•**Pre-filtering** macro content improved dataset quality and downstream IE performance.
•**The method** generalizes to other domains mixing templated and instance-specific text.

UW Medicine
UW SCHOOL OF MEDICINE

An NLP Method to Identify Macro Guideline Sentences in Radiology Reports

Zhaoyi Sun, MS[1], Namu Park, MS[1], Özlem Uzuner, PhD[2], Martin Gunn, MBChB[1], Meliha Yetişgen, PhD[1]
[1]University of Washington, Seattle, WA; [2]George Mason University, Fairfax, VA

BACKGROUND

Radiology reports summarize imaging findings and provide diagnostic impressions or recommendations to guide clinical care. These reports often combine patient-specific information with system-generated *macro guidelines* – standardized template sentences that describe general management recommendations. These guidelines can reference similar clinical concepts as the patient findings, which presents a challenge for information extraction (IE) systems designed to capture patient-specific details[1]. Without filtering macro guidelines, IE models may incorrectly interpret general recommendations as actual clinical findings. For example, in Figure 1, a cyst size mentioned in a guideline could be mistakenly linked to a specific patient lesion.

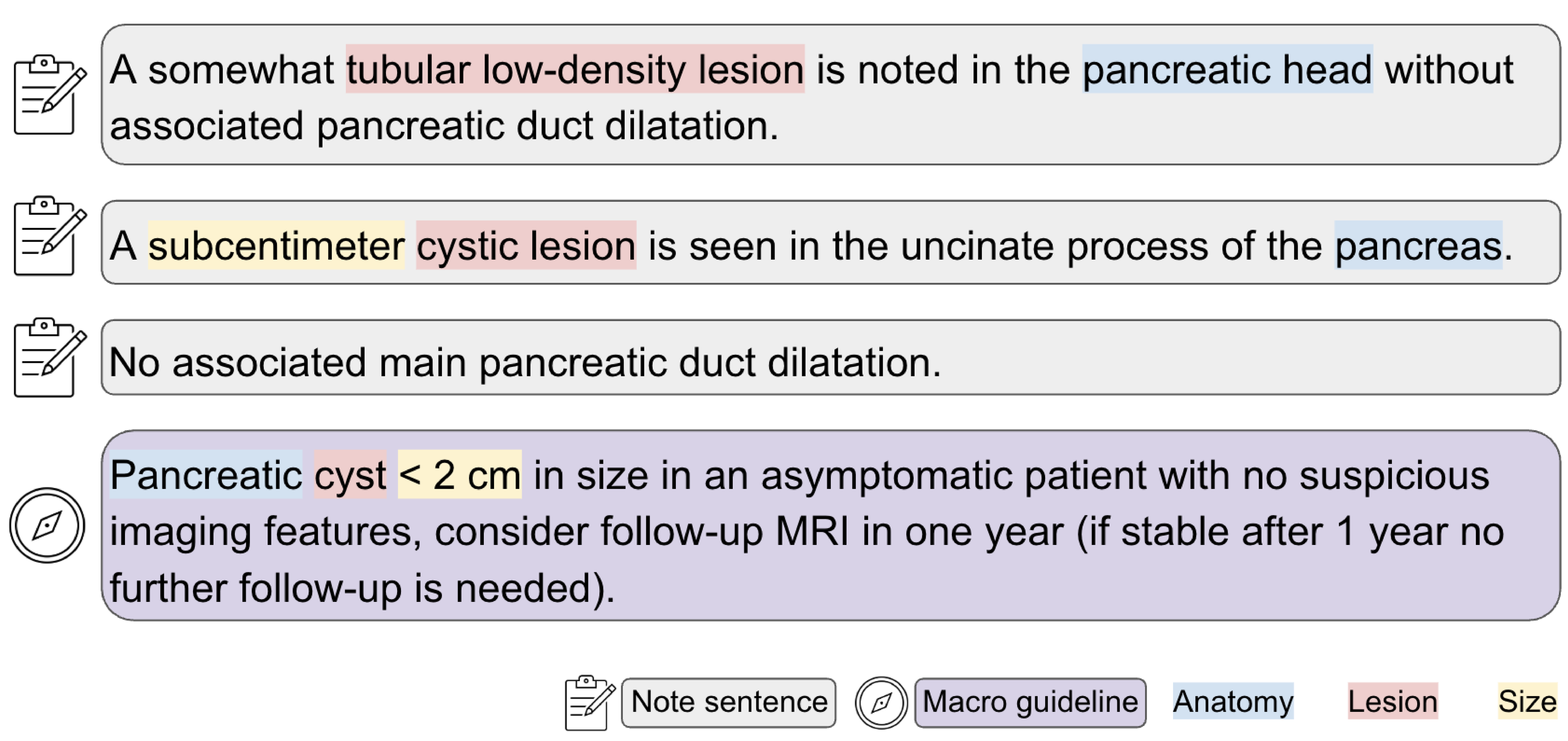

Figure 1. Example note sentences and macro guideline in a radiology report

METHODS

Task & Data We treat macro-guideline detection as a sentence classification task. Sentences were tokenized with NLTK from 150 reports (8,860 sentences). Three radiologists annotated 777 macro sentences and 8,083 non-macro sentences. Macros commonly span multiple sentences (median length=5). We use 5-fold **report-level** cross-validation (80/20 split).

Model

- **TF-IDF + SVM**: baseline
- **ClinicalBERT[2]**: fine-tuned on our dataset; Adam (lr=1e-5), batch size=32, 10 epochs, incorporating class weighting for imbalance; input concatenates **previous + current sentences**
- **gpt-oss-20B (0-shot)**: Instruction-only prompt
- **gpt-oss-20B (Static, 10-shot)**: The base instruction + 10 manually curated examples from the training set
- **gpt-oss-20B (Dynamic, 10-shot)**: The base instruction + top-10 retrieved examples per test sentence using BM25 over the training corpus in each fold

Base prompt You are a radiology domain expert. Your task is to decide whether a sentence in a radiology report is a macro guideline (i.e., a general template or standard recommendation not specific to the current patient). Output only one word: "yes" if the sentence is a macro guideline, or "no" if it is patient-specific content.

RESULTS

- ClinicalBERT achieved the highest overall performance
- Among gpt-oss-20B variants, performance improved with context examples (dynamic 10-shot > static 10-shot > 0-shot)
- Using previous + current sentences yielded the best ClinicalBERT performance

Table 1. Classification performance of the models

Method	Category	P	R	F
TF-IDF + SVM	note sentence	0.9946	0.9968	0.9957
	macro guideline	0.9657	0.9434	0.9544
ClinicalBERT	note sentence	0.9976	0.9969	0.9973
	macro guideline	0.9681	**0.9755**	**0.9718**
gpt-oss-20B (0-shot)	note sentence	0.9525	0.7148	0.8167
	macro guideline	0.1913	0.6543	0.2961
gpt-oss-20B (static, 10-shot)	note sentence	0.9854	0.7295	0.8383
	macro guideline	0.2544	0.8951	0.3962
gpt-oss-20B (dynamic, 10-shot)	note sentence	0.9954	0.9720	0.9836
	macro guideline	0.7789	0.9568	0.8587

Table 2. Classification performance by input sentence length (ClinicalBERT)

Input sentence length	Category	P	R	F
current	note sentence	0.9952	0.9923	0.9938
	macro guideline	0.9208	0.9484	0.9345
previous + current	note sentence	0.9976	0.9969	0.9973
	macro guideline	0.9681	**0.9755**	**0.9718**
previous + current + following	note sentence	0.9968	0.9967	0.9968
	macro guideline	0.9669	0.9648	0.9657

DISCUSSION AND CONCLUSIONS

- ClinicalBERT achieved the highest accuracy and efficiency for macro guideline identification.
- Pre-filtering macro content can improve dataset quality and downstream IE performance.
- The method can generalize to other domains mixing templated and instance-specific text.
- Fine-tuned ClinicalBERT performed best on local institutional data but may overfit when applied to data from other institutions.
- Error analysis
  - False negatives: Commonly involved introductory macro captions (e.g., "Incidental findings and recommendations:") or single-sentence macro recommendations.
  - False positives: Often included non-macro clinical recommendations (e.g., "Recommend follow-up imaging.").

REFERENCES

1. Reichenpfader D, Müller H, Denecke K. A scoping review of large language model based approaches for information extraction from radiology reports. NPJ Digit Med. 2024 24 Aug;7(1):222.
2. Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. arXiv [csCL]. 2019.