Capability of Multi-Modal Large Language Models for Matching Findings in Longitudinal CT Studies

Tejas Sudharshan Mathai, PhD, Staff Scientist, National Institutes of Health (NIH) Clinical Center
Boah Kim, PhD; Praveen Thoppey Srinivasan Balamuralikrishna, MBBS; Ronald Summers, MD, PhD;

Introduction

Radiologists routinely compare findings between the prior and follow-up CT exams, and then assess interval changes. However, this task is currently manually performed, and it can become cumbersome when comparing multiple time points.

Hypothesis

To evaluate the capability of Multi-Modal Large Language Models (MLLM) for matching findings between two longitudinal exams (prior vs. follow-up) using report sentences and CT images.

Methods

In this retrospective study, the public CT-RATE dataset containing longitudinal non-contrast chest CT studies was used. CT volumes and reports from the prior and follow-up visits of 67 patients were included. Findings in the reports (e.g., nodules, pleural/pericardial effusion) were automatically extracted, and the slice in the CT volume containing the respective finding was manually identified. Given a finding and CT image from the follow-up study, the MLLM identified the matched finding in the prior study. Two MLLMs (GPT-4o and Gemini-1.5-Pro) were evaluated, and the use of report text alone (i.e., Gemini-R) was compared against the combined use of both images and text (i.e., Gemini-C). Agreement with a rater was measured using Cohen's κ.

Results

Longitudinal CT studies and reports from 67 patients (M/F ratio: 44/23, ages: 24 - 89 years, 134 CT volumes, 134 reports) were used. Gemini-R obtained the best results with 98.7% precision, 98.4% specificity, 79.6% sensitivity, with substantial agreement (κ = 0.75) with the rater. GPT-4o-C achieved 92.7% precision, 90.8% specificity, 82.6% sensitivity with substantial agreement (κ = 0.72). No significant differences were observed (p> .05) between GPT-4o-C vs. GPT-4o-R, GPT-4o-C vs. Gemini-C, GPT-4o-R vs. Gemini-R, respectively. However, there was a significant difference between Gemini-R and Gemini-C (p= .03).

Conclusion

In this pilot study, the Gemini-R MLLM (using report text alone) matched findings across longitudinal CT studies and showed potential for interval change assessment.

Figure(s)

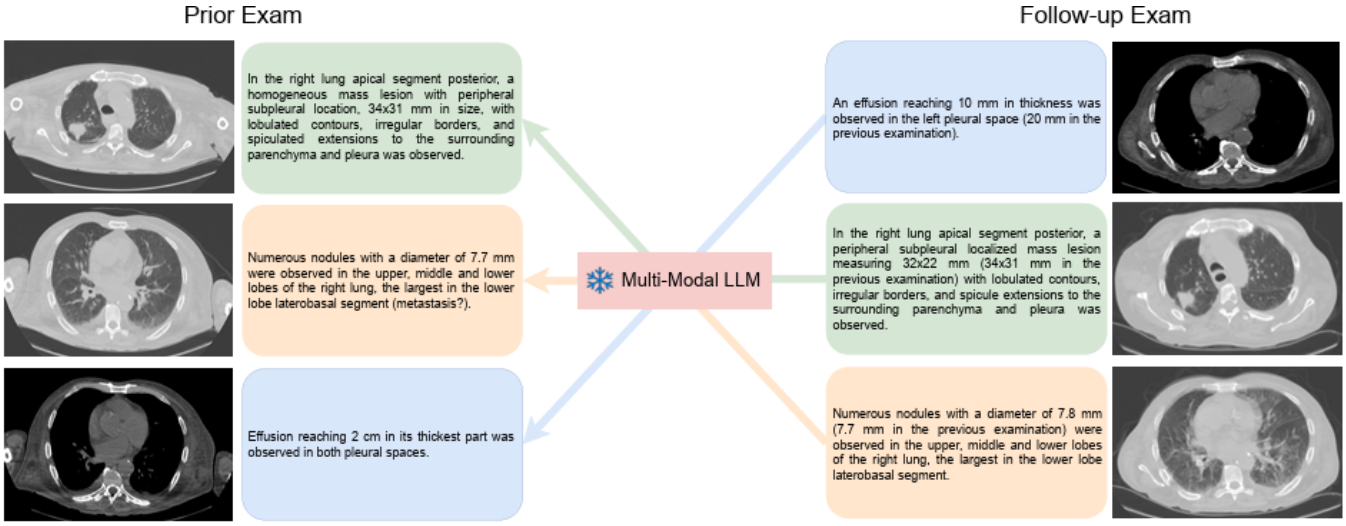

Figure 1. Multi-Modal Large Language Models (MLLMs) were used to match findings between prior and follow-up non-contrast chest CT studies in the public CT-RATE dataset. Given a report sentence containing a finding and the corresponding CT slice from the follow-up exam, an MLLM (e.g., GPT-4o) was tasked with identifying the option in the prior exam that best matched the finding in the follow-up.

Keywords

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