Application of Large Language Models to Mammography Rad-Path Correlation for Automated Data Extraction and Concordance Prediction

Shawn H Sun, Kellie Cho, Dillon Sommer, Julia Tran, Gillean Cortes, Jasmine Zhao, Roozbeh Houshyar, Peter Chang, Jennifer Young, Irene Tsai

Department of Radiological Sciences, University of California, Irvine Medical Center, Orange, CA, USA | Contact: shsun@hs.uci.edu

Key takeaway: Prompt-engineered gpt-oss extracted key mammography/pathology findings with near-perfect accuracy, supporting scalable automated database labeling from routine reports.

BACKGROUND + OBJECTIVES

Clinical need

Manual rad-path review and retrospective **database labeling** are labor-intensive.

Key data are trapped in **unstructured reports**.

Study aim

Evaluate **gpt-oss** for automated extraction of mammography/pathology findings.

Secondary: exploratory concordance prediction.

Methods at a glance

- De-identified mammography and pathology reports processed by gpt-oss
- Structured output extraction performed to identify key rad-path descriptors
- Base vs prompt-engineered concordance prediction compared against manual review

Study cohort

211 patients from 2023,
215 stereotactic biopsy targets,
38 augmented discordant cases

MODEL WORKFLOW

gpt-oss pipeline for report-level extraction and concordance assessment

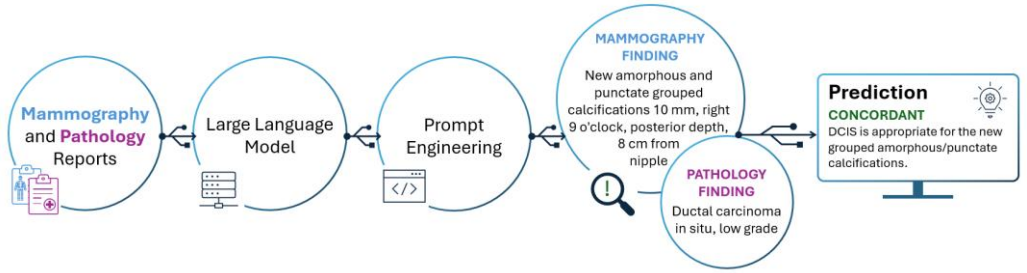


Figure 1. Single-case workflow: raw reports -> open-source LLM -> key findings -> concordance prediction.

KEY RESULTS

99.5% **mammography** extraction

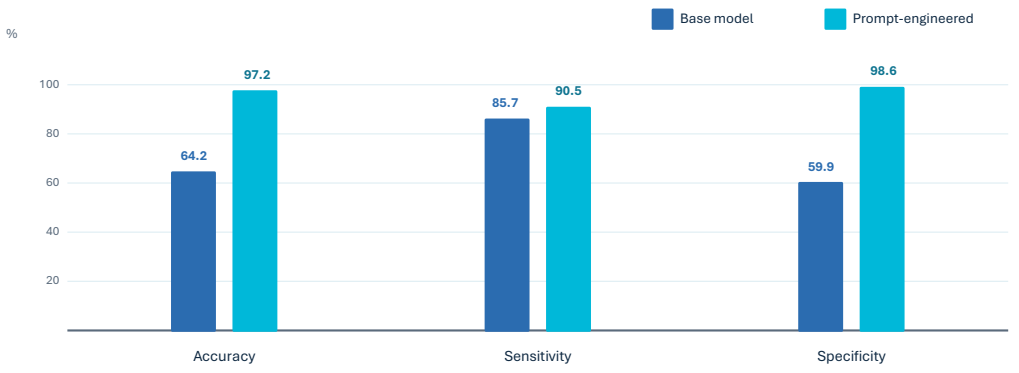
100% **pathology** extraction

97.2% **prompt-engineered** accuracy | base: 64.2%

p < 0.001 **prompted vs base** McNemar test

Prompt engineering improved concordance prediction, while extraction performance was already near-perfect.

CONCORDANCE PREDICTION PERFORMANCE



Prompt engineering improved accuracy and specificity without model fine-tuning; concordance prediction remains an exploratory use case requiring validation.

CONCLUSION + CLINICAL RELEVANCE

Database labeling at scale

Extraction supports structuring of routine mammography and pathology reports for retrospective research.

Prompt engineering matters

Prompt-engineered gpt-oss markedly improved concordance prediction performance without model fine-tuning.

Concordance prediction remains exploratory

The signal is promising, but clinical decision-support use requires larger, external validation before adoption.

Future work: validate on larger multi-institutional cohorts enriched for discordant cases.

REFERENCES

1. Le Guellec B, et al. Performance of an Open-Source Large Language Model in Extracting Information from Free-Text Radiology Reports. Radiol Artif Intell. 2024. 2. Lee A, Curpen B, Alikhassi A. Performance of ChatGPT-4o in Determining Radiology-Pathology Concordance and Management Recommendations Following Image-Guided Breast Biopsies. Diagnostics. 2025;15(19):2536. 3. Reichenpfader D, Muller H, Denecke K. A scoping review of large language model based approaches for information extraction from radiology reports. npj Digit Med. 2024;7:222.