs and that of 849 Israeli resident physicians across five core medical specialities. GPT-4 demonstrated superior performance in psychiatry, ranking above the median physician with a median 75th percentile among physicians. In internal medicine and general surgery, GPT-4 closely approached the median physician, while in paediatrics and OB/GYN, its performance was less impressive but still noteworthy. Conversely, GPT-3.5 consistently lagged behind both GPT-4 and the physicians across all specialties except psychiatry.

Building on Progress: Extending LLM Evaluation to Real Examination Settings

This study builds upon previous research that primarily focused on simulated medical scenarios, now extending to real examination settings. By including a large cohort of resident physicians and openly sharing examination data, the study provided a robust basis for comparing LLMs with human performance. While the integration of LLMs into clinical practice may still be in its early stages, their potential for enhancing medical education, simulations, personal assessment, and feedback evaluation methods is promising. Collaborative efforts between AI and physicians, as demonstrated in superior results in diagnosing radiologic imaging, highlight the synergy between AI and human expertise.

However, the study acknowledges limitations such as the exclusion of image-based questions and potential biases in translation from Hebrew to

English. Nevertheless, the impressive performance of GPT-4 suggests that the adoption of LLMs in clinical practice is on the horizon, paving the way for transformative changes in physician training and capabilities in conjunction with AI advancements.

Source & Image Credit: **New England Journal of Medicine AI**

Published on : Tue, 16 Apr 2024