

Accuracy of a Large Language AI Model for MRI Brain Interpretation

Dr Saranya Baleswaran, Dr Rahul Khamar, Dr Rachelle Rachmat, Dr Kasim Alimohamed
Guy’s and St Thomas’ NHS Foundation Trust

Introduction

Artificial intelligence is driving major changes across radiology, including opportunities to streamline report production and improve consistency. However, large language models used for radiology reporting have recognised limitations (including hallucinations and other reliability and governance concerns), reinforcing the need for careful validation and clinician oversight. [1] This study assessed the performance of an AI large language model system to generate conclusions for MRI head reports.

Methods

The sample consisted of 35 pre-existing MRI head reports. Using the findings from each report as input, conclusions were generated by an AI large language system. Output accuracy was rated on a 1–10 scale by comparison with the radiologist-authored conclusions (gold standard). Summary statistics were calculated, including mean accuracy, standard deviation (SD), and a 95% confidence interval (CI).

Results

The mean accuracy score for AI-generated conclusions was 7.09 (SD = 2.41, 95% CI 6.26–7.91), suggesting moderate agreement with radiologists’ conclusions. Qualitatively, the AI system showed limitations in maintaining the concise, prioritized style expected of radiology conclusions. Outputs tended to be more verbose, with variable emphasis on key positive findings and less consistent “clinical framing” (eg, explicitly recommending correlation with symptoms, follow-up pathways, escalation to the relevant specialty team, discussion in an appropriate multidisciplinary meeting, or consideration of patient risk factors and preventive management where applicable).

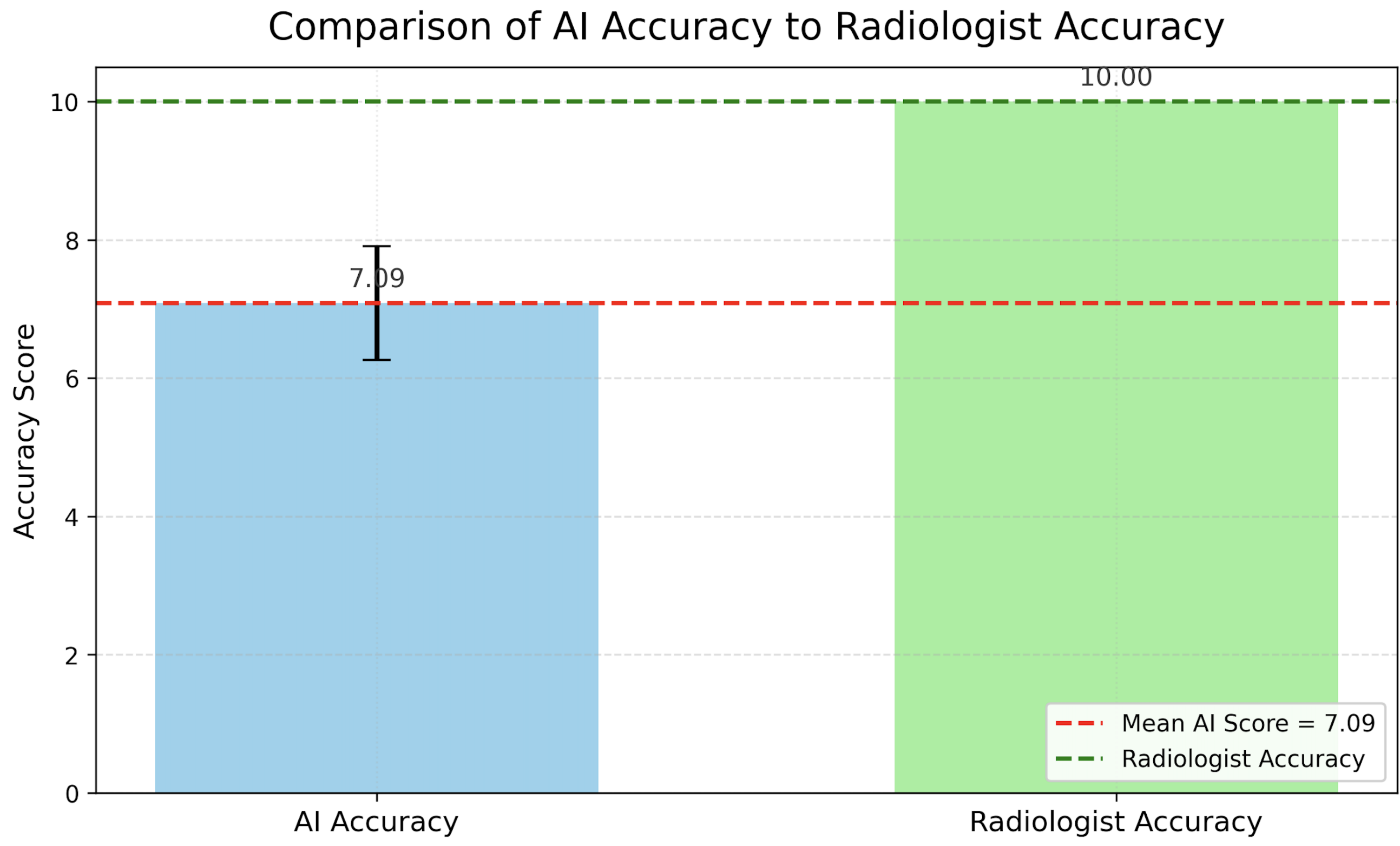


Figure 1:
AI Accuracy vs Radiologist Accuracy

Conclusion

AI generated generally accurate summaries of MRI head findings; however, the resulting conclusions often lacked clear clinical direction and were typically more verbose than those produced by reporting radiologists. These results support a role for LLMs as an assistive drafting tool, but highlight the need for further refinement (eg, stronger constraints on length/structure, improved prioritization of salient abnormalities, and better integration of clinical context). Continued expert radiologist input and oversight will remain essential to ensure safe, clinically useful outputs.

References

1. Parillo M, Vaccarino F, Beomonte Zobel B, Mallio CA. ChatGPT and radiology report: potential applications and limitations. Radiol Med. 2024 Dec;129(12):1849-1863. doi: 10.1007/s11547-024-01915-7. Epub 2024 Nov 7. PMID: 39508933.

Contact

Dr Saranya Baleswaran
Guy’s and St Thomas NHS Foundation Trust
Email: Saranya.baleswaran@nhs.net
Website: <http://www.guysandstthomas.nhs.uk>

