Comparing the Multimodal Performance of ChatGPT and Gemini Pro in Retinal Image Interpretation

David Mikhail[1], MD (C), Andrew Mihalache[1], MD (C), Ryan S. Huang[1], MD (C), Marko M. Popovic[2], MD MPH, Daniel Milad[3], MD,

[1]Temerty of Medicine, University of Toronto
[2]Department of Ophthalmology and Vision Sciences, University of Toronto
[3]Department of Ophthalmology and Vision Sciences, University of Montreal

Introduction: This paper aims to assess the performance of two new large language models (LLMs), ChatGPT-4 and Google's Gemini Pro, on retinal multimodal imaging interpretation. Specifically, this study compares the LLMs' diagnostic accuracy on a public dataset of retinal cases containing ophthalmic images and clinical data.

Methods: We systematically prompted ChatGPT-4 and Gemini Pro with a public dataset of 73 retinal cases, of which 64 cases were included, from the ophthalmology education website OCTCases.com from December 22, 2023 to December 24, 2023. Using the entire clinical case and ophthalmic images, we asked the LLMs: "What is the most likely diagnosis?" We developed a prompting algorithm to compare the factor(s) implicated in the LLMs' correct and incorrect diagnoses. We recorded the case characteristics, LLMs' responses to initial and follow-up prompting, response length, and the factors contributing to their responses. We reported the diagnostic accuracy of ChatGPT-4 and Gemini Pro in each case by comparing the LLMs' outputs with the answer key on OCTCases. Accuracy was the primary outcome and was measured as the proportion of correctly diagnosed cases from the total number of cases. The clinical characteristics that were contributory to decision-making of the LLMs was considered a secondary endpoint. Proportions of accuracies and contributory factors were compared between LLM models using a chi-squared (χ^2) test. Differences in performance were considered statistically significant at a p value of < 0.05.

Results: ChatGPT-4 achieved 39.0% diagnostic accuracy, while Gemini Pro achieved 20.3% diagnostic accuracy (χ^2, $p<0.05$). In correctly answered cases, imaging findings were the primary factor identified as most contributory to the decision-making of both ChatGPT-4 (40%) and Bard (53.8%) ($p>0.05$). In incorrectly answered cases, patients' age (39.2%) and imaging findings (43.6%) were most commonly implicated in decision-making by Gemini Pro and ChatGPT-4, respectively. ChatGPT-4 and Gemini Pro self-identified a mean of 5.2 and 3.8 factors contributing to their decision per case (Mann-Whitney U, $p>0.05$).

Conclusions: While the performance of both LLMs was overall poor, ChatGPT-4 outperformed Gemini Pro on multimodal analysis of clinical retinal cases. After further prompting, ophthalmic images were most frequently cited as the key factor in achieving correct diagnoses. Future research may consider testing LLMs on larger datasets to improve generalizability of results, and to compare LLMs with traditional ML models on image analysis and predictions of treatment outcomes.