LLM Showdown: Benchmarking Performance, Cost, and Speed of Granular Ground Truth Extraction in Radiology Reports for Post-Deployment Monitoring of AI Models

Aawez Mansuri, MS, Systems Software Engineer, Emory University
Theo Dapamede, MD, PhD; Hanssen Li, MD; Wasif Bala, MD; John Moon, MD; Bardia Khosravi, MD, MPH, MHPE; Chad Robichaux, MPH; Frank Li, PhD; Mohammedreza Chavoshi, MD; Beatrice Brown-Mulry; Rohan Isaac, MS; Dan Cohen, MD; Ninad Salastekar, MD; Judy Gichoya,

Introduction

Large language models (LLMs) show promise for extracting granular clinical detail from radiology reports, yet their performance, cost, and inference speed relative to smaller models remain unclear. We evaluated large and small open-source and proprietary LLMs—including GPT-4o, GPT-4o-mini, Meta Llama 3.1 8B, Llama 3.1 70B, Llama 3.3 70B, Microsoft Phi 3.5-mini, and Phi 3.5-moe—on extracting granular labels of clinically relevant subtypes from intracranial hemorrhage (ICH) and pulmonary embolism (PE) reports. This includes information on acuity, location/depth, size, and presence of complications. By comparing model outputs to human-annotated ground truths and assessing cost-to-performance trade-offs, we provide insights to guide clinical NLP deployment.

Hypothesis

We hypothesized that proprietary models (e.g., GPT-4o, GPT-4o-mini) would surpass open-source counterparts and that smaller variants (e.g., GPT-4o-mini) would achieve near-equivalent performance to larger models at lower cost.

Methods

We selected 600 radiology reports (300 ICH and 300 PE) each annotated by three radiology residents for ground truth. Models received identical prompts, and outputs were compared to annotations. We measured inference times and evaluated cost-to-performance for proprietary models using token-based pricing.

Results

GPT-4o excelled in extracting detailed ICH labels with variable performance for PE. GPT-4o-mini delivered comparable accuracy at lower cost and faster inference. Among open-source models, Llama 3.3 70B emerged as top performer, exceeding smaller open-source variants (Llama 3.1 8B, Phi 3.5-mini), but at a high computational cost. Overall, GPT-4o-mini offered a strong balance of accuracy, speed, and cost, while most smaller models did not match their larger counterparts' performance.

Conclusion

This benchmarking shows that proprietary models like GPT-4o lead in label extraction, but smaller, more cost-effective options like GPT-4o-mini achieve nearly the same accuracy at lower cost. Although Llama 3.3 70B performs well among open-source models, its high computational demands limits practical use. Ultimately, selecting an LLM for post-deployment radiology AI monitoring should consider accuracy, cost, and speed, leveraging these insights for more balanced, resource-conscious decisions.

Figure(s)

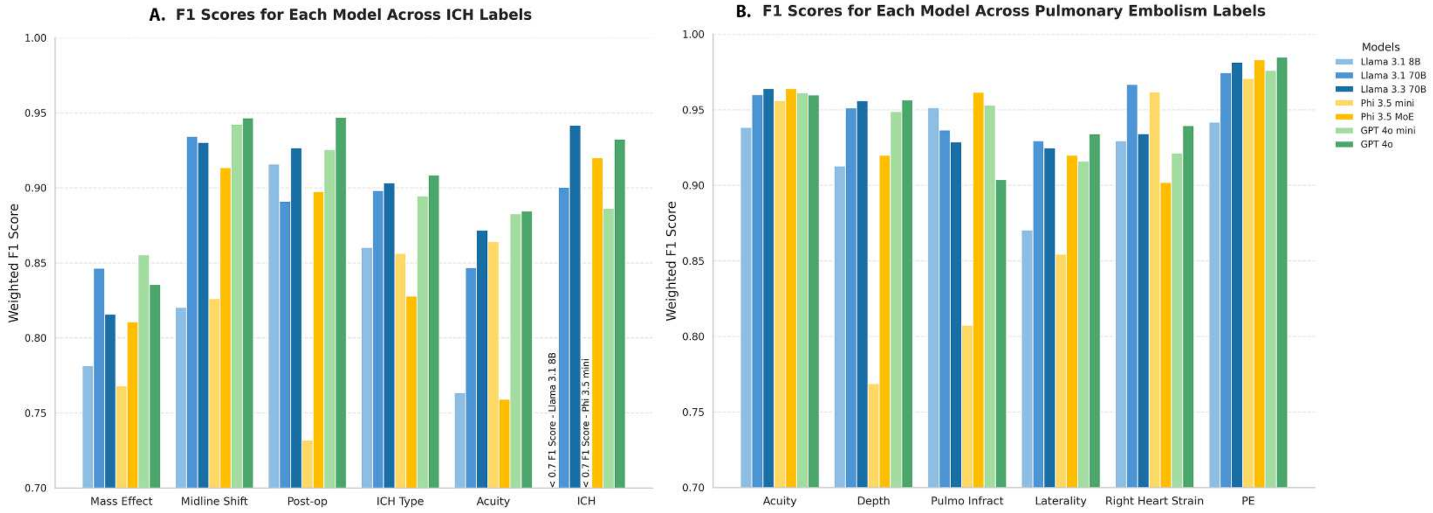
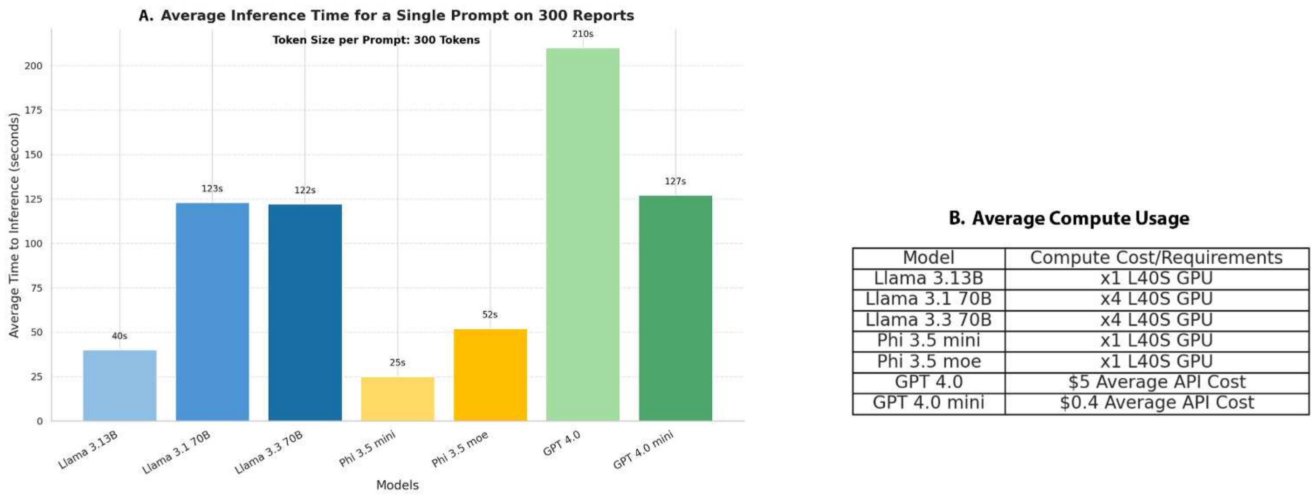


Figure 1. Bar plots display the Weighted F1 Scores for various labels across different LLMs for (A) ICH cases and (B) PE cases.



Model	Compute Cost/Requirements
Llama 3.13B	x1 L40S GPU
Llama 3.1 70B	x4 L40S GPU
Llama 3.3 70B	x4 L40S GPU
Phi 3.5 mini	x1 L40S GPU
Phi 3.5 moe	x1 L40S GPU
GPT 4.0	$5 Average API Cost
GPT 4.0 mini	$0.4 Average API Cost

Figure 2A presents the average inference time per single prompt across 300 reports (inference time for a single label) for different LLMs, while B shows the average compute usage.

Keywords

Artificial Intelligence/Machine Learning; Clinical Workflow & Productivity; Quality Improvement & Quality Assurance

Scientific Research & Applied Informatics Posters and Demonstrations