Enhancing Imaging Appropriateness in Acute Care: Aligning Large Language Models with the American College of Radiology Appropriateness Criteria

Hersh Sagreiya, MD, Assistant Professor of Radiology, Radiology, University of Pennsylvania
Michael Yao; Charles Kahn, MD, MS, FACR; Walter Witschey, PhD; James Gee, PhD; Osbert Bastani, PhD; Allison Chae

Introduction

Diagnostic imaging plays an essential role in managing acute patient care, yet many ordered studies are misaligned with established medical guidelines, such as the American College of Radiology (ACR) Appropriateness Criteria (AC). As a result, many ordered imaging studies lead to unnecessary costs, patient risk, and a burden to the overloaded healthcare system. This study explores the potential of large language models (LLMs) as clinical decision support tools to help clinicians order more appropriate imaging studies in acute healthcare settings.

Hypothesis

Prior work has primarily focused on leveraging language models to directly assign imaging studies to input patient case descriptions. In contrast, our method instead asks an LLM to predict the most appropriate ACR AC guideline title, or "Topic," to describe an input patient scenario (Fig. 1A). Separately, we then parse through the textual ACR guidelines to determine the most appropriate, evidence-based imaging study(s) based on the predicted ACR AC Topic. We hypothesize that this inference strategy will enable LLMs to more accurately recommend appropriate imaging studies for acute patient presentations.

Methods

To experimentally validate our novel LLM inference strategy, we empirically evaluate six state-of-the-art LLMs on their ability to predict the correct ACR AC Topic label for an input patient case (Fig. 1B). We then further improve the performance of LLMs using zero-shot prompting techniques such as retrieval-augmented generation and chain-of-thought prompting.

Results

Our results demonstrate that our novel inference strategy can improve the accuracy of LLMs in predicting the most appropriate diagnostic imaging study by up to 50%. In a retrospective study with real patient case descriptions, autonomous LLM agents ordered more accurate imaging studies while simultaneously reducing the rate of unnecessary imaging orders compared with physicians (Fig. 2).

Conclusion

Overall, our results underscore the potential of AI-driven support to improve clinical workflows. Future work will explore how similar LLM inference strategies can be applied to other areas of evidence-based medicine, where adherence to guidelines is critical for quality care.

Figure(s)

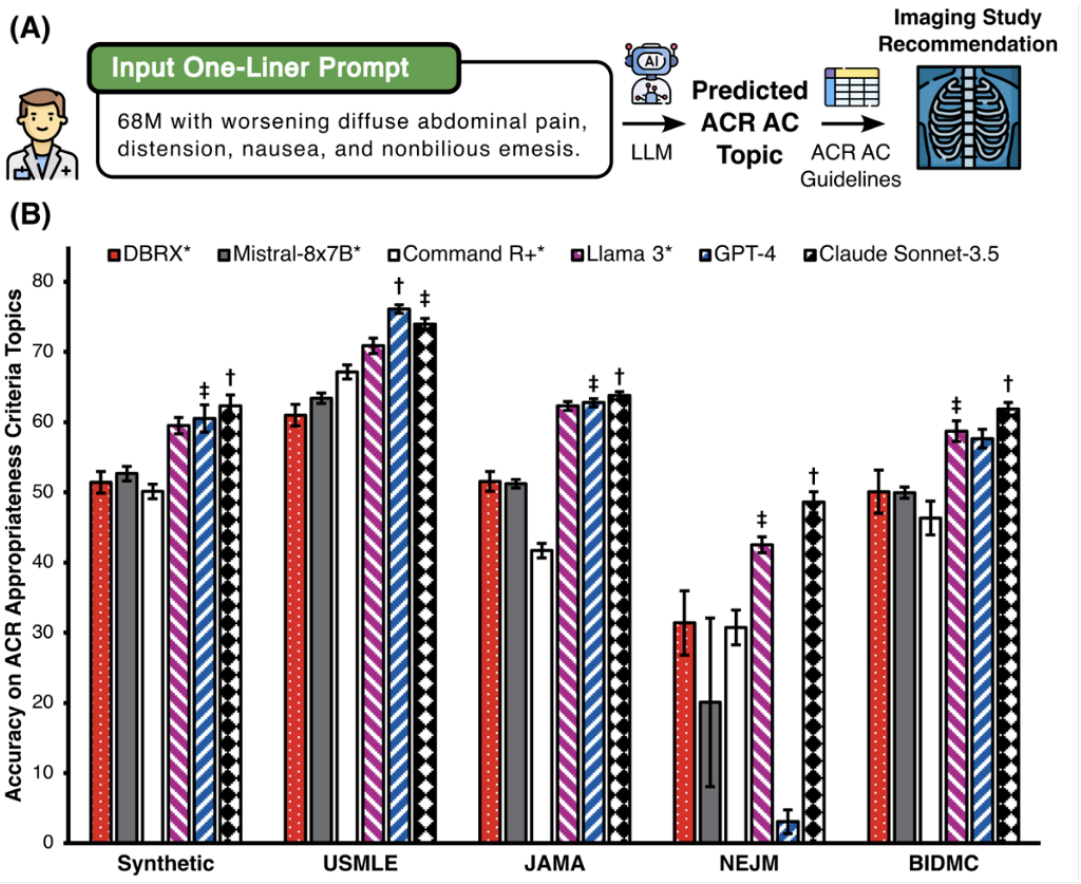

Figure 1. LLM Performance on the RadCases Dataset. (A) To align LLMs with the evidence-based ACR Appropriateness Criteria (AC), we query a language model to return the most relevant ACR AC Topic (a 224-way classification task) given an input patient one-liner description. We then programmatically query the ACR AC to deterministically return the most appropriate diagnostic imaging study (or lack thereof) given the predicted topic. (B) We evaluate six state-of-the-art language models on their ability to correctly identify the ACR AC Topic most relevant to a patient one-liner. Open-source models are identified by an asterisk, and the best (second best) performing model for a RadCases dataset partition is identified by a dagger (double dagger). Error bars represent a ± 95% CI over n = 5 independent experimental runs.

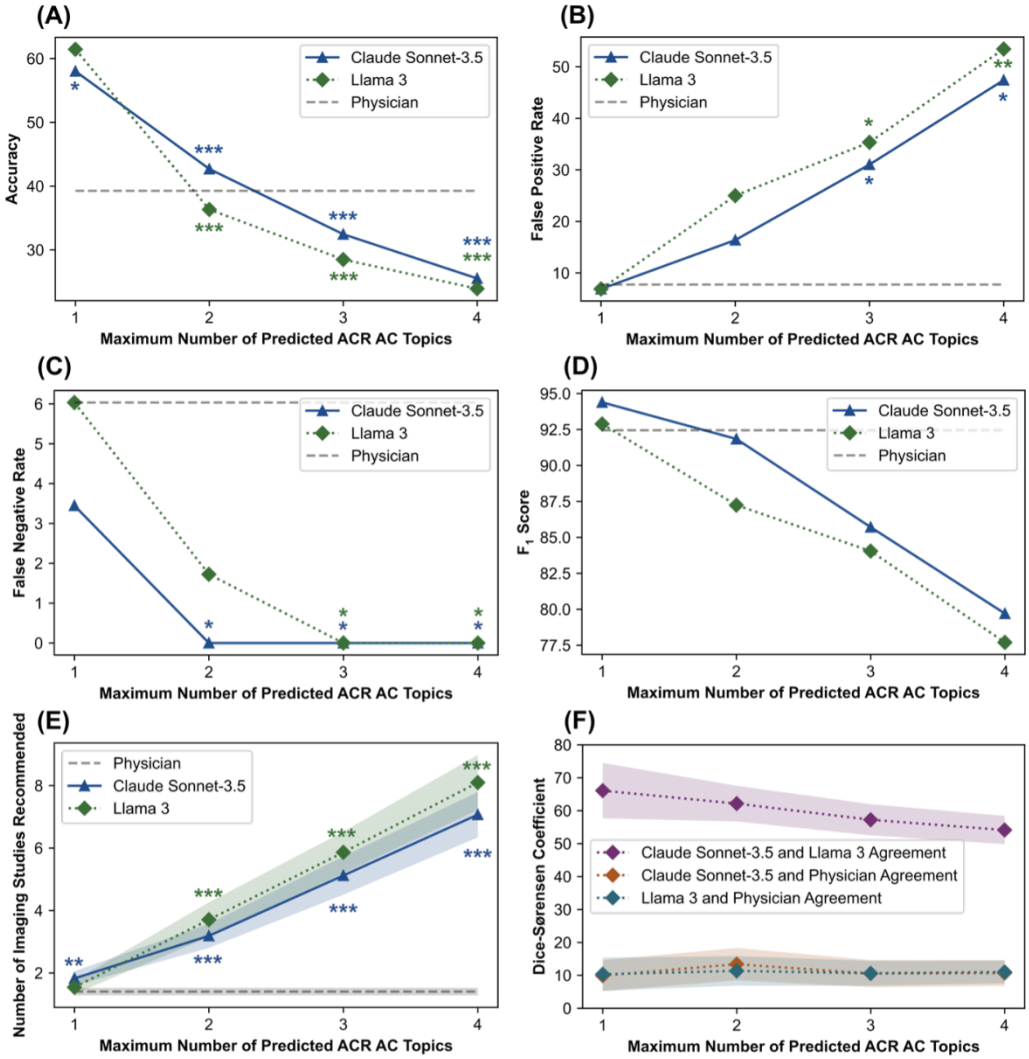

Figure 2. Retrospective Study of Clinician-Ordered versus LLM-Ordered Imaging Studies. We compare the diagnostic imaging studies ordered by the prompt-optimized LLMs Claude Sonnet-3.5 and Llama 3 against those ordered by clinicians in a retrospective study. We vary the maximum number of ACR AC Topic predictions requested from each language model on the x-axis. Compared with clinicians, Claude Sonnet-3.5 and Llama 3 achieve better (A) accuracy scores; and (C) false negative rates (i.e., the rate at which a patient should have received an imaging workup but did not); and also achieve (B) false positive rates (i.e., the rate of unnecessary imaging studies); (D) F1 scores (when x = 1); and (E) number of recommended imaging studies that are noninferior to that of clinicians. (F) According to the Dice-Sørensen Coefficient (DSC) metric, Claude Sonnet-3.5 and Llama 3 order imaging studies that are more similar to one another than to clinicians across all values of x (two sample, two-tailed homoscedastic t-test; $p < 0.0001$). Single, double, and triple asterisks represent significance values of $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively.

Keywords

Applications; Artificial Intelligence/Machine Learning; Clinical Workflow & Productivity