Evaluating Large Language Models for Multi-Institutional Radiology Report Annotation: A Prompt-Engineering Approach

Mana Moassefi, MD, Postdoctoral Research Fellow, Radiology, Mayo Clinic
Les Folio, MD; Ghulam Rasool, MD; Sina Houshmand, MD; Peter Chang, MD; Katherine Andriole, PhD, FSIIM; Maryellen Giger, PhD; Jessyca Wagner, PhD; Judy Gichoya, MS, MD, FSIIM

Introduction

The rapid evolution of large language models (LLMs) offers promising opportunities for radiology report annotation, aiding in determining the presence of specific findings. This study evaluates the effectiveness of a human-optimized prompt in labeling radiology reports across multiple institutions using LLMs.

Hypothesis

A human-optimized prompt can enable LLMs to accurately and consistently annotate radiology reports across multiple institutions, regardless of variations in report structures and institutional practices.

Methods

A multi-institutional dataset was curated, comprising 500 radiology reports per site from Mayo Clinic, University of California, San Francisco(UCSF), Massachusetts General Hospital(MGH), Unniversity of California-Irvine (UCI) and Emory. The reports' findings included five categories: liver metastases (CT abdomen), subarachnoid hemorrhage (CT brain), pneumonia (chest X-ray), cervical spine fracture (CT), and glioma progression (MRI brain). A standardized Python script was distributed to participating sites, allowing the use of different locally executed LLMs and a human-optimized prompt(Figure 1). The script executed the LLM's analysis for each report, using a predefined answer set (e.g., ['Yes', 'No'] or ['Progression', 'Stable', 'Improved']), and compared predictions to ground truth labels provided by local investigators. Models' performance using accuracy were calculated and results were aggregated centrally.

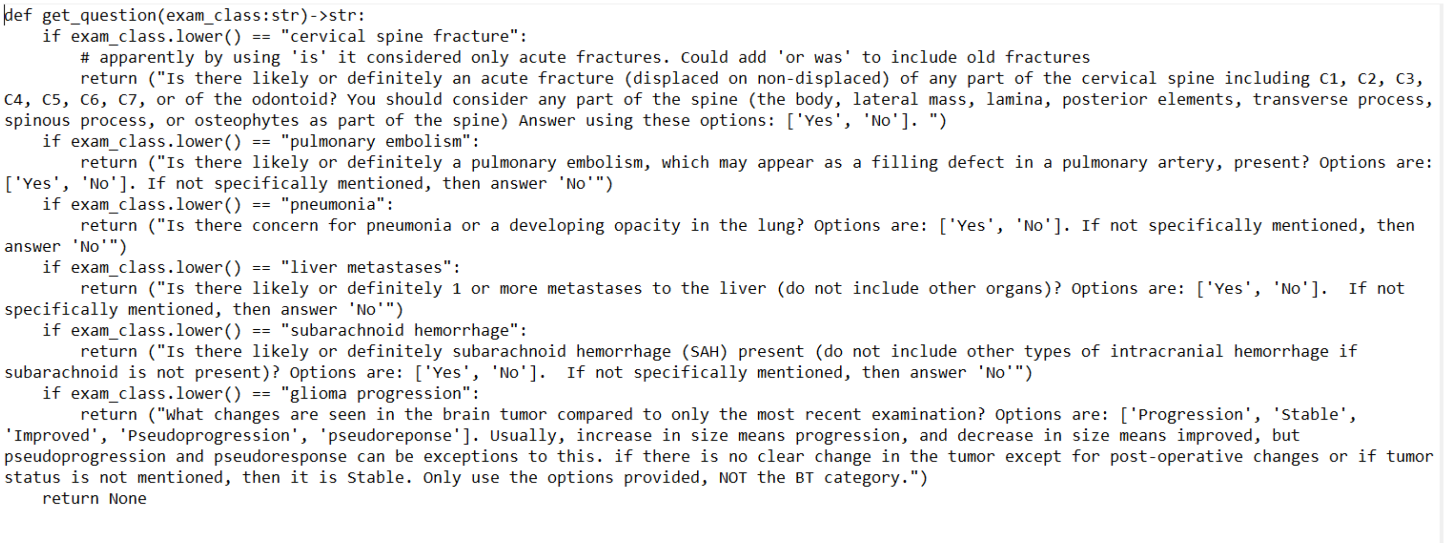
Results

The human-optimized prompt demonstrated high consistency across sites and pathologies, with overall performance surpassing initial expectations(table-1). Preliminary analysis indicates significant agreement between the LLM's outputs and investigator-provided ground truths across multiple institutions. At Mayo Clinic, eight LLMs were systematically compared, with Llama 3.1 70b achieving the highest performance in accurately identifying the specified findings. Comparable performance with Llama 3.1 70b was observed at two additional centers, demonstrating the model's robust adaptability to variations in report structures and institutional practices. We also note that for a small percentage of cases, the LLMs did not respond with the required short answer (e.g. 'Yes' or 'No') but provided a long explanation that usually was correct. However, these were counted as incorrect because the LLM did not follow the prompt.

Conclusion

Our findings illustrate the potential of optimized prompt engineering in leveraging LLMs for cross-institutional radiology report annotation. By eliminating the need for federated learning, this approach simplifies implementation while maintaining high accuracy and adaptability. Future work will explore model robustness to diverse report structures and further refine prompts to improve generalizability.

Figure(s)



```
def get_question(exam_class:str)->str:
    if exam_class.lower() == "cervical spine fracture":
        # apparently by using 'is' it considered only acute fractures. Could add 'or was' to include old fractures
        return ("Is there likely or definitely an acute fracture (displaced on non-displaced) of any part of the cervical spine including C1, C2, C3, 
C4, C5, C6, C7, or of the odontoid? You should consider any part of the spine (the body, lateral mass, lamina, posterior elements, transverse process, 
spinous process, or osteophytes as part of the spine) Answer using these options: ['Yes', 'No']. ")
    if exam_class.lower() == "pulmonary embolism":
        return ("Is there likely or definitely a pulmonary embolism, which may appear as a filling defect in a pulmonary artery, present? Options are: 
['Yes', 'No']. If not specifically mentioned, then answer 'No'")
    if exam_class.lower() == "pneumonia":
        return ("Is there concern for pneumonia or a developing opacity in the lung? Options are: ['Yes', 'No']. If not specifically mentioned, then 
answer 'No'")
    if exam_class.lower() == "liver metastases":
        return ("Is there likely or definitely 1 or more metastases to the liver (do not include other organs)? Options are: ['Yes', 'No'].  If not 
specifically mentioned, then answer 'No'")
    if exam_class.lower() == "subarachnoid hemorrhage":
        return ("Is there likely or definitely subarachnoid hemorrhage (SAH) present (do not include other types of intracranial hemorrhage if 
subarachnoid is not present)? Options are: ['Yes', 'No'].  If not specifically mentioned, then answer 'No'")
    if exam_class.lower() == "glioma progression":
        return ("What changes are seen in the brain tumor compared to only the most recent examination? Options are: ['Progression', 'Stable', 
'Improved', 'Pseudoprogression', 'pseudoreponse']. Usually, increase in size means progression, and decrease in size means improved, but 
pseudoprogression and pseudoresponse can be exceptions to this. if there is no clear change in the tumor except for post-operative changes or if tumor 
status is not mentioned, then it is Stable. Only use the options provided, NOT the BT category.")
    return None
```

Figure 1. This code snippet represents part of a Python function designed to generate prompts for querying the LLMs about specific findings in radiology reports. Each finding, such as cervical spine fractures, cervical spine fracture, pneumonia, liver metastases, subarachnoid hemorrhage, or glioma progression, is associated with a tailored question format to ensure precise responses based on the report content.

Configuration	Center	Cervical Spine Fracture	Glioma Progression	Subarachnoid hemorrhage	Pneumonia	Liver Metastasis	Average Accuracy
meta-llama/Met a-Llama- 3-70B-Instruct	Emory	0.657	0.56	0.41	0.88	0.535	0.9
GPT-4o-2024-05-13	UCSF	0.926	0.77	0.9	0.98	0.99	0.99

llama3	UCSF	0.844	0.93	0.66	0.92	0.86	0.85
Llama3.1 70b- instruct	UCSF	0.91	0.78	0.8	0.99	0.99	0.99
llama3.17 0b- instruct	MGH	0.932	0.82	0.89	0.99	0.99	0.97
llama3.17 0b instruct simple	MGH	0.912	0.97	0.93	1.0	1.0	0.66
llama3.17 0b instruct	Mayo Clinic	0.936	0.96	0.95	0.93	0.95	0.89
Llama3.1 7b chatqa	Mayo Clinic	0.95	0.95	0.95	0.98	0.91	0.94
Llama3.1 7b	Mayo Clinic	0.62	0.95	0.52	0.89	0.7	0.04
Llama3.1 70b instruct	Mayo Clinic	0.904	0.94	0.93	0.88	0.92	0.85
Llama3.3 70b	Mayo Clinic	0.937	0.891	0.903	0.899	0.912	0.907

Table 1. Performance metrics of various large language models (LLMs) in identifying findings across different radiology report categories, including cervical spine fractures, Glioma, subarachnoid hemorrhage, pneumonia, and liver metastases, as observed at multiple institutions. (University of California, San Francisco(UCSF), Massachusetts General Hospital(MGH))

Keywords

Applications; Artificial Intelligence/Machine Learning; Clinical Workflow & Productivity; Emerging Technologies; Standards & Interoperability