SIIM CAIMI25
Conference on Artificial Intelligence in Medical Imaging
OCT 20-21 | UNIVERSITY OF CALIFORNIA, IRVINE

Revolutionizing AI in Radiology — Evaluating Open-source versus Proprietary LLMs: Deepseek versus GPT-4o on Radiological Images

Christopher Z. Liu, UCI School of Medicine; Shawn H. Sun, MD; Divya Balchander, MD; Roozbeh Houshyar, MD

Introduction/Background

Proprietary large-sized models like GPT-4o have demonstrated success in high-level radiological reasoning but have concerns with heavy resource usage and patient privacy concerns. DeepSeek Janus Pro-7B is a novel open-source vision-capable AI model that is smaller in size with similar reported performance. We aim to assess the vision capabilities of open-source DeepSeek and proprietary GPT-4o models, which allow analysis of image inputs in addition to textual data, on radiology cases.

Methods/Intervention

40 excerpts of radiology cases were utilized for analysis. 20 cases related to nuclear medicine and 20 neuroradiology cases were selected. The images from these cases were extracted with a standardized prompt to assess the DeepSeek and GPT-4o models on their analysis. The models were queried for the modality of imaging, type of radiotracer/contrast, body part and plane of imaging, specific findings, and the overall suggested diagnosis. A Likert Scale (1-5) was utilized to assess the model's findings and diagnosis, and binary correct/incorrect values were assigned to the other components of the output. McNemar's test was used to compare the binary categories, and the two-sample t-test was used for comparing Likert scores.

Results/Outcome

GPT-4o substantially outperformed DeepSeek for binary, lower-level analysis (Figure 1, $p < 0.01$). This trend continued for higher-level analysis — on average, GPT-4o achieved a score of 2.63 out of 5 for describing the findings in the images, whereas DeepSeek achieved 1.58 out of 5 ($p < 0.0001$). Both GPT-4o and DeepSeek improved their performance on the diagnosis portion of the output when provided with textual data containing patient history (GPT: 2.10 vs. 2.68 out of 5, $p = 0.011$; DeepSeek: 1.48 vs. 2.13, $p = 0.0006$).

Conclusion

GPT-4o outperformed DeepSeek across different radiology specialties and for both low-level and higher order analysis. The vision capabilities of open-source models currently are limited for both nuclear medicine and neuroradiology applications. Further studies on training open-source AI may improve clinical efficacy.

Statement of Impact

We provide the first head-to-head evaluation of open-source and proprietary vision-capable LLMs on radiological image interpretation, highlighting critical performance gaps that currently limit clinical viability of open-source models. These findings offer valuable insight for future development of open-source AI tools.

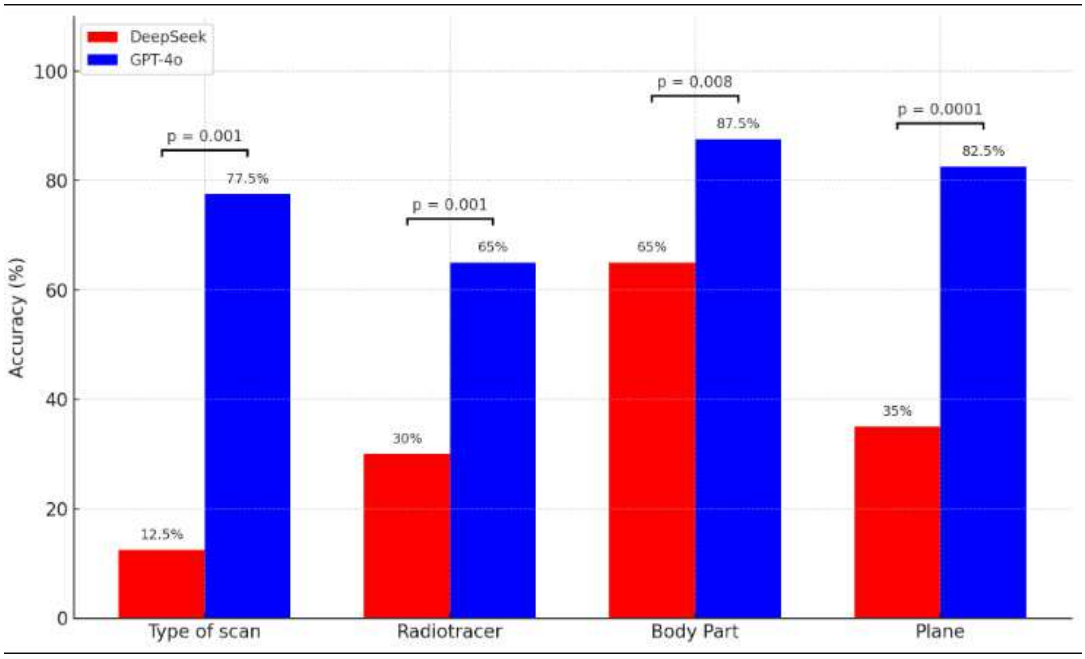

Figure 1: Graph shows accuracy for DeepSeek Janus Pro 7b vs GPT-4o for distinguishing modality of imaging, radiotracer used, body part imaged, and plane of imaging. McNemar's test was used to compare the binary accuracy categories.

Keywords

Large Language Models (LLM)**;** Open-source AI**;** Vision-capable Models**;** Comparative Model Evaluation**;** Radiology**;** Medical Imaging Analysis