Identifying Pulmonary Embolism from Radiology Reports Using Large Language Models

Farzad Ahmed, BS[1], Namu Park, MS[2], Barbara D. Lam, MD[2,3], Meliha Yetisgen, PhD[2], Özlem Uzuner, PhD[1]

[1]George Mason University, Fairfax, VA, USA, [2]University of Washington, Seattle, WA, USA, [3]Fred Hutch Cancer Center, Seattle, WA, USA

Introduction

Problem
Pulmonary embolism (PE) is a major cause of preventable in-hospital deaths. Accurate identification of PE in radiology reports is essential for timely care and quality monitoring. Traditional methods such as manual chart review and ICD-code–based identification are labor-intensive and prone to error.

Current Challenges in PE Detection
Detection remains difficult due to variability in report structure and terminology, as well as ambiguous or uncertain phrasing (e.g., *"may represent embolism"*). These factors reduce consistency across automated systems and hinder large-scale, standardized surveillance.

Study Goal
This study evaluates two large language models (LLMs)—**Llama 3.1 8B** and **GPT-4o**—for detecting **acute pulmonary embolism** in radiology reports. We compare zero-shot, few-shot, and fine-tuned approaches to assess their performance and potential to support scalable, automated PE identification.

Methods

We used the **MIMIC-IV-Ext-PE** dataset with **19,942 radiology reports** (1,591 positive; 18,351 negative).
Models: Llama 3.1 8B and **GPT-4o** were evaluated in three setups:
Zero-shot: no examples.
Few-shot: three labeled examples (1 positive, 2 negative).
Fine-tuned: 70% train, 15% validation, 30% test.
All models used the **same prompt** to extract PE-related text spans and classify reports.
Evaluation used **precision**, **recall**, and **F1-score**, focusing on the **positive PE class**

FINDINGS: Small, centralized, nonocclusive filling defects are present in the segmental and subsegmental branches of the medial basal right lower lobe (__, 401b:24), consistent with acute embolus. There is no main or lobar pulmonary embolus. The thoracic aorta is normal in caliber throughout. The aortic arch has a normal three-vessel branch morphology. Wall calcifications and possible stent are noted in the left anterior descending coronary artery. The heart is normal in size, with trace pericardial effusion, and no evidence of right heart strain.

Figure 1: Radiology report excerpt with PE evidence in bold

Prompt Structure

You are a physician reviewing a radiology report. Please isolate any sentences that describe pulmonary embolism.
If the radiology report describes an acute pulmonary embolism, label the report as 'positive'.
All other reports are labeled 'negative'. If there is no pulmonary embolism, chronic pulmonary embolism only, or the findings are equivocal, label the report as 'negative'.
Other types of findings such as other thrombi or masses are labeled 'negative'.
Output 'positive' or 'negative' at the very end of the response, in a single line that starts with 'Label:'.

Figure 2: Prompt Structure for LLM models

Results

- **Table 1** and **Figure 3** summarize model performance on PE detection.
- All models achieved an **F1-score of 0.99** on the **negative class**, reflecting strong accuracy for non-PE reports.
- **GPT-4o (zero-shot)** obtained the **highest positive-class F1 = 0.96**, showing effective performance without fine-tuning.
- **Llama 3.1 8B** showed steady improvement: **zero-shot (0.91) → few-shot (0.92) → fine-tuned (0.93)**.
- The **best recall (0.99)** for the positive class was achieved by both **zero-shot Llama 3.1 8B** and **GPT-4o**.

Model	Class	Precision	Recall	F1-Score
Llama 3.1 8B (Zero-shot)	Positive	0.84	0.99	0.91
	Negative	1.00	0.98	0.99
Llama 3.1 8B (Few-shot)	Positive	**0.94**	0.91	0.92
	Negative	0.99	0.99	0.99
Llama 3.1 8B (Fine-tuned)	Positive	0.93	0.92	0.93
	Negative	0.99	0.99	0.99
GPT-4o (Zero-shot)	Positive	0.93	**0.99**	**0.96**
	Negative	1.00	0.99	0.99

Table 1: Precision, recall, and F1-scores for Llama 3.1 8B and GPT-4o in detecting pulmonary embolism (PE) from radiology reports.

- **Few-shot** and **fine-tuned** configurations increased **precision** slightly but reduced recall.
- All models achieved **0.99 F1** on the **negative class**.
- **Error analysis** showed most **false negatives** arose from **ambiguous or uncertain phrasing** (e.g., *"may represent embolism," "cannot exclude embolism"*) and **complex negations** (*"no evidence of right lower lobe embolism"*).
- **Figure 3** illustrates the **F1-score trend**, with GPT-4o leading overall and Llama 3.1 8B improving with additional supervision.

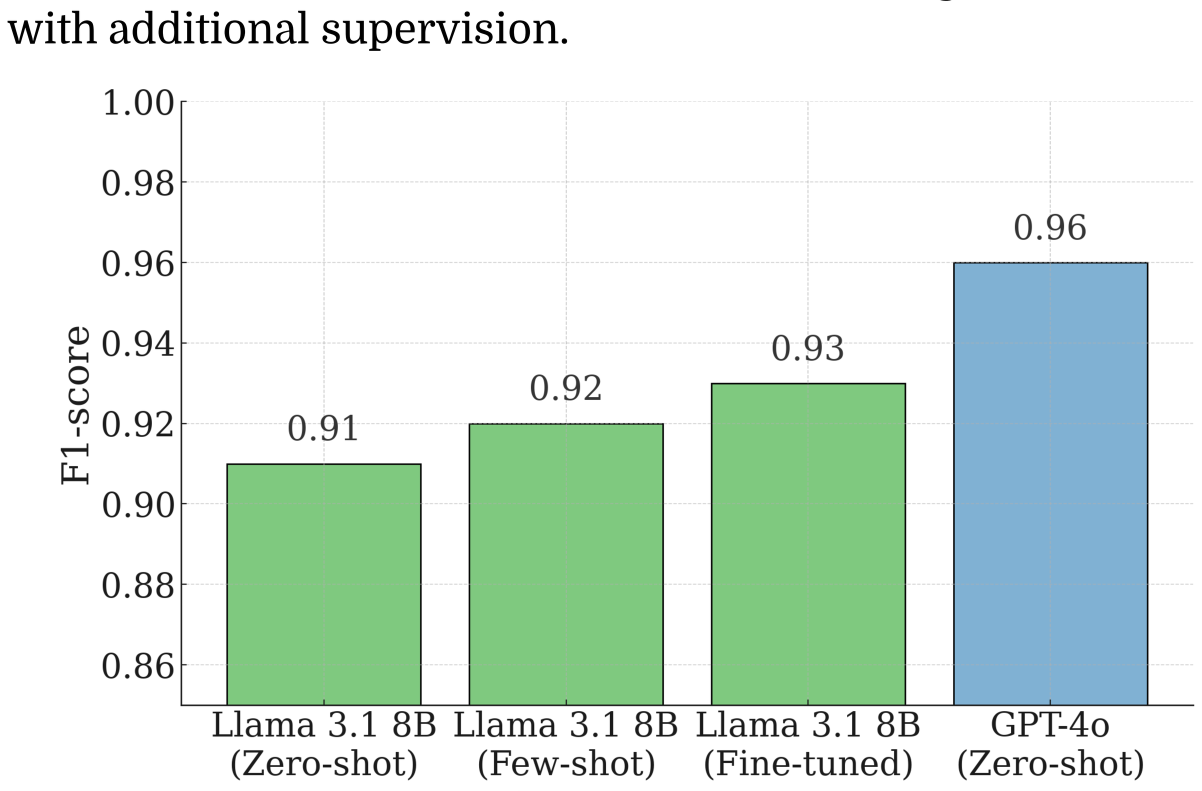

Figure 3: F1-scores for positive pulmonary embolism (PE) detection across model configurations

Conclusions, Limitations and Future work

- **LLMs such as GPT-4o and Llama 3.1 8B** effectively detect **acute pulmonary embolism (PE)** from radiology reports, achieving high precision and recall.
- **GPT-4o (zero-shot)** reached the best positive-class F1 (0.96), while **Llama 3.1 8B** improved with few-shot and fine-tuned setups.
- Performance was evaluated on a **single dataset (MIMIC-IV-Ext-PE)**, and results may vary across institutions and reporting styles.
- **Future work:** external validation, improved prompt design, model interpretability, and integration into clinical quality-monitoring workflows.

Acknowledgements

- This work was supported in part by **National Institutes of Health - National Cancer Institute (NCI)** and **National Library of Medicine (NLM),** (Grant Nos. 1R01CA248422-01A1), (Grant No. 2R15LM013209-02A1). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.