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

Validating GPT-4 for Automated Protocoling in Diagnostic Imaging

Kartik Gupta, MSc, University of Western Ontario; Jaron Chong, MD

Introduction/Background

The growing volume of radiology exams, especially CT scans, necessitates more efficient workflows. Protocoling, taking up to 6% of a radiologist's time, is an opportunity for automation. Traditional machine learning methods need large datasets and are hard to adapt across institutions. Large language models (LLMs) demonstrate performance in medical question-answering and protocoling. This study evaluates the zero-shot prediction of OpenAI's GPT-4 in automating Chest CT scan protocoling, using prompts with institution-specific rules.

Methods/Intervention

A dataset of 796 labelled Chest CT Thorax imaging requests and protocols from Victoria Hospital, London Ontario, was analyzed. One data sample contains a requisition with a provided clinical indication from the ordering physician, and the assigned imaging protocol. There were 4 different classes of protocols; Chest CT “with contrast”, “without contrast”, “interstitial”, and “low-dose contrast”. Four prompts were tested with GPT-4: a baseline 'Control' prompt, a 'Classification Rules' (CR) prompt with specific guidelines, an 'Ablated' version with fewer guidelines, and a 'Refined' version (CR-V2) with improved rules. Performance was measured using accuracy, precision, recall, and F1 score. Statistical significance was assessed using McNemar's test where p-values less than 0.05 were significant.

Results/Outcome

The CR prompt significantly outperformed the 'Control' (accuracy: 0.88 vs 0.79, precision: 0.77 vs 0.61, recall: 0.89 vs 0.79, F1 score: 0.82 vs 0.66; $P < 0.001$). The 'Ablated' model showed reduced performance to CR yet superior performance to the 'Control' (accuracy: 0.85, $P = 0.002$ vs Control). The CR-V2 model achieved the highest metrics (accuracy: 0.9, precision: 0.8, recall: 0.89, F1 score: 0.84), significantly outperforming both the 'Control' ($P < 0.001$) and 'Classification Rules' ($P = 0.014$).

Conclusion

Providing specific instructions for GPT-4 can markedly improve the accuracy of protocol predictions in radiology. The use of specific prompting also improves performance of protocoling compared to no prompt (“Control”). The study demonstrates the potential of large language models in zero-shot protocol prediction for enhancing radiological workflow across institutions by adapting a set of protocoling rules.

Statement of Impact

By introducing custom prompts, institutions can tailor their automated pipelines with LLMs according to their own rules and improve protocoling accuracy. The zero-shot performance ensures that large training datasets are not required.

Figure 1. A) Table breaking down total dataset size per class. 4 classes were evaluated, with “Chest with Contrast” being the majority class, making up 76% of the dataset. **B)** Performance metrics for all prompts. P-values are provided that compare the Control and Classification Rules (CR) prompts using McNemar’s test, which compare the accuracy of each prompt. Significant differences are noted by p-values <0.05. **C)** Bar plots of all metrics across all prompts demonstrating improved performance with CR and CR-Enhanced.

A)

Chest CT Class	Chest with Contrast	Chest without Contrast	Interstitial	Low Dose CT	Total
Number of Samples	605	97	27	67	796

B)

	Accuracy	Precision	Recall	F1 Score	P value vs. Control	P value vs. CR
Control	0.79	0.61	0.79	0.66	-	-
Classification Rules (CR)	0.88	0.77	0.89	0.82	<0.001	-
CR-Ablation	0.85	0.73	0.77	0.7	<0.001	0.002
CR-Enhanced	0.9	0.8	0.89	0.84	<0.001	<0.001

C)

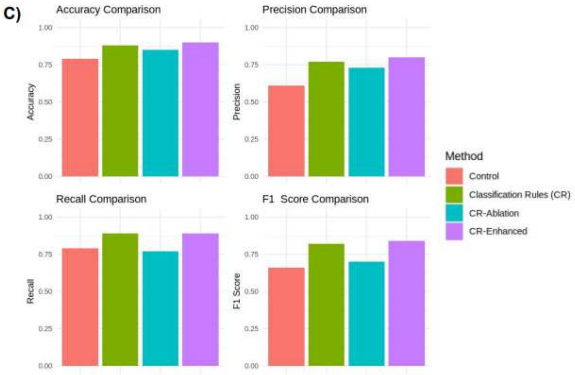

Figure 1. A) Table breaking down total dataset size per class. 4 classes were evaluated, with “Chest with Contrast” being the majority class, making up 76% of the dataset. B) Performance metrics for all prompts. P-values are provided that compare the Control and Classification Rules (CR) prompts using McNemar’s test, which compare the accuracy of each prompt. Significant differences are noted by p-values

Keywords

Large Language Models; Protocols; Zero-shot prediction