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

Prompt-Induced Bias in Vision Language Models: Implications for Pneumonia Detection in Pediatric Chest Radiographs

David Li, MD, London Health Sciences Center; Jaron Chong, MD, MHI

Introduction/Background

Vision language models (VLMs) have the potential to revolutionize medical imaging. However, the effects of text prompts on visual tasks are not well understood. This study investigates how variations in text prompts influence the diagnostic accuracy of GPT-4 Turbo in detecting pneumonia in pediatric chest radiographs. We hypothesize that subtle differences in prompt context can lead to biased predictions in visual diagnostic tasks.

Methods/Intervention

This retrospective study utilized publicly available data and was exempt from institutional review board approval. 5856 pediatric chest radiographs were obtained from the Guangzhou Women and Children's Medical Center. A test set of 200 radiographs, including 100 pneumonia cases and 100 normal cases, was randomly selected from patients aged 1 to 5 years. The latest version of GPT-4 Turbo with Vision was used to classify each radiograph as either pneumonia or normal, employing four prompt variations: neutral, query positive, clinically symptomatic, and leading answer. VLM performance was evaluated using sensitivity, specificity, and area under the receiver operating characteristic curve (AUROC). Statistical analysis was performed using McNemar's tests, with a significance threshold of $p < 0.05$.

Results/Outcome

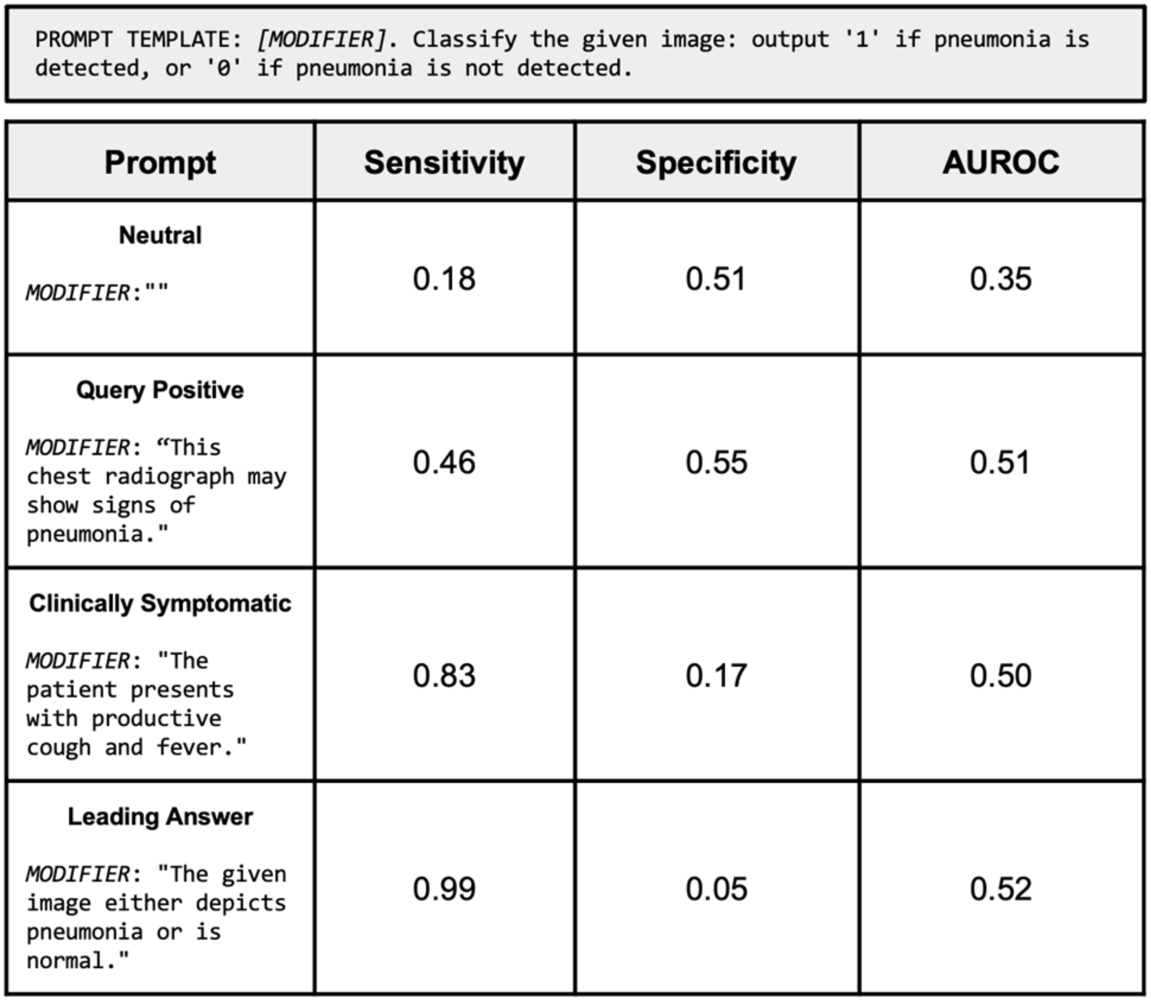
The AUROC for the four prompts ranged from 0.35 to 0.53 (Table 1). Subgroup analysis showed that sensitivity increased progressively with greater prompt bias, ranging from 0.18 for neutral prompts to 0.99 for leading answer prompts (Fig. 1). Significant differences were observed in pairwise comparisons between the neutral and clinically symptomatic prompts ($p = 0.026$) and between the neutral and leading answer prompts ($p < 0.001$).

Conclusion

This study highlights that prompt-induced bias significantly impacts GPT-4 Turbo's performance in detecting pneumonia in pediatric chest radiographs. Moreover, VLM performance was lower compared to previously published benchmarks for convolutional neural networks in chest radiograph interpretation. Further research is needed to identify and address prompt-induced bias to ensure reliable clinical deployment.

Statement of Impact

Currently, VLMs without specialized medical fine-tuning demonstrate limited accuracy in interpreting chest radiographs. Prompt-induced bias significantly affects diagnostic performance in visual tasks. To enhance the clinical effectiveness of VLMs, it is crucial to conduct rigorous validation studies using neutral prompts to minimize bias and avoid overestimating results.



PROMPT TEMPLATE: *[MODIFIER]*. Classify the given image: output '1' if pneumonia is detected, or '0' if pneumonia is not detected.

Prompt	Sensitivity	Specificity	AUROC
Neutral *MODIFIER*:""	0.18	0.51	0.35
Query Positive *MODIFIER*: "This chest radiograph may show signs of pneumonia."	0.46	0.55	0.51
Clinically Symptomatic *MODIFIER*: "The patient presents with productive cough and fever."	0.83	0.17	0.50
Leading Answer *MODIFIER*: "The given image either depicts pneumonia or is normal."	0.99	0.05	0.52

Table 1: Evaluation of pediatric chest radiographs for pneumonia detection using GPT-4 Turbo with Vision. The vision language model’s diagnostic accuracy was evaluated with four text prompts, each introducing varying degrees of contextual bias.

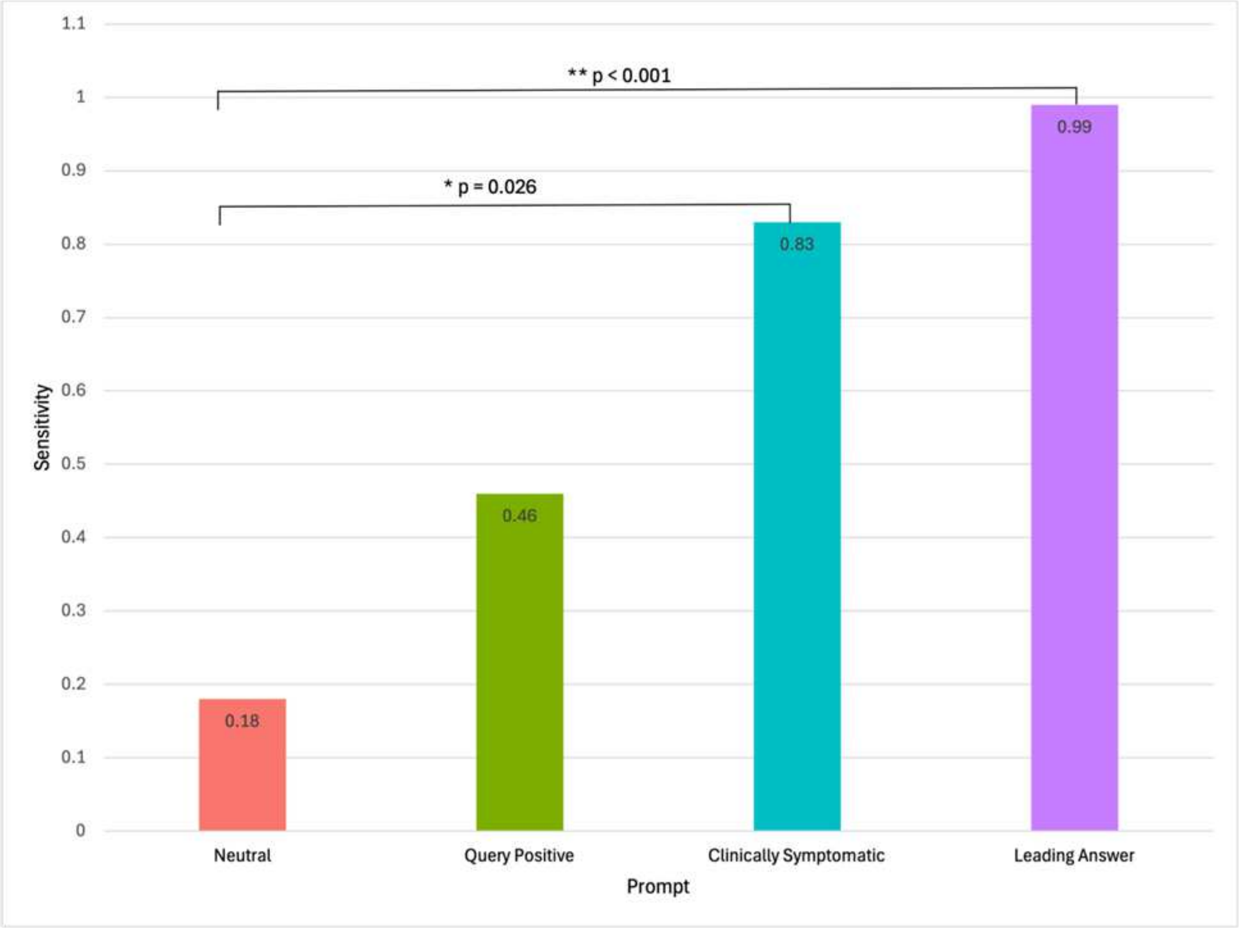


Fig. 1: Sensitivity of GPT-4 Turbo for pneumonia detection in pediatric chest radiographs by prompt type.

Keywords

Vision language models; Large language models; Multimodality; Generative pre-trained transformers; Bias; Fairness