RADIUS: Community-Driven Radiology AI Large Language Model and Vision Language Model Leaderboard

David Li, MD; Jaron Chong, MD, MHI

Schulich
MEDICINE & DENTISTRY

Western

Disclosures

- DL: No relevant relationships.

- JC: Chair of the Canadian Association of Radiologists Artificial Intelligence Standing Committee and board member for AMS Healthcare.

Introduction

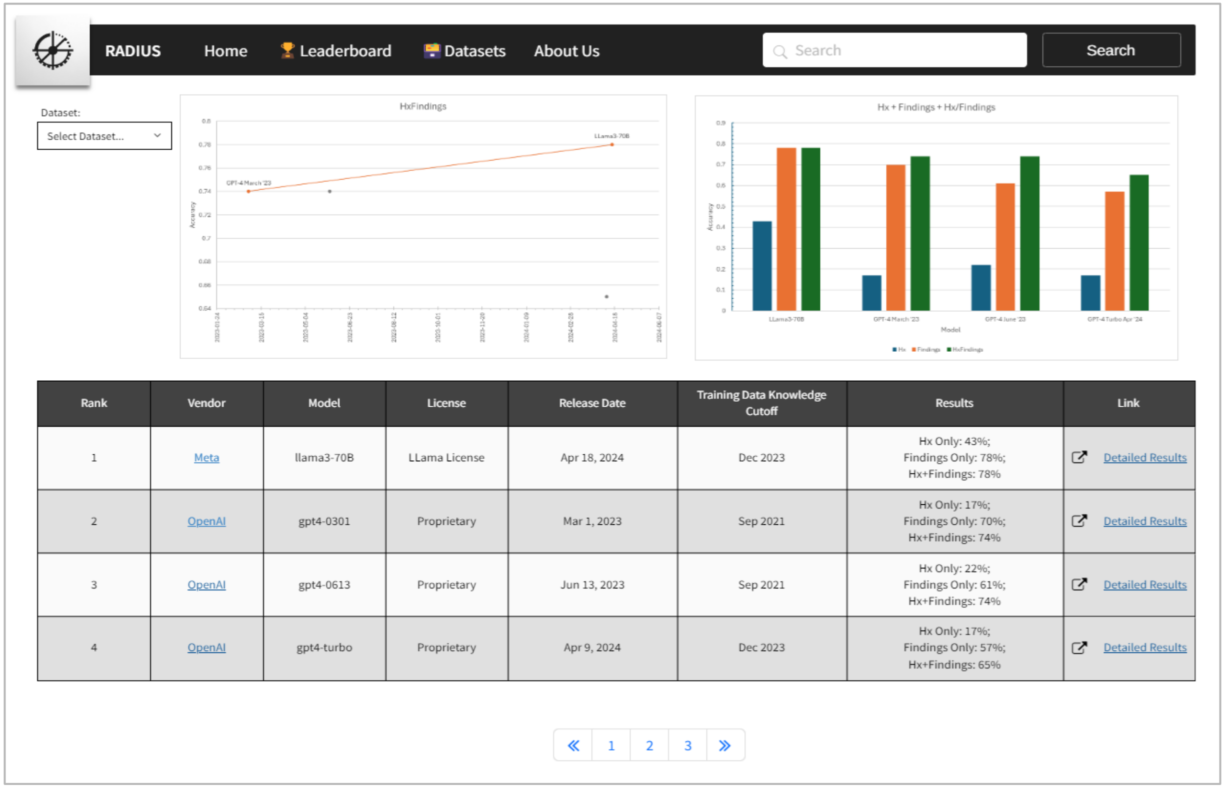
- Exponential growth in large language models (LLMs) and vision-language models (VLMs) presents tremendous potential to transform radiology

- Yet, evaluating and comparing model performance in radiology remains challenging

Methods

- Standardized platform for evaluating and comparing models across diverse radiological tasks and datasets, integrating both text and images
- Evaluation framework developed in collaboration with radiologists featuring domain-specific criteria
- Initialized using published results and is open to contributions from both academic and industry partners

RADIUS



Rank	Vendor	Model	License	Release Date	Training Data Knowledge Cutoff	Results	Link
1	Meta	llama3-70B	LLama License	Apr 18, 2024	Dec 2023	Hx Only: 43%; Findings Only: 78%; Hx+Findings: 78%	Detailed Results
2	OpenAI	gpt4-0301	Proprietary	Mar 1, 2023	Sep 2021	Hx Only: 17%; Findings Only: 70%; Hx+Findings: 74%	Detailed Results
3	OpenAI	gpt4-0613	Proprietary	Jun 13, 2023	Sep 2021	Hx Only: 22%; Findings Only: 61%; Hx+Findings: 74%	Detailed Results
4	OpenAI	gpt4-turbo	Proprietary	Apr 9, 2024	Dec 2023	Hx Only: 17%; Findings Only: 57%; Hx+Findings: 65%	Detailed Results

Figure 1: RADIUS website for evaluation and comparison of LLMs on differential diagnoses generation task. Models are ordered by dataset-specific performance metrics plotted over time and by model.

RADIUS

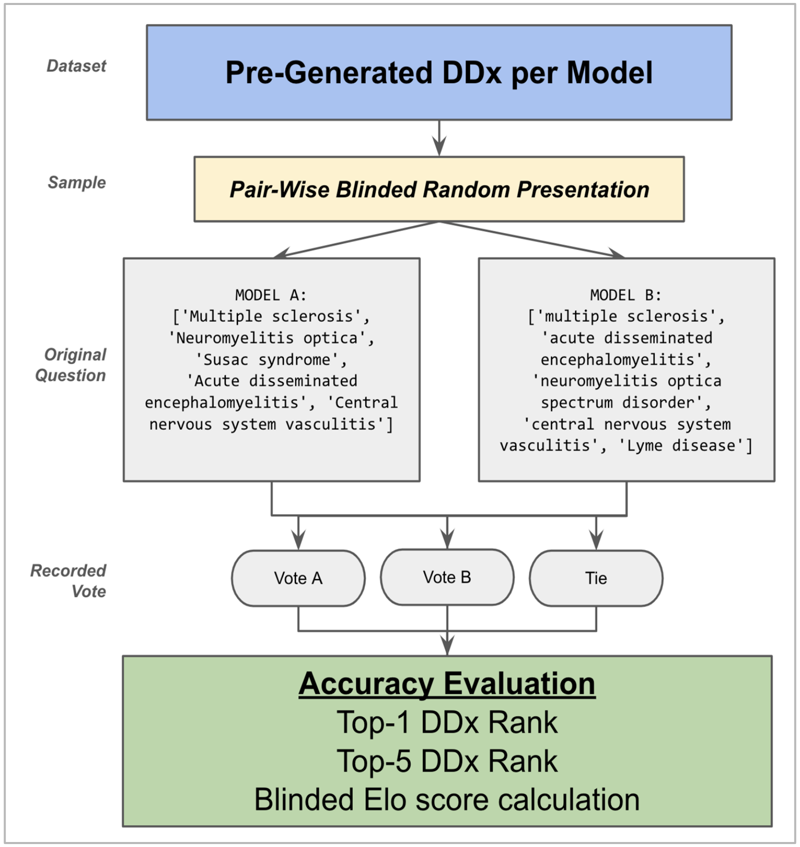

Dataset
Pre-Generated DDx per Model
Sample
Pair-Wise Blinded Random Presentation
Original Question
MODEL A:
['Multiple sclerosis', 'Neuromyelitis optica', 'Susac syndrome', 'Acute disseminated encephalomyelitis', 'Central nervous system vasculitis']
MODEL B:
['multiple sclerosis', 'acute disseminated encephalomyelitis', 'neuromyelitis optica spectrum disorder', 'central nervous system vasculitis', 'Lyme disease']
Recorded Vote
Vote A
Vote B
Tie
Accuracy Evaluation
Top-1 DDx Rank
Top-5 DDx Rank
Blinded Elo score calculation

Figure 2: Blinded voting system for radiologists to evaluate model performance on subjective interpretive tasks.

Results

- Radiologists evaluated multiple proprietary and open-source models using a blinded voting system
- Model performance was transparently reported and ranked, accompanied by longitudinal analysis
- RADIUS offers a viable and effective approach for addressing the challenge of evaluating and comparing model performance in radiology

Discussion

- RADIUS promotes fairness, transparency, and collaboration within the radiology and AI communities
- This represents a step towards standardizing model evaluation in radiology
- Ongoing development will support further quality improvement in clinical applications of generative AI

Thank you