Practical Use of LLMs to Adjudicate Radiology Reports to Assess Performance of a Clinical AI Triage Tool

Adam Flanders, MD, CIIP, FSIIM, Vice-Chair Imaging Informatics, Radiology, Thomas Jefferson University
Paras Lakhani, MD; Prahlad Menon, PhD; Robyn Ball, PhD; George Shih, MD MS; Luciano Prevedello, MD

Introduction

The majority of commercial computer vision vendors provide limited data to customers to evaluate local performance of deployed solutions. Legislation mandates that customers monitor performance of these systems for drift/bias yet the effort needed to monitor these systems at scale is not trivial and can require substantial resources. Large language models (LLMs) applied to the diagnostic report may be useful in automating this process.

Hypothesis

To determine whether an ensemble of LLMs can be used effectively to monitor a commercial triage AI system.

Methods

The AI inference results for a commercial CT brain hemorrhage detector and the diagnostic reports were retrospectively collected on 16,172 ED and outpatient exams derived from 18 hospitals and 35 CT scanners in two states. The diagnostic reports were parsed to extract only the impression section and were presented to an ensemble of five LLMs (llama3.2:1b, llama3.2:3b, codellama:7b, llama3.1:8b, granite3-dense:2b) employing a single-shot prompt to confirm if presence of hemorrhage was documented in the report. The results returned presence of hemorrhage and if hemorrhage was present, the hemorrhage subtype. Each LLM was compared to the consensus (majority vote) across LLMs and to a subset of manually reviewed exams. Finally, the LLM consensus was compared to the manually reviewed exam set to assess overall performance.

Results

Agreement among the five LLMs varied considerably, with kappas ranging from 0.1 to 0.79. The smallest LLM (nlpheme1) performed substantially worse compared to both the LLM consensus and the manually reviewed exam set. Similarly, there was a wide range of Cohen's kappa (0.16-0.9) comparing each LLM to the consensus with the smaller sized model as the outlier. Comparison of the consensus to a random subset of 390 manually reviewed reports showed agreement of 0.68 which was augmented to 0.75 after removal of the two low performing LLMs. There was no substantial difference in F1 score for the ICH AI model using the two consensus schemes or the single best performing LLM.

Conclusion

A consensus ensemble of LLMs reviewing reports may have promise in verifying AI model performance in an automated fashion. Model selection, custom prompt engineering and manual verification are critical in ensuring useful results.

Figure(s)

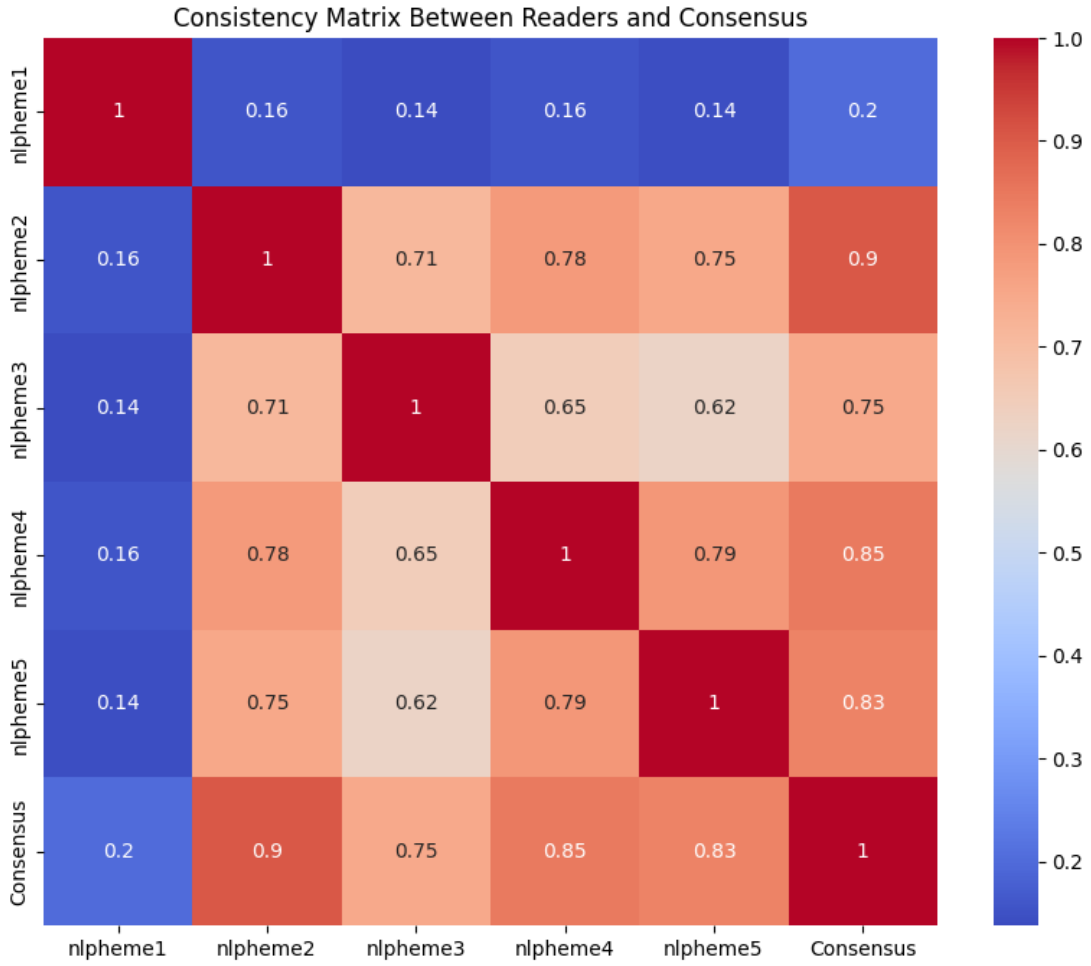

Figure 1. Confusion matrix showing Cohen Kappa values for each of the five LLMs (nlpheme1 to nlpheme5) compared to consensus where three or more LLMs were concordant.

Keywords

Administration & Operations; Applications; Artificial Intelligence/Machine Learning; Clinical Workflow & Productivity; Enterprise Imaging; Quality Improvement & Quality Assurance