The Effects of Prompt Engineering and Consensus on Identification of Incidental Breast Findings by Large Language Models from Radiology Reports

Benjamin Rush, PhD, MPH, Informatics Data Scientist, Department of Radiology, University of Wisconsin-Madison
Thanh Nguyen; Kayla Berigan, MD; Ryan Woods, MD, MPH; John Garrett, PhD

Introduction

Radiology reports of imaging scans are unstandardized, leading to about 36.6% of incidental findings not receiving follow-up within 1 year. In addition, 7% of chest CT scans have incidental findings, of which 28% are malignant. Large language models (LLMs) can analyze reports and potentially identify cases for follow-up. Experiments testing LLMs' capabilities of analyzing reports exist yet test few types of LLMs and prompts. The effects of incremental prompts, LLM size, and LLM training data remain largely unexplored. We compared the performance of identifying incidental breast findings from radiology reports by using multiple prompts for individual LLMs and by consensus on cases between LLMs.

Hypothesis

Optimizing prompting for incidental breast findings in radiology reports detection using consensus approaches can improve sensitivity and specificity.

Methods

We randomly selected 500 exams with "breast" in the radiology report from chest CTs obtained at our institution between 2015-2017 from female patients ages 40-72. We compared the performance of 126 combinations from 7 LLMs, 9 incremental prompts, and 2 LLM roles when identifying incidental breast findings in reports compared to a breast imaging fellow reader. Combinations and consensus between LLMs were evaluated by sensitivity, positive predictive value (PPV), specificity, negative predictive value (NPV).

Results

The reader identified 31 (6%) cases with incidental breast findings while individual LLM combinations ranged from identifying 98-478 cases with 0.67-1.00 sensitivity, 0.05-0.84 specificity, PPVs not exceeding 0.23, and NPVs above 0.95. Consensus on case identification reduced false positives: the consensus of 3 highly sensitive combinations identified 86 (17%) cases with 0.87 sensitivity, 0.31 PPV, 0.87 specificity, and 0.99 NPV.

Conclusion

Consensus from highly sensitive LLM, prompt, and role combination generally increased performance, though an optimization algorithm selecting high sensitivity combinations with dissimilar positive case labelling would likely further increase performance.

Figure(s)

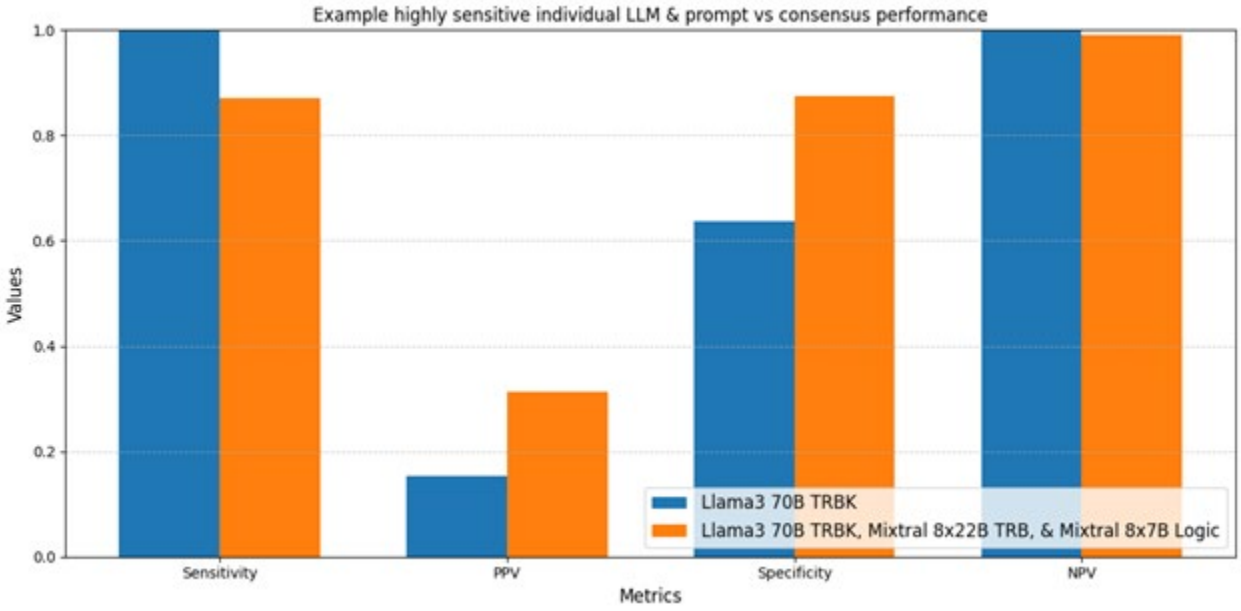

Figure 1. Incidental breast finding identification performance of Llama3 70B with a prompt that included the task to be completed (T), radiology report (R), background on incidental findings (B), and keywords that could indicate incidental findings (K) with a role of a highly skilled radiologist. This combination identified 201 (40%) of cases having incidental findings compared to 31 (6%) by the reader. The LLMs and prompts in the consensus combination were: 1. Llama3 70B with a prompt that included the task to be completed (T), radiology report (R), background on incidental findings (B), and keywords that could indicate incidental findings (K) with a role of a highly skilled radiologist; 2. Mixtral 8x22B with a prompt that included the task to be completed (T), radiology report (R), background on incidental findings (B)) with a role of a highly skilled radiologist; 3. Mixtral 8x7B with a prompt that was a summarized set of logical rules to follow to identify incidental findings with a role of a highly skilled radiologist. This combination identified 86 (17%) of cases having incidental findings compared to 31 (6%) by the reader. PPV=positive predictive value, NPV= negative predictive value.

Keywords

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