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

Evaluation of Llama2 and Llama3 for Automated Extraction of Ground Truth from Radiology Reports for Post-Deployment Monitoring of Pulmonary Embolism and Intracranial Hemorrhage Detection AI Models

Theo Dapamede, MD, PhD, Emory University; Bardia Khosravi, MD, MPH, MHPE ; Chad Robichaux, MPH; Aawez Mansuri, MS; Mohammadreza Chavoshi, MD; Alex Belov; Angela Udongwo, MPH; Chinonyelum Igwe; Frank Li, PhD; Beatrice Brown-Mulry; Hanssen Li, MD; John Moon, MD; Judy Gichoya, MD, MS, FSIIM; Hari Trivedi, MD

Introduction/Background

Clinical use of AI models requires post-deployment monitoring for performance and potential drift. However, this requires comparison of model outputs to ground-truth radiologist interpretations which can be laborious. We evaluate the performance of 2 generations of open-source large language models (LLM) for label extraction tasks for pulmonary embolism (PE) and intracranial hemorrhage (ICH) against human annotated ground truths.

Methods/Intervention

We identified 4,668 CT PE exams and 74,394 non-contrast CT head exams from 2020-2022 and randomly sampled 250 reports for each exam type for manual annotation. PE labels were: PE, acuity, laterality, largest depth, right heart strain, and pulmonary artery hypertension. ICH labels were: ICH, acuity, laterality, subtype, midline shift, and mass effect. Reports were annotated by 6 human annotators using a browser-based interface and difficult cases were flagged for review by a senior radiologist. Multiple prompt styles were tested in preliminary analysis using Llama 2 7B. The top performing prompting style was selected and used to evaluate Llama2 (7B, 13B, and 70B) and Llama3 (8B and 70B) models.

Results/Outcome

Llama3 8B had the highest overall performance for both PE (sensitivity: 1.0; specificity: 1.0) and ICH (sensitivity: 0.93; specificity: 1.0). Across all models, performance for PE depth (accuracy range: 0.25-0.61) and ICH acuity (accuracy range: 0.63-0.74) were lowest. Llama2 performance improved with increasing parameters for most classes. However, Llama3 8B and 70B performance was similar across all categories. Llama3 8B significantly outperformed Llama2 7B for all labels, despite similar parameter sizes.

Conclusion

This study evaluated Llama2 and Llama3 models to extract labels for PE and ICH against human annotated ground truths. Llama3 8B had the highest performance with significant improvements over Llama2. Model performance for extracting binary PE and ICH labels was robust, however no model was able to successfully extract subgroup labels for PE or ICH to acceptable accuracy.

Statement of Impact

LLMs are a promising tool for post-deployment monitoring of AI models and can successfully extract binary ground

truth from ICH and PE radiology reports for comparison to AI model predictions. If properly tuned, these models may also allow for robust subgroup evaluation to deliver further insights into model performance.

Figures

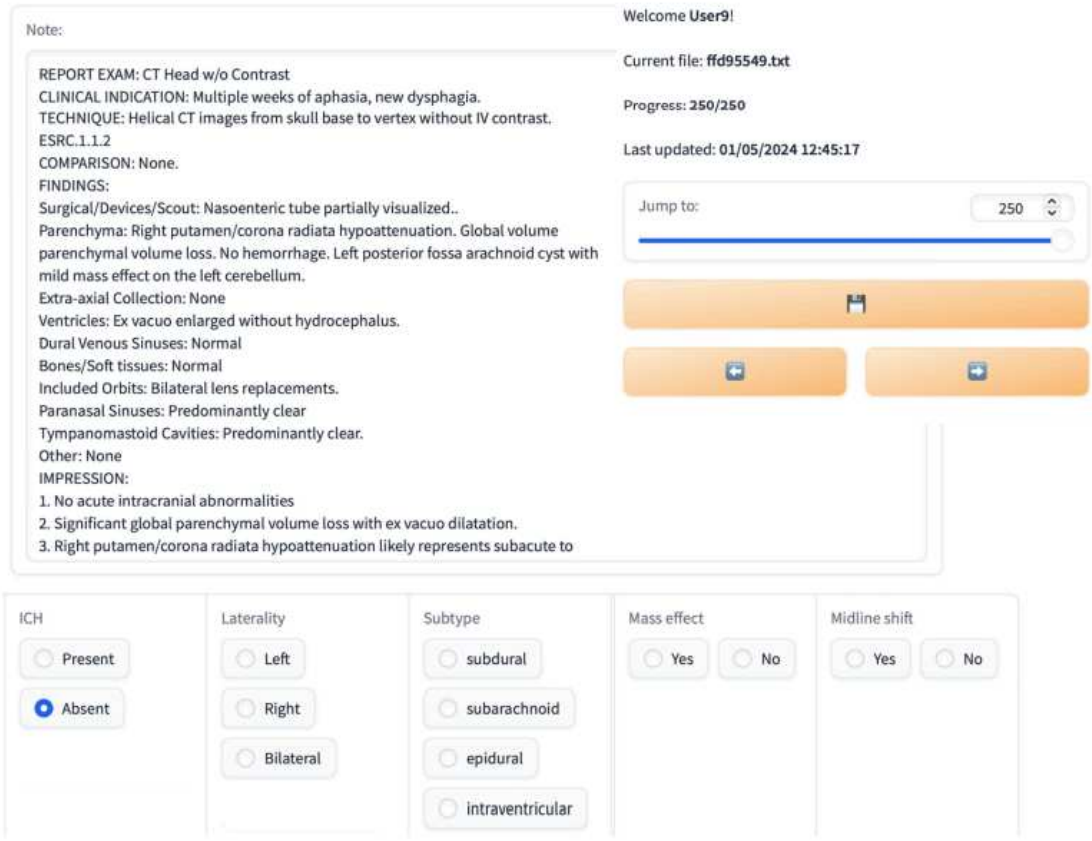

Figure 1. Browser-based annotation interface for ICH. Not all subgroups and labels are shown due to size constraints.

Browser-based annotation interface for ICH

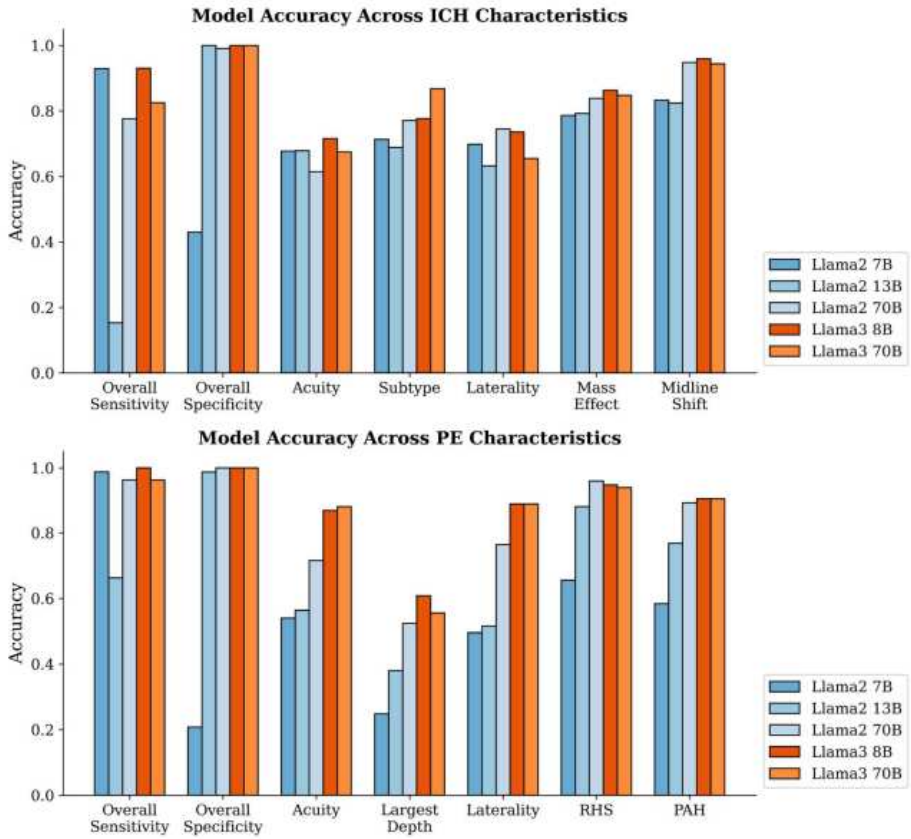

Figure 2. Overall and subgroup performance of label extraction for PE and ICH ground truths across Llama2 and Llama3 models.

Overall and subgroup performance of label extraction for PE and ICH ground truths across Llama2 and Llama3 models

Keywords

Llama; Pulmonary Embolism; Intracranial Hemorrhage