Identifying Imaging Follow-Up in Radiology Reports: A Comparative Analysis of Traditional ML and LLM Approaches

Namu Park[1*], Giridhar Kaushik Ramachandran[2*], Kevin Lybarger[2]
Fei Xia[3], Özlem Uzuner[2], Martin Gunn[4], Meliha Yetisgen[1]
[1]Department of Biomedical Informatics and Medical Education, School of Medicine, University of Washington, Seattle, WA, USA
[2]Department of Information Sciences and Technology, George Mason University, Fairfax, VA, USA
[3]Department of Linguistics, University of Washington, Seattle, WA, USA
[4]Department of Radiology, School of Medicine, University of Washington, Seattle, WA, USA
*The first two authors contributed equally to this work

Abstract

Large language models (LLMs) have shown considerable promise in clinical natural language processing, yet few domain-specific datasets exist to rigorously evaluate their performance on radiology tasks. In this work, we introduce an annotated corpus of 6,393 radiology reports from 586 patients, each labeled for follow-up imaging status, to support the development and benchmarking of follow-up adherence detection systems. Using this corpus, we systematically compared traditional machine-learning classifiers—logistic regression (LR), support vector machines (SVM), Longformer, and a fully fine-tuned Llama3-8B-Instruct—with recent generative LLMs. To evaluate generative LLMs, we tested GPT-4o and the open-source GPT-OSS-20B under two configurations: a baseline (Base) and a task-optimized (Advanced) setting that focused inputs on metadata, recommendation sentences, and their surrounding context. A refined prompt for GPT-OSS-20B further improved reasoning accuracy. Performance was assessed using precision, recall, and F1 scores with 95% confidence intervals estimated via non-parametric bootstrapping. Inter-annotator agreement was high (F1 = 0.846). GPT-4o (Advanced) achieved the best performance (F1 = 0.832), followed closely by GPT-OSS-20B (Advanced; F1 = 0.828). LR and SVM also performed strongly (F1 = 0.776 and 0.775), underscoring that while LLMs approach human-level agreement through prompt optimization, interpretable and resource-efficient models remain valuable baselines.

Keywords: Radiology NLP, Large Language Models, Comparative Analysis, Clinical Decision Support

1. Introduction

Follow-up recommendations are frequently included in radiology reports, varying depending on the modality and body region, with an average of 19.9% of reports containing such recommendations (Lau et al., 2020). These often, but not always, specify an imaging modality, reason, and timeframe. The referrer is typically responsible for ensuring patient communication and follow-up (Larson et al., 2014), but radiology departments usually do not track recommendations and may be unaware if follow-ups are ordered or completed. Accordingly, follow-up adherence is suboptimal, with over a third of recommendations not being followed up as reported in a recent publication (Mabotuwana et al., 2019). Although undoubtedly a sizeable proportion of follow-up imaging is not ordered because it is felt unnecessary by the patient or clinician, one study found that more than 35% of recommendations were not followed up simply due to the referring clinician not acknowledging them (Callen et al., 2012). Reasons related to failure to follow-up, especially on incidental findings, include referring clinician missing the recommendations or losing track of them while addressing a more acute illness, loss of information during handover between care teams, the recommendation not being communicated to the patient, and the patient failing to schedule or show up for the follow-up appointment (Kulon, 2016).

Missing recommended follow-ups can have serious clinical consequences, particularly when incidental findings represent early manifestations of malignancy or other progressive disease. Beyond individual patient harm, inadequate follow-up adherence also contributes to inefficiencies in healthcare delivery and increased system costs (Kadom et al., 2022). Accordingly, improving the identification and tracking of follow-up recommendations is essential for ensuring timely care and reducing preventable morbidity and mortality. However, progress in this area has been limited in part by the lack of publicly available, well-annotated corpora tailored to this clinical problem.

To address this gap, we developed a corpus of 6,393 radiology reports from 586 patients, each manually annotated for the follow-up imaging status. Using this resource, we systematically evaluated multiple machine learning and large language model (LLM) approaches for the task of follow-up identification in free-text radiology reports—a domain characterized by high linguistic variability, implicit clinical reasoning, and long-context dependen-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7498–7510
11–16 May 2026.

cies. The evaluated models included logistic regression (LR), support vector machines (SVM)(Hearst et al., 1998), Longformer (Beltagy et al., 2020), and a fully fine-tuned Llama3-8B-Instruct (Grattafiori et al., 2024), alongside generative models GPT-4o (Hurst et al., 2024) and GPT-OSS-20B. By introducing this annotated corpus and benchmarking across both traditional and generative paradigms, this study provides a valuable resource and empirical foundation for advancing real-world clinical NLP applications in radiology.

2. Related Work

Given the critical importance of identifying follow-up recommendations to prevent potentially fatal outcomes, numerous NLP-based methods have been developed for radiology reports. For instance, Yetisgen et al. introduced a supervised text classification approach to detect recommendation sentences, leveraging features beyond simple unigram tokens (Yetisgen-Yildiz et al., 2011). Conditional Random Fields (CRF) (Lafferty et al., 2001) have also been utilized to capture temporal information within follow-up recommendations (Xu et al., 2012).

Other studies have compared traditional machine learning models, such as Support Vector Machines (SVMs) (Hearst et al., 1998) and Random Forests (RFs) (Breiman, 2001), with deep learning approaches like Recurrent Neural Networks (RNNs) (Schuster and Paliwal, 1997) for identifying follow-up recommendations (Carrodeguas et al., 2019). Lau et al. created a corpus of radiology reports annotated with entities related to follow-up recommendations, trained a neural model on this dataset, and applied it at scale to assess adherence rates (Lau et al., 2020).

In addition, other research has focused on building tools that evaluate follow-up compliance by extracting essential elements such as the recommended time frame and imaging modality from radiology reports (Mabotuwana et al., 2018). Dalal et al. tackled the challenge of tracking follow-up adherence with an Extremely Randomized Trees model (Geurts et al., 2006) that integrates various clinical features, including recommendation attributes, metadata, and text similarity (Dalal et al., 2020). This model achieved an F1 score of 0.807, closely approaching the inter-annotator agreement of 0.853 F1, demonstrating its effectiveness in tracking follow-up adherence and its potential utility in real-world clinical settings.

However, the application of generative LLMs to this specific task is still understudied. Therefore, using our annotated corpus as a benchmark resource, we aim to evaluate both traditional machine learning models and generative LLMs for the task of follow-up identification, providing insights into model performance, generalizability, and practical suitability for clinical deployment.

3. Methods

This retrospective study was approved by the Institutional Review Board (IRB) at the authors' institution and satisfies the waiver of patient informed consent. All radiology reports were de-identified prior to annotation. Python code used for the experiments is available on our GitHub repository: https://github.com/uw-bionlp/ReportMatcher.

3.1. Task Description

To track completion of follow-up imaging tests, we define two types of radiology reports. Reports that contained a finding for which an imaging follow-up recommendation was present were termed "index reports." Follow-up recommendations are often associated with potential malignancy and typically (though not invariably) specify the recommended follow-up timeframe, imaging modality, and anatomy or pathology to be re-imaged (e.g., "recommend follow-up with chest CT in 6 months to assess stability of the lung nodule"). For a given index report, all subsequent imaging reports for the same patient in the radiology information system (RIS) were considered—referred to hereafter as "candidate reports." Thus, each patient is represented by a single index report and its associated candidate reports. Figure 1 provides a visual overview of the follow-up identification task.

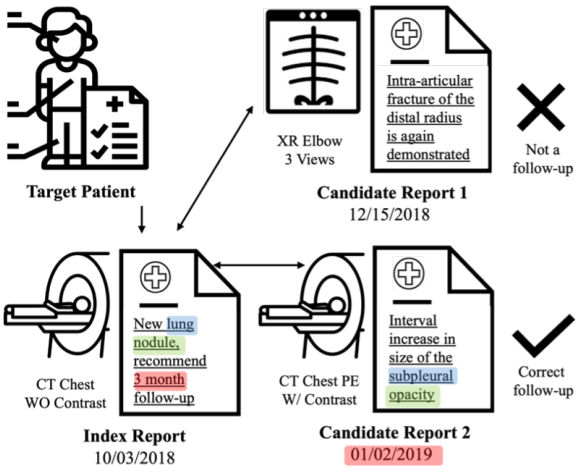

Figure 1: Follow-up identification task

3.2. Dataset

3.2.1. Sampling Process

We used a clinical database containing 7 million radiology reports from 959,382 patients, covering

the years 2007–2020. All reports were de-identified using a neural de-identifier (Lee et al., 2021). During this period, each patient underwent an average of 7.45 radiology examinations, with 29.9% having only a single report.

We focused on selecting the most relevant index reports for our task using the approaches described in Figure 2. We limited index report modality to CT and MRI, as these frequently include follow-up recommendations.

To ensure rapid and accurate sampling, we combined automated and manual methods. First, to increase the prevalence of recommendation sentences among candidate index reports, we developed a lightweight SVM-based recommendation sentence classifier using an existing annotated corpus of 500 CT radiology reports (Lau et al., 2022), split into training, development, and test sets (80% / 10% / 10%). Trained on binary sentence labels (*has recommendation* vs. *no recommendation*), the classifier evaluated each sentence and marked it as a recommendation when appropriate. The sentence-level performance was 0.87 F1, and it identified 181,020 CT and 52,809 MRI reports with at least one recommendation sentence—12.8% and 9.8% of notes per modality in our dataset, respectively.

From reports with at least one follow-up recommendation, we next identified those including a mass lesion. After randomly selecting 25,000 reports per modality, we applied a BERT-based entity and relation extractor from prior work (Park et al., 2024) to extract detailed clinical findings (mass lesions, problems, attributes). The number of randomly selected reports (n = 25,000 per modality) was determined based primarily on the size of the MRI subset. As shown in Figure 2, 52,809 MRI reports contained at least one follow-up recommendation. We aimed to sample approximately 50% of this eligible MRI population to obtain a sufficiently large and representative cohort for the subsequent information extraction stage. The final MRI sample size (25,000) corresponds to 47.3% of eligible MRI reports. To maintain a balanced design across imaging modalities during downstream processing, we matched the CT sample size to the same number (n = 25,000), despite the larger pool of eligible CT reports. Afterwards, We targeted cases with newly discovered lesion findings in six anatomies (lung, kidney, liver, adrenal gland, pancreas, and thyroid) yielding 2,095 CT reports and 464 MRI reports, each with at least one new lesion in a target anatomy and at least one recommendation sentence.

From this refined pool of 2,559 eligible reports (2,095 CT and 464 MRI), 720 index reports were randomly selected. Based on initial pilot annotation rounds, we determined that sampling 720 reports would provide an appropriate balance between dataset size, annotation feasibility, and statistical robustness, given available annotation resources and time constraints. Two medical student annotators then excluded potentially incorrect samples from automated filtering, removing reports with excessively vague recommendations (e.g., "*attention on follow-up is recommended*," "*clinical correlation is recommended*") and those recommending or leading to same-day characterization of acute findings. The combined filtering resulted in 586 index reports, with 134 inappropriate samples removed. For each of the 586 index reports, corresponding candidate reports were sampled without restriction on imaging modality, resulting in 5,807 candidate reports (mean 9.9 per index; range [1, 168]).

3.2.2. Data Annotation

All reports were annotated by two medical students, who identified the earliest qualifying follow-up report among the candidate reports, if it existed. A candidate report was considered a follow-up if it addressed the same anatomical region as the lesion in the index report and explicitly referenced or negated the prior finding (e.g., *"multiple pulmonary nodules unchanged from the previous examination," "redemonstration of hypodense mass measuring 2.6 $\times$ 2.4 cm, previously 3.2 $\times$ 3.0 cm"*). In multi-lesion cases, identifying the most suitable follow-up exam was challenging; such complex cases were flagged and reviewed with a board-certified radiologist. The annotation process comprised 16 rounds. In the first 11 rounds, samples were doubly annotated, disagreements were resolved in weekly meetings, and any controversial cases were adjudicated with a board-certified radiologist. After 11 rounds, inter-annotator agreement was 0.846 F1. The subsequent 5 rounds were single annotated to increase volume, resulting in 347 single-annotated report pairs and 239 double-annotated report pairs.

Among the 586 index–candidate pairs, a qualifying follow-up report was identified in 417 pairs (71.3%), while 169 index reports (28.7%) had no follow-up identified. For 54% of index reports, the first or second candidate chronologically was labeled as the follow-up. While index reports included only CT and MRI, candidate reports spanned modalities: computed/digital radiography (n=2,139, 36.8%), CT (n=2,003, 34.5%), MRI (n=520, 8.9%), ultrasound (n=516, 8.8%), nuclear medicine (n=171, 2.9%), angiography (n=127, 2.2%), PET-CT (n=99, 1.7%), mammography (n=65, 1.1%), etc. Index reports averaged 437.5 tokens (30.4 sentences) versus 252 tokens (17.3 sentences) for candidate reports. We used the BERT tokenizer (Devlin et al., 2019), which converts text into subword tokens.

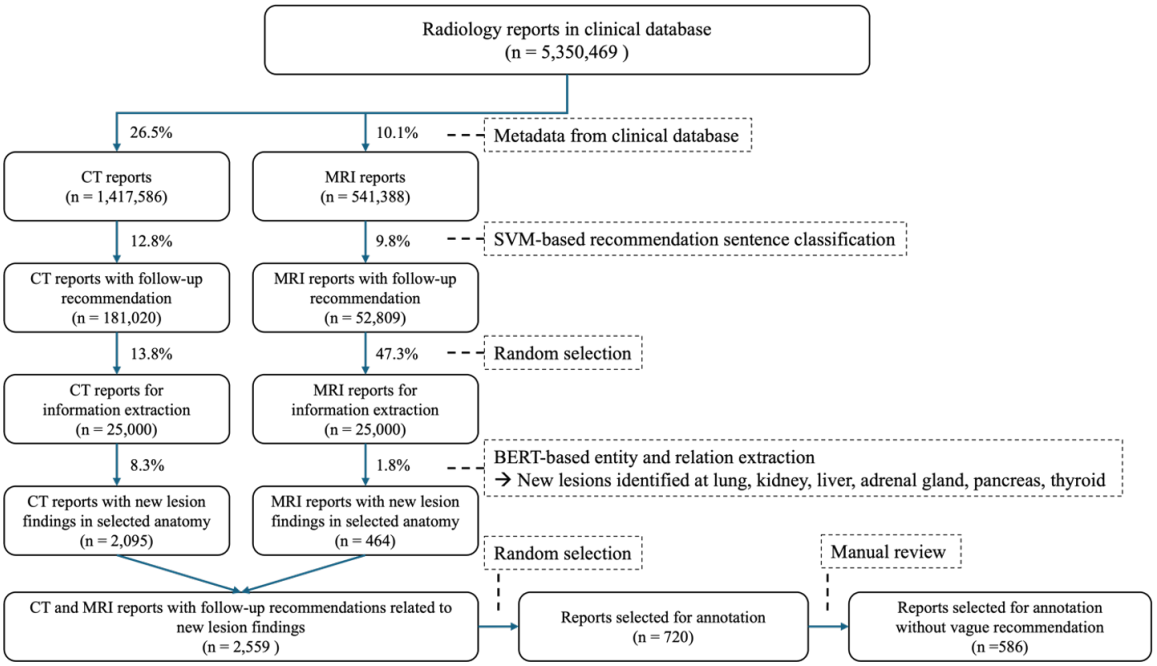

Figure 2: Sampling process for index reports

3.3. Models

Prior work (Mabotuwana et al., 2018) in follow-up imaging identification applied discrete machine learning models combining text and engineered linguistic features to predict the likelihood of a correct candidate report. We framed follow-up identification as a binary classification task by constructing index–candidate pairs and labeling each pair positive if the candidate report was the correct follow-up for the index report.

3.3.1. Feature-based Traditional Models

We evaluated Logistic Regression (LR) and Support Vector Machines (SVM) in a supervised learning setting, with inputs consisting of index and candidate report text plus metadata (imaging modality and report time gap in days). Separate vector representations for index and candidate reports were created using TF-IDF (Sparck Jones, 1972). Metadata were encoded using binary indicator vectors. A binary vector representing words shared by index and candidate reports was concatenated to form the final representation vector. The SVM used a sigmoid kernel and L2 regularization; both SVM and LR used class-balanced loss functions. Figure 3 illustrates the SVM and LR pipelines. Individual vectors with separate encodings for metadata and text represented each of the index (V_{index}) and candidate ($V_{\text{candidate}}$) reports; the final pair vector (V_{pair}) concatenated index, candidate, and shared-word vectors.

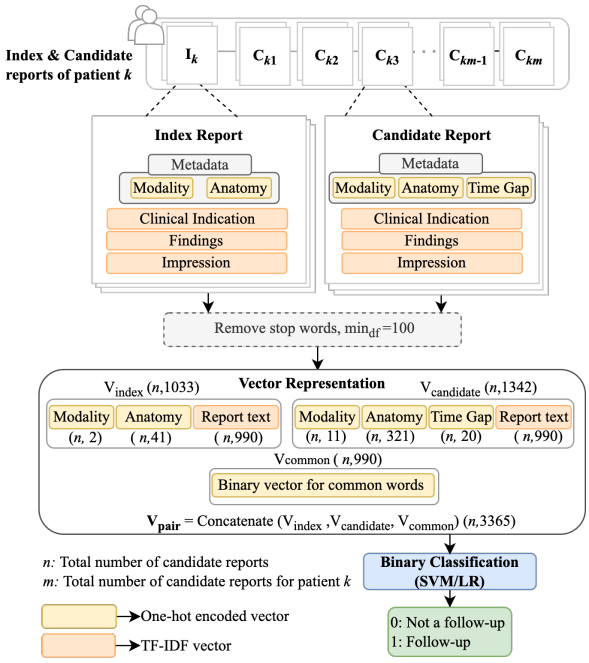

Figure 3: Follow-up identification using feature-based traditional models – Support Vector Machine and Logistic Regression

3.3.2. Supervised Learning using Transformer-based Models

We also investigated supervised learning with Transformer-based models capable of extended inputs, including Longformer (Beltagy et al., 2020), BioClinicalModernBERT (Sounack et al.,

2025), Llama3-8B-Instruct and Llama3.1-8B-Instruct (Grattafiori et al., 2024), with model selection guided by compute constraints. We report results for Longformer and Llama3-8B-Instruct, which demonstrated the highest performance. Each input concatenated index and candidate report text separated by special tokens. Longformer used a linear classification layer. Both Longformer and Llama3-8B-Instruct were trained on the complete index and candidate text plus metadata. For Llama3-8B-Instruct, we tested several prompt designs; the best included a brief instructional prefix preceding the report pair (Figure 4). The model was then instruction-tuned via full supervised fine-tuning (SFT) using this engineered prompt. Additional implementation details are available in our GitHub repository.

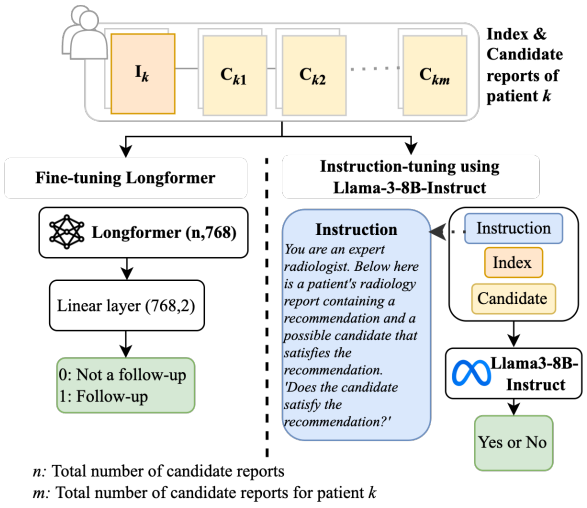

Figure 4: Transformer-based follow-up identification models – Longformer and Llama3-8B-Instruct

3.4. Generative Large Language Model

To evaluate generative LLMs for identifying follow-up imaging studies, we assessed GPT-4o (Hurst et al., 2024) (version 2024-05-13) and GPT-OSS-20B (Agarwal et al., 2025) in our HIPAA-compliant environment using two strategies: a baseline setting and an advanced, task-optimized setting (Figure 5). GPT-4o provides strong instruction-following and clinical reasoning, and GPT-OSS-20B is a recent open-source model suitable for secure institutional deployment. Larger variants such as GPT-OSS-120B were not used, as they offered only marginal gains in our task at substantially higher computational cost.

In the baseline setting, the model received the full index and candidate reports including metadata, accompanied by minimal task-specific instruction, and was asked to determine whether the candidate report represented an appropriate follow-up. In contrast, the advanced setting restricted input to metadata and the recommendation sentence—identified using the SVM-based sentence classifier in Section 3.2.1—along with its immediate preceding sentence (green box in Figure 5), emphasizing clinically relevant content for follow-up determination.

To optimize prompt engineering, we randomly selected 60 index reports and their corresponding candidate reports for iterative refinement during development. All prompt optimization was conducted using GPT-4o, and the final prompt that achieved the best development performance was adopted as the standard prompt for both the baseline and advanced settings. This optimized prompt was then directly applied to GPT-OSS-20B to assess cross-model transferability. The 60 development samples used for prompt optimization were excluded from final evaluation, which followed the performance criteria described in the next section.

Following initial error inspection on the development set, we observed that GPT-OSS-20B tended to judge candidate reports as incorrect when not all findings in the recommendation were re-addressed. To account for this behavior, we introduced an additional instruction on top of the GPT-4o prompt clarifying that a follow-up remains valid even if only a subset of the recommended findings is addressed (Figure 5).

3.5. Evaluation

We evaluated performance at the index report-level chronologically (Figure 6) using the following categories: 1) True Positive (TP) - correctly predicted the follow-up candidate report, if it existed. If multiple positive predictions existed, only the first one was considered for comparison because this is the clinically most important examination to ensure that follow-up did occur. 2) False Positive (FP) - incorrectly identified a follow-up when it was not a follow-up; 3) False Negative (FN) - did not predict the correct follow-up which existed; and 4) True Negative (TN) - correctly predicted the absence of a follow-up when there was no matching follow-up examination.

The evaluation set consisted of 526 index reports and their corresponding candidate reports, excluding 60 index reports and their associated candidate reports that were used for prompt tuning of GPT-4o. We evaluated our models in two settings: (1) five-fold cross-validation (CV) for supervised learning models and (2) in-context learning for generative LLMs. For CV, we set train-validation-test splits and tune the hyperparameters at every fold. We reported precision, sensitivity, F1, and specificity for all our approaches, along with their 95% Confidence Intervals (CI). We utilized non-parametric bootstrap test to compare the models' F1 scores with a significance threshold of 0.05.

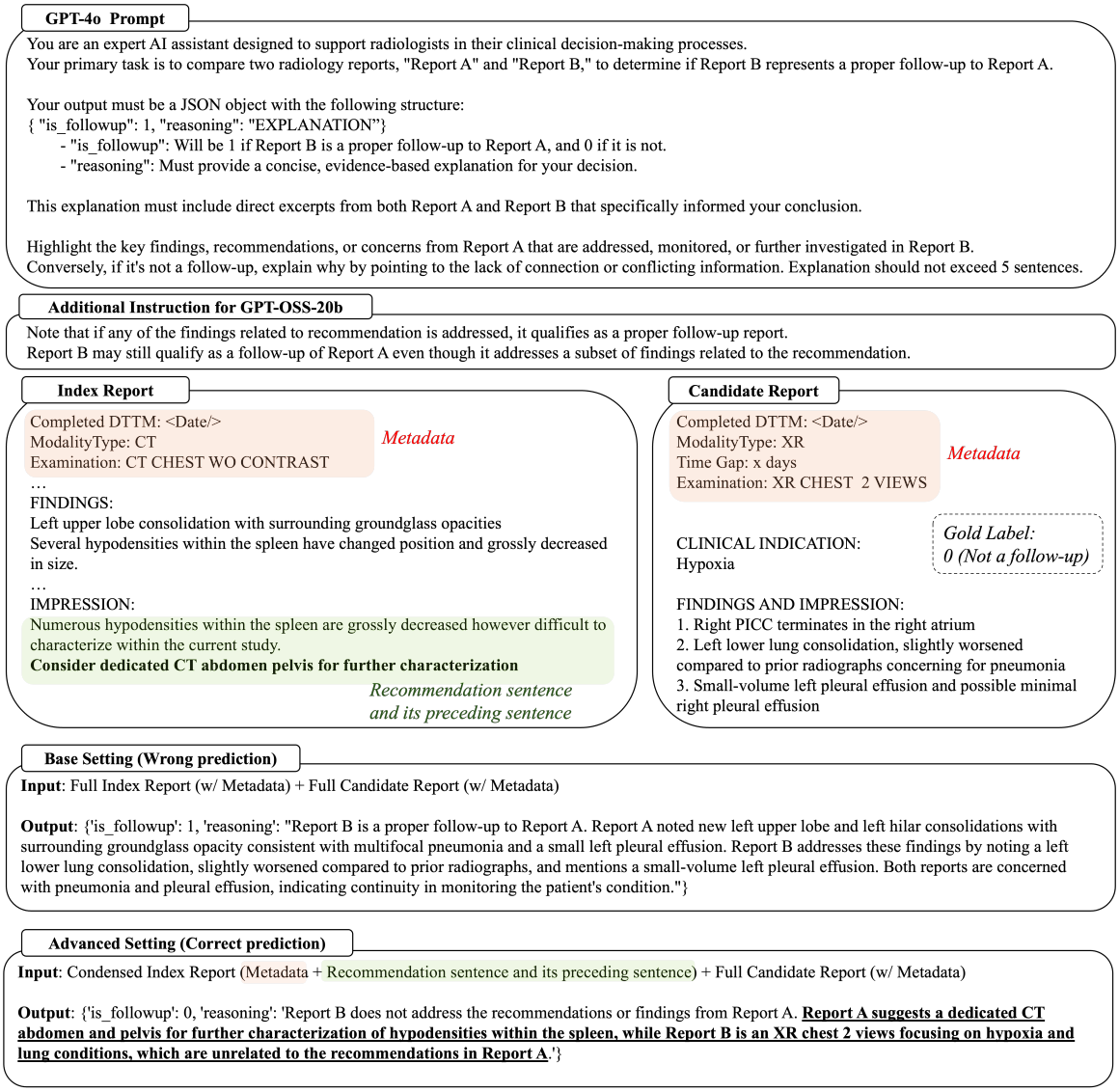

Figure 5: Follow-up identification using generative LLMs. GPT-OSS-20B prompt is created by adding 2-sentence instructions on top of GPT-4o prompt. Outputs for base and advanced settings are actual predictions from GPT-4o.

4. Results

Table 1 summarizes the aggregated performance of all evaluated models. Among them, GPT-4o (Advanced) achieved the highest overall performance, with an F1 score of 0.832 (95% CI: 0.802–0.860)—closely approximating the inter-annotator agreement score of 0.846. The next best performer was GPT-OSS-20B (Advanced), which achieved an F1 of 0.828 (0.799–0.856), demonstrating nearly equivalent performance to GPT-4o despite being a smaller, fully open-source model. Both advanced configurations markedly outperformed their respective base counterparts, underscoring the benefit of the optimized input design that emphasized follow-up recommendation cues within the radiology reports. Specifically, GPT-4o (Advanced) achieved a precision gain of Δ+0.101, a recall increase of Δ+0.018, and a corresponding F1 improvement of Δ+0.060 relative to its base setting. Similarly, GPT-OSS-20B (Advanced) improved precision by Δ+0.090, recall by Δ+0.058, and F1 by Δ+0.075 compared to its base configuration. Statistical significance testing (Table 2) confirmed that these gains were significant ($p < 0.05$) for each model relative to their baselines. However, no statistically significant difference was observed between GPT-4o (Advanced) and GPT-OSS-20B (Advanced), suggesting that GPT-OSS-20B can serve as a viable and complementary alternative to GPT-4o—particularly in resource-constrained or privacy-sensitive environments where closed-source APIs are less feasible.

The fully fine-tuned Llama3-8B-Instruct model

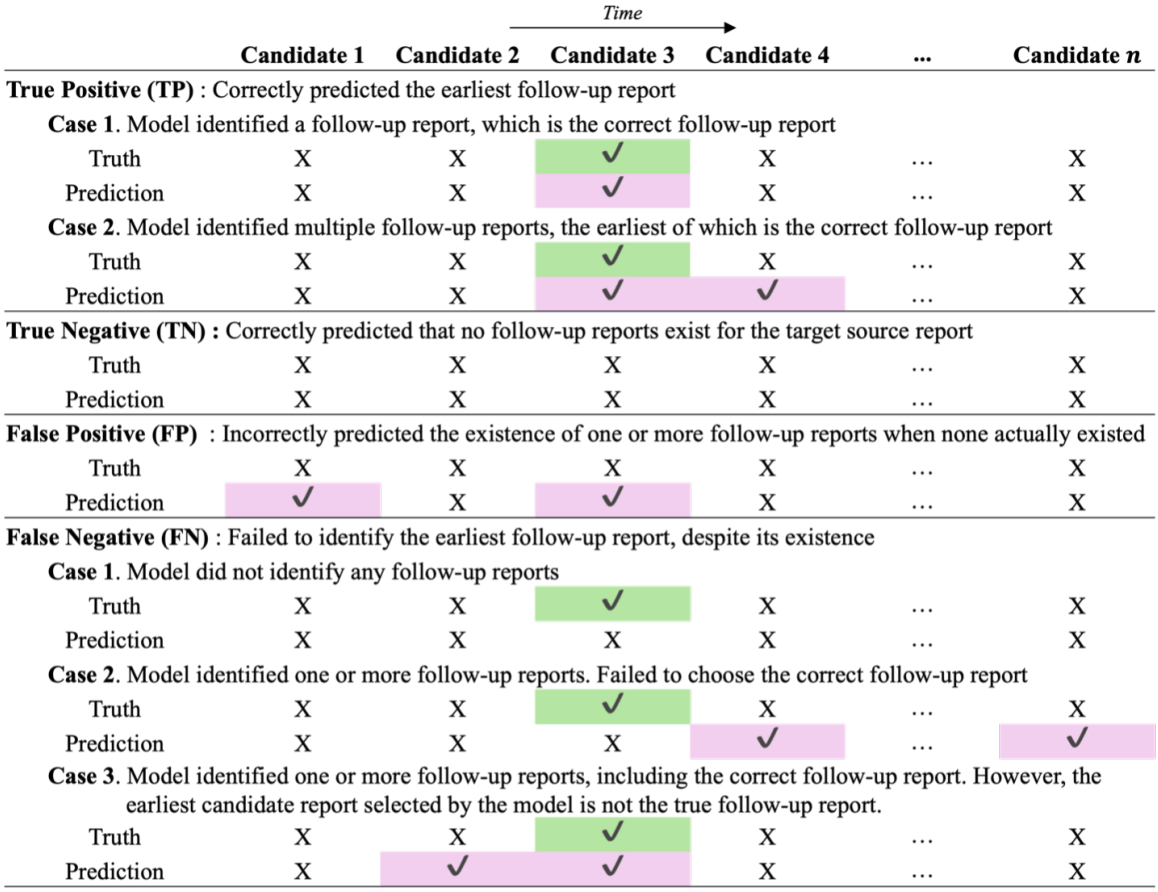

Figure 6: Evaluation method for follow-up identification task. The check mark indicates that the target candidate report has been identified as a follow-up report either by the model (Prediction) or by the annotators (Truth). Boxes in purple show the prediction by the model and boxes in green are the true labels

Table 1: Model performance metrics with 95% confidence intervals

Model	Precision	Recall	F1
SVM	0.769 (0.727, 0.812)	0.785 (0.743, 0.827)	0.777 (0.743, 0.809)
LR	0.792 (0.750, 0.833)	0.759 (0.715, 0.802)	0.775 (0.740, 0.808)
Longformer	0.741 (0.694, 0.788)	0.639 (0.590, 0.687)	0.686 (0.646, 0.724)
Llama3-8B-Instruct	**0.907 (0.872, 0.943)**	0.623 (0.573, 0.670)	0.738 (0.698, 0.775)
GPT-4o (Base)	0.740 (0.698, 0.783)	0.807 (0.767, 0.846)	0.772 (0.738, 0.803)
GPT-4o (Advanced)	0.841 (0.803, 0.878)	0.825 (0.786, 0.863)	**0.832 (0.802, 0.860)**
GPT-OSS-20B (Base)	0.734 (0.690,0.777)	0.775 (0.733,0.818)	0.753 (0.720,0.787)
GPT-OSS-20B (Advanced)	0.824 (0.786,0.862)	**0.833 (0.795,0.870)**	0.828 (0.799,0.856)

achieved the highest precision among all systems (0.907, 95% CI: 0.872–0.943), yet its recall remained substantially lower (0.623, 0.573–0.670), resulting in a reduced F1 score compared to most other approaches. This imbalance may reflect overfitting during fine-tuning, where the model captured training-specific patterns at the expense of generalization to unseen cases. Alternatively, the limited recall could stem from the pronounced class imbalance in the dataset—only 417 candidate reports (7.2% of total pairs) represented true follow-ups. Such skewed distributions can bias models toward negative predictions, diminishing sensitivity to the minority class. Future work could explore class-balanced sampling or focal loss optimization to mitigate this effect and improve model robustness on rare follow-up cases.

5. Discussion

During prompt engineering using GPT-4o, we observed cases where the model exhibited a tendency to classify all candidate reports as non–follow-ups. Prompts that included overly detailed instructions

Table 2: Pairwise statistical significance comparisons for our models using the non-parametric bootstrap test (n=253 patients drawn with replacement, 10,000 repetitions). P-values are indicated in parentheses.

	GPT-4o (Advanced)	GPT-OSS-20B (Advanced)	GPT-4o (Base)	GPT-OSS-20B (Base)	SVM	LR	Llama3-8B	Longformer
F1 (95% CI)	0.832 (0.802, 0.860)	0.828 (0.799, 0.856)	0.772 (0.738, 0.803)	0.753 (0.720, 0.787)	0.777 (0.743, 0.809)	0.775 (0.740, 0.808)	0.738 (0.698, 0.775)	0.686 (0.646, 0.724)
GPT-4o (Advanced)	N/A	No (0.3141)	**Yes (0.0001)**	**Yes (0.0001)**	**Yes (0.0001)**	**Yes (0.0003)**	**Yes (0.0001)**	**Yes (0.0001)**
GPT-OSS-20B (Advanced)		N/A	**Yes (0.0002)**	**Yes (0.0001)**	**Yes (0.0005)**	**Yes (0.0001)**	**Yes (0.0001)**	**Yes (0.0001)**
GPT-4o (Base)			N/A	**Yes (0.0454)**	No (0.3839)	No (0.4432)	**Yes (0.0462)**	**Yes (0.0001)**
GPT-OSS-20B (Base)				N/A	No (0.0833)	No (0.1024)	No (0.2433)	**Yes (0.0002)**
SVM					N/A	No (0.4567)	**Yes (0.0376)**	**Yes (0.0007)**
LR						N/A	**Yes (0.0371)**	**Yes (0.0005)**
Llama3-8B							N/A	**Yes (0.0107)**
Longformer								N/A

often led to incorrect predictions, highlighting LLMs' sensitivity to prompt complexity. Even with optimized prompting, GPT-4o (Base) achieved an F1 score of 0.772 (95% CI: 0.738–0.803), which was not statistically different from SVM (F1 = 0.777, 95% CI: 0.743–0.809) or LR (F1 = 0.775, 95% CI: 0.740–0.808). This finding suggests that effective prompt engineering is essential for GPT-4o, while more sophisticated, task-tailored approaches may still be required for it to consistently outperform simpler, more efficient traditional ML models.

The examples shown in Figure 5 demonstrate the benefit of using task-specific, condensed input for GPT-4o (Advanced). The index report documents multiple findings, including *"left upper lobe consolidation"* and *"hypodensities within the spleen,"* but the follow-up recommendation specifically concerns the splenic hypodensities and calls for a CT abdomen and pelvis. The candidate report, however, is not a CT of the recommended anatomy and does not address the splenic findings, making it unqualified as an appropriate follow-up. In the Base setting, which provided the full index report as input, the model incorrectly predicted this case as a valid follow-up by focusing only on lexical overlap. In contrast, the Advanced setting—by restricting input to the recommendation sentence and metadata—correctly identified the mismatch in anatomy and modality. We anticipate that this targeted input design reduced false positives, leading to improved precision (0.740 vs. 0.841).

Extending this analysis, the open-source GPT-OSS-20B model provided additional insight into model-specific prompt sensitivity. When using the same prompt as GPT-4o, GPT-OSS-20B exhibited lower performance (F1 = 0.799, 95% CI: 0.767–0.831), primarily due to over-strict reasoning that rejected valid follow-ups when not all findings from the recommendation were addressed. After refining the prompt to clarify that *"a follow-up remains valid even if only a subset of the recommended findings is discussed"*, GPT-OSS-20B (Advanced) improved substantially to 0.828 F1 (95% CI: 0.799–0.856), achieving performance statistically indistinguishable from GPT-4o (Advanced) (p = 0.3141). This demonstrates that small, conceptually meaningful adjustments to task framing can harmonize LLM reasoning with clinical logic, particularly for open-source models whose instruction-following behavior may differ from proprietary counterparts.

It is noteworthy that feature-based models such as LR and SVM performed comparable or better than GPT-4o (Base) and GPT-OSS-20B (Base), underscoring their value as interpretable and resource-efficient alternatives for deployment in clinical applications. These models are especially advantageous in settings where computational resources for model development, inference, and validation are limited. In our study, analysis of the LR feature weights revealed that the most influential predictors combined metadata (e.g., time gap, anatomy) with terms describing findings and their characteristics. Words such as "*nodule,*" "*lesion,*" *"unremarkable,"* and *"benign"* were among the strongest positive contributors—closely aligning with the reasoning processes employed by radiologists.

6. Conclusion

In this study, we present a newly annotated corpus of 6,393 radiology reports from 586 patients, designed to support the development and benchmarking of models for identifying imaging follow-ups. Using this resource, we comprehensively evaluated a spectrum of methods ranging from traditional feature-based classifiers and transformer-based encoders to recent generative LLMs. Among all evaluated systems, GPT-4o (Advanced) achieved the highest performance (F1 = 0.832), closely matching inter-annotator agreement (F1 = 0.846). The open-

source GPT-OSS-20B (Advanced), when guided by a refined prompt incorporating task-specific clarification, achieved comparable results (F1 = 0.828) without a statistically significant difference.

These results underscore the importance of task-aware prompt design and input curation in optimizing LLM reasoning for clinical applications. While closed-source models such as GPT-4o demonstrate strong out-of-the-box performance, open-source systems like GPT-OSS-20B achieve comparable accuracy with greater flexibility for fine-tuning and integration within secure institutional environments. Despite the advances of LLMs, traditional models such as logistic regression (LR) and support vector machines (SVM) remain valuable, offering interpretable and computationally efficient baselines. Collectively, this corpus and evaluation provide a reproducible foundation for future research, highlighting a complementary landscape where open and closed LLMs, alongside interpretable classical models, together advance the robustness and transparency of clinical NLP systems.

7. Limitations

Accurate identification of follow-up imaging has important implications for clinical practice, including reducing unnecessary scans and improving the management of incidental findings. To enhance model generalizability, future work should leverage multi-institutional datasets encompassing a broader range of imaging modalities and clinical scenarios. Multi-modal approaches that integrate imaging data with radiology text, as well as the use of domain-adapted vision-language models (e.g., MedGemma (Sellergren et al., 2025), MedVLM-R1 (Pan et al., 2025)), represent promising avenues for advancing performance and clinical relevance.

An additional limitation relates to class imbalance in the constructed index–candidate pairs. Only 417 candidate reports (approximately 7% of total pairs) represented true follow-up examinations, resulting in a highly skewed label distribution. Although this prevalence reflects realistic clinical workflows, such imbalance may bias models toward negative predictions and reduce sensitivity to rare but clinically important follow-up cases. Addressing imbalance-aware training strategies or data-level augmentation approaches remains an important direction for future work.

Furthermore, evaluating LLM outputs should extend beyond standard performance metrics such as precision, recall, and F1. Our current study was limited to these measures, but future evaluations should incorporate radiologist review, clinical reasoning assessments, and systematic evaluations of reasoning quality, factual consistency, and clinical validity. Establishing such robust evaluation frameworks will be essential to ensure the safe and effective deployment of LLMs in real-world healthcare settings.

Finally, although this work establishes technical feasibility, we did not assess real-world clinical impact, workflow integration, or production deployment considerations, which warrant prospective evaluation in operational clinical settings.

8. Ethics Statement

We obtained approval from our institution's Institutional Review Board (IRB) with a waiver of informed consent for the use of clinical text data. Radiology reports may contain patient Protected Health Information (PHI), including names, contact information, and other identifiers. To ensure patient privacy, all reports were automatically de-identified using a neural de-identification model, followed by manual review and secondary de-identification by trained medical student annotators to verify that no residual PHI remained. Both the original and de-identified reports were securely stored on a Health Insurance Portability and Accountability Act (HIPAA)-compliant server. All researchers and annotators completed human subjects training and were authorized to handle data containing PHI.

The annotated corpus used in this study was randomly sampled from the general population of patients undergoing imaging examinations at a single academic medical center. The dataset includes a diverse range of imaging modalities and clinical contexts, and was developed to support the task of identifying imaging follow-ups. Demographic variables were not used during sampling, and the patient population may not be representative of other institutions or the broader population. Differences in report style, structure, and documentation practices across sites may therefore affect the generalizability of the developed models.

9. Acknowledgments

This work was supported in part by the National Institutes of Health and the National Cancer Institute (NCI) (GrantNr. 1R01CA248422-01A1). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

10. Bibliographical References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

Joanne L. Callen et al. 2012. Failure to follow-up test results for ambulatory patients: a systematic review. *J Gen Intern Med*, 27(10):1334–1348.

Emmanuel Carrodeguas, Ronilda Lacson, Whitney Swanson, and Ramin Khorasani. 2019. Use of machine learning to identify follow-up recommendations in radiology reports. *Journal of the American College of Radiology*, 16(3):336–343.

Sandeep Dalal, Vadiraj Hombal, Wei-Hung Weng, Gabe Mankovich, Thusitha Mabotuwana, Christopher S. Hall, Joseph Fuller, Bruce E. Lehnert, and Martin L. Gunn. 2020. Determining Follow-Up Imaging Study Using Radiology Reports. *J Digit Imaging*, 33(1):121–130.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine learning*, 63:3–42.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Nadja Kadom, Arjun K Venkatesh, Samantha A Shugarman, Judy H Burleson, Christopher L Moore, and David Seidenwurm. 2022. Novel quality measure set: closing the completion loop on radiology follow-up recommendations for noncritical actionable incidental findings. *Journal of the American College of Radiology*, 19(7):881–890.

Michal E. Kulon. 2016. Lost to Follow-Up : Automated Detection of Patients Who Missed Follow-Ups Which Were Recommended on Radiology Reports.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 1(2):3.

Paul A. Larson et al. 2014. Actionable findings and the role of IT support: report of the ACR Actionable Reporting Work Group. *J Am Coll Radiol*, 11(6):552–558.

Wilson Lau, Kevin Lybarger, Martin L Gunn, and Meliha Yetisgen. 2022. Event-based clinical finding extraction from radiology reports with pre-trained language model. *Journal of Digital Imaging*, pages 1–14.

Wilson Lau, Thomas H Payne, Ozlem Uzuner, and Meliha Yetisgen. 2020. Extraction and analysis of clinically important follow-up recommendations in a large radiology dataset. *AMIA Summits on Translational Science Proceedings*, 2020:335.

Kahyun Lee, Nicholas J Dobbins, Bridget McInnes, Meliha Yetisgen, and Özlem Uzuner. 2021. Transferability of neural network clinical deidentification systems. *Journal of the American Medical Informatics Association*, 28(12):2661–2669.

Thusitha Mabotuwana, Christopher S Hall, Vadiraj Hombal, et al. 2019. Automated tracking of follow-up imaging recommendations. *American Journal of Roentgenology*, 212(6):1287–1294.

Thusitha Mabotuwana, Christopher S Hall, Joel Tieder, and Martin L. Gunn. 2018. Improving Quality of Follow-Up Imaging Recommendations in Radiology. *AMIA Annu Symp Proc*, 2017:1196–1204.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. 2025. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634*.

Namu Park, Kevin Lybarger, Giridhar Kaushik Ramachandran, Spencer Lewis, Aashka Damani, Ozlem Uzuner, Martin Gunn, and Meliha Yetisgen. 2024. A novel corpus of annotated medical imaging reports and information extraction results using bert-based language models. *arXiv preprint arXiv:2403.18975*.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Thomas Sounack, Joshua Davis, Brigitte Durieux, Antoine Chaffin, Tom J Pollard, Eric Lehman, Alistair EW Johnson, Matthew McDermott, Tristan Naumann, and Charlotta Lindvall. 2025. Bioclinical modernbert: A state-of-the-art long-context encoder for biomedical and clinical nlp. *arXiv preprint arXiv:2506.10896*.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Yan Xu, Junichi Tsujii, and Eric I-Chao Chang. 2012. Named entity recognition of follow-up and time information in 20 000 radiology reports. *Journal of the American Medical Informatics Association*, 19(5):792–799.

Meliha Yetisgen-Yildiz, Martin L Gunn, Fei Xia, and Thomas H Payne. 2011. Automatic identification of critical follow-up recommendation sentences in radiology reports. In *AMIA Annual Symposium Proceedings*, volume 2011, page 1593. American Medical Informatics Association.

Appendix A: Annotation Guidelines for Follow-up Report Identification

A.1. Definitions

- **Index Report:** Radiology report with the follow-up recommendation.
- **Candidate Report:** Radiology reports following the index/source report. All modalities are included.

Each index/source report will be presented through OneDrive and contains metadata to provide details about the radiology report including modality, description of exam, and date/time.

Candidate reports will be presented through brat and include the same metadata as well as the number of dates in between the source report and the candidate report.

A.2. Annotation Task

For each index/source report, the annotators will review the candidate notes in chronological order and label the **first** report that qualifies as a follow-up.

For a candidate exam to meet the criteria for follow-up, it must include the same body region (i.e. anatomy) as the lesion on the index report.

A.2.1. Surveillance and Incidentalomas

Most of the reports in the cohort do not contain incidentalomas, even though follow-up may occur on some patients. For example, a patient with known cancer such as lung cancer may be undergoing a series of CT scans to monitor the progress of their disease whether they are on therapy or not. This is called **surveillance imaging**. The time intervals between such imaging are usually in the order of months (occasionally as short as 1 month). Radiologists don't typically recommend follow-up in these cases because the clinicians follow standard guidelines for the timeframe and modality for surveillance imaging, or they link the timeframes to the therapies they are using.

Even for incidentalomas, the index exam might not be the first discovery of the incidental lesion. By chance, it might be part of the follow-up (surveillance) chain of imaging for the incidentaloma. For example, for small lung nodules identified at time 0 months, radiology used to recommend follow-up at 3, 6, 12, 24 months. If the index report you are presented with is 3 months after it became an incidentaloma, then a candidate report at 6 (or 12) months would meet the criteria for follow-up, even though the index report appears to be a follow-up report anyway. Nevertheless, the **first reasonable candidate exam after the index should be marked as follow-up**.

A.2.2. Opportunistic Follow-up

Follow-up imaging may be deliberate, or 'opportunistic' (i.e. a patient gets imaging for another reason, but in doing so, the imaging meets criteria for follow-up). We should mark a candidate report as a follow-up report even if the exam appears to be opportunistic. For example, if the index report contains an incidental lung nodule, and a CT of the chest is done 3 months later because the patient had a car crash, then the candidate follow-up CT would still qualify.

Beware of opportunistic follow-up that appears to be very soon after the index report. Occasionally an opportunistic exam might actually mean that further follow-up is unnecessary (e.g. the radiologist characterizes the lesion as benign).

A.2.3. Characterization

Sometimes an incidental finding identified on one modality will undergo **characterization** with a different modality (or variation of the same modality) soon after the initial finding to determine whether it is likely benign or malignant. If you see a scan performed to characterize a lesion that was identified on the index scan, it meets criteria for follow-up. Examples of characterization include:

- A thyroid nodule identified on a CT may undergo ultrasound soon afterwards (weeks)
- An adrenal lesion identified on a CT may undergo either MRI or CT to determine whether it is an adrenal adenoma or not
- A nodule seen in the liver on a CT may undergo MRI or multi-phase CT to determine whether it is malignant or not

- A nodule seen on a CXR may undergo a CT for characterization (e.g. Is it real? Is it likely to be a cancer?)
- A lung nodule found on CT may undergo PET-CT for characterization
- A pancreatic or renal lesion seen with ultrasound or CT may undergo CT (or MR, sometime ultrasound) to figure out whether it is likely to be a cancer or not.

A.3. Timing and Intervals

For the most part, a follow-up exam of the same modality will be at least **six weeks** after the index exam. There are a few exceptions:

- When the incidental finding is found on a less specific imaging modality (e.g. CXR), and a follow-up CT is done rapidly to determine whether the finding is real or not (and to figure out whether it is likely cancer or not).
- If the radiologist thinks believes, in the index report, that the finding is unlikely to be cancer, and is likely to go away quite quickly (e.g. lung nodules that look like infection)

There are standard follow-up intervals for certain types of incidentalomas:

- Initial incidental lung nodule follow-up imaging will usually occur months after the index exam, e.g. 3, 6, 12 months. Non-large lung nodules are usually followed up by CT
- Incidental thyroid nodule imaging usually occurs months or years (1–24 months) after the initial exam, and thyroid nodules are often followed up by ultrasound.

Often CT, MRI or PET-CT will show multiple lesions in multiple organs. Deciding which is the best candidate exam in this case will be a challenge. Do your best and ask for the radiologist's advice.