# High Performance Prompting for LLM Extraction of Findings from Radiology Reports

**Mohammed M. Kanani, MS,** Medical Student, University of Washington School of Medicine
Arezu Monawer, MD; Lauryn Brown, MD; William King, MD, PhD; Zach Miller, MD; Nitin Venugopal, MD; Patrick Heagerty, PhD; Jeffrey Jarvik, MD, MPH; Trevor Cohen, PhD; Nathan Cross, MD, MS

## Introduction

Extracting information from radiology reports can provide critical data to empower clinical, QA/QI, research, and educational radiology workflows. For spinal compression fractures, this data can identify at risk populations and facilitate evidence-based follow up and treatment. Manual extraction from free-text radiology reports is laborious, and prone to errors. Large language models (LLMs) have shown promise. Fine tuning strategies are time and resource intensive, but a variety of prompting strategies have achieved similar results with less resources, data and annotation. Our study pioneers the use of Meta's Llama 3.1 for automated extraction from free-text radiology reports to detect spinal compression fractures, outputting structured data without model training.

## Hypothesis

An open-source LLM can extract imaging findings on specific pathologies from free-text radiology reports using prompt-based strategies, eliminating the need for specific model training.

## Methods

We tested performance on a time-based sample of CT exams covering the spine from 2/20/2024 to 2/22/2024 acquired across our healthcare enterprise (637 anonymized reports, age 18-102, 47% Female). Ground truth annotations were manually generated by a group of attending/fellow/resident physicians, and medical students. These annotations were compared against the performance of three models (Llama 3.1 70B, Llama 3.1 8B, and Vicuna 13B) with nine different prompting configurations for a total of 27 model/prompt experiments (Figure-1).

## Results

Among the 637 reports, 49 were classified as true by annotators (prevalence: 7.69%). The highest F1 score (0.91) was achieved by the 70B Llama 3.1 model when provided with a radiologist-written background, with similar results when the background was written by a separate LLM (0.86). The addition of few-shot examples had variable impact on these prompts (0.89, 0.84 respectively). Comparable ROC-AUC and PR-AUC performance was observed (Figure-2).

## Conclusion

An open-source LLM excelled at extracting diagnostic information from free-text radiology reports using prompt-based techniques without model training.

## Figure(s)

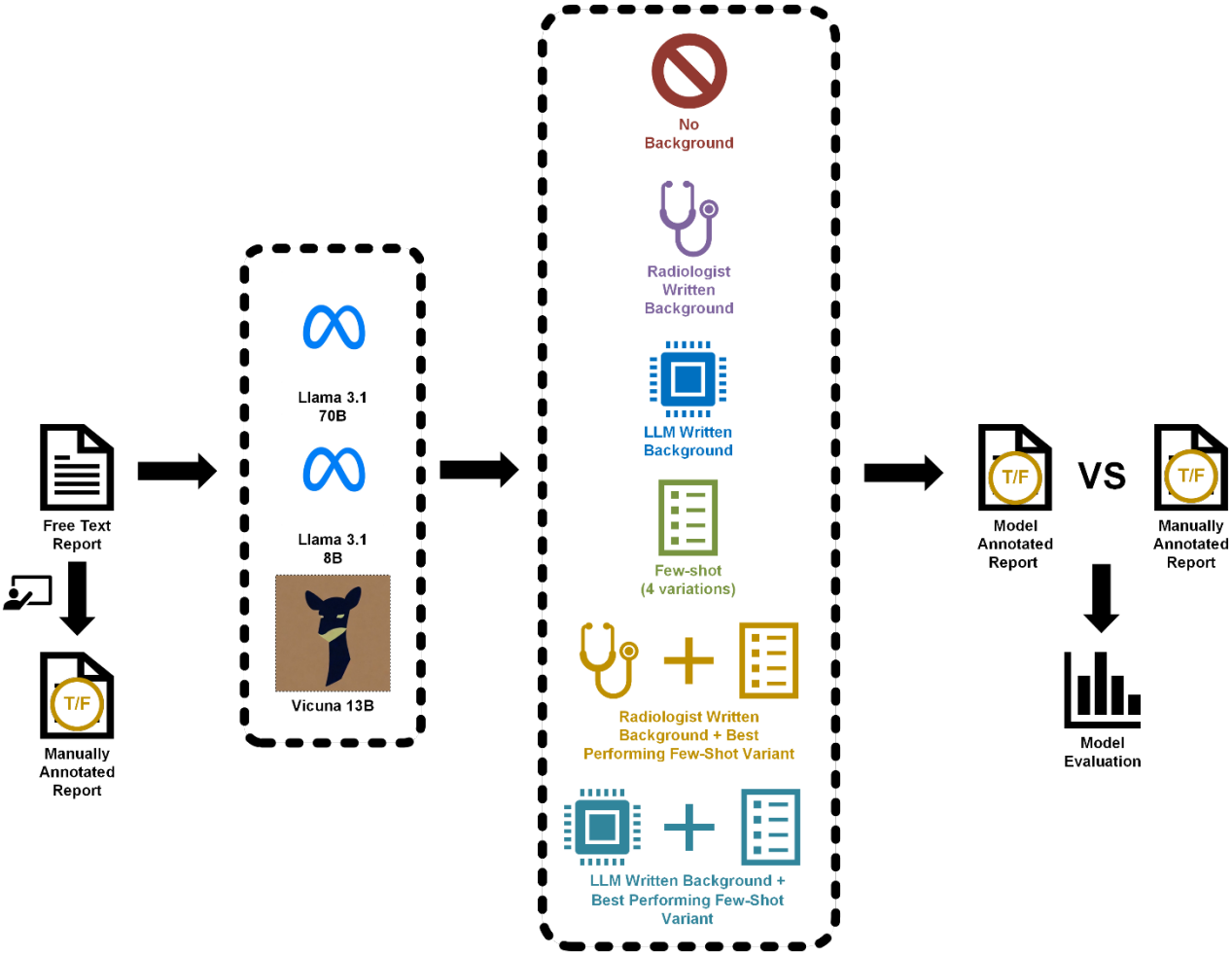


**Figure 1.** Experimental workflow illustrating the evaluation of the performance of three models (Llama 3.1 70B, Llama 3.1 8B, and Vicuna 13B) across nine prompting configurations, totaling 27 model/prompt experiments.

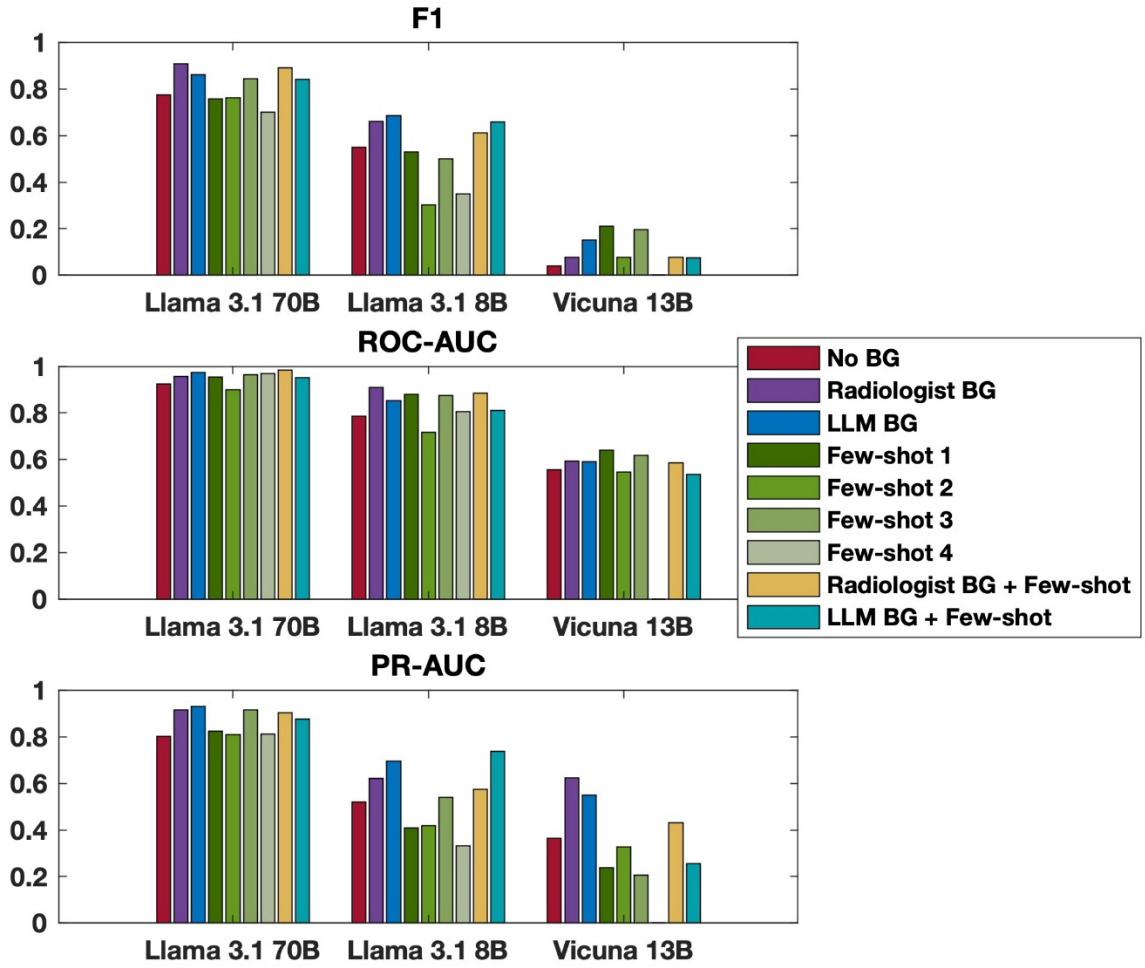


**Figure 2.** Performance comparison of model/prompt experiments using F1 score, ROC-AUC, and PR-AUC metrics. Vicuna 13B was not tested against Few-shot 4 due to context restrictions.


## Keywords

Artificial Intelligence/Machine Learning; Emerging Technologies; Imaging Research; Quality Improvement & Quality Assurance