SIIM CMIMI24
Oct 21-22 | Boston, MA
Conference On Machine Intelligence In Medical Imaging

Radiology AI Leaderboard: An Evaluation Platform for Large Language and Vision Language Models

David Li, MD, London Health Sciences Center; Jaron Chong, MD, MHI

Introduction/Background

Rapid advancements in large language models (LLMs) and vision language models (VLMs) hold great promise for transforming radiology. However, assessing and comparing the performance of these models in radiology remains challenging due to the lack of standardized, transparent benchmarks. To address this gap, we created a comprehensive platform designed to evaluate and compare LLM and VLM performance in radiology tasks.

Methods/Intervention

The platform features an evaluation and voting framework with domain-specific criteria to ensure accurate performance assessment. It supports both public and proprietary datasets, including multimodal datasets. Visualization tools enable radiologists to easily compare model performance across various metrics, datasets, and tasks over time. Researchers and vendors are encouraged to submit their models for evaluation.

Results/Outcome

The platform has demonstrated both feasibility and effectiveness in evaluating LLMs and VLMs for radiology tasks. Models are assessed across a range of radiology-specific tasks and datasets, with performance transparently reported and ranked. While initial results were based on academic research, we have also evaluated 19 models with board-certified radiologists. The latest proprietary models, such as GPT-4o and Claude 3.5 Sonnet, as well as open-source models like LLaMA 3.1 405B, have been benchmarked. Preliminary results indicate that model performance on radiology-specific tasks differs substantially from general-purpose benchmarks, highlighting the need for radiology-specific benchmarks.

Conclusion

The Radiology AI Leaderboard represents a major advancement in standardizing the evaluation of LLMs and VLMs within radiology. It addresses a critical gap by introducing specialized benchmarks tailored to radiology, setting new standards for transparency and collaboration. The platform not only improves the accuracy of performance evaluations but also establishes a robust foundation for the safe and effective integration of AI into clinical practice.

Statement of Impact

This platform advances the evaluation of LLMs and VLMs in radiology by providing standardized and transparent benchmarking. By ensuring rigorous and equitable assessments, it facilitates the integration of generative AI into clinical practice.

Keywords

Large language models; Vision language models; Model validation; Benchmark; Leaderboard