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

The Effect of Prompt Elements on Labelling Incidental Breast Findings by Llama3-8B in Radiology Reports

Benjamin Rush, PhD, MPH, University of Wisconsin-Madison; John Garrett, PhD; Thanh Nguyen;
Ryan W. Woods, MD, MPH

Introduction/Background

Breast cancer is a leading cause of mortality among women, and early detection can improve survival probability. Incidental breast abnormalities are identified in approximately 7% of chest CT scans, of which about 28% are malignant. However, the radiology reports from CT scans are often lengthy, unstandardized text where incidental findings might be overlooked by physicians. We propose using large language models (LLMs) to label CT radiology reports for incidental breast findings, which could flag for additional diagnostic imaging.

Methods/Intervention

We selected 17752 routine chest CTs from female patients ages 40-72 obtained at UW-Health between 2015-2017. We subselected 3226 exams with “breast” in the radiology report and randomly sampled 500 exams for evaluation. We compared the performance of Llama3-8B incidental breast findings labelling with varying prompts to a human reader. The LLM was tasked with labeling “Yes” or “No” for incidental breast findings with the role of a radiologist or annotator. Prompt elements included the task and radiology report, and varying combinations of background, keywords, and examples. Each prompt was run 30 times to evaluate consistency. We conducted sensitivity, specificity, and Fleiss’ Kappa consistency analyses to compare the human reader and LLM.

Results/Outcome

The human reader identified 125 (25.0%) of reports having incidental breast findings. The LLM’s average positively labelled cases ranged from 236.1 (47.2%) to 412 (82.4%) of reports. Sensitivity ranged from 0.76 to 0.99, though the highest average positive predictive value was 0.50. Specificity ranged from 0.23 to 0.71, with the lowest negative predictive value at 0.86. While sensitivity generally decreased with more prompt elements, specificity increased with more detailed prompts. Fleiss’ Kappa indicated high agreement among prompt iterations with the at $\kappa=0.94$.

Conclusion

The LLM and prompts labelled many false positives but had high negative predictive values with high consistency across all prompts. Future work will evaluate the parameter size of models on metric performance.

Statement of Impact

LLMs remain as a possible flagging system for missed details and prevention system, however larger models or fine-tuning might be required to match human performance.

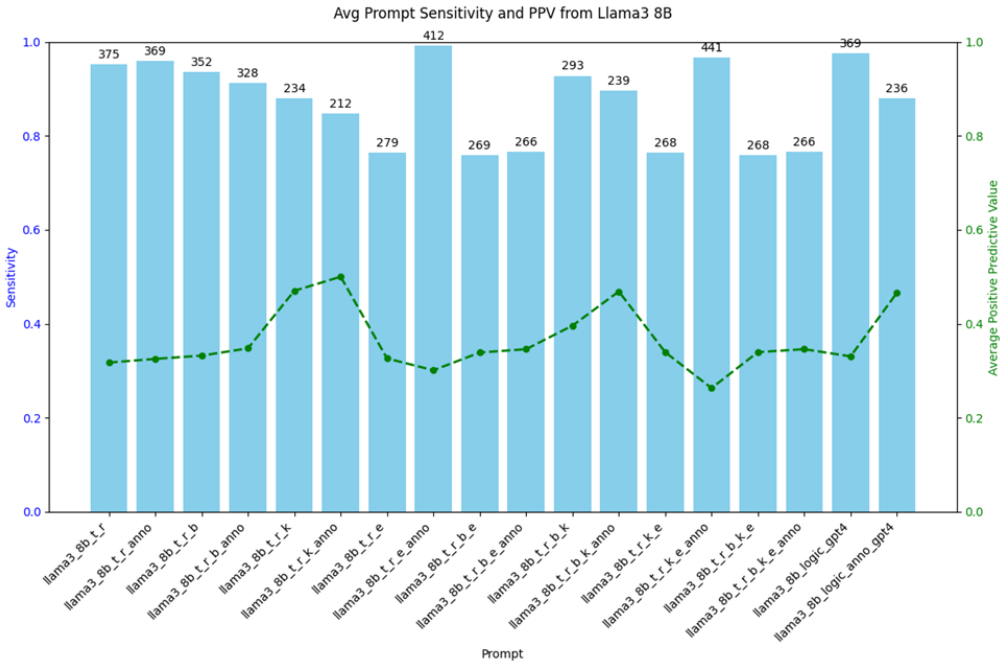

Figure 1. Average sensitivity (blue) and positive predictive value (green) of Llama 3 8B labelling incidental breast findings with different prompt combinations. The number of average positive cases for each prompt is at the top of each bar. Prompts all contained the radiology report (r) and the task (t) to label the radiology report as having or not having an incidental breast finding, and then combinations of background (b), keywords (k), examples (e). A logic prompt was created using GPT-4 reducing the prompt with background, keywords, and examples to a logic flow to follow. Prompts with “anno” indicate the medical annotator role for the LLM to take on versus a radiologist role.

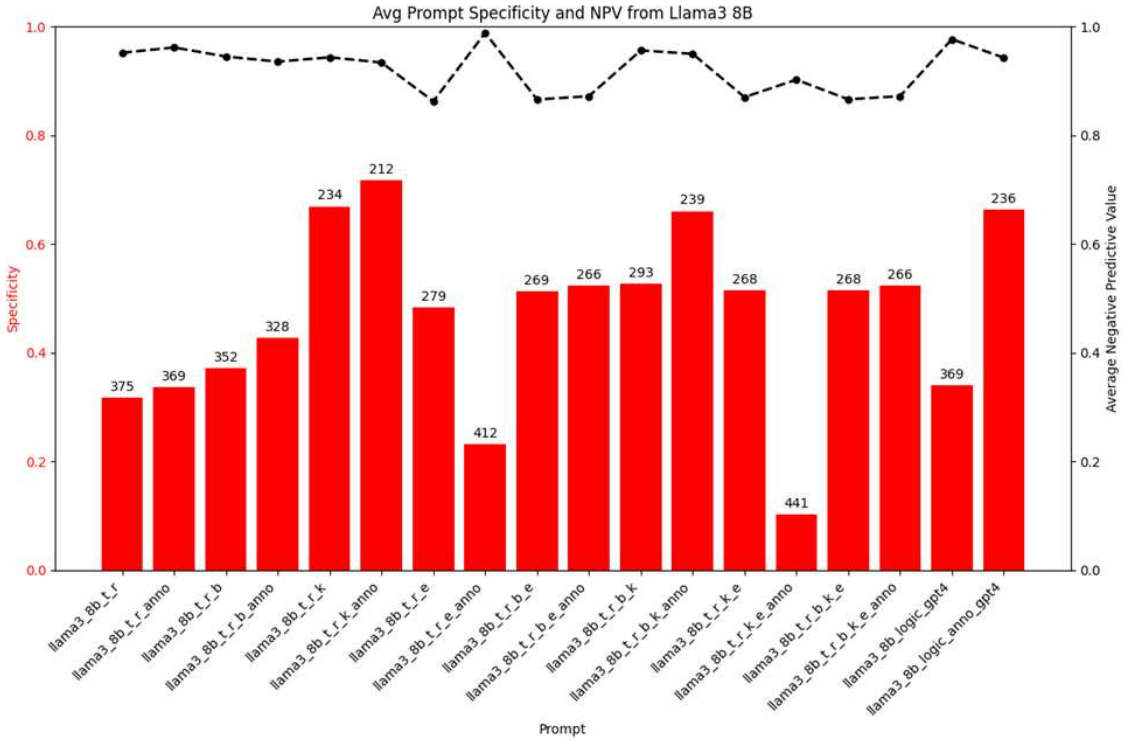

Figure 2. Average specificity (red) and negative predictive value (black) of Llama 3 8B labelling incidental breast findings with different prompt combinations. The number of average positive cases for each prompt is at the top of each bar. Prompts all contained the radiology report (r) and the task (t) to label the radiology report as having or not having an incidental breast finding, and then combinations of background (b), keywords (k), examples (e). A logic prompt was

created using GPT-4 reducing the prompt with background, keywords, and examples to a logic flow to follow. Prompts with “anno” indicate the medical annotator role for the LLM to take on versus a radiologist role.

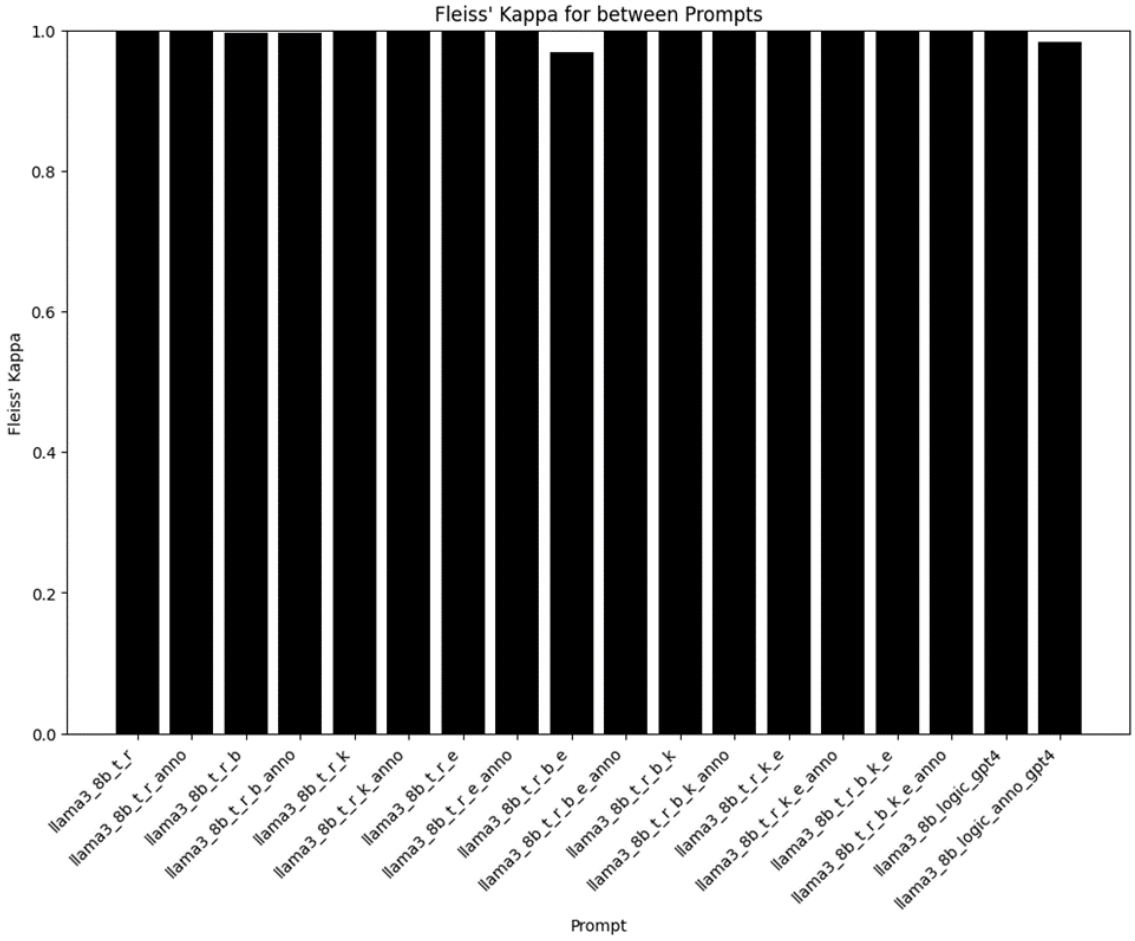

Figure 3. Fleiss’ Kappa showing consistency between the 30 iterations of the same prompt among the 500 cases. Prompts all contained the radiology report (r) and the task (t) to label the radiology report as having or not having an incidental breast finding, and then combinations of background (b), keywords (k), examples (e). A logic prompt was created using GPT-4 reducing the prompt with background, keywords, and examples to a logic flow to follow. Prompts with “anno” indicate the medical annotator role for the LLM to take on versus a radiologist role.

Keywords
Large Language Models; Incidental Findings; Breast Health; Radiology Reports; Prompt Engineering