Can GPT-4 Do Your Systematic Review? Large Language Models as a Research Assistant for Literature Reviews in Radiology

Dana Alkhulaifat, MD, Postdoctoral Research Fellow, Children's Hospital of Philadelphia
Satvik Tripathi; Suhani Dheer; Allison Brea; Yohan Kim; Dania Daye, MD, PhD; Tessa Cook, MD, PhD, CIIP, FSIIM

Background/Problem Being Solved

The increasing adoption of large language models (LLMs) in radiology research underscores the need to evaluate their utility in research workflows, particularly in evidence synthesis. This study focuses on assessing GPT-4's ability to perform data exploration and visualization tasks, critical components of systematic reviews, while emphasizing its potential to streamline research processes in radiology.

Intervention(s)

A systematic search was conducted across five databases: PubMed, EMBASE, SCOPUS, Web of Science, and IEEE Xplore. Boolean operators and targeted keywords, including "Large Language Models," "Radiology," and specific LLMs like GPT-4 and LLaMA, were used. Only original research articles published from 2022 onwards were included, excluding reviews, commentaries, editorials, and preprints. Two independent reviewers conducted the title and abstract screening, with adjudication by a third reviewer as needed. Data were extracted on application domains, specific LLMs used, and publication year. GPT-4 was employed to assist with data synthesis, and visualization in the context of systematic reviews, showcasing its ability to enhance efficiency and accuracy in these tasks.

Barriers/Challenges

Ensuring the accuracy and reliability of LLM outputs and validating the suitability of generated visualizations for scientific reporting remain key challenges.

Outcome

GPT-4 was evaluated for its performance in processing extracted data and generating figures such as bar plots, trend analyses, pie charts, and word clouds. The resulting visualizations accurately represented key trends, including adoption rates of LLMs, domain-specific applications, and keyword frequencies.

Conclusion/Statement of Impact/Lessons Learned

GPT-4 enhances systematic reviews by streamlining data exploration and visualization. However, limitations include potential biases, occasional inaccuracies, and inability to perform critical appraisal. Human oversight is essential, making GPT-4 a useful but complementary tool.

Figure(s)

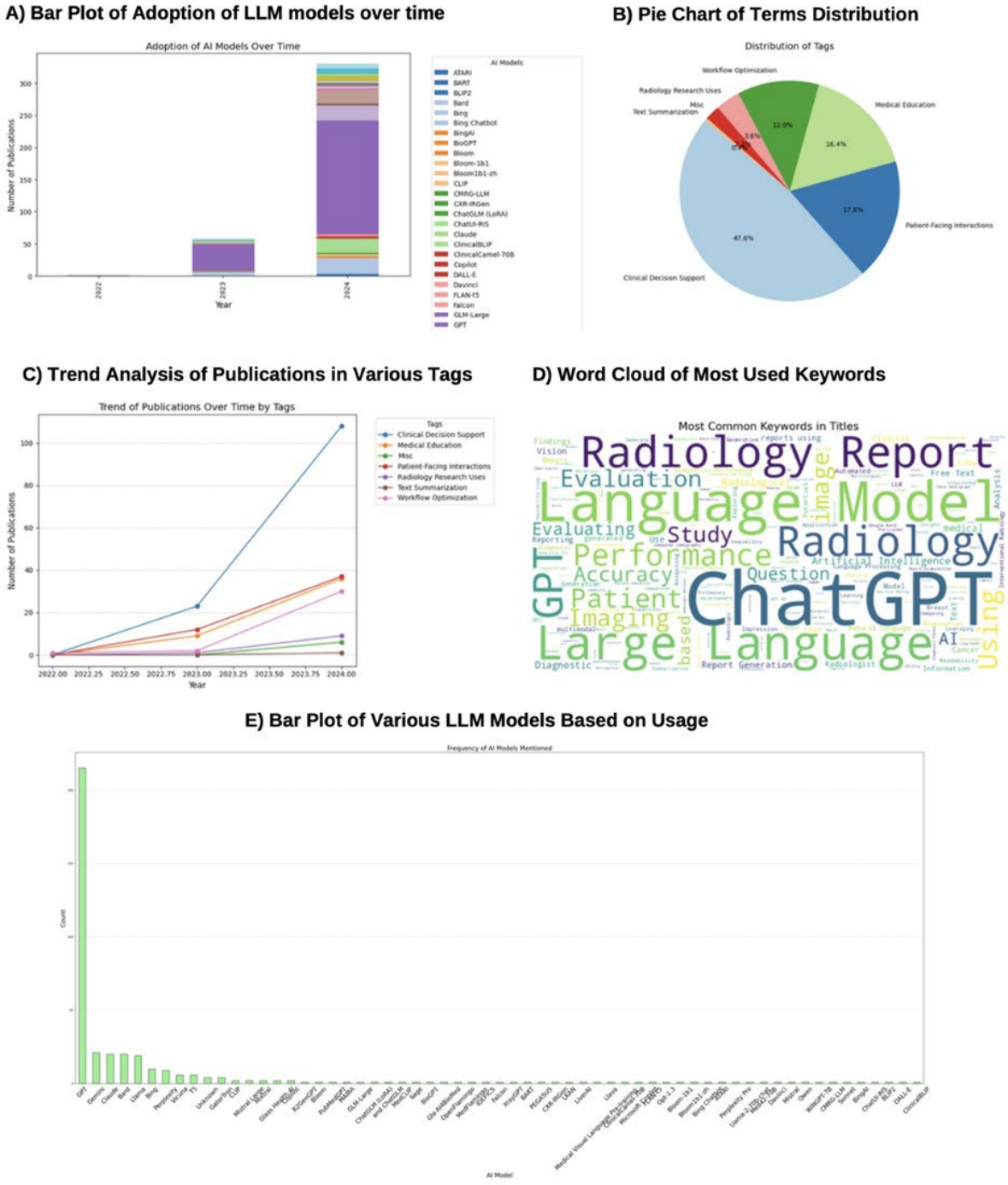

Figure 1. Example of various illustration figures generated using GPT-4 based on the results of the data extraction.

Keywords

Artificial Intelligence/Machine Learning