Context-Aware Filtering of Unstructured Radiology Reports by Anatomical Region

Zakk Heile ZAKK.HEILE@DUKE.EDU
Department of Computer Science, Duke University

Pranav Manjunath PRANAV.MANJUNATH@DUKE.EDU
Department of Biomedical Engineering, Duke University

Brian Lerner BRIAN.LERNER@DUKE.EDU
Department of Electrical and Computer Engineering, Duke University

Samuel Berchuck SAMUEL.BERCHUCK@DUKE.EDU
Department of Biostatistics & Bioinformatics, Duke University

Monica Agrawal MONICA.AGRAWAL@DUKE.EDU
Department of Computer Science, Duke University

Timothy W. Dunn TIMOTHY.DUNN@DUKE.EDU
Department of Biomedical Engineering, Duke University

Abstract

Radiology reports contain essential clinical information but often remain in unstructured, free-text formats. Notably, multiple imaging examinations performed simultaneously (such as CT head, facial bones, and cervical spine in trauma cases) may be bundled into a single report that consolidates findings from all studies into one free-text document, written jointly. Because individual sentences may reference ambiguous or overlapping anatomy (e.g., "there is a fracture"), sentence-level anatomic classification—filtering a report to retain only findings relevant to a specific anatomical region—is essential for downstream tasks such as structured label extraction and for creating clean, bijective training data for radiology report generation models. While formatting differs across reports, the clinical language remains precise. Using that fact, we develop context-aware classical models with feature engineering that surpass trained neural networks and pre-trained language models. We show that the learned model weights generalize effectively to MIMIC-IV radiology reports and that our approach achieves near-optimal performance with only a small amount of labeled training data. Together, these results make our approach practical and reproducible for new settings.

Keywords: radiology reports, information extraction, anatomical region, context-aware models, clinical NLP, report filtering

Data and Code Availability We make use of two datasets in this work. The first dataset comprises de-identified radiology reports from the Duke Health System (Manjunath et al., 2025). We additionally used the publicly available MIMIC-IV-Note dataset (Johnson et al., 2023; Goldberger et al., 2000), which contains de-identified free-text clinical notes, including radiology reports. Code is available at https://github.com/zakk-h/ContextAwareFiltering and https://doi.org/10.5281/zenodo.17575412.

Institutional Review Board (IRB) This study was approved by the Duke Health Institutional Review Board (Protocol Pro00103826).

1. Introduction

Radiology reports are central to clinical care, serving as the primary record of imaging findings and recommendations. They are increasingly used in machine learning as a source of clinical information, even though they are written in free-text form. This narrative format aids in conveying clinical nuance, but it also results in wide variability in how information is organized and presented. This leads to an inconsistent structure that complicates automated analy-

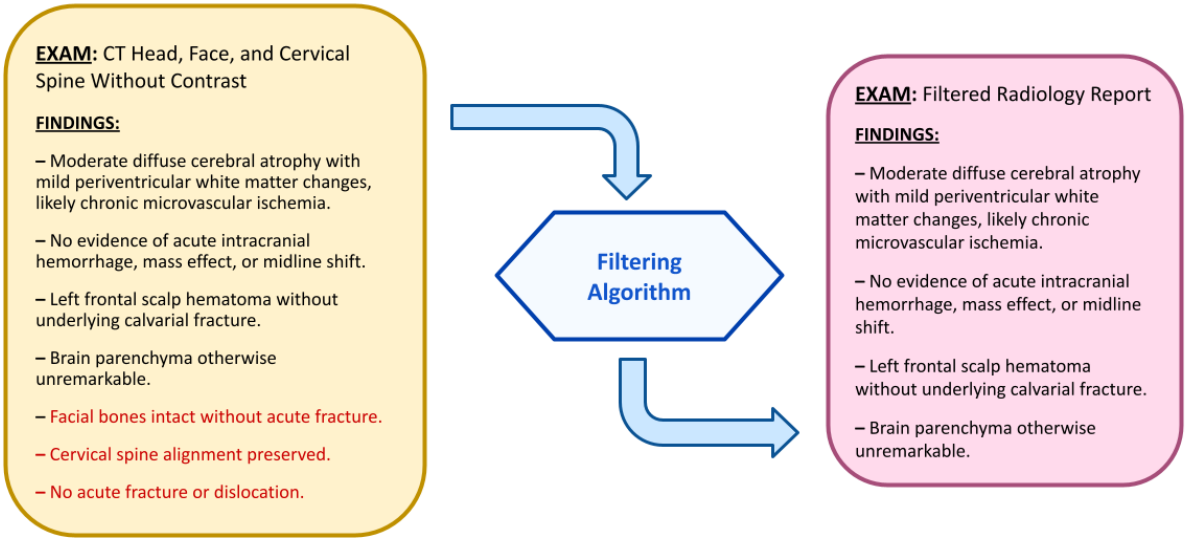

Figure 1: Simplified view of radiology report and task (preserving only the information needed about a head CT). In our algorithm, we handle more than the findings section and much more variability in formatting.

sis and many downstream tasks. For instance, due to various clinical and operational reasons, reports may combine information from several different imaging studies into a single jointly written document, bundling the information together. These and related issues hinder direct use without preprocessing.

For example, automated extraction tasks typically require substantial data cleaning before machine learning models or rules-based algorithms can be applied (Sugimoto et al., 2021; Hu et al., 2024; Zhou et al., 2023; Reichenpfader et al., 2024). Structured labelers like CheXbert and RadGraph expect reports to focus on a single anatomical region; mixed or bundled content breaks this assumption. Additionally, radiology report generation models require a clean, bijective mapping between an image and its corresponding text. Given the anatomical orientation of radiology reporting and the reasons described above, it is natural to organize reports by anatomical regions for easier extraction and downstream use.

Our study focuses on filtering radiology reports to retain only findings from head CTs, though our approaches naturally generalize to any number of anatomical classes. Because radiology reports are unstructured and anatomical relevance often varies at the sentence level, classification should not be performed at a coarser granularity such as the section or report level. It is also unlikely that a single sentence would contain findings from distinctly different regions, making sentence-level classification by anatomical region the most natural approach. While large language models have recently been applied to radiology text processing (Liu et al., 2023; Zhao et al., 2024; Liu et al., 2025), deploying such systems can be impractical in many clinical settings due to substantial compute requirements and potential privacy risks. Ideally, this task could be achieved without reliance on large language models.

In this work, we developed and evaluated classical and neural modeling approaches for sentence-level anatomic classification of radiology reports, in which the model determines whether each sentence pertains to the head CT scan or to non-head content (e.g., the cervical spine or facial bones). To address this task, we design context-aware classical models, including an autoregressive logistic regression approach and conditional random fields, that effectively incorporate neighboring sentences. Our experiments show that these classical models not only outperform trained neural networks and pre-trained language models on reports from our own institution but also retain high performance when their trained weights are applied directly to reports from another institution. We further examine the generalizability and robustness of our models through systematic feature ablation studies and variation of training data size.

Our contributions are as follows:

1. We develop three computationally accessible approaches that achieve test AUROCs above 0.997, including a novel sequence-labeling architecture that incorporates radiology-specific contextual features. All three models outperform both the

prior state-of-the-art method for this task and pre-trained language models.

2. We evaluate our models on an external dataset from another institution and time period, including both individual reports and merged variants designed to mimic mixed formats, and show they generalize significantly better than both the prior state of the art and pre-trained language models.

3. We vary the amount of training data and show that our models are highly data-efficient, outperforming pre-trained language models with only a fraction of the data, making them especially practical and desirable for real-world use like downstream labeling and further machine-learning.

2. Related Work

2.1. Named Entity Recognition

Extracting structured information from radiology reports has long been a goal, with approaches ranging from rule-based systems to machine learning and modern deep learning. Early tools such as MetaMap focused on recognizing biomedical entities and mapping them to a standardized vocabulary (UMLS) using rules and heuristics (Aronson and Lang, 2010).

Following this, statistical learning approaches began to address more complex relationships in text. For example, Roberts et al. (2012) used support vector machines to not only detect anatomical entities but also resolve the anatomical site most closely associated with a clinical finding, working at a highly granular level (Hearst et al., 1998). With the rise of deep learning, systems shifted toward learning context directly from data. Cornegruta et al. (2016) and Casey et al. (2021) built on the previous approaches, labeling parts of the input with entity types and whether they are negated using approaches like BiLSTMs (Graves and Schmidhuber, 2005).

The previous methods are all general-purpose information extraction tools for medical text, but there also exist task-specific labeling systems for radiology reports. RadGraph, for example, uses machine learning to extract entities and relations from chest X-ray reports, while CheXpert (and its BERT-based successor CheXbert) generate predefined pathology labels at the report level, serving primarily as weak supervision for vision models (Jain et al., 2021a; Irvin et al., 2019; Draelos et al., 2021; Smit et al., 2020). These tools do not work in opposition to our approach, for they expect well-defined, homogeneous radiology reports. Our approach enables the use of such tools in broader clinical contexts, given the variability of this data type. These labelers are important for downstream machine learning use as they can be used to train vision models to predict those labels based on only the image (Jain et al., 2021b).

2.2. Radiology Report Generation

Much recent work in machine learning for health has focused on extracting useful information from images. This has come in the form of both visual question answering and radiology report generation (Moor et al., 2023; Sloan et al., 2025; Tanno et al., 2023; Chen et al., 2020; Ostmeier et al., 2024). However, this adds a new source of error for downstream tasks, because generated reports can hallucinate content, generalize poorly, or miss patient-specific details that influence how clinicians write real reports (Huang et al., 2025; Xu et al., 2025). As Nguyen et al. (2023) argues, radiology report generation should consider why the scan was ordered (the clinical indication), not just what is in the image. In other words, the standard image-to-text pipeline overlooks the communicative role of radiology reports, which is to address specific clinical questions for referring physicians. Consequently, access to real radiology reports, even if unstructured, remains very desirable—provided they can be processed and filtered into structured forms that better support clinician-facing tools and downstream machine learning applications. Filtered reports can also serve as high-quality training data for report-generation models, since they contain only information that is actually visible in the scan.

2.3. Anatomic Classification

A closely related study by Nishigaki et al. (2023) seeks to classify the *Findings* section of radiology reports written in Japanese. Their best performing approach fine-tuned a BERT model pretrained on a large Japanese clinical corpus (UTH-BERT) to classify each sentence. Importantly, their method treated sentences independently, without leveraging contextual information from neighboring sentences. Also, in contrast to Nishigaki et al. (2023), we choose to filter the entirety of each report to ensure general applicability, as inconsistency in formatting can make

section-specific approaches unreliable, and to capture clinically relevant information that may appear outside the *Findings* section, such as in *Impressions*.

3. Data

3.1. Duke Health System

Annotators labeled 4,228 sentences from 249 radiology reports with a sentence-level anatomical category indicating whether each sentence described head CT findings or not. These labeled reports come from a larger corpus of de-identified radiology reports sourced from the Duke Health System. This dataset contains 56,659 reports from patients meeting two of three inclusion criteria: head CT performed in the emergency department, chief complaint associated with traumatic brain injury, or traumatic brain injury diagnosis. Because many reports in this cohort contain findings from multiple imaging studies, sentences may reference different anatomical regions. Filtering is required to preserve the clinically relevant head-CT content and to enable downstream analyses with region-specific labelers or when aligning reports to imaging data. In this corpus, 76% of labeled sentences belong to the head CT class. Additionally, in our subset of labeled reports, each report corresponds to a distinct patient, though this need not hold in general.

We define a sentence as being i) between two punctuation marks, ii) at the beginning of the report through the first punctuation mark, or iii) after the last punctuation mark in a report. We also require sentences to have more than two words to avoid trivial consequences of numbered reports (consider ". 2.").

Additionally, reports feature boilerplate language that can be removed as well via rules. The language is standard and contributes no information to the specific findings of the report. This reduces the dataset size to 3786 sentences. The labeled dataset contains 76% brain-related findings, with the remaining 24% being findings from cervical spine and facial bones exams.

3.2. MIMIC

We also test our models on MIMIC-IV notes to assess generalization (Johnson et al., 2023; Goldberger et al., 2000). We assembled an additional test set of 300 reports—100 brain, 100 facial bones, and 100 spine—containing 10,629 sentences solely for testing our models. Due to the clean format of the radiology reports in MIMIC-IV (multi-scan reports are effectively absent in this dataset), we devised methods to simulate bundled reports. We evaluate our models both on the individual reports and on merged variants that simulate bundled and unstructured formats.

3.3. Data Division

We split the Duke Health System dataset into 3,066 training sentences (199 reports) and 720 test sentences (50 reports), ensuring that all sentences from a given report remained within the same partition. An additional 10,629 sentences (300 reports) from the external dataset were used exclusively for out-of-distribution evaluation and testing.

For neural models, a validation set was carved out from the training data for hyperparameter tuning and early stopping; after tuning, part of this validation set was returned to the training set, leaving a smaller portion for early stopping. For classical models, no validation set was needed after tuning, so the entire training set is used.

4. Methods

We contribute three approaches to this task and compare them to pre-trained language models as well as the current state-of-the-art architecture: (1) a novel autoregressive logistic regression framework for sequence labeling with optional feature engineering, (2) a classical conditional random field (CRF) that can incorporate the same domain-informed features, and (3) a neural CRF architecture that integrates clinical embeddings and compartmentalizes information into distinct components while enabling structured interactions between these components.

4.1. Baseline Models

SOTA: The state-of-the-art architecture described in Nishigaki et al. (2023) for this task, outperforming all their other approaches, is to feed a sentence into UTH-BERT (a Japanese clinical-domain BERT) and extract the [CLS] token's output vector to represent the sentence, then feed it into a linear layer and softmax to classify. The mentioned approach fine-tunes BERT and the classification head jointly. We use ClinicalBERT as an analog (Alsentzer et al., 2019).

SLMs: We also compare our performance to small language models (approximately 7B parameters,

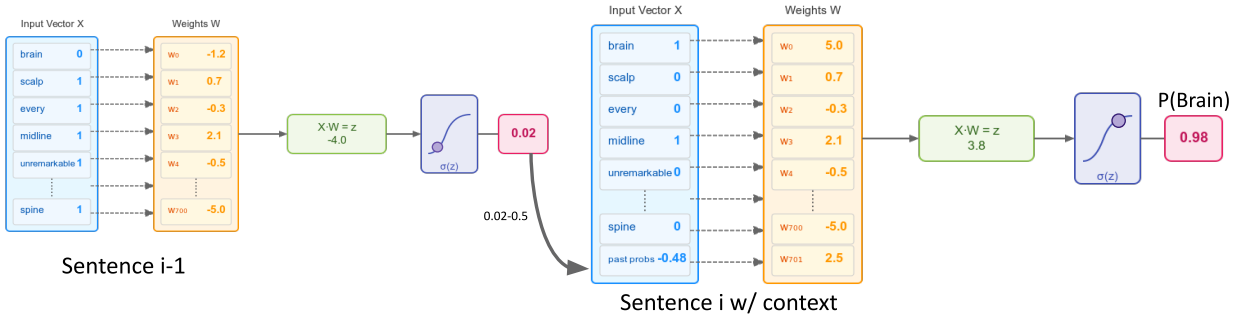

Figure 2: Simplified depiction of the autoregressive logistic regression model. The prediction from the previous sentence is concatenated with the current sentence's embedding to provide contextual information. In practice, we use a context window of three sentences and aggregate past probabilities using simple summary statistics, without altering the scalability.

which may be considered LLMs in some texts), which can be run locally on a single GPU without substantially slow inference. Such models are attractive in healthcare settings where data privacy, regulatory constraints, and limited compute prohibit the deployment of larger LLMs. Notably, our environment does not permit access to HIPAA-compliant LLMs. We evaluated several instruction-tuned and biomedical SLMs, including Qwen 2.5-7B Instruct, Qwen 1.5-7B Chat, and BioMistral-7B (Jiang et al., 2023; Labrak et al., 2024; Team et al., 2025; Bai et al., 2023). Qwen 2.5-7B was consistently the strongest performer, and we report its results in the main experiments.

We conducted experiments with multiple prompting strategies, including a system prompt containing 0–3 labeled examples and sequential sentence-level prediction within a single chat session per report. For a report, we initialized a new chat and fed the model the sentences chronologically, one at a time, with its answer in between, and the full chat history as context for future predictions. We evaluated both providing the full report at the start for context at the beginning versus only supplying the sentences as you go, analogous to the autoregressive logistic regression framework. In all cases, inference was performed with greedy deterministic decoding (temperature = 0), and we additionally explored varying the decision threshold by computing the ratio $p(\text{brain} \mid \text{chat history})/p(\text{not brain} \mid \text{chat history})$ with adjustments to accommodate multi-token outputs.

4.2. Autoregressive Logistic Regression

First, we introduce a logistic regression architecture that incorporates features from earlier sentences' model outputs—an approach we refer to as autoregressive in the sense that each prediction depends on summary statistics of past predictions. As such, it is also an online algorithm.

We embed each sentence with a frequency-invariant binary bag of words embedding, possibly excluding rare words. There are two components of the model: a base model *BaseLogReg*, and the model that outputs the final predictions, *MetaLogReg*.

Consider the base logistic regression model. *BaseLogReg* will give a probability that the sentence should be included (it is in the desired anatomical region), given its embedding.

MetaLogReg will predict on sentence i in the report using its embedding, but also past context in the form of functions of *BaseLogReg* (a simplified view is shown in Figure 2). We take the average, minimum, and maximum of the last $k = 3$ *BaseLogReg* probabilities (and those k probabilities themselves)

As shown in our feature ablation studies, there are diminishing returns after the first addition. We subtract 0.5 from the aggregation features to improve interpretability, effectively shifting their range from $(0, 1)$ to $(-0.5, 0.5)$, so that the threshold for decision-making aligns with 0. Where a prediction is undefined because j is negative, we substitute 0 (or 0.5 before shifting) as our uninformative prior.

Formally, this transformation does not change the optimal feature weights of the logistic regression model. However, it does shift the intercept by an amount equal to

$$\Delta b = 0.5 \times \sum_{j \in S} w_j,$$

where S is the set of feature indices that are shifted, and w_j denotes the weight assigned to feature j.

We also explored augmenting the bag-of-words embeddings with additional feature engineering de-

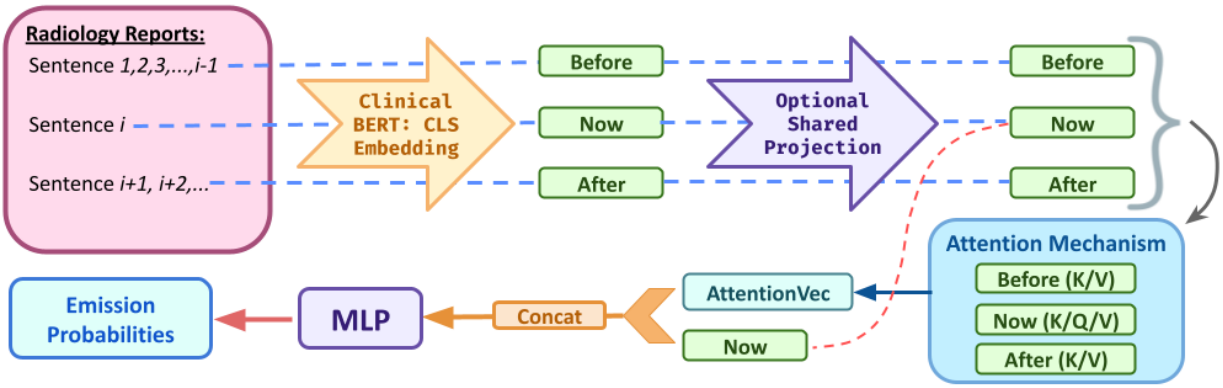

Figure 3: Architecture for computing learned emission probabilities in our Neural CRF. All components except BERT are trained end-to-end.

signed to capture global report structure and resolve ambiguous sentences. While the models trained with these features heavily depend on them, we found that they have little effect on performance. Thus, for accessibility and ease of adoption, we therefore report all main results without them; see Appendix A for details and comparisons.

We note several appealing properties of this architecture: each component is globally optimal, and the overall model captures report-wide structure, including interactions and past probabilities. We could even allow direct interactions between past probabilities and the current sentence embedding—something not possible in architectures like CRFs.

Formally, let $\mathbf{x}_{1:T} = (x_1, \ldots, x_T)$ and $\hat{\mathbf{p}}^{(0)}_{1:T} = (\hat{p}^{(0)}_1, \ldots, \hat{p}^{(0)}_T)$ denote the sequence of sentence features and the sequence of stage-0 (base model) probabilities, respectively. The model then predicts

$$\Pr(y_t = 1 \mid \mathbf{x}_{1:T}, \hat{\mathbf{p}}^{(0)}_{1:T}) = f\left(\mathbf{x}_{1:T}, \hat{\mathbf{p}}^{(0)}_{1:T}, t\right).$$

We choose to condition only on past inputs and predictions; this formulation is compatible with incremental processing and real-time clinical workflows.

We train our two logistic regression models with default regularization. To prevent data leakage from the base model to the meta model on the training set, we train the base model on all but one fold, make predictions on that fold to be used by *MetaLogReg*, and rotate what data it was trained on to make predictions for the whole training set.

4.3. Classical Conditional Random Field

We can directly apply all of the feature engineering, except for prior probabilities and (5), to a classical linear-chain conditional random field (CCRF), which compartmentalizes the prediction into transition and emission probabilities (Lafferty et al., 2001).

4.4. Neural Conditional Random Field

Thirdly, we can replace our feature-engineered CRF with a neural version. We highlight all architecture options we consider.

For sentence i, we construct three different embeddings using (frozen) BioClinicalBERT: past, current, and future sentence embeddings. Past and future may either be derived from just sentence $i-1$ or $i+1$, or may have all prior or all future context. We consider embedding with CLS, component-wise max per token, or average token embedding. Each of these embeddings is optionally projected through a shared linear layer to a lower-dimensional space. These three context vectors are then stacked, and a self-attention mechanism may be applied to facilitate interactions. To prevent the loss of direct signal from the current sentence, we concatenate this attention vector with the original "now" embedding. If we don't apply the attention mechanism, then we just concatenate all three vectors. This combined representation is then passed through an MLP to produce emissions for a linear-chain CRF.

We created a validation set using 30% of the training data and performed hyperparameter tuning over the main architectural features (see Appendix A for details). The best configuration was: `proj_dim` = (no projection), `attn_dim` = 256, `mlp_dim` = 128, `dropout` = 0.6, `weight_decay` = 0.05, giving it just over 700,000 parameters. This also utilizes CLS embeddings and full past and future history.

5. Evaluations

We evaluate models trained and tested on Duke data, their robustness under reduced training data, and their generalization to MIMIC-IV with and without retraining. Across all experiments, the task is binary

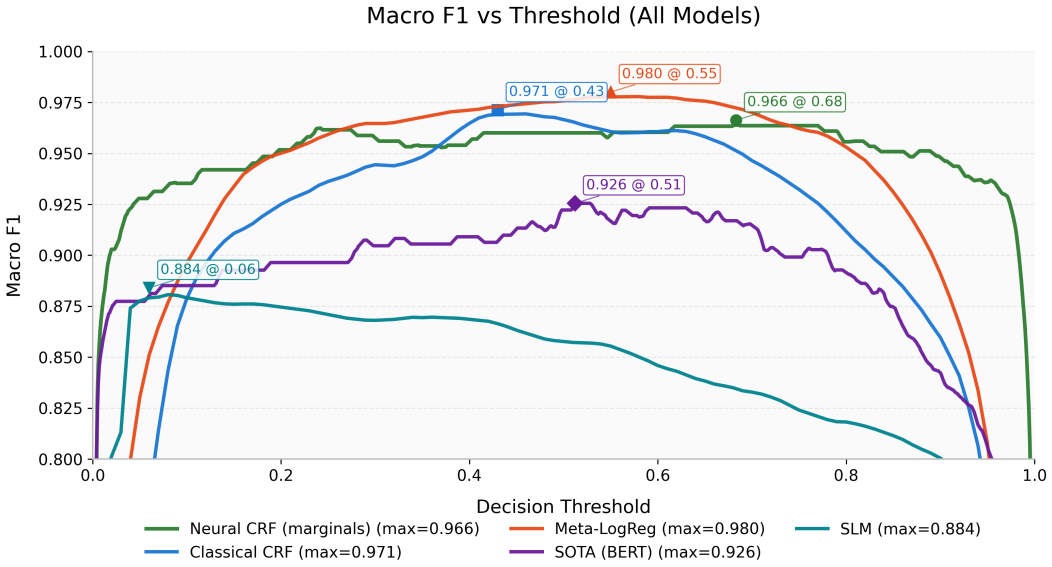

Figure 4: Macro F1 on the Duke Health System test set across different thresholds.

sentence-level classification: determining whether a sentence describes head CT findings. AUROC and Macro F1 use per-sentence probabilities (model outputs or CRF marginals), while hard predictions use a 0.5 threshold for logistic-regression models and MAP decoding for CRFs.

5.1. Performance on Duke Dataset

As seen in Figure 4, our three modeling approaches achieve substantially higher Macro F1 than the state-of-the-art model and Small Language Models.

Model	AUROC
Meta-LogReg	0.997
Neural CRF	0.997
Classical CRF	0.997
SOTA	0.984
SLM	0.972

Table 1: Duke Health System test set, averaged over random seeds if training was non-deterministic.

Additionally, our three proposed approaches perform nearly equally in terms of AUROC (Table 1).

5.2. Required Training Data

We ask the question of how much data our classical models need to perform well to establish how accessible they are to use over a pre-trained language model. We fix the test set and sample a proportion p of the training set reports and display the test accuracy averaged across random seeds in the plot.

Figure 5 shows that the autoregressive logistic regression model consistently outperforms all other modeling approaches. Notably, the CRF models underperformed all other modeling approaches with limited data, possibly because of poorly calibrated transition probabilities. Additionally, even with just 25% of the training data, the classical modeling approaches perform nearly as well. This shows that these approaches are accessible to use with only a small amount of labeled data. Feature engineering provides clear benefits in the low-data regime, yet hinders performance once more training data is available (see Appendix A).

5.3. Generalizability to MIMIC-IV

Motivated by the need for models that remain reliable under distributional shift, we next assess external generalization. We fix our model weights and evaluate the models (trained on the Duke Health System dataset) on the 300 reports derived from MIMIC-IV notes described previously (Johnson et al., 2023; Goldberger et al., 2000).

Given that MIMIC-IV does not contain the bundled, multi-scan reports that motivate our task, we evaluate our models in three different ways: (1) testing on individual reports directly, (2) combining multiple reports by simple concatenation, and (3) constructing synthetic bundled reports using rules-based reasoning to extract and reorganize sections to emulate how they may appear jointly in practice. See

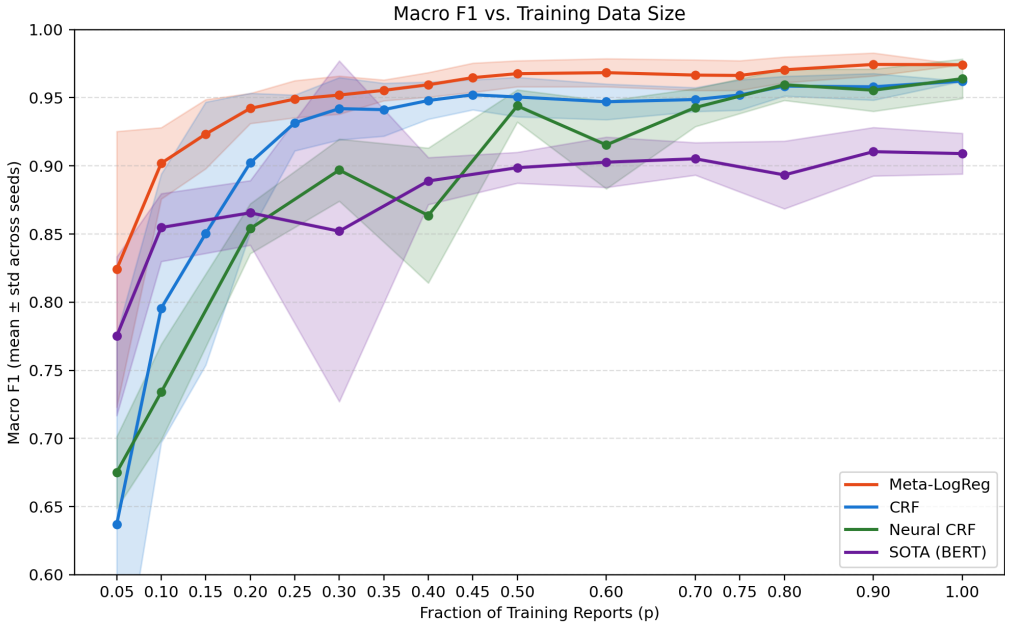

Figure 5: Model Test Performance on the Duke Health System dataset when training on some proportion p of the training reports (averaged across 10 random seeds). ± Standard deviation is shaded.

Appendix A for more information on how these joint reports were constructed.

Model	Indiv.	Concat	Bundled
Meta-LogReg	0.912	0.901	0.937
CCRF	0.965	0.977	0.956
Neural CRF	0.903 ± 0.009	0.893 ± 0.012	0.897 ± 0.016
SOTA	0.850 ± 0.022	0.861 ± 0.020	0.886 ± 0.016

Table 2: AUROC across three evaluation settings with mean ± std reported for models that have non-deterministic training. These models were trained on the Duke Health System dataset and tested on MIMIC.

Across the three report-merging methods shown in Table 2—individual reports, simple concatenations, and more complex rule-based bundles—our methods remain robust. This demonstrates that our models perform consistently, independent of report structure. In all cases, our context-aware approaches outperform the prior state-of-the-art.

5.4. Training on MIMIC-IV

Because our goal is not only to demonstrate generalizability of interpretable modeling approaches but also to show that these architectures remain reliable and effective when trained on a small amount of institution-specific data, we additionally train each model using only 10% of the available MIMIC-IV bundled reports.

Table 3: Training and testing on the created MIMIC Bundled Reports (10% Train, 90% Test), averaged across train–test splits.

Model	Mean ± Std (AUROC)
Classical CRF	0.977 ± 0.002
Meta-LogReg	0.974 ± 0.005
Neural CRF	0.931 ± 0.011
SOTA	0.951 ± 0.003

As shown in Table 3, the classical models again have superb discrimination capability even with training on a small fraction of the labeled reports. They also achieve AUROCs above 0.97 with extremely small variance across splits, indicating that these models are not only accurate but also stable under limited supervision.

This mirrors the behavior we observe on the Duke Health System dataset and highlights a practical advantage: institutions can train reliable, lightweight filters using only a modest number of locally labeled reports, without requiring large corpora, resource-

intensive fine-tuning, or on-premise LLM deployments.

6. Discussion

Our results clearly show that carefully engineered classical models achieve stellar performance on anatomical region classification in unstructured radiology reports. Our autoregressive logistic-regression model is novel in sequence labeling and performs comparably to the classical CRF on held-out reports from the same institution. Importantly, it offers practical advantages: it operates at the sentence level—making individual predictions directly interpretable—and it yields straightforward per-sentence probability estimates. It is globally optimal at each step and can consider context arbitrarily ahead or behind. Notably, limiting the context window to earlier sentences allows for live integration into clinical dictation environments.

With just $20-25\%$ of the training data, both classical models perform essentially just as well as they do with all the data. This is significant because it shows that these approaches remain practical without demanding large amounts of annotation time. Additionally, they are vastly superior to using small pretrained language models, or even bigger ones. However, this is not strictly required because our models generalize to mixed MIMIC-IV radiology reports with up to 0.977 AUROC.

In contrast, our Neural CRF performed well on the Duke test set, but both it and the SOTA model failed to generalize when their weights were applied to MIMIC, each dropping by roughly 0.1 AUROC. This motivates considering why simpler models behaved differently. Theoretical and empirical results show that interpretable models often match or exceed the performance of more complex methods on tabular datasets (Rudin et al., 2024; Semenova et al., 2023). This makes it plausible that converting clinical text into a tabular representation preserved the essential predictive structure. When this structure is retained, the problem effectively becomes tabular—and in that domain, simple interpretable models are known to perform exceptionally well.

6.1. Limitations and Future Work

Although our models perform well on the given task, we only sought to extract information from one anatomical region. However, all of our methods extend naturally to multiclass settings. In most downstream tasks, the primary goal remains isolating one anatomical region with minimal false inclusions.

While we test our models on data from two different institutions, with reports from different time periods, publicly available datasets like MIMIC-IV have processed radiology reports without the unstructured and bundled issues we described. Although we attempted to inject information from individual MIMIC reports to form a multi-anatomy report with variable formatting, it may not fully capture the distributional characteristics of naturally occurring clinical text. Even so, these single-scan and synthetic-merge results are encouraging, implying that our earlier architectures trained on large datasets could generalize across institutions with great performance.

In addition to training these architectures on larger and more general datasets, one could incorporate a human-in-the-loop correction mechanism to improve labeling consistency. For example, if a clinician overrides the predicted label of a specific sentence, both the conditional random field (CRF) and the autoregressive logistic regression (MetaLogReg) models can be modified to condition future predictions on the known label. However, we view these extensions as complementary rather than central to the work: our primary contribution is providing a reliable filtering mechanism that enables downstream labeling and future model development.

While large language models are powerful, deploying them on protected clinical text is often limited by institutional privacy requirements and the cost of maintaining secure, on-premise models. When an institution has access to such a system, it could be used to pseudolabel a modest number of reports, after which a lightweight model can be trained and deployed efficiently. In this sense, LLM-assisted labeling and classical approaches are complementary: LLMs generate labels when available, while classical models yield inexpensive and deployable inference.

7. Conclusion

Our work demonstrates that simple, interpretable, and data-efficient classical sequence models can robustly filter unstructured radiology reports to preserve anatomy-specific content. The approach is computationally lightweight, generalizes across institutions, and integrates naturally into both real-time clinical workflows as well as to obtain clean training data for image-to-text models.

Acknowledgments

We gratefully acknowledge the support of Bass Connections at Duke University. This research was supported in part by the National Institutes of Health (NIH) under award R01NS123275. We would additionally like to thank Dr. Bradley Kolls and Dr. Audrey Verde for valuable discussions throughout this project.

References

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. Publicly available clinical bert embeddings, 2019. URL https://arxiv.org/abs/1904.03323.

Alan R Aronson and François-Michel Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, may 2010. doi: 10.1136/jamia.2009.002733.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report, 2023. URL https://arxiv.org/abs/2309.16609.

Abigail Casey, Emma Davidson, Michael Poon, Honghan Dong, Daniel Duma, Alexandros Grivas, Claire Grover, Víctor Suárez-Paniagua, Richard Tobin, William Whiteley, Honghan Wu, and Beatrice Alex. A systematic review of natural language processing applied to radiology reports. *BMC Medical Informatics and Decision Making*, 21(1):179, jun 2021. doi: 10.1186/s12911-021-01533-7.

Lihu Chen and Gaël Varoquaux. What is the role of small models in the llm era: A survey, 2025. URL https://arxiv.org/abs/2409.06857.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1979–1994. Association for Computational Linguistics, November 2020. URL https://aclanthology.org/2020.emnlp-main.727.

Savelie Cornegruta, Robert Bakewell, Samuel Withey, and Giovanni Montana. Modelling radiological language with bidirectional long short-term memory networks. In Cyril Grouin, Thierry Hamon, Aurélie Névéol, and Pierre Zweigenbaum, editors, *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, pages 17–27, Austin, TX, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-6103. URL https://aclanthology.org/W16-6103/.

Rachel Lea Draelos, David Dov, Maciej A. Mazurowski, Joseph Y. Lo, Ricardo Henao, Geoffrey D. Rubin, and Lawrence Carin. Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. *Medical Image Analysis*, 67:101857, January 2021. ISSN 1361-8415. doi: 10.1016/j.media.2020.101857. URL http://dx.doi.org/10.1016/j.media.2020.101857.

A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. C. Ivanov, R. Mark, and H. E. Stanley. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. RRID:SCR_007345.

Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610, jun 2005. doi: 10.1016/j.neunet.2005.06.042.

M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28, 1998. doi: 10.1109/5254.708428.

Danqing Hu, Bing Liu, Xiaofeng Zhu, Xudong Lu, and Nan Wu. Zero-shot information extraction from radiological reports using chatgpt. *International Journal of Medical Informatics*, 183:105321,

March 2024. ISSN 1386-5056. doi: 10.1016/j.ijmedinf.2023.105321. URL http://dx.doi.org/10.1016/j.ijmedinf.2023.105321.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, January 2025. ISSN 1558-2868. doi: 10.1145/3703155. URL http://dx.doi.org/10.1145/3703155.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.3301590. URL https://doi.org/10.1609/aaai.v33i01.3301590.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports (version 1.0.0), 2021a. URL https://doi.org/10.13026/hm87-5p47. RRID:SCR_007345.

Saahil Jain, Akshay Smit, Andrew Y. Ng, and Pranav Rajpurkar. Effect of radiology report labeler quality on deep learning models for chest x-ray interpretation, 2021b. URL https://arxiv.org/abs/2104.00793.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

A. Johnson, T. Pollard, S. Horng, L. A. Celi, and R. Mark. Mimic-iv-note: Deidentified free-text clinical notes (version 2.2), 2023. URL https://doi.org/10.13026/1n74-ne17. RRID:SCR_007345.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. BioMistral: A collection of open-source pretrained large language models for medical domains. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.348. URL https://aclanthology.org/2024.findings-acl.348/.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-778-1. URL http://dl.acm.org/citation.cfm?id=645530.655813.

Zhengliang Liu, Tianyang Zhong, Yiwei Li, Yutong Zhang, Yi Pan, Zihao Zhao, Peixin Dong, Chao Cao, Yuxiao Liu, Peng Shu, Yaonai Wei, Zihao Wu, Chong Ma, Jiaqi Wang, Sheng Wang, Mengyue Zhou, Zuowei Jiang, Chunlin Li, Jason Holmes, Shaochen Xu, Lu Zhang, Haixing Dai, Kai Zhang, Lin Zhao, Yuanhao Chen, Xu Liu, Peilong Wang, Pingkun Yan, Jun Liu, Bao Ge, Lichao Sun, Dajiang Zhu, Xiang Li, Wei Liu, Xiaoyan Cai, Xintao Hu, Xi Jiang, Shu Zhang, Xin Zhang, Tuo Zhang,

Shijie Zhao, Quanzheng Li, Hongtu Zhu, Dinggang Shen, and Tianming Liu. Evaluating large language models for radiology natural language processing, 2023. URL https://arxiv.org/abs/2307.13693.

Zhengliang Liu, Yiwei Li, Peng Shu, Aoxiao Zhong, Hanqi Jiang, Yi Pan, Longtao Yang, Chao Ju, Zihao Wu, Chong Ma, Cheng Chen, Sekeun Kim, Haixing Dai, Lin Zhao, Lichao Sun, Dajiang Zhu, Jun Liu, Wei Liu, Dinggang Shen, Quanzheng Li, Tianming Liu, and Xiang Li. Radiology-gpt: A large language model for radiology. *Meta-Radiology*, 3(2):100153, 2025. ISSN 2950-1628. doi: https://doi.org/10.1016/j.metrad.2025.100153. URL https://www.sciencedirect.com/science/article/pii/S2950162825000219.

Pranav Manjunath, Brian Lerner, and Timothy W Dunn. Trust Your Neighbors: Multimodal Patient Retrieval for TBI Prognosis. working paper or preprint, September 2025. URL https://hal.science/hal-05285969.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. 225:353–367, 10 Dec 2023. URL https://proceedings.mlr.press/v225/moor23a.html.

Dang Nguyen, Chacha Chen, He He, and Chenhao Tan. Pragmatic radiology report generation. 225:385–402, 10 Dec 2023. URL https://proceedings.mlr.press/v225/nguyen23a.html.

Daichi Nishigaki, Yuki Suzuki, Takuma Wataya, Kazuyuki Kita, Kenta Yamagata, Junya Sato, Shoichiro Kido, and Noboru Tomiyama. Bert-based transfer learning in sentence-level anatomic classification of free-text radiology reports. *Radiology: Artificial Intelligence*, 5(2):e220097, feb 2023. doi: 10.1148/ryai.220097.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and Jean-Benoit Delbrouck. Green: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, page 374–390. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.21. URL http://dx.doi.org/10.18653/v1/2024.findings-emnlp.21.

Daniel Reichenpfader, Jonas Knupp, André Sander, and Kerstin Denecke. Radex: A framework for structured information extraction from radiology reports based on large language models, 2024. URL https://arxiv.org/abs/2406.15465.

Kirk Roberts, Bryan Rink, Sanda M Harabagiu, Richard H Scheuermann, Sean Toomay, Todd Browning, Tom Bosler, and Ronald Peshock. A machine learning approach for identifying anatomical locations of actionable findings in radiology reports. *AMIA Annual Symposium Proceedings*, 2012:779–788, nov 2012.

Cynthia Rudin, Chudi Zhong, Lesia Semenova, Margo Seltzer, Ronald Parr, Jiachang Liu, Srikar Katta, Jon Donnelly, Harry Chen, and Zachery Boner. Position: amazing things come from having many good models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Lesia Semenova, Harry Chen, Ronald Parr, and Cynthia Rudin. A path to simpler models starts with noise. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 3362–3401. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/0a49935d2b3d3342ca08d6db0adcfa34-Paper-Conference.pdf.

Phillip Sloan, Philip Clatworthy, Edwin Simpson, and Majid Mirmehdi. Automated radiology report generation: A review of recent advances. *IEEE Reviews in Biomedical Engineering*, 18: 368–387, 2025. ISSN 1941-1189. doi: 10.1109/rbme.2024.3408456. URL http://dx.doi.org/10.1109/RBME.2024.3408456.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert, 2020.

Kento Sugimoto, Toshihiro Takeda, Jong-Hoon Oh, Shoya Wada, Shozo Konishi, Asuka Yamahata,

Shiro Manabe, Noriyuki Tomiyama, Takashi Matsunaga, Katsuyuki Nakanishi, and Yasushi Matsumura. Extracting clinical terms from radiology reports with deep learning. *Journal of Biomedical Informatics*, 116:103729, 2021. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2021.103729. URL https://www.sciencedirect.com/science/article/pii/S1532046421000587.

Ryutaro Tanno, David G. T. Barrett, Andrew Sellergren, Sumedh Ghaisas, Sumanth Dathathri, Abigail See, Johannes Welbl, Karan Singhal, Shekoofeh Azizi, Tao Tu, Mike Schaekermann, Rhys May, Roy Lee, SiWai Man, Zahra Ahmed, Sara Mahdavi, Yossi Matias, Joelle Barral, Ali Eslami, Danielle Belgrave, Vivek Natarajan, Shravya Shetty, Pushmeet Kohli, Po-Sen Huang, Alan Karthikesalingam, and Ira Ktena. Consensus, dissensus and synergy between clinicians and specialist foundation models in radiology report generation, 2023. URL https://arxiv.org/abs/2311.18260.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models, 2025. URL https://arxiv.org/abs/2401.11817.

Xingmeng Zhao, Tongnian Wang, and Anthony Rios. Improving expert radiology report summarization by prompting large language models with a layperson summary, 2024. URL https://arxiv.org/abs/2406.14500.

Sitong Zhou, Meliha Yetisgen, and Mari Ostendorf. Building blocks for complex tasks: Robust generative event extraction for radiology reports under domain shifts. In Tristan Naumann, Asma Ben Abacha, Steven Bethard, Kirk Roberts, and Anna Rumshisky, editors, *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 344–357, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.clinicalnlp-1.38. URL https://aclanthology.org/2023.clinicalnlp-1.38/.

Appendix A. Appendix

A.1. LoRA

In addition to employing the state-of-the-art architecture described by Nishigaki et al. (2023), we also explore parameter-efficient finetuning using LoRA (Hu et al., 2022). The results with this approach on the Duke Health System dataset test set are shown in Table 4.

Rank	Class	Precision	Recall	Accuracy
4	0	0.7623	0.8857	0.9354
4	1	0.9766	0.9453	0.9354
8	0	0.7895	0.8571	0.9386
8	1	0.9712	0.9547	0.9386
12	0	0.9149	0.8190	0.9575
12	1	0.9649	0.9849	0.9575
16	0	0.8725	0.8476	0.9543
16	1	0.9700	0.9755	0.9543

Table 4: Test set precision, recall, and accuracy for LoRA-BERT models at different ranks. Performance improves with larger LoRA rank updates but still falls short of full BERT finetuning (though it is not significantly different).

A.2. Feature Engineering

We also consider augmenting the bag-of-words (BOW) embeddings with additional feature engineering. Many of these capture global structure of the report outside of just the current sentence.

1. Sentence position flag (`Sent_Num` $\leq$ `3`)
2. `spine_earlier`, `facial_bones_earlier` indicators
3. Each word × `spine_earlier`
4. Each word × `facial_bones_earlier`
5. (Optional) Each word × `Prior_1`
6. Top $\ell = 300$ most frequent word × word pairs

(1) is intended to detect the header, (2) and (3), and extensions are to detect any header information (if in the report), or mention of specific regions that would tend to reveal the presence of information about a specific scan. Interaction terms in (3), (4), (5), and (6) seek to give clarity to ambiguous words. However, we find that these additional features are not required to achieve strong performance, so we omit them in our main experiments for simplicity. Although models trained with them place substantial weight on these engineered features, they do not yield measurable performance gains, as demonstrated later in this section.

Table 5: Without Feature Engineering

Class	Precision	Recall
0	0.900 / 0.688	0.611 / 0.155
1	0.929 / 0.857	0.983 / 0.996

Table 6: With Feature Engineering

Class	Precision	Recall
0	0.895 / 0.835	0.722 / 0.417
1	0.948 / 0.898	0.981 / 0.991

We evaluated these features in an extremely low-data regime, using just 5% of the training reports. As shown in the Table 5 and Table 6, feature engineering led to a substantial improvement in performance—particularly for the classical CRF, which was otherwise unstable on minority-class examples. For the MetaLogReg model, gains were more modest but still noticeable.

Without feature engineering, the CRF struggled to recall non-brain sentences (15.5% recall for class 0), whereas feature engineering improved this to 41.7%. The MetaLogReg model already performed well without features, but still saw an increase in precision and recall for both classes with their inclusion.

However, as shown in Figure 6 and Figure 7, the benefits of feature engineering disappear once the model is trained on at least half of the labeled data. At $p = 0.5$ and above, both MetaLogReg and CRF perform nearly identically with or without engineered features. This suggests that while domain-informed features are helpful in low-data regimes, they are ultimately unnecessary for models trained on more substantial datasets.

Following Occam's razor, we ran all of our classical model experiments in the main text without these additional features. However, just because there exist equally good models that can choose between all the features and the restricted feature set does not mean

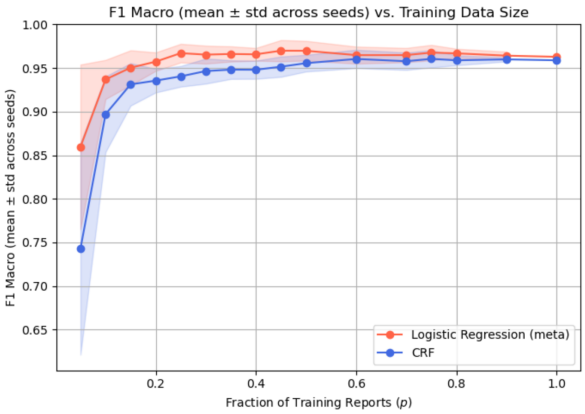

Figure 6: (a) With Feature Engineering

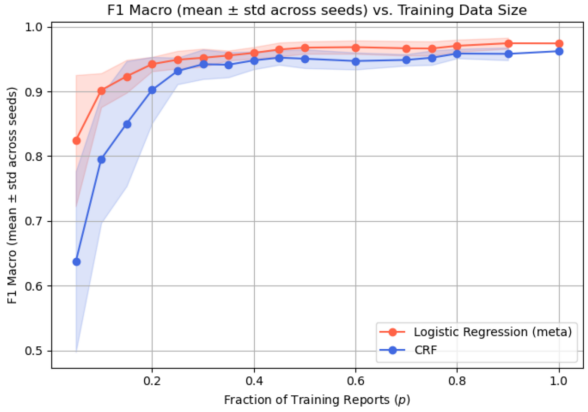

Figure 7: (b) Without Feature Engineering

that the six categories of features we proposed adding were not useful to a good model.

For completeness, we also show in Figure 8 how the feature-engineered models would perform on the MIMIC-IV test data. Here, the two classical approaches use feature engineering and perform worse than they did without feature engineering in Table 2. This shows that the feature engineering (which did not impact performance on the Duke Health System dataset) did not generalize sufficiently well to MIMIC-IV reports. Despite this fact, however, they still outperform our NeuralCRF, the state-of-the-art model, and pre-trained language models.

A.3. Need for Prior Features

One of the defining features of our novel autoregressive logistic regression approach is the aggregation of prior predictions. In this section, we study how the size of our context window (which we restrict to only earlier sentences to allow for applications in streaming) impacts model performance. Likewise, we will also explore the impact of the aggregation function used to summarize the context window.

First, we establish that some form of aggregation is needed. To do this, we train *MetaLogReg* without any prior features and display the results below in Figure 9. Notably, for the first time, autoregressive logistic regression is severely underperforming the classical conditional random field. Compared to the results in Figure 7, we see that the model without prior information is strictly worse at differentiating between the classes.

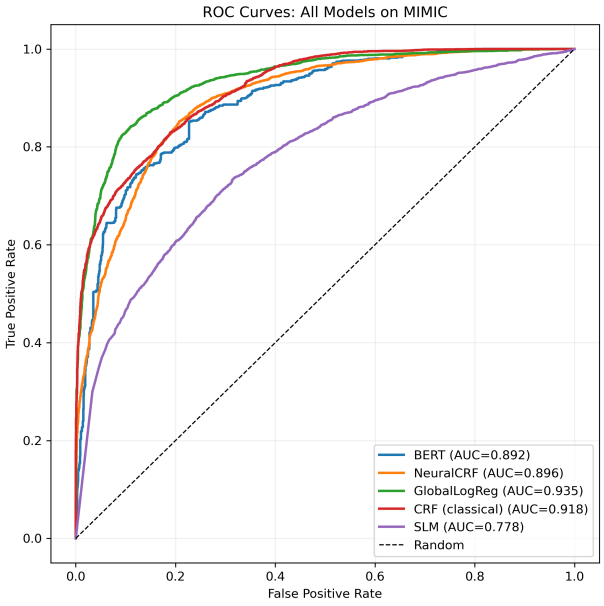

Figure 8: AUC for all models on MIMIC-IV synthetically merged reports.

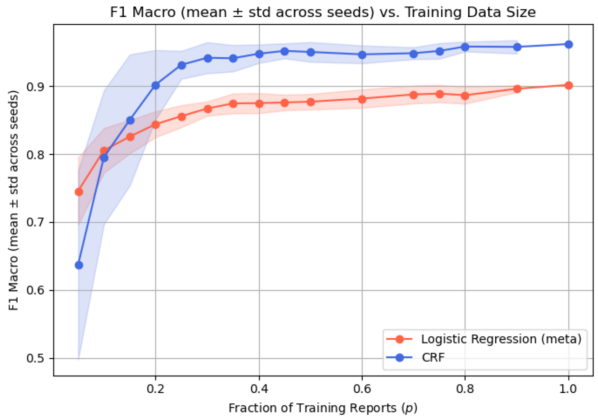

Figure 9: No Feature Engineering for Classical Models and LogReg without Prior Features

Next, we evaluate the effect of varying the context window size k, where only predictions from the k previous sentences are considered. We constrain the model to use past-only context to allow deployment in real-time or streaming settings. As shown in Table 7, performance improves notably when moving from k=1 (AUC 0.9902) to k=3 (AUC 0.9971), but plateaus thereafter. This suggests that modest context is sufficient for most sentences, and that large windows yield diminishing returns.

Table 7: Sentence-level AUC for the autoregressive logistic regression model when varying context window sizes k and ablations at $k = 3$.

Model Variant	**AUC**
Context window $k = 5$	0.9962
Context window $k = 4$	0.9970
Context window $k = 3$	0.9971
Context window $k = 2$	0.9963
Context window $k = 1$	0.9902
$k = 3$, without average pooling	0.9969
$k = 3$, without max pooling	0.9970
$k = 3$, without min pooling	0.9971
$k = 3$, without any pooling	0.9971

Finally, we perform an ablation study to assess the importance of each aggregation function at $k = 3$. The model computes the average, maximum, and minimum of prior probabilities across the context window. Removing any one of these has minimal effect to no effect, indicating mild redundancy. Removing all of them also does not affect model performance. Thus, while the choice of aggregation functions can be simplified with minimal performance loss, maintaining a reasonably sized context window remains important.

A.4. Feature Importances

To better understand what drives the predictions of our autoregressive logistic regression model, we examine the most influential features by inspecting the learned coefficients. Since our model is linear and additive in its input features, the magnitude and sign of each coefficient provide a direct measure of its impact on the output probability. This interpretability is one of the primary advantages of our approach over more complex models.

We compare two variants of MetaLogReg: one trained with domain-informed feature engineering, and one trained without it (relying only on word tokens and prior predictions).

Despite achieving similar overall performance, the MetaLogReg models trained with and without feature engineering rely on substantially different sets of features. Among the top 25 features ranked by absolute coefficient magnitude, only 7 features overlap between the two models—including intracranial, brain, head, globe, and certain prior-based predictions. This low overlap (Jaccard index ≈ 0.16) reflects a sharp shift in what the model considers most predictive once structured, domain-informed features are introduced. This has many interesting impacts on interpretability (Rudin et al., 2024).

A.5. Hyperparameter Tuning

Many of the models we evaluated require minimal hyperparameter tuning. The small language models are used in a zero- or few-shot setting with no trainable parameters. The two classical models involved only a single regularization parameter. For fine-tuning BERT with a linear classification head, we used Adam without weight decay, so as not to impose strong regularization that could disrupt the initialization inherited from the pre-trained weights. The only model with a substantial hyperparameter space was the Neural CRF, for which we conducted a grid search; the explored ranges and selected configuration are summarized in Table 10.

Table 8: Top 25 Features by Absolute Coefficient Magnitude (MetaLogReg with Feature Engineering)

Rank	Feature	Coefficient
1	intracranial	1.9359
2	Prior_123on_Max	1.8694
3	SentNumLEQ3	1.7594
4	Prior_1_Prediction	1.7300
5	spine_earlier	-1.6949
6	Facial_Bones_earlier	-1.6472
7	spine_earlier_x_FacialBones	1.6210
8	acute_x_intracranial	1.3583
9	Prior_123on_Min	1.3087
10	acute_x_fracture	-1.2211
11	face	-1.2086
12	Prior_123on_Max_x_FacialBones	-1.1795
13	globe	-1.1643
14	brain	1.1609
15	Prior_123on_Avg	1.1176
16	Prior_2_Prediction	1.0333
17	normal_x_paranasal	1.0262
18	cervical	-0.9898
19	gas	0.9606
20	normal_x_paranasal_x_FacialBones	0.9481
21	head	0.9399
22	Prior_3_Prediction	0.9317
23	frontal_x_left	0.9043
24	measuring	0.9015
25	frontal	0.8833

Table 9: Top 25 Features by Absolute Coefficient Magnitude (MetaLogReg without Feature Engineering)

Rank	Feature	Coefficient
1	intracranial	3.2557
2	brain	2.3872
3	Prior_123on_Min	2.3538
4	head	2.2441
5	deid_block_reports	1.8397
6	cervical	-1.7497
7	recesses	-1.7090
8	facial	-1.6994
9	globe	-1.5468
10	cranium	1.5402
11	appreciated	-1.5289
12	nondisplaced	1.5284
13	Prior_1_Prediction	1.5277
14	extraaxial	1.5126
15	performed	1.5077
16	electronically	-1.5012
17	concur	-1.5012
18	comparison	1.4228
19	abnormalities	1.4148
20	spine	-1.3936
21	subluxation	-1.3434
22	joints	-1.2308
23	calvarium	1.2227
24	nasal	-1.2188
25	optic	-1.2034

Table 10: Hyperparameter ranges explored in grid search and final chosen configuration.

Hyperparameter	**Search Range**	**Chosen Value**
Learning rate	fixed (5e-4)	5e-4
Weight decay (L_2)	{0.005, 0.01, 0.05, 0.1}	0.05
Dropout	{0.1, 0.2, 0.4, 0.5, 0.6}	0.6
Projection dim (d_{proj})	{128, 256, 384, 512, 768}	768 (no projection)
Attention used	{True, False}	True
Attention dim (d_{attn})	{32, 64, 128, 256}	256
Concat w/ attention	fixed (true)	True
Value bottleneck	fixed (false)	False
MLP interaction	fixed (true)	True
MLP output dim	{32, 64, 128}	128

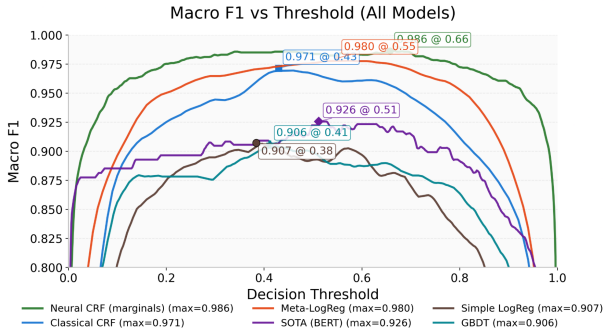

Figure 10: 2 Additional Baselines on the Academic Health Center dataset

A.6. Additional Baselines

We additionally train a logistic regression model on the bag-of-words embeddings and a gradient-boosted decision tree classifier to provide common tabular-data baselines. These models allow us to assess how much predictive signal is captured by simple, context-free classifiers and how much value the additional past prediction features add. The results can be observed in Figure 10.

A.7. Performance on Challenging Report

To motivate the varying generalizability of models and the need to be robust to out-of-distribution samples, we consider a radiology report that contains information: CT brain, face, C-spine, chest, abdomen, and pelvis, and bony pelvis.

Our neural model provided substantially worse performance than our classical models, with even the best seed still only scoring 63%.

Table 11: Accuracy on Rare 7-Scan Radiology Report

Model	**Accuracy (%)**
Classical CRF	87.06
MetaLogReg	84.71
Neural CRF (avg over 100 seeds)	35.49

A.8. Statistical Results

To obtain statistically robust comparisons between models, we compute 95% confidence intervals (CIs) for each model's AUC as well as for the pairwise differences in AUC (AUC) between model pairs. Unlike sentence-level bootstrap procedures—which violate independence because sentences within a report are correlated—we resample entire reports.

We additionally modify the synthetic bundling procedure so that each original report appears in exactly one bundled report, ensuring that all bundled samples are independent. Under this setup, only a single CI for the pairwise differences overlaps with zero; importantly, none of the comparisons between our three modeling approaches and the SOTA baseline include zero.

Across models trained on the Duke Health System dataset, the Classical CRF achieves the highest AUC (0.9565), followed closely by MetaLogReg (0.9266). The AUC confidence intervals show that both classical models significantly outperform the SOTA baseline and the Neural CRF, with intervals that do not include zero.

Model	AUC (95% CI)	ΔAUC vs Meta LogReg (95% CI)	ΔAUC vs Classical CRF (95% CI)
Classical CRF	0.9565 (0.9403–0.9704)	+0.0299 (0.0174–0.0425)	—
Meta LogReg	0.9266 (0.9113–0.9409)	—	−0.0299 (−0.0425–−0.0174)
Neural CRF	0.9211 (0.8991–0.9417)	−0.0053 (−0.0275–−0.0157)	−0.0352 (−0.0572–−0.0146)
SOTA	0.8906 (0.8755–0.9054)	−0.0359 (−0.0479–−0.0238)	−0.0659 (−0.0812–−0.0505)

Table 12: AUC and pairwise DeLong ΔAUC comparisons with 95% confidence intervals. These were models trained on the Duke Health System dataset and evaluated on bundled MIMIC-IV reports that are independent from one another.

A.9. Example Radiology Reports

We created an example radiology report to show what most radiology reports look like in the Duke Health System dataset.

Radiology Report Example:

EXAM: Noncontrast CT head, CT facial bones without contrast, noncontrast CT cervical spine

INDICATION: Patient fell.

TECHNIQUE: Axial images through the head, facial bones and cervical spine were acquired without contrast. Dose modulation performed with AEC or manually adjusted parameters based on patient size.

CONTRAST DOSE: none administered.

COMPARISON: Prior head CT and sinus CT from [[DATES]].

FINDINGS: The calvarium is intact without fracture. No subgaleal collections. Cerebral sulci and ventricles are prominent consistent with diffuse parenchymal volume loss. No midline shift. No acute hemorrhage. Patchy periventricular hypodensity consistent with chronic microvascular disease. Right orbital soft tissue swelling is present with overlying air locules suggesting open laceration. There is a depressed fracture of the anterior wall of the right maxillary sinus measuring approximately 4–5 mm in displacement. Additional fracture noted involving posterior wall with trace hemorrhage within the maxillary sinus. Inferior orbital wall shows linear fracture without significant displacement. The right globe is intact. Minimal fluid in right ethmoid air cells. Frontal sinuses are underdeveloped. No abnormality in the sphenoid sinus. Mastoid air cells are well aerated. Mild reversal of cervical lordosis centered at C3-C4. Vertebral body heights preserved. There is mild anterolisthesis at C4 on C5. No acute fracture. Severe uncovertebral and facet hypertrophy noted from C2 through C7. Greatest narrowing at C5-C6 and C6-C7 due to disc osteophyte complexes. Right neural foramen at C4-C5 appears nearly obliterated. Mild canal narrowing at C6-C7 is observed. No abnormal prevertebral soft tissue. Limited view of lung apices is unremarkable. Small reactive lymph nodes are present bilaterally. Airways grossly normal.

IMPRESSION: 1. No evidence of acute intracranial abnormality. 2. Acute displaced fracture of right anterior and posterior maxillary sinus walls with associated hemorrhage and orbital floor involvement. 3. Soft tissue injury consistent with right orbital trauma. 4. Chronic multilevel degenerative changes of the cervical spine with severe right neural foraminal narrowing at C4-C5, no acute fracture identified. Results were preliminarily discussed with [[Provider]] in ED by [[Radiologist]] ([[Vendor]]) at 12:47 AM on [[DATE]]. Electronically Signed by: [[Radiologist]] at [[Hospital]] Radiology Department Electronically Signed on: [[DATE]] 7:38 AM

A.10. Creating Synthetic Bundled Reports

To evaluate generalization on longer, mixed-anatomy inputs, we constructed 400 concatenated reports and 400 synthetic bundled reports using 300 single-scan reports from MIMIC-IV: 100 brain-related, 100 face-related, and 100 spine-related. We evaluated models

on these synthetic reports in addition to testing on the original reports, as shown in Table 2.

For concatenation, we combine a brain report with either nothing, a face report, a spine report, or both. In this setting, we preserve all content from each source report — nothing is excluded or modified.

For bundling, the goal is to create a joint report that retains the original structure by merging the relevant sections from each source report. However, we cannot do this for all sections — particularly background sections like history or indication — as doing so would introduce conflicting or redundant patient information. To avoid this, we focus on three clinically meaningful sections: EXAMINATION, FINDINGS, and IMPRESSIONS. For each bundled report, we extract these sections from the component reports and merge them within-section. For example, all FINDINGS from the individual reports are concatenated into a single FINDINGS section in the bundled report. This approach preserves all the important content from anatomical regions while avoiding artifacts in merging.

In total, we generated 400 bundled reports across the four combinations described, preserving sentence-level labels from their source reports to enable evaluation of anatomical region filtering within a realistic multi-exam structure.

An example of what a synthetically merged bundled report could look like is shown below.

Representative Example of Merged Bundled Report:

> EXAMINATION: CT HEAD WITHOUT CONTRAST CT FACIAL BONES AND SINUSES WITHOUT CONTRAST CT CERVICAL SPINE WITHOUT CONTRAST
>
> FINDINGS: The patient has a history of right-sided craniotomy with metallic surgical clips visualized along the right supraclinoid internal carotid artery. There is no acute intracranial hemorrhage, edema, or mass effect. Ventricular size and sulcal pattern are within normal limits. A small region of encephalomalacia is again seen in the right frontal lobe, stable compared to previous imaging. Basilar cisterns are preserved. No midline shift. The skull is intact without acute calvarial fracture.
>
> The paranasal sinuses demonstrate mild mucosal thickening in the ethmoid and sphenoid sinuses. There are several tiny retention cysts noted in the maxillary sinuses. The right nasal bone is fractured with mild lateral displacement and surrounding soft tissue swelling. The orbits are symmetric and intact. Extraocular muscles and intra- and extraconal fat planes are preserved.
>
> Within the cervical spine, vertebral alignment is normal. No acute fracture is identified. Degenerative disc changes are most notable at C5–C6, with narrowing of the disc space and anterior osteophyte formation. The spinal canal and neural foramina are patent throughout. A small focus of calcification is noted along the nuchal ligament. There is no prevertebral soft tissue swelling. The upper airway is unremarkable.
>
> Visualized portions of the lung apices are clear. The thyroid gland appears within normal limits.
>
> IMPRESSION:
>
> No acute intracranial process. Stable post-surgical findings following right craniotomy.
>
> Right nasal bone fracture with mild displacement and adjacent soft tissue swelling.
>
> Mild chronic sinus disease without evidence of acute sinusitis.
>
> Multilevel cervical spondylosis, most pronounced at C5–C6. No acute fracture or canal stenosis.

A.11. Small Large Language Models

We cast sentence-level anatomy filtering as a constrained next-token prediction task. For each sentence, we build a fresh prompt that includes (i) a short instruction with few-shot examples, (ii) the entire report as context, and (iii) the current sentence to classify, ending with Answer:. We then compute probabilities for the verbalizers "yes" (brain) vs "no" (not brain) and a score where a threshold yields the binary label. Each sentence is evaluated independently (no cross-sentence memory), but the model can still use full-report context to disambiguate vague sentences.

We evaluated instruction-tuned and biomedical 7B models, including Qwen 2.5-7B Instruct, Qwen 1.5-7B Chat, and BioMistral-7B (Jiang et al., 2023; Labrak et al., 2024; Team et al., 2025; Bai et al.,

2023). While models in the 7B parameter range, such as those we evaluate, may be classified as either small or large depending on context, we refer to them as small language models (SLMs) to distinguish them from frontier-scale LLMs with more than hundreds of billions of parameters. As discussed by Chen and Varoquaux (2025), the definition of "small" remains fluid in the literature.

Qwen 2.5-7B was the strongest small language model: Maximum Macro-F1 = 0.88 and AUC = 0.9716 on our test set. Qwen 1.5-7B lagged substantially: Macro-F1 $\approx$ 0.50.

BioMistral-7B is an open-source biomedical large language model built upon the Mistral 7B-Instruct architecture and further pre-trained on PubMed Central to imbue it with domain-specific knowledge. However, it only achieved AUC = 0.8096 and Macro-F1 of 0.6828 under the same setup.

Generally, when we remove report context and feed only the sentence, AUC drops by up to 0.05, highlighting the importance of context.

Below, we present an illustrative sketch of this conversational setup, followed by representative prompt variants that we considered during development.

Representative Prompt:

> You are a medical AI assistant. For each sentence from a radiology report, respond with only one word: `yes` if the content is from a brain or head CT scan, and `no` if it is from any other type of CT scan (such as facial bones, spine, neck, sinuses, etc). Respond only with that one word.
>
> Example of a brain CT report:
>
> Sentence: EXAM: Head CT
> `yes`
> Sentence: INDICATION: Auditory and hallucinations. Confusion.
> `yes`
> Sentence: There is generalized atrophy with changes of chronic white matter microvascular disease.
> `yes`
> Sentence: No intracranial hemorrhage or signs of an acute infarction.
> `yes`
> Sentence: Ventricles and CSF spaces: There is no evidence of obstructive hydrocephalus.
> `yes`
> Sentence: The paranasal sinuses are clear.
> `yes`
> Sentence: IMPRESSION: Atrophy and chronic white matter changes.
> `yes`
>
> Now, I will paste the entire radiology report, and then specify one sentence from it for you to classify.

Conversation-style application:

> You are a medical AI assistant... [SYSTEM-STR above]
>
> Full report: [All sentences concatenated into one string]
>
> Sentence: [Current sentence to classify] With the full report as context, answer 'yes' or 'no': Is this sentence brain-related? Answer:

We include several alternative prompt variants below. While some emphasize direct classification, others provide contrastive examples across scan types.

Additional Prompts:

> You are a medical AI assistant. For each sentence from a radiology report, respond with only one word:
> `brain` if the sentence refers to findings from a brain or head CT scan, and `not brain` if it refers to any other type of CT scan (e.g., facial bones, spine, neck, or sinuses). Respond only with `brain` or `not brain`.
>
> Example sentences from a brain CT scan:
> Sentence: *EXAM: Head CT*
> `brain`
> Sentence: *The ventricles are symmetric and midline.*
> `brain`
> Sentence: *There is no acute hemorrhage or mass effect.*
> `brain`
> Sentence: *Paranasal sinuses are clear.*
> `brain`
> Sentence: *IMPRESSION: Mild cerebral atrophy, no acute findings.*
> `brain`
>
> Classify each following sentence as `brain` or `not brain`.

Classify each sentence from a radiology report with only one word: `brain` if it describes findings from a brain or head CT scan, or `not brain` for any other type of scan. Only respond with `brain` or `not brain`.

You are a medical AI assistant. For each sentence, respond with only one word: `brain` if it is from a head CT scan, `not brain` otherwise.

Example from a brain CT:
Sentence: *INDICATION: Confusion and memory loss.*
`brain`
Sentence: *There is age-appropriate volume loss.*
`brain`
Sentence: *Old infarct in left frontal lobe.*
`brain`

Example from a facial CT:
Sentence: *Bony orbits are intact.*
`not brain`
Sentence: *Mild mucosal thickening in maxillary sinus.*
`not brain`
Sentence: *Mandible is unremarkable.*
`not brain`

Classify each following sentence as `brain` or `not brain`.