Leveraging Large Language Models for Clinical Information Extraction in Radiology Reports

Namu Park

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:
Meliha Yetişgen, Chair
Trevor A. Cohen
Kevin Lybarger

Program Authorized to Offer Degree:
Biomedical Informatics and Medical Education

University of Washington

Abstract

Leveraging Large Language Models for Clinical Information Extraction in Radiology Reports

Namu Park

Chair of the Supervisory Committee:
Meliha Yetişgen
Department of Biomedical and Health Informatics

Medical imaging plays a central role in diagnosing, monitoring, and managing a wide spectrum of diseases, including cancer, cardiovascular disorders, neurological conditions, and musculoskeletal abnormalities. Radiologists interpret complex imaging data and summarize their findings in narrative reports, which remain largely unstructured. The rapid expansion of imaging utilization has led to an overwhelming volume of such reports, posing significant challenges for clinical decision support. Their unstructured format limits automated analysis, secondary use, and integration into downstream clinical workflows. This dissertation addresses two major barriers to the effective use of radiology reports in data-driven clinical systems: (1) the absence of publicly available, large-scale annotated corpora of radiology reports with detailed clinical findings suitable for training supervised models, and (2) the limited application of machine learning approaches, particularly large language models (LLMs), to real-world clinical tasks at scale.

To overcome these challenges, the research is organized around three core aims: (1) developing a corpus of radiology reports annotated with detailed clinical findings and designing an advanced information extraction framework optimized for radiologic text; (2) evaluating the performance of diverse machine learning approaches, with emphasis on LLMs, for the practical task of identifying follow-up imaging recommendations; and (3) constructing a large-scale repository of incidental findings (incidentalomas) derived from the model outputs

and proposing an NLP-based framework for automated incidentaloma detection to enhance clinical decision-making.

Collectively, this work contributes a high-quality annotated dataset for radiologic text analysis and demonstrates the feasibility and utility of LLM-based approaches for transforming unstructured radiology reports into structured clinical intelligence, advancing the integration of medical imaging data into precision healthcare.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

ACTIVE LEARNING: A semi-supervised learning approach in which a model selectively queries the most informative or uncertain samples for annotation, improving dataset efficiency and diversity.

ANATOMY-AWARE PROMPTING: A prompting strategy for large language models that explicitly encodes anatomical context (e.g., *lung*, *liver*, *adrenal*) to improve lesion-specific reasoning and classification accuracy.

ANNOTATION SCHEMA: A structured framework defining how entities, relations, and events are labeled in text, ensuring consistency across annotators and supporting standardized downstream analysis.

BERT (BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS): An encoder-only transformer language model that learns contextualized word embeddings by jointly conditioning on both left and right context. It serves as the foundation for many clinical NLP models, including BioClinicalBERT, ClinicalBERT, and ModernBERT variants, widely used for information extraction and classification tasks.

BIOCLINICALBERT: A transformer-based model pre-trained on PubMed abstracts and MIMIC-III clinical notes, designed for biomedical and clinical NLP tasks such as concept extraction and relation detection.

BOOTSTRAPPED SIGNIFICANCE TESTING: A non-parametric resampling method used to estimate confidence intervals and assess statistical significance in model performance comparisons.

CAMIR (CORPUS OF ANNOTATED MEDICAL IMAGING REPORTS): A large-scale, event-structured corpus developed in this dissertation, containing radiology reports annotated with lesion-level findings, anatomical sites, and contextual attributes.

CLINICAL DECISION SUPPORT (CDS): A class of systems that assist healthcare professionals by delivering context-aware, evidence-based recommendations to enhance diagnostic and treatment accuracy.

CLINICAL NATURAL LANGUAGE PROCESSING (CLINICAL NLP): The subfield of NLP dedicated to analyzing unstructured medical text such as radiology or pathology reports, discharge summaries, and clinical notes.

CONDITIONAL RANDOM FIELD (CRF): A probabilistic sequence model used for token-level labeling tasks such as named entity recognition or event extraction.

ENTITY EXTRACTION: The process of identifying and categorizing key medical concepts (e.g., anatomy, finding, diagnosis) within unstructured text.

EVENT EXTRACTION: An NLP task that captures structured semantic relationships among entities, such as a radiologic finding (*lesion*) and its associated attributes (e.g., size, location, characterization).

FINE-TUNING: A supervised training process that adapts a pre-trained model (e.g., BERT, Llama) to a specific domain or task using labeled data.

FOLLOW-UP IMAGING RECOMMENDATION: A directive in a radiology report suggesting future imaging to monitor or evaluate a finding. Correct identification and adherence are central to patient safety and outcome tracking.

GENERATIVE LARGE LANGUAGE MODEL (LLM): A decoder-only transformer model capable of producing free-form text and performing reasoning tasks through in-context learning (ICL) or instruction following.

IN-CONTEXT LEARNING (ICL): A capability of LLMs to perform a new task by conditioning on examples provided directly in the prompt, without parameter updates.

INCIDENTAL FINDING / INCIDENTALOMA: An unexpected lesion or abnormality detected during imaging performed for unrelated reasons. Its identification and risk classification are critical for appropriate clinical management.

INFORMATION EXTRACTION (IE): The process of transforming unstructured text into structured representations (entities, relations, and events) suitable for automated analysis and decision support.

INSTRUCTION FINE-TUNING: A process where an LLM is fine-tuned on datasets of task instructions and responses to enhance its ability to generalize across diverse natural-language tasks.

INTER-ANNOTATOR AGREEMENT (IAA): A measure of label consistency among multiple human annotators, commonly evaluated using the F1-score or Cohen's κ, indicating dataset reliability.

LARGE LANGUAGE MODEL (LLM): A transformer-based model with billions of parameters trained on large corpora to perform diverse NLP tasks, including reasoning, summarization, and extraction.

LESION-LEVEL INFORMATION EXTRACTION: A fine-grained form of IE focusing on individual imaging findings within a report (e.g., nodules, cysts, or masses), capturing attributes such as anatomy, size, and clinical impression.

MODERNBERT / BIOCLINICALMODERNBERT: Long-context transformer architectures that extend BERT's maximum sequence length and incorporate biomedical vocabulary for improved clinical document understanding.

NAMED ENTITY RECOGNITION (NER): An NLP task that identifies and classifies textual spans representing domain-specific entities such as anatomical sites or clinical findings.

NON-PARAMETRIC BOOTSTRAP: A statistical technique that estimates confidence intervals by repeatedly resampling with replacement from the dataset, used in this dissertation to compare model F1-scores.

PL-MARKER++: An improved event-based information extraction architecture that extends PL-Marker with relation reasoning and multi-attribute lesion tagging for radiology text.

PROMPT ENGINEERING: The design and optimization of input prompts that guide LLMs toward accurate and interpretable outputs in specific domains.

RADIOLOGY REPORT: A narrative document written by a radiologist that summarizes imaging findings, interpretations, and recommendations; the primary textual source for radiology NLP research.

RELATION EXTRACTION (RE): The task of identifying semantic relationships between entities, such as linking a lesion to its anatomical location or follow-up recommendation.

RETRIEVAL-AUGMENTED GENERATION (RAG): An LLM approach that supplements model generation with externally retrieved context (e.g., guidelines or prior studies) to improve factual accuracy.

SUPERVISED LEARNING: A paradigm in which a model is trained on labeled data to predict outputs such as entity types, relations, or classification categories.

TRANSFORMER ARCHITECTURE: A neural network architecture based on self-attention mechanisms that capture contextual dependencies among tokens, forming the foundation of modern LLMs.

ZERO-SHOT AND FEW-SHOT LEARNING: Evaluation settings where a model performs a task without explicit fine-tuning (*zero-shot*) or with only a few examples provided in-context (*few-shot*).

ACKNOWLEDGMENTS

When I began my PhD, I often questioned whether I could grow into an independent researcher and contribute meaningfully to the field of clinical NLP. Completing a dissertation felt like an overwhelming challenge. I was able to reach this point because of the guidance, encouragement, and support I received from many people throughout this journey.

I would like to express my deepest gratitude to my advisor, Meliha Yetisgen. From my first day in the UW BioNLP Lab, she provided steady mentorship and unwavering support. Her guidance helped me develop a rigorous approach to research design, academic writing, and presentation skills. She also helped me settle smoothly in Seattle when I first arrived, which made the early stages of my PhD much easier. I am grateful for the opportunity to work on her research grants in radiology information extraction, which laid the foundation of this dissertation. Her thoughtful feedback and consistent encouragement played a central role in my growth as a researcher.

I am thankful to my committee members, Trevor Cohen, Aylin Caliskan, and Kevin Lybarger. Trevor's guidance in my first-year coursework and his insightful comments on my dissertation strengthened the way I think about research. Aylin generously agreed to serve as my GSR and provided valuable input as I shaped this dissertation. Kevin, who mentored me when he was a postdoctoral researcher in the lab, influenced my development through our collaborations and many meaningful discussions.

I am also grateful for the opportunity to work with collaborators such as Martin Gunn, Ozlem Uzuner, and Fei Xia. Their expertise and feedback during our meetings shaped many aspects of my work and helped me grow as a clinical NLP researcher.

My PhD experience was made much more enjoyable because of the supportive community

around me. I would like to thank my lab mates Nic Dobbins, Sitong Zhou, Weipeng Zhou, Arjun Chakraborty, Velvin Fu, Zhaoyi Sun, Avery Yu, and Sihang Zeng, as well as my BIME cohort Bhargav Vemuri, Ehsan Alipour, Kevin Li, Ashmitha Rajendran, and Serena Xie. I am grateful for the many coffee chats with my Korean friends in South Lake Union, which provided comfort and encouragement throughout the program. I also appreciate the friends who shared countless tennis games with me, creating some of my favorite and most relaxing memories during these years. I am especially thankful to my cousin, Jung Lee, whose support in many aspects of daily life in Seattle allowed me to stay focused on my research.

Finally, I thank my parents and sister for their unconditional love and support. Their encouragement has always been my foundation and made it possible for me to pursue and complete this long academic journey.

Chapter 1

INTRODUCTION

1.1 Context and Motivation

Radiology reports capture radiologists' interpretations of medical images through detailed narrative text. Although structured reporting systems using Common Data Elements (CDE) have been introduced to standardize interpretations with predefined medical concepts [110, 117], the majority of reports continue to be in free-text format [143]. Information extraction (IE) techniques offer a pathway to automatically convert these unstructured narratives into structured semantic formats, enabling their integration into secondary applications for clinical decision support and research. Such applications include cohort discovery [22], image retrieval [46], automated follow-up tracking [85], integration with computer vision systems [152], clinical decision support [32], impression section generation [83, 129], and report summarization [142].

Despite notable advancements in radiology IE, especially with the rise of Large Language Models (LLMs), much of this research remains narrowly focused on specific clinical tasks [22, 37] or limited to individual imaging modalities [35]. Furthermore, many existing studies use annotation schemas that do not fully capture the wide range of information in radiology reports. Effective model development depends on high-quality annotated datasets that reflect clinically relevant insights and capture radiologists' nuanced language. For robust and generalizable models, diverse datasets encompassing various imaging modalities and conditions are essential. However, the lack of publicly available, general-purpose radiology datasets hinders the ability to compare different IE models. Although some studies have employed large-scale radiology datasets for training [146, 121], large-scale IE in real-world radiology reports remains largely unexplored. Recent findings show that GPT-4 [96] performs comparable to or better than specialized radiology models [76, 77]; however, its applicability

to practical clinical tasks remains largely unexamined.

This study addresses these gaps with two main objectives: (1) the development of a general-purpose corpus of radiology reports annotated with detailed clinical findings, and (2) the application of scalable, LLM-based IE in real-world clinical tasks, leveraging a large-scale clinical database. By pursuing these goals, we aim to identify optimal IE strategies for secondary applications of radiology reports, ultimately improving patient outcomes and reducing costs associated with clinical misunderstandings that stem from unstructured, verbose radiology narratives.

1.2 Objectives and Contributions

The overarching objective of this dissertation is to advance the automated understanding and structured representation of radiology reports to support clinical decision-making and large-scale data analysis. Radiology reports contain rich clinical information, yet their predominantly unstructured narrative format poses major challenges for computational processing, integration, and secondary use. To address these limitations, this dissertation proposes a unified framework that combines dataset construction, model development, and large-scale application through three specific aims.

Aim 1: Development of an Annotated Dataset of Radiology Reports with Clinically Significant Findings and BERT-based Information Extraction Model We developed a richly annotated dataset of radiology reports, incorporating detailed annotations of clinically significant findings. The reports are sampled from an existing clinical database spanning 2007–2020, encompassing a broad patient population from the University of Washington Medical System Hospitals. This dataset serves as a training resource for fine-tuning an information extraction model designed to function as a foundational model for extracting information from radiology reports. We conduct baseline experiments to compare our proposed architecture with existing IE models in radiology and then apply it on a larger scale, extracting clinically relevant information from all available radiology reports within our clinical database.

Aim 2: Identification of Medical Imaging Follow-up through Longitudinal Report-Level Analysis

While existing NLP studies in radiology address a range of tasks, limited research has focused on analyzing sequences of radiology reports for individual patients over time. To bridge this gap, we address the task of identifying follow-up imaging reports from each patient's chronologically ordered series of radiology reports, a critical real-world task with direct implications for patient health outcomes. Using the same database as in Aim 1, we create a dataset comprising a series of reports for each patient. This dataset enables us to evaluate both traditional machine learning approaches and the capabilities of recent LLMs for this task. Through a detailed analysis of LLM performance, we offer insights into effective prompt design for radiology report-related tasks, enhancing understanding of LLM utility in real-world clinical settings.

Aim 3: Automatic Detection of Incidental Findings in Radiology Reports with Enhanced Explainability

Identifying incidental findings in real-world clinical settings is a complex task that demands clinical expertise, as it requires consideration of multiple facets of a patient's health status and medical history. Despite its clinical importance, the application of NLP techniques for the automatic detection of incidental findings remains underexplored. With input from medical experts, we develop a dataset with detailed annotations necessary for accurately identifying incidental findings and capturing their clinical context. We evaluate the performance of supervised learning models alongside state-of-the-art generative LLMs to assess their accuracy and explainability. Finally, we apply this integrated framework to a selected cohort within our institutional database, enabling retrospective analyses of follow-up adherence related to incidentalomas and establishing a foundational resource for large-scale incidentaloma identification in clinical settings.

Summary of Contributions

Collectively, these objectives and corresponding aims establish a cohesive framework that integrates dataset creation, model development, and real-world clinical evaluation. The

primary contributions of this dissertation are as follows: (1) development of a large-scale annotated corpus of radiology reports with detailed clinical findings; (2) systematic evaluation of machine learning and LLM-based methods for identifying follow-up imaging recommendations using longitudinal report sequences; and (3) creation of an NLP-based framework for detecting incidental findings at scale to support clinical decision-making.

Together, these contributions provide essential methodological and practical advances for transforming unstructured radiology text into structured, actionable knowledge and lay the groundwork for scalable, interpretable, and clinically meaningful applications of NLP in medical imaging.

1.3 Guide for the Reader

The remainder of this thesis is organized as follows:

Chapter 2: Background and Motivations. This chapter reviews the foundational literature and technical context that underpin this work. It summarizes the structure and content of radiology reports, existing annotated corpora for radiology NLP, and major categories of information extraction (IE) models. The chapter also surveys prior studies on follow-up imaging recommendation identification and incidentaloma detection, highlighting critical limitations in current methods. These discussions provide the conceptual and methodological basis for the data creation and modeling approaches described in later chapters.

Chapter 3: CAMIR – Corpus of Annotated Medical Imaging Reports. This chapter presents the construction of a richly annotated corpus of radiology reports (CAMIR), which serves as a large-scale benchmark for radiologic IE. It describes the data curation and annotation pipeline, the event-based schema that captures lesion-level and contextual attributes, and evaluation results for both baseline and advanced IE models, including PL-Marker++ [101]. This chapter establishes the foundational dataset and extraction framework upon which subsequent tasks build, enabling generalizable and anatomically detailed representations of radiologic findings.

Chapter 4: Identification of Imaging Follow-Up in Radiology Reports. This

chapter extends the scope of information extraction to longitudinal report-level analysis by addressing the clinically important task of identifying follow-up imaging recommendations. It outlines the dataset design, sampling and annotation procedures, and comparative evaluation of traditional machine learning, transformer-based encoders, and generative LLMs. Quantitative and qualitative analyses, including statistical significance testing and error characterization, demonstrate the advantages and limitations of LLM-based reasoning for real-world clinical tracking of follow-up adherence.

Chapter 5: Automatic Identification of Incidentalomas in Radiology Reports. This chapter focuses on the automated detection of incidental findings (incidentalomas) across multiple anatomies, integrating lesion-level information extraction with interpretability and clinical validation. It introduces a unified incidentaloma classification framework evaluated across both supervised and generative LLM architectures, accompanied by pairwise bootstrap significance testing and detailed error analyses. The discussion sections highlight ensemble effects, anatomy-specific performance, and clinical implications for follow-up adherence monitoring, contributing a scalable NLP-based methodology for incidentaloma surveillance.

Chapter 6: Case Study: Adrenal Incidentaloma. This chapter integrates the methods developed in Chapters 3–5 and applies them to a focused clinical use case: adrenal incidentalomas. Using CAMIR-based lesion extraction, longitudinal follow-up identification, and LLM-based incidentaloma classification, we construct an end-to-end pipeline for large-scale adrenal lesion analysis. The chapter summarizes the adherence rate and severity distribution of adrenal incidentalomas and relates these findings to recent evidence in adrenal imaging, aiming to provide empirical data that inform and contextualize evolving clinical recommendations.

Chapter 7: Conclusion and Future Work. The final chapter synthesizes the key contributions of this dissertation, emphasizing the creation of a large-scale annotated resource, the systematic evaluation of LLMs for radiologic text understanding, and the demonstration of clinical utility through incidentaloma and follow-up detection. It concludes by discussing methodological limitations, potential applications in multimodal modeling and

clinical deployment, and future directions for advancing LLM interpretability and real-world generalization.

1.4 References

The chapters of this dissertation are adapted from my peer-reviewed, first-author publications listed below:

- Chapter 3 : adapted from Park et al. [101];
- Chapter 4 : adapted from Park et al. [103];
- Chapter 5 : adapted from Park et al. [102];

Chapter 2 synthesizes and contextualizes findings from all of the above works.

Chapter 2

BACKGROUND AND MOTIVATIONS

This chapter reviews key developments in applications of NLP to the domain of radiology, emphasizing report structure, annotated corpora, and methods for clinical information extraction. It outlines the evolution from rule-based systems to transformer and LLM-based models and summarizes prior research on extracting follow-up recommendations and detecting incidental findings, providing the foundation for the methods introduced in later chapters.

2.1 Structure and Content of Radiology Reports

Radiology reports serve as the primary means of communicating diagnostic imaging findings between radiologists and referring clinicians. As the formal documentation of a radiologist's interpretation, the report must be clear, comprehensive, and clinically precise. It not only describes imaging observations but also conveys diagnostic reasoning, clinical context, and recommendations for further management. Given that downstream clinical decisions, such as follow-up imaging, biopsies, or treatment planning, often depend on the contents of these reports, the accuracy and structure of radiology documentation are critical for ensuring patient safety and continuity of care.

While there is no single universally mandated format for radiology reports, most follow a broadly similar structure. A conventional report is typically divided into sections covering key information such as the patient and study identifiers, the clinical history or indication for the exam, the imaging technique or protocol used, the description of findings, and an impression or conclusion [6]. This standardized sectional format helps ensure that all essential information is included and easy to locate. For example, the American College of Radiology (ACR) recommends that radiology reports include specific components such as procedure

details, clinical indication, comparison to prior studies when available, detailed findings, and a clear conclusion or impression. Certain subspecialties have even more formalized reporting structures. Breast imaging reports, for instance, often use the Breast Imaging Reporting and Data System (BI-RADS) format to ensure consistent content and terminology [124].

The information content of each section serves a distinct and complementary role (Figure 2.1). The findings section typically provides a detailed narrative of all significant imaging observations, often organized by anatomic region or pathology, and may include quantitative measurements or characterization of abnormalities. The impression (or conclusion) section distills these findings into a concise summary that addresses the clinical question and emphasizes the most relevant results. This section is often prioritized by referring physicians, as it conveys the actionable conclusions or diagnoses derived from the imaging study. In addition, high-quality reporting practice integrates relevant clinical information (such as patient history or presenting symptoms) and compares current findings with prior imaging studies to assess longitudinal changes. Including such contextual details enhances the interpretative value of the report for clinicians and facilitates downstream management decisions.

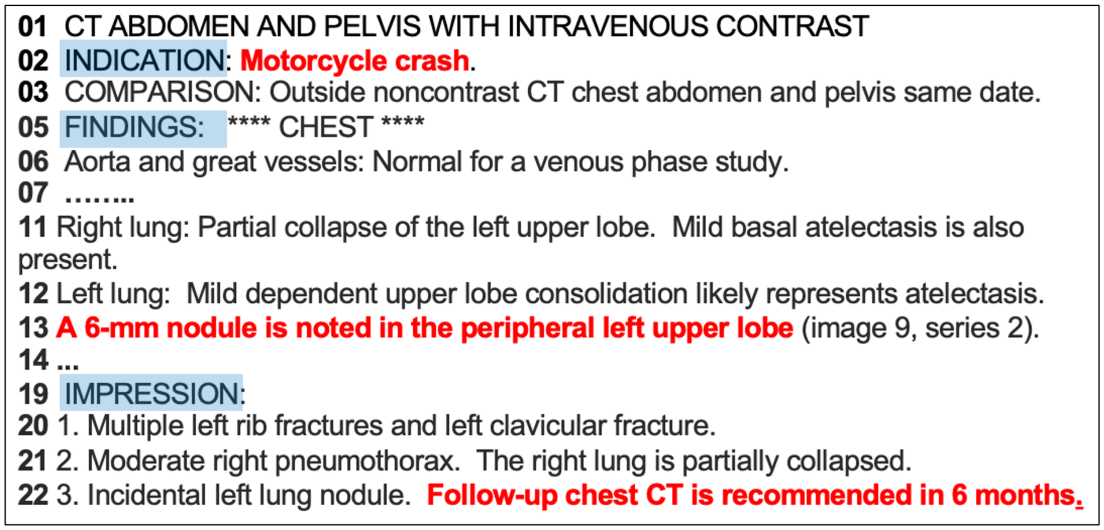

Figure 2.1: Example of a radiology report with structured reporting

Despite general guidelines, studies have noted substantial variation in how radiology reports are structured and in the completeness of the information they contain. For instance, a study examining hundreds of radiology reports found wide differences in reporting style, largely influenced by the radiologist's subspecialty, institutional culture, and the nature of findings such as nodule size or lesion type [48]. To address these inconsistencies, structured reporting has emerged as a major focus in radiology informatics. The European Society of Radiology (ESR) identified three main benefits of structured radiology reports: (1) quality improvement, (2) quantification, and (3) accessibility [95]. Structured reporting enhances consistency by enforcing standardized templates, ensuring all clinically relevant questions are addressed, and employing controlled terminologies that improve clarity and interoperability. Moreover, structured formats enable the integration of quantitative imaging biomarkers and clinical metadata, facilitating personalized medicine, decision support, and artificial intelligence (AI) applications through standardized vocabularies such as RadLex [69]. Regarding accessibility, structured reports, when encoded in interoperable formats, support secondary uses such as research, clinical audit, and quality assurance, although legacy systems and inconsistent adoption still limit full integration across institutions.

In recent years, there has been a concerted effort within the radiology community to promote structured reporting through the use of standardized templates and software-assisted authoring tools [93]. Structured reporting involves predefined sections or phrase libraries that prompt radiologists to provide complete and consistent information, thereby reducing variability and omissions. Multiple studies suggest that structured reports improve report completeness, comparability, and communication with both clinicians and patients [132, 135]. However, implementation challenges remain. Structured formats can restrict the expressive flexibility of narrative text [44] and may face resistance due to workflow disruptions or radiologists' preference for free-text reporting [107].

Beyond the human-centered perspective, the structure and content of radiology reports have become increasingly important in the era of data-driven healthcare. The linguistic regularities and standardized sections of radiology reports make them a valuable substrate

for developing natural language processing (NLP) and large language model (LLM) methods aimed at clinical information extraction. By systematically analyzing report sections such as "Findings" and "Impression," computational models can identify clinical entities, infer diagnostic intent, and detect follow-up recommendations or incidental findings. As such, understanding the organization and information flow of radiology reports is essential for designing robust NLP pipelines that can support automated decision-making and enhance clinical workflow efficiency.

2.2 Annotated Corpora for Radiology NLP Tasks

In existing radiology corpora, document-level or sentence-level annotations link relevant text spans to normalized values, covering various label categories such as metastasis descriptions [115, 35, 147, 62] and incidental findings [108, 134, 90]. Entity annotations are used to identify key phrases representing concepts such as anatomical sites [138] or tumor characteristics [150]. More advanced annotation frameworks include relation and event structures that capture multi-attribute descriptions of medical conditions [72]. Several corpora have also integrated standardized radiological terminologies for anatomy [81, 28, 92] and other domain-specific vocabularies such as RadLex [29], which support semantic normalization and interoperability.

Despite these advances, current radiology corpora exhibit several limitations related to diversity, coverage, and scalability. Many datasets are constrained by narrow clinical focus, emphasizing specific diseases or pathologies such as hepatocellular carcinoma [150] or appendicitis [109], which limits their generalizability across radiology subspecialties. Others are restricted to a single imaging modality [72, 128] or rely on small sample sizes ($n<200$) [51], reducing the robustness of downstream models trained on them. Furthermore, relation extraction corpora often lack normalization of extracted entities to standardized ontologies [62], which hinders consistent concept representation. Although some efforts [81] have successfully aligned extracted findings with normalized anatomical structures, these annotations frequently omit fine-grained contextual attributes such as lesion size, characterization, or associated clinical recommendations.

Overall, there remains a need for large-scale, diverse, and richly annotated corpora that capture the breadth of radiological findings across anatomies, modalities, and clinical contexts. Developing such datasets with comprehensive annotation schemas is essential for advancing robust, generalizable information extraction models that can effectively support clinical decision-making and downstream applications in radiology informatics.

2.3 Information Extraction Models in Radiology NLP

Early approaches to radiology information extraction (IE) primarily used discrete machine learning models with handcrafted linguistic and lexical features. Support Vector Machines (SVMs) were employed to detect findings related to appendicitis [109], while Conditional Random Fields (CRF) [130] were applied to extract anatomical details from the findings sections of reports [51]. These approaches relied heavily on domain-specific feature engineering and limited contextual understanding. Subsequently, neural network architectures, including Convolutional Neural Networks (CNNs) [97] and Recurrent Neural Networks (RNNs) [112], achieved higher accuracy in capturing semantic and sequential dependencies within clinical text. Neural models have been successfully applied to a range of IE tasks such as recommendation extraction [21, 126], identification of clinical concepts [154], and spatial or relational information extraction [30].

The advent of transformer-based language models [136] has fundamentally advanced clinical NLP by providing contextualized embeddings derived from large-scale biomedical and clinical corpora. Models such as BioClinicalBERT [5], trained on MIMIC-III [63] and PubMed texts, have shown strong performance in concept extraction and relation detection tasks. Variants including ClinicalBERT [58] and other transfer learning approaches [104] improved robustness across various clinical domains. To accommodate longer medical narratives, Clinical Longformer and Clinical BigBird introduced sparse attention mechanisms for efficient modeling of extended sequences, including radiology reports [75]. More recent architectures, such as ModernBERT [139] and BioClinicalModernBERT [123], incorporate long-context adaptations and biomedical vocabulary alignment to better capture domain-specific information.

Generative Large Language Models (LLMs) have expanded the scope of clinical NLP beyond supervised classification and extraction. Models such as GPT-4 [2] and GPT-4o [59] demonstrate advanced reasoning and comprehension capabilities, achieving competitive performance on medical question answering and diagnostic reasoning benchmarks [94, 59]. Similarly, Med-PaLM [118], an instruction-tuned version of PaLM [24], performs strongly on clinically focused question answering and reasoning tasks. Open-source models like LLaMA [133] have become the backbone of domain-specific fine-tuning through lightweight adaptation techniques such as Low-Rank Adaptation (LoRA) [56], while instruction-tuned variants like MedAlpaca [50] demonstrate that general-purpose models can be effectively aligned with medical tasks using specialized instruction datasets.

In the radiology domain, both transformer-based and generative models have been applied to diverse NLP tasks including report classification, lesion entity extraction [55, 73], and malignancy detection [61, 45]. Comparative evaluations highlight their advantages in document-level classification but also underscore challenges in fine-grained multi-class tasks, where false negatives may lead to missed actionable findings [140, 86]. Generative approaches further extend these capabilities to tasks such as report summarization [77], structured entity extraction [101], and clinical reasoning from free-text imaging reports [137]. Recent work has also explored multimodal IE that integrates textual and imaging data for more comprehensive modeling of radiological context [60, 8, 105, 106]. GPT-based systems have demonstrated particular promise in structuring information from narrative reports [43, 89, 3], with GPT-4 [96] showing performance comparable to expert-written impression sections [76].

Despite these advancements, large-scale, real-world deployment of LLMs in radiology remains limited due to challenges in scalability, domain adaptation, and interpretability. Addressing these issues, this dissertation develops a scalable IE framework for radiology that integrates both supervised transformer-based models and generative LLMs. The approach is systematically evaluated for incidentaloma detection across multiple anatomies, introducing various domain-specific prompting techniques. Through this unified design, the work advances radiology NLP from document-level prediction to lesion-level understanding, bridging the

gap between automated information extraction and clinically actionable insight.

2.4 Follow-Up Imaging Adherence Identification

2.4.1 Follow-up Imaging Characteristics

Radiology reports frequently contain recommendations for follow-up imaging, which play a critical role in patient management and longitudinal disease monitoring. Prior research estimates that between 4% and 26% of radiology reports include at least one recommendation for additional imaging, depending on the clinical context and modality of the initial examination [119]. These recommendations may specify a preferred imaging modality, the clinical rationale, and a suggested time frame for reassessment, although the degree of detail varies considerably across reports and radiologists. In routine clinical workflows, radiologists communicate their recommendations to the referring clinician, who is then responsible for informing the patient and coordinating subsequent care [70]. However, many radiology departments lack comprehensive systems for tracking the completion of these follow-up actions, which often prevents radiologists from knowing whether recommended imaging was ever scheduled or performed.

Ensuring appropriate follow-up is particularly crucial in cases involving indeterminate or potentially malignant findings. Despite the clinical importance, adherence to follow-up recommendations remains inconsistent and often suboptimal. While some follow-ups are intentionally deferred for valid clinical reasons, a substantial proportion are missed due to communication and process breakdowns. Studies indicate that more than one-third of follow-up recommendations are never acted upon, frequently because the referring clinician did not acknowledge the recommendation [20]. Another investigation found that over 10% of cases involving potentially malignant findings lacked appropriate follow-up imaging [120]. Contributing factors include inadequate communication between radiologists and clinicians, competing clinical priorities, fragmented care transitions, insufficient patient notification, and logistical barriers that prevent patients from completing recommended imaging [67]. These

failures represent not only operational inefficiencies but also potential patient safety risks, as delayed follow-ups can lead to missed or late diagnoses.

Recent studies have sought to characterize and quantify the nature of follow-up recommendations documented in radiology reports. White et al. performed a detailed manual review of radiology reports from a large academic medical center to analyze recommendation patterns and assess "loop closure" rates—the proportion of recommended follow-up studies that were completed [141]. Their analysis revealed that among 532 reports containing one or more actionable recommendations, 370 (61.9%) lacked a specified follow-up time frame. The study also highlighted the frequent use of ambiguous or conditional language (e.g., "if clinically indicated" or "consider repeat imaging"), which further complicates follow-up adherence tracking. Other research has explored variability in follow-up practices, examining how patient demographics, clinical settings, imaging modalities, and even individual radiologists influence the frequency and specificity of recommendations [25].

Together, these studies underscore a persistent gap between radiologic recommendation generation and downstream follow-up completion. From an informatics perspective, this gap reflects a broader need for automated systems capable of identifying, extracting, and monitoring follow-up recommendations from free-text radiology reports at scale. Leveraging NLP and LLMs to detect these recommendations could enable real-time surveillance of adherence and facilitate communication between radiology departments, clinicians, and patients.

2.4.2 Follow-up Recommendation Identification and Adherence Tracking

Given the clinical importance of identifying follow-up recommendations to prevent delayed or missed diagnoses, many NLP-based approaches have been developed to automatically extract this information from radiology reports. Early work by Yetisgen et al. introduced a supervised text classification framework that identified recommendation sentences using linguistic and semantic features beyond simple unigram tokens [149]. Conditional Random Fields (CRFs) [68] were later applied to capture temporal cues and contextual dependencies within follow-up

recommendation statements [145]. Other studies compared traditional machine learning algorithms such as Support Vector Machines (SVMs) [52] and Random Forests (RFs) [17] with deep learning models including Recurrent Neural Networks (RNNs) [112] for detecting and classifying follow-up recommendations [21].

More recent work has shifted toward corpus-driven neural architectures trained on annotated datasets. Lau et al. constructed a large corpus of radiology reports labeled with entities and attributes related to follow-up recommendations, trained a neural model on this dataset, and applied it at scale to assess adherence rates [71]. Similarly, other studies have developed automated tools that extract the key components of follow-up recommendations, such as the recommended time interval, imaging modality, and anatomical focus, to evaluate compliance with clinical guidelines [84]. These systems have improved detection accuracy and facilitated institutional assessments of follow-up practices.

Although detection of follow-up recommendation sentences has been studied extensively, relatively few investigations have addressed the more complex task of determining whether the recommended follow-up was completed. Some studies have attempted to estimate adherence by using imaging modality type alone [71], but this method only approximates true follow-up events because identical modalities may be used for unrelated clinical purposes. Other systems, such as those described by Cook et al., attempted to automate tracking of follow-up completion [26], but they depended on predefined templates and institution-specific lexicons, which limited their generalizability across clinical settings.

Dalal et al. advanced this area by developing an Extremely Randomized Trees model [47] that incorporated multiple clinical features, including recommendation metadata, text similarity measures, and report-level information, to determine follow-up completion status [27]. The model achieved an F1 score of 0.807, approaching the inter-annotator agreement of 0.853, and demonstrated strong potential for practical deployment in radiology workflows. Despite these advances, scalable and generalizable systems that can reason over heterogeneous report formats and variable linguistic expressions remain limited.

Recent progress in LLMs provides new opportunities to overcome these limitations. LLMs

can integrate contextual reasoning, temporal relationships, and implicit clinical cues that conventional models often miss. In this dissertation, we extend prior research by systematically evaluating both traditional machine learning methods and state-of-the-art LLMs for the task of identifying and verifying follow-up imaging adherence within large radiology corpora. Our approach aims to create a scalable, interpretable, and clinically applicable framework for automated follow-up tracking.

2.5 Automated Detection of Incidentalomas

2.5.1 Incidentaloma Concepts and Challenges

Incidental findings, also referred to as *incidentalomas*, are unexpected abnormalities detected during imaging studies that were performed for unrelated clinical reasons. Their increasing recognition as a critical issue in modern medicine reflects the rapid rise in advanced imaging utilization and improvements in image resolution, particularly in emergency and outpatient settings [122]. Despite their clinical significance, the systematic identification and characterization of incidentalomas remain relatively understudied, especially in the context of NLP and automated information extraction. Determining the potential malignancy of incidental findings requires considerable clinical expertise, which makes the annotation and interpretation processes complex. Moreover, because incidentalomas are relatively rare compared with other findings, constructing representative datasets for model development is challenging. As a result, limited research has directly addressed this problem, even though accurate detection has substantial implications for timely diagnosis, follow-up adherence, and overall patient outcomes.

The prevalence of incidentalomas varies widely across imaging modalities and clinical contexts. Recent studies report that incidental findings are observed in more than one-third of cardiac MRI scans, chest CT examinations (which may reveal thoracic, abdominal, spinal, or cardiac incidental findings), and CT colonography studies where extra-colonic abnormalities are frequently identified [98, 53]. In contrast, prevalence rates fall below 5% for chest CTs

performed to detect incidental pulmonary embolisms and for whole-body PET or PET/CT scans in both cancer and non-cancer populations [33, 88, 19]. MRI scans of the brain and spine demonstrate intermediate rates, with prevalence estimates of approximately 18% and 22%, respectively [65, 114].

Malignancy rates among incidentalomas also vary considerably depending on the organ system involved. For example, 48% of breast incidentalomas have been reported as malignant [14], 13.8% of colorectal incidentalomas were malignant with an additional 40.4% categorized as premalignant [66], and 19.8% of thyroid incidentalomas were confirmed to be malignant [91]. Although many incidentalomas are ultimately benign, the potential for malignancy, particularly in specific anatomical regions, creates a meaningful association with elevated mortality risk. Consequently, early identification and appropriate follow-up of malignant incidentalomas are essential for improving patient outcomes. However, adherence to follow-up recommendations is often suboptimal. For instance, Maher et al. found that among 245 patients with adrenal incidentalomas, only 88 (36%) received follow-up care within the same institution [87].

The observed variability in both prevalence and malignancy risk underscores the importance of developing targeted and scalable methods for incidentaloma detection. Automated approaches that leverage NLP and large language models can help identify these findings consistently and facilitate timely clinical intervention. As incidental findings continue to increase with the growth of diagnostic imaging, systematic detection frameworks are essential for reducing missed follow-ups and supporting evidence-based patient management.

2.5.2 Identification of Incidentalomas

With several studies analyzing the prevalence and clinical characteristics of incidentalomas [23, 15, 80, 99, 10], substantial research has focused on management strategies for these unexpected findings, often tailored to specific anatomical regions [151, 9, 31, 131]. However, relatively few studies have explored the use of NLP for identifying incidentalomas directly from radiology reports. Early approaches relied on traditional machine learning and rule-based

systems. For example, Random Forest (RF) classifiers [17] and keyword-driven methods were applied to detect incidental findings [38, 40, 64]. Schumm et al. developed a Structured Query Language (SQL)-based keyword search system to identify adrenal incidentalomas, achieving a detection rate of 68.7% [111]. However, this method required manual clinician validation, which limited scalability and generalizability. Similarly, Bala et al. employed a Convolutional Neural Network (CNN) [97] to predict whether radiology reports contained adrenal incidentalomas, using a corpus of 4,090 manually annotated reports [7]. Although the model demonstrated promising accuracy, the low prevalence of incidentalomas (9.9%) within the dataset reduced the availability of positive samples for effective supervised learning.

Recent advancements in LLMs have introduced new opportunities for incidentaloma detection. Bhayana et al. [16] evaluated GPT-4's capability to identify incidental adrenal nodules, pancreatic cystic lesions, and vascular calcifications, performing document-level binary classification using single-shot prompts. Their results were comparable to those of specialized NLP systems, indicating the feasibility of using GPT-4 for incidental finding detection. Building upon this work, Woo et al. investigated GPT-4's performance in identifying actionable incidental findings within Emergency Department (ED) radiology reports [144]. Their study employed a zero-shot prompt refined on 50 reports and subsequently applied it to 430 additional reports. Although the model showed promising results, only 108 reports contained true incidental findings, and potential sampling bias may have influenced performance due to institutional report templates and keyword-based selection.

Despite these advances, current research on incidentaloma detection exhibits several important limitations. Most existing studies focus on document-level classification of incidental findings, which does not provide the granularity required for lesion-level understanding. Lesion-specific details such as anatomical location, malignancy assessment, and associated follow-up recommendations are essential for guiding individualized patient care. Furthermore, while GPT-4 has been explored in this domain, other advanced language models, including BERT-based architectures, Llama3 [1], and GPT-4o [59], have not yet been systematically evaluated for incidentaloma detection.

To address these gaps, this dissertation aims to create a large, annotated dataset of incidentalomas that captures detailed lesion-level information, including attributes such as malignancy status and follow-up recommendations. This resource will serve as a foundation for training and evaluating multiple NLP models, encompassing both supervised transformer-based approaches and generative LLMs. Finally, the developed models will be applied across the entire institutional radiology report database, enabling large-scale, population-level analyses of incidentaloma prevalence, clinical management patterns, and downstream health outcomes.

Chapter 3

CAMIR: CORPUS OF ANNOTATED MEDICAL IMAGING REPORTS

In this chapter, we present our novel radiology report dataset, **C**orpus of **A**nnotated **M**edical **I**maging **R**eports (**CAMIR**), which includes 609 annotated radiology reports from three imaging modality types: Computed Tomography, Magnetic Resonance Imaging, and Positron Emission Tomography-Computed Tomography. CAMIR uniquely combines a granular event structure and concept normalization. Reports were annotated using an event-based schema that captures important clinical aspects including clinical indications, lesions, and medical problems. To extract CAMIR events, we explored two BERT (Bi-directional Encoder Representation from Transformers)-based architectures, including an existing architecture (mSpERT [82]) that jointly extracts all event information and a multi-step approach (PL-Marker++ [101]) that we augmented for the CAMIR schema. Annotation guidelines and code used for all experiments from this chapter are available in our GitHub Repository[1].

3.1 Methods

In this section, we will describe the details of our dataset curation and information extraction model development.

3.1.1 Corpus Creation

CAMIR was developed using a clinical database comprising 1,417,586 CT, 541,388 MRI, and 39,150 PET-CT reports from 2007-2020, representing a diverse patient population from

[1]https://github.com/uw-bionlp/CAMIR

four hospitals within the University of Washington Medical System. For this dataset, we randomly sampled 203 CT, 202 MRI, and 204 PET-CT reports, which were then de-identified using a neural de-identifier [74]. BRAT rapid annotation tool [127] was used throughout the annotation process (Figure 3.1). This study was approved by the University of Washington Institutional Review Board (IRB).

Annotation Schema

The annotation schema in CAMIR uses an event-based structure, where each event includes a trigger and arguments that provide detailed characterization (Table 3.1). CAMIR captures three event types: (1) *Indication* - the reason for imaging (e.g., "cancer" in line 1 of Figure 3.1); (2) *Lesion* - mass-occupying pathological structures (e.g., "metastasis" in line 3 of Figure 3.1); and (3) *Medical Problem* - non-mass-like abnormalities, defined as findings not classified as potential masses (e.g., "scarring" in line 1 of Figure 3.1). Additionally, CAMIR specifies two types of arguments: (1) *span-only* arguments, where text spans are labeled with an argument type (e.g., "focal" assigned the *Characteristic* argument in line 2 of Figure 3.1); and (2) *span-with-value* arguments, where text spans are assigned both an argument label and a subtype label (e.g., "New" assigned the *Size Trend* argument with a subtype value of *new* in line 2 of Figure 3.1). These structured annotations, illustrated in Figure 3.1 and

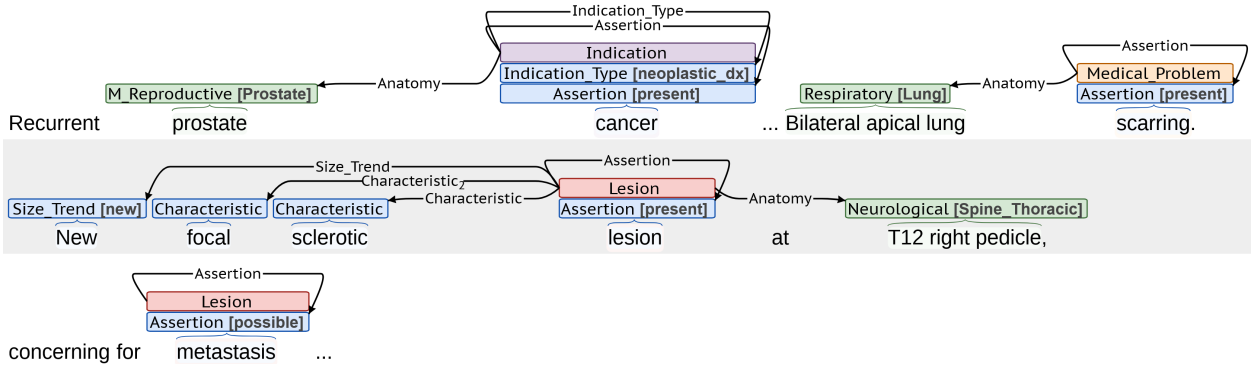

Figure 3.1: Example annotations from CAMIR using BRAT rapid annotation tool

summarized in Table 3.1, provide CAMIR with the granularity required for in-depth analysis and secondary applications. To improve the granularity of our annotation schema, anatomy arguments are normalized to a set of hierarchical anatomical SNOMED-CT [36] concepts, including 16 *Anatomy Parent* and 71 *Anatomy Child* labels listed in Table 3.2 (e.g., "Bilateral apical lung" assigned *Anatomy Parent* - *Respiratory* and *Anatomy Child* - *Lung* in line 1 of Figure 3.1).

Table 3.1: Summary of the CAMIR event schema. * indicates the argument is required. + *Anatomy Parent* and *Anatomy Child* are listed in Table 3.2.

Event	**Trigger/ Argument**	**Argument subtypes**	**Span examples**
Indication	Trigger*	–	"hemorrhage," "sepsis"
	Type*	{trauma, symptom, neoplastic diagnosis, non-neoplastic diagnosis}	"seminoma," "sarcoid"
	Assertion*	{present, absent, possible}	"r/o," "concern"
	Anatomy	Anatomy Parent and Child labels+	"abdominal," "alveolar"
Lesion	Trigger*	–	"'lymhpadenopathy"
	Assertion*	{present, absent, possible}	"most likely,"
	Anatomy	Anatomy Parent and Child labels+	"lower back"
	Size	{current, past}	"up to 5mm"
	Size Trend	{new, disappear, increasing, decreasing, no-change}	"increasing in size"
	Count	–	"multiple," "numerous"
	Characteristic	–	"peripheral," "enlarged"
Medical Problem	Trigger*	–	"dilation," "calcification"
	Assertion*	{present, absent, possible}	"possibly"
	Anatomy	Anatomy Parent and Child labels+	"mucosal," "supraagger"

Table 3.2: Anatomy Parent–Child hierarchy used for anatomy normalization. All 16 Parent and 71 Child labels correspond to SNOMED-CT concepts. Count reflects the number of Parent-level annotations.

Anatomy Parent	**Anatomy Children**	**Count**
Abdomen (113345001)	Abdominal Wall, Adrenal Gland, Mesentery, Peritoneal Sac, Retroperitoneal, Spleen, Undetermined	512
Cardiovascular System (59820001)	Arterial, Coronary Artery, Heart, Pericardial Sac, Pulmonary Artery, Venous, Undetermined	770
Digestive System (49596003)	Esophagus, Intestine, Large Intestine, Small Intestine, Stomach, Undetermined	425
Female Reproductive System (27436002) & Obstetric (308762002)	Adnexal, Breast, Extra-embryonic, Female Genital Structure, Fetus, Ovary, Placenta, Umbilical Cord, Uterus, Undetermined	272
Head & Neck (774007)	Ear, Eye, Laryngeal, Mouth, Nasal Sinus, Neck, Pharynx, Thyroid, Undetermined	1096
Hepato-Biliary System (34707002, 122489005)	Bile Duct, Gallbladder, Liver, Pancreas, Undetermined	609
Lymphatic (91688001)	Undetermined	559
Male Reproductive System (90264002)	Epididymis, Prostate, Testis, Undetermined	49
Miscellaneous	Adipose Tissue, Biomedical Device, Connective Tissue, Undetermined	59
Musculoskeletal (312717002)	Bone/Joint, Skeletal or Muscle, Undetermined	1811
Neurological System (25087005)	Brain, CSF Pathway, Cerebrovascular System, Extraaxial, Nerve, Pituitary, Spine (Cervical, Thoracic, Lumbar, Sacral, Unspecified), Undetermined	3235
Other Body Regions (272625005)	Entire Body, Lower Limb, Pelvis, Upper Limb, Undetermined	887
Respiratory System (714323000)	Lung, Pleural Membrane, Tracheobronchial, Undetermined	1200
Skin (400199006)	Skin or Mucous Membrane, Subcutaneous, Undetermined	58
Thoracic (51185008)	Mediastinum, Undetermined	772
Urinary System (122489005)	Kidney, Ureter, Urinary Bladder, Undetermined	378

Annotation Process

The annotation process involved four medical students who received guidance from a senior radiology resident and an experienced board-certified radiologist, both of whom contributed to the creation and refinement of annotation guidelines. The guidelines were carefully revised to capture relevant clinical information comprehensively, especially for research on cancer and incidental findings. The hierarchical normalization schema for anatomy was developed with input from the radiologist, leveraging widely used SNOMED-CT concepts [36].

Annotation was performed in rounds, with two pairs of medical students initially double-annotating 357 reports to reach consistent inter-annotator agreement (IAA) after five rounds of double annotation. The remaining 252 reports were then single-annotated to expedite the process. Domain experts adjudicated disagreements, and the guidelines were updated as necessary. CAMIR includes training, validation, and test splits (70%:10%:20%), with 41% of the training set doubly annotated, and the validation and test sets fully doubly annotated for reliable evaluation. To assess consistency, we compared the average frequency of labels between singly and doubly annotated reports. The doubly annotated reports had an average of 2.65±0.48 *Indication*, 10.15±1.31 *Medical Problem*, and 9.77±0.99 *Lesion* triggers per report, while the singly annotated reports averaged 2.14±0.26 *Indication*, 9.91±2.58 *Medical Problem*, and 8.78±1.06 *Lesion* triggers. Although singly annotated reports showed slightly lower trigger frequencies, the doubly annotated test set ensures a robust evaluation, capturing any potential annotation noise from singly annotated training examples.

3.1.2 Information Extraction using CAMIR

To extract events from the CAMIR dataset, we employed two advanced BERT-based language models: (1) **mSpERT** [82], an extension of SpERT [39] designed for multi-label entity classification, and (2) **PL-Marker**++ [101], an enhanced version of PL-Marker [148] tailored for the CAMIR schema. For both models, we decomposed events into entities and relations, with each relation consisting of a trigger (head) and an argument (tail).

mSpERT

The mSpERT [82] model builds on SpERT [39] that jointly extracts entities and relations at a span-level. mSpERT introduces additional layers that allow for multi-label span predictions, making it well-suited for handling CAMIR's complex event structures. As shown in Figure 3.2, the architecture includes three main output layers: Entity Type, Entity Subtype, and Relation. The Entity Type classifier (ϕ_e) is a linear layer that uses the sentence representation (e_{CLS}), max-pooled hidden states for each span ($e(s_i)$), and span width embeddings (w_{k+1}) to predict entity types. The Entity Subtype layer, specific to mSpERT, allows multi-label span predictions for each *span-with-value* argument type, leveraging both entity type predictions and span features. Finally, the Relation classifier (ψ_r) predicts relationships by examining pairs of max-pooled spans, their widths, and pooled hidden states between spans, thus generating structured event representations for CAMIR.

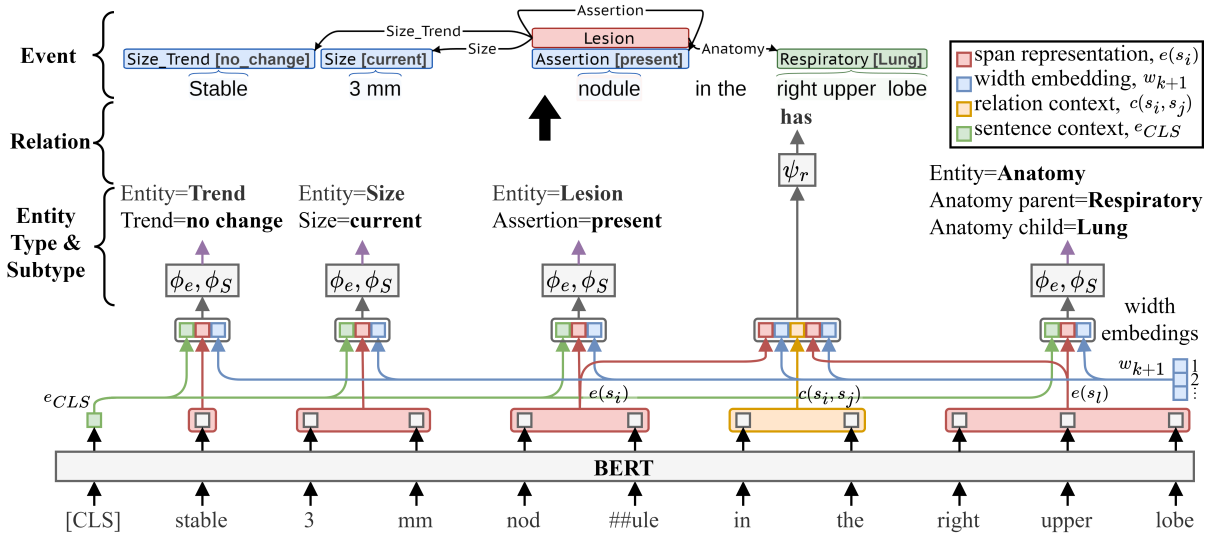

Figure 3.2: mSpERT framework with CAMIR annotation examples

PL-Marker++

PL-Marker [148] is a multi-step extraction model that first identifies entities, then resolves relations. We enhanced this model to create PL-Marker++, adding a third stage specifically for classifying *span-with-value* subtype labels. The modified architecture (illustrated in Figure 3.3) maintains the original PL-Marker structure for Entity Type and Relation stages while introducing a final classification layer. In the Entity Type stage, spans are grouped and processed with a packing strategy that models interdependencies efficiently. The Relation stage uses subject-oriented packing to represent each relation head with its associated tails, capturing dependencies among pairs of spans. In the third stage, subtype classification, each entity is reprocessed within its sentence context using BERT, where typed markers identify the entity in focus. The CLS token from this entity-specific sentence is then used in a multi-label classifier to predict subtype labels for each *span-with-value* argument.

3.2 Evaluation

Model hyperparameters were optimized using the CAMIR training and validation sets, and final performance was evaluated on the held-out CAMIR test set. Evaluation followed the overlap span equivalence criterion, in which two spans are considered equivalent if they share any overlapping tokens. For example, when extracting anatomy mentions (as shown in line 2 of Figure 3.1), the span "right pedicle" would be treated as equivalent to "T12 right pedicle" because of the shared overlap. Trigger spans are defined as equivalent when their event types are identical and their text spans overlap. Span-only arguments are considered equivalent when their argument types match, their argument spans overlap, and the connected triggers are equivalent. Span-with-value arguments follow the same rule, with the additional requirement that subtype labels also match. The overlap span equivalence criterion aligns with the CAMIR annotation schema and extraction task, since most arguments in the dataset are normalized to predefined medical concepts. This criterion is also appropriate for downstream secondary-use applications, where partial matches often remain clinically meaningful and we

conducted extensive error analyses to validate the appropriateness of this criterion in the following sections. Model performance is reported in terms of precision, recall, and F1-score, and statistical significance was assessed using a non-parametric bootstrap test [12].

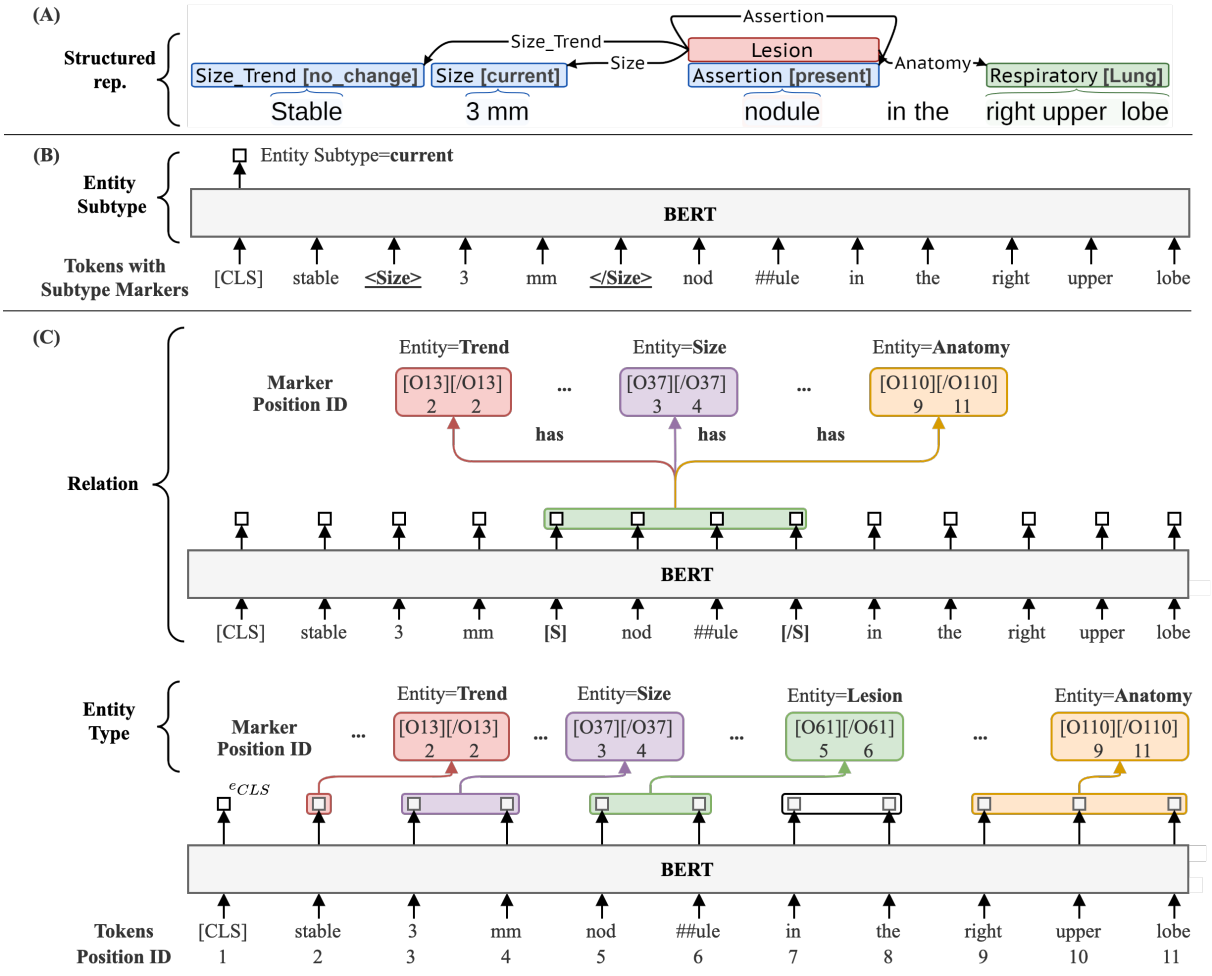

Figure 3.3: PL-Marker++ framework with CAMIR annotation examples

3.3 Results

3.3.1 Annotation Statistics

Table 3.3 shows the distribution of annotated events and arguments within the CAMIR dataset, along with the IAA measured by the F1 score, a widely accepted metric for assessing agreement between annotators in information retrieval tasks [54]. IAA was calculated using an overlap span equivalence criterion, where two spans are considered equivalent if they overlap. For example, in line 2 of Figure 3.1, the anatomy span *"right pedicle"* would be deemed equivalent to "*T12 right pedicle"* due to overlapping spans. For triggers to be equivalent, both the event types and spans must match. Argument types fall into two categories: *span-only* arguments are considered equivalent if their argument types, spans, and associated triggers match; *span-with-value* arguments require an additional match on subtype labels.

Overall, CAMIR achieved an IAA of 0.762 F1 for triggers and arguments in doubly annotated reports. For triggers, the agreement rates were highest at 0.856 F1, 0.805 F1, and 0.854 F1 for *Indication*, *Lesion*, and *Medical Problem*, respectively. Most disagreements involved cases of ambiguity between event types. For instance, the term "disease" could refer to either a *Lesion* or *Medical Problem* depending on the context ("residual disease" vs. "small vessel disease"), while "recurrence" could also be classified as either, depending on the nature of the recurring finding. Argument-specific IAA was generally high, although certain arguments such as *Size Trend*, *Count*, and *Characteristic* achieved lower agreement. *Size Trend* and *Count* were relatively rare in the dataset, making them more challenging for annotators to capture consistently. The *Characteristic* argument, designed as an inclusive category, presented linguistic variability that led to more frequent false negatives.

The distribution of events is consistent across modalities, with 2.4-2.5 *Indication* triggers per report on average, largely attributed to imaging for neoplastic diagnosis. For each report, the average number of *Lesion* and *Medical Problem* events was similar across modalities, with approximately 9.2-9.7 *Lesion* and 10.2-10.4 *Medical Problem* triggers per report. Most *Lesion* and *Medical Problem* triggers were assigned the *Assertion* label *present*.

Table 3.3: Distribution of annotated event types and arguments across imaging modalities in CAMIR. Values in parentheses indicate the average number of triggers per report. Inter-Annotator Agreement (IAA) is reported at the argument type level, calculated using the F1 score.

Event Type	**Argument Type**	**Argument Subtype**	**Modality Type**			**IAA F1**
			CT (n=203)	**MR (n=202)**	**PET-CT (n=204)**	
Indication	Trigger	-	507 (2.510)	496 (2.443)	491 (2.407)	0.856
	Assertion	present	449	435	436	0.820
		absent	11	1	5	
		possible	47	60	50	
	Anatomy	all	276	263	278	0.797
	Indication Type	neoplastic	184	181	193	0.804
		non-neoplastic	112	102	91	
		symptom	149	150	134	
		trauma	23	32	21	
Lesion	Trigger	-	1855 (9.183)	1967 (9.690)	1887 (9.250)	0.805
	Assertion	present	1190	1302	1222	0.762
		absent	547	531	539	
		possible	118	134	126	
	Anatomy	all	2321	2536	2378	0.710
	Size	current	303	364	349	0.715
		past	46	63	36	
	Size Trend	decreasing	26	38	36	0.560
		disappear	22	18	26	
		increasing	35	61	32	
		new	64	58	46	
		no change	109	142	130	
	Count	-	762	841	921	0.564
	Characteristic	-	119	112	132	0.481
Medical Problem	Trigger	-	2063 (10.213)	2111 (10.399)	2080 (10.196)	0.854
	Assertion	present	1217	1294	1189	0.815
		absent	607	592	631	
		possible	239	225	260	
	Anatomy	all	2197	2316	2083	0.761

3.3.2 Information Extraction Performance

Table 3.4 presents extraction performance for the held-out CAMIR test set, evaluated using precision, recall, and F1 scores, with statistical significance determined via a non-parametric bootstrap test [12]. Overall, PL-Marker++ outperformed mSpERT, achieving an F1 score of 0.759 compared to mSpERT's 0.736. While the two models performed similarly on *Indication* and *Medical Problem* triggers and arguments, PL-Marker++ showed significantly better performance on *Lesion* triggers and nearly all argument types, with notable gains of $\Delta 0.05$ F1 for *Characteristic*, *Size*, and *Size Trend* arguments within *Lesion* events. This improvement can likely be attributed to PL-Marker++'s use of trigger and argument location information across all BERT layers, enhancing the model's sensitivity to nuanced clinical information.

To validate the span overlap criterion, we further evaluated PL-Marker++'s trigger extraction performance using a stricter, exact match criterion. Under this criterion, PL-Marker++ achieved F1 scores of 0.749 for *Indication*, 0.681 for *Lesion*, and 0.765 for *Medical Problem* triggers. In total, 279 triggers met the overlap criterion but did not satisfy the exact match criterion. Manual review of these discrepancies confirmed that all predicted triggers still captured the clinically relevant information. A closer examination of these 279 cases revealed that 203 of the predictions were shorter than the reference spans, often omitting modifiers (e.g., reference: "Mild FDG activity" vs. predicted: "FDG activity"; reference: "hypodense lesions" vs. predicted: "lesions"). The remaining 76 predictions were longer than the reference, typically incorporating additional modifiers (e.g., reference: "lesion" vs. predicted: "mass lesion"; reference: "carcinoma" vs. predicted: "renal cell carcinoma"). This analysis underscores PL-Marker++'s effectiveness in identifying key clinical triggers, even when exact span alignment with reference annotations is not achieved.

Table 3.4: Event extraction performance for mSpERT and PL-Marker++ evaluated using overlap criteria on the held-out test set. Higher F1-scores are bolded. † indicates statistical significance ($p < 0.05$).

Event	**Argument**	**Count**	**mSpERT**			**PL-Marker++**		
			P	**R**	**F1**	**P**	**R**	**F1**
Indication	Trigger	285	0.818	0.758	**0.787**	0.878	0.705	0.782
	Assertion	285	0.816	0.730	**0.770**	0.852	0.684	0.759
	Anatomy Parent	157	0.696	0.554	0.617	0.711	0.580	**0.639**
	Anatomy Child	157	0.675	0.529	0.593	0.711	0.580	**0.639**
	Type	262	0.783	0.687	**0.732**	0.782	0.683	0.729
Lesion	Trigger	1169	0.859	0.846	0.853	0.880	0.888	**0.884**†
	Assertion	1169	0.840	0.810	0.825	0.863	0.870	**0.866**†
	Anatomy Parent	1448	0.753	0.620	0.680	0.769	0.673	**0.718**†
	Anatomy Child	1448	0.720	0.586	0.646	0.733	0.642	**0.684**†
	Characteristic	652	0.654	0.420	0.512	0.776	0.477	**0.591**†
	Count	75	0.833	0.800	**0.816**	0.902	0.733	0.809
	Size	294	0.761	0.670	0.713	0.890	0.691	**0.778**†
	Size Trend	206	0.720	0.587	0.647	0.795	0.714	**0.752**†
Medical Problem	Trigger	1271	0.897	0.832	0.863	0.886	0.866	**0.875**
	Assertion	1271	0.878	0.802	0.839	0.854	0.834	**0.844**
	Anatomy Parent	1349	0.792	0.623	**0.697**	0.752	0.633	0.688
	Anatomy Child	1349	0.725	0.563	**0.633**	0.687	0.578	0.628
OVERALL		12847	0.798	0.684	0.736	0.805	0.718	**0.759**†

3.3.3 Large-scale Information Extraction using PL-Marker++

Leveraging PL-Marker++ and its extraction capability of event-based schema used for CAMIR, we tried large-scale information extraction in our clinical database. Among 7 million reports of various modality types, we focused on extracting findings from CT, MR, PT which are modality types included in CAMIR, and CR (Computed Radiography) which was not included in the dataset used to train PL-Marker++.

Table 3.5 summarizes the distribution of identified findings using PL-Marker++. In addition to the extraction of triggers (indication, medical problem, lesion), all other attributes in the CAMIR event schema has been extracted. We post-processed the extracted information in our database to allow efficient secondary-use of these clinical findings.

Extracted information is stored in two tables within the database: 1) Entity table and 2) Relation table (Figure 3.4). Patients with specific clinical findings can be discovered by referring to radiology reports that mention particular clinical findings, by leveraging basic SQL queries to join both tables. Without the need to manually review all radiology reports, clinicians can have quciker access to the radiology reports of their interests.

Table 3.5: Distribution of extracted findings by modality type in the clinical database

Modality	**# Total**	**Indication (Avg)**	**Med Prob (Avg)**	**Lesion (Avg)**
CR	3,222,384	2,974,330 (0.92)	13,424,499 (4.17)	1,207,996 (0.37)
CT	1,313,911	2,430,491 (1.85)	15,855,630 (12.07)	7,184,523 (5.47)
MR	486,988	1,012,361 (2.08)	5,949,379 (12.22)	2,530,281 (5.20)
PT	38,610	78,504 (2.03)	289,661 (7.50)	769,818 (19.94)
Overall	5,061,893	6,495,686	35,519,169	11,692,618

(A)

Accession Number	ID	Sentence Index	Entity Type	Entity Text	Char Start Index	Char End Index	Token Start Index	Token End Index	SubType Assertion	SubType Anatomy Parent	SubType Anatomy Child	SubType Indication Type
Patient1	Patient1_0_ent0	0	Anatomy	Thoracic	21	29	4	5	NULL	Thoracic	Undetermined	NULL
Patient1	Patient1_0_ent1	0	Indication	pain	30	34	5	6	present	NULL	NULL	symptom
Patient1	Patient1_1_ent0	1	Medical Problem	fractures	212	221	3	4	absent	NULL	NULL	NULL
Patient1	Patient1_4_ent0	4	Characteristic	lytic	273	278	2	3	NULL	NULL	NULL	NULL
Patient1	Patient1_4_ent1	4	Characteristic	sclerotic	282	291	4	5	NULL	NULL	NULL	NULL
Patient1	Patient1_4_ent2	4	Lesion	lesions	292	299	5	6	absent	NULL	NULL	NULL
Patient2	Patient2_0_ent0	0	Anatomy	low back	29	32	5	7	NULL	Body_Regions	Undetermined	NULL
Patient2	Patient2_0_ent1	0	Indication	pain	38	42	7	8	present	NULL	NULL	symptom
Patient2	Patient2_0_ent2	0	Anatomy	right side	57	62	10	12	NULL	Body_Regions	Undetermined	NULL
Patient2	Patient2_0_ent3	0	Indication	pain	68	72	12	13	present	NULL	NULL	symptom
Patient2	Patient2_2_ent0	2	Anatomy	L4 - 5	223	225	5	8	NULL	Neurological	Spine_Lumbar	NULL
Patient2	Patient2_2_ent1	2	Medical Problem	retrolisthesis	228	242	8	9	present	NULL	NULL	NULL

(B)

HeadEntityID	TailEntityID
Patient1_0_ent1	Patient1_0_ent0
Patient1_4_ent2	Patient1_4_ent0
Patient1_4_ent2	Patient1_4_ent1
Patient2_0_ent1	Patient2_0_ent0
Patient2_0_ent3	Patient2_0_ent2
Patient2_2_ent1	Patient2_2_ent0

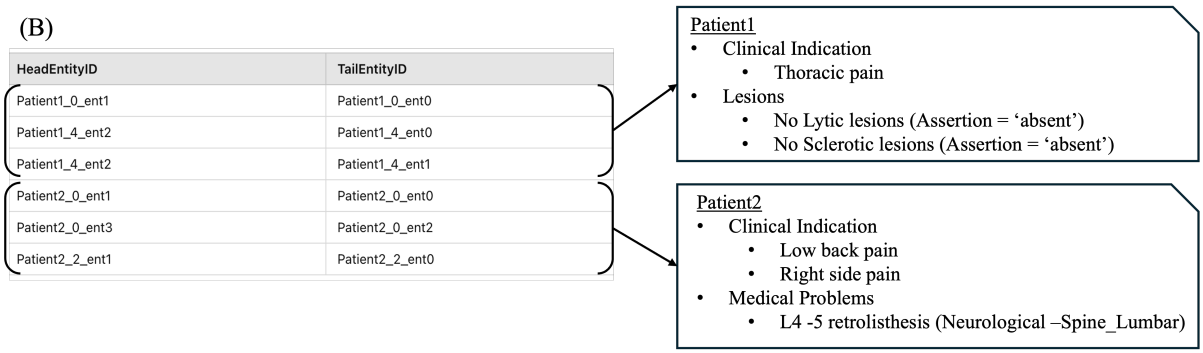

Figure 3.4: (A) Entity table including all attributes from CAMIR event schema. *Span-with-value* arguments are also stored in this table, (B) Relation table where head entities refer to triggers and tail entities refer to all other attributes.

3.4 Conclusion

Our annotation guidelines provide a robust and adaptable framework for identifying clinical findings in radiology reports without depending on institution-specific templates or formatting. This flexibility allows our approach to accommodate variations in report structure across different imaging modalities and healthcare institutions, as we anticipate that the core clinical

findings will be described in ways that align with our guidelines. Although our current focus is on three primary imaging modalities, the annotation schema itself was designed with versatility in mind and is not limited to any specific modality. Consequently, we expect only minimal adjustments, if any, would be needed to apply these guidelines across various institutions or to other imaging types, enabling broader dataset creation.

3.5 Limitations

To the best of our knowledge, CAMIR provides a level of detailed annotation that enhances information extraction models' ability to automatically identify intricate clinical findings. Notably, PL-Marker++, trained on CAMIR, demonstrates performance on par with human annotators in many respects. Despite these promising results, several limitations remain. CAMIR is based on data from a single urban hospital system and focuses exclusively on three imaging modalities. Although the dataset includes over 13,000 clinical events, it is limited to 609 reports, raising questions about the generalizability of the annotated corpus and extraction models to other healthcare settings and additional imaging modalities. Future research will address this by evaluating the effectiveness of larger generative language models (e.g., GPT-5, GPT-4o) in fine-tuning and in-context learning scenarios, potentially expanding the model's adaptability and performance across diverse clinical contexts.

Chapter 4

IDENTIFICATION OF IMAGING FOLLOW-UP IN RADIOLOGY REPORTS

Radiology reports often contain recommendations for follow-up imaging to monitor indeterminate or potentially significant findings. Accurately identifying these follow-up recommendations and determining whether the recommended imaging was performed are essential for closing the communication loop between radiologists, referring clinicians, and patients. Missed or delayed follow-ups can lead to adverse outcomes, including delayed diagnosis of malignancies and preventable disease progression. Despite their clinical importance, systematic tracking of follow-up imaging remains limited in most health systems because radiology reports are typically stored as unstructured text and lack standardized documentation of follow-up recommendations.

This chapter presents a comprehensive framework for the automated identification of imaging follow-up in radiology reports. Building on prior research that applied NLP to extract actionable recommendations, this work extends the problem to a longitudinal, report-level task that evaluates whether a recommended follow-up was subsequently completed. The chapter details the dataset construction process, model development, and comparative evaluation of both traditional machine learning and LLM approaches. It further provides quantitative and qualitative analyses that highlight model strengths, common error patterns, and implications for clinical deployment.

4.1 Methods

4.1.1 Task Formulation and Study Design

This retrospective study was conducted using de-identified radiology reports under Institutional Review Board (IRB) approval, with a waiver of informed consent in accordance with institutional policy. The objective of this work is to automatically identify imaging follow-up events in free-text radiology reports, a task that requires both understanding of clinical context and temporal analysis across multiple patient encounters.

We conceptualize the task as a longitudinal report-pair classification problem, where each sample consists of an *index report* and one or more *candidate reports*. The index report represents the initial imaging study that contains a finding for which a follow-up imaging recommendation was made. These recommendations are often linked to potentially malignant or indeterminate findings and typically specify the target anatomy, imaging modality, and recommended interval (e.g., "follow-up chest CT in 6 months to evaluate stability of the lung nodule") [149, 145]. For each index report, all subsequent imaging reports from the same patient within the radiology information system were retrieved as candidate reports. The system then determines whether any candidate report fulfills the follow-up recommendation from the corresponding index report.

The study design enables evaluation of both sentence-level and report-level reasoning by incorporating temporal proximity, imaging modality, and linguistic similarity between reports. The formulation reflects a clinically relevant workflow where radiologists make recommendations and referring clinicians are responsible for ensuring completion of follow-up imaging. Figure 4.1 illustrates this setup, showing the progression from index to candidate reports in a typical follow-up identification scenario.

This framing also allows the problem to be approached as a supervised binary classification task, where the model predicts whether a given index–candidate pair represents a completed follow-up. The approach facilitates benchmarking across diverse modeling paradigms, from feature-based traditional machine learning to transformer-based and generative large lan-

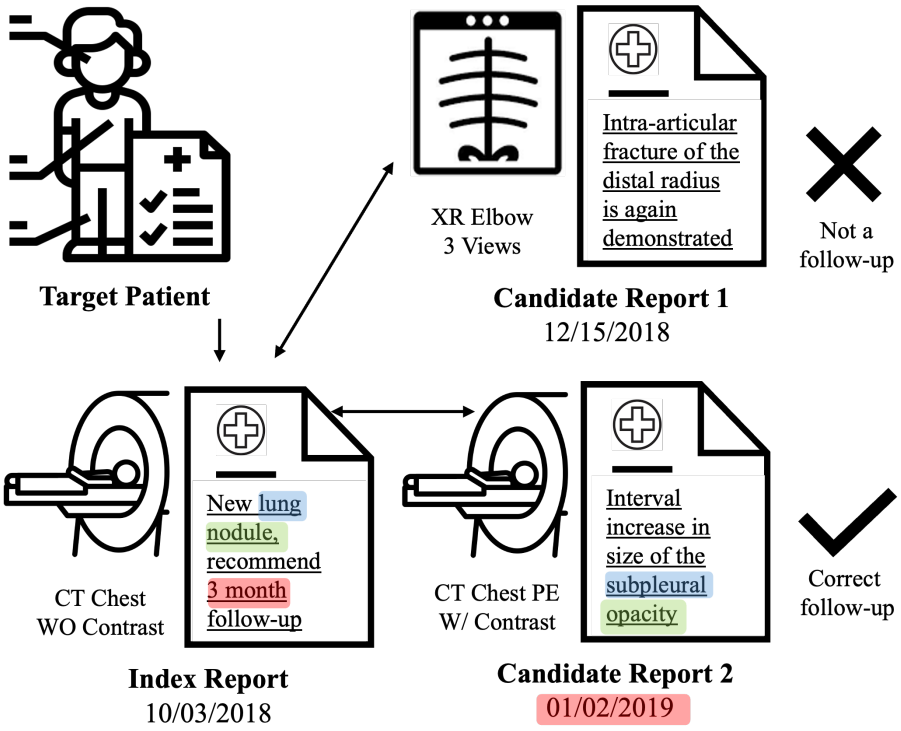

Figure 4.1: Illustration of the imaging follow-up identification task using radiology reports. Highlighted text indicates that Candidate Report 2 is a valid follow-up to the target index report, based on relevant clinical information.

guage models (LLMs), providing a unified structure for comparison and subsequent clinical deployment.

4.1.2 Sampling Process

We used a large clinical radiology database containing approximately seven million reports from 959,382 patients collected between 2007 and 2020. All reports were de-identified using a neural de-identification system prior to analysis [74]. During this period, each patient underwent an average of 7.45 imaging examinations, with nearly 30% of patients having only a single report. To create a representative and well-balanced dataset for the imaging follow-up

identification task, we employed a multistage sampling and filtering process designed to increase the prevalence of reports containing actionable follow-up recommendations.

Index reports were defined as those containing explicit follow-up recommendations related to a radiologic finding. We limited our scope to computed tomography (CT) and magnetic resonance imaging (MRI) modalities, as these are most frequently associated with follow-up recommendations. To rapidly identify candidate index reports, we combined automated filtering with manual verification, as outlined in Figure 4.2. First, we applied a lightweight Support Vector Machine (SVM)-based sentence classifier trained on an existing annotated corpus of recommendation statements [71]. The classifier labeled each sentence as either containing or not containing a follow-up recommendation. It achieved an F1 score of 0.87 at the sentence level and identified 181,020 CT reports and 52,809 MRI reports containing at least one recommendation sentence, corresponding to 12.8% and 9.8% of all reports per

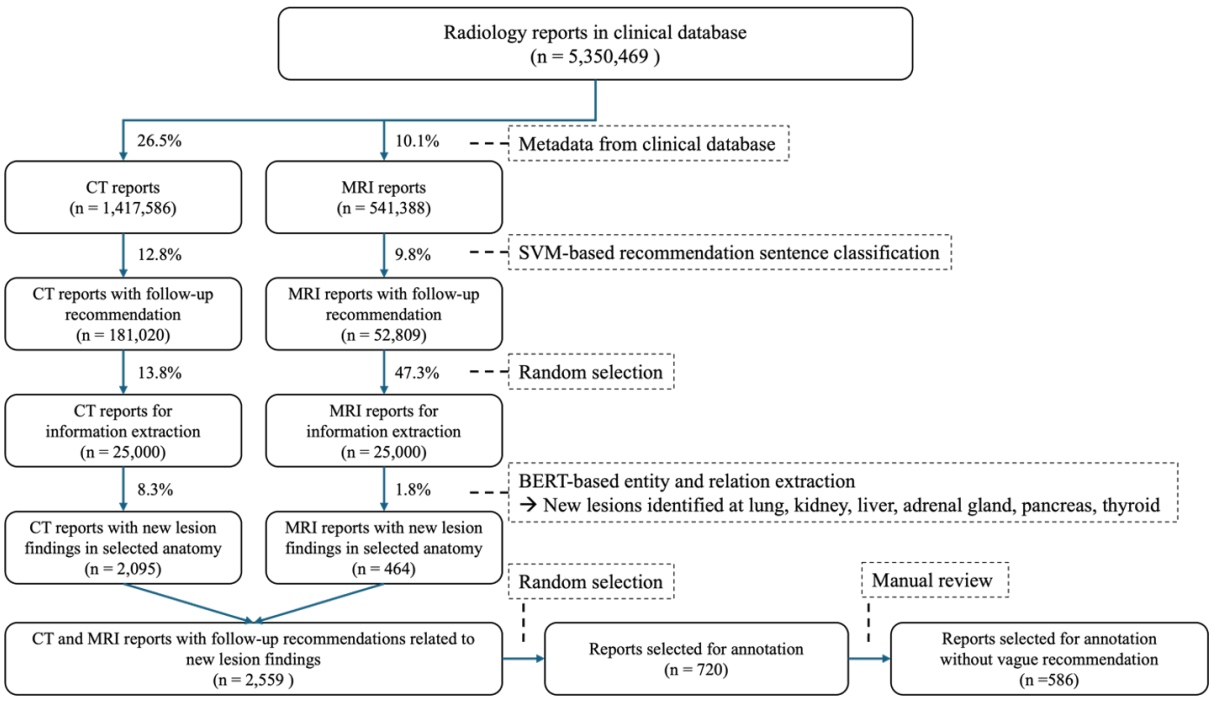

Figure 4.2: Sampling process for index reports

modality, respectively.

From the subset of reports containing follow-up recommendations, we identified those describing a mass lesion, since lesion-based findings are more likely to warrant follow-up imaging. We randomly sampled 25,000 CT and 25,000 MRI reports and applied PL-Marker++, a BERT-based entity and relation extraction model from Chapter 3, to identify clinically relevant entities such as mass lesions, anatomical sites, and their attributes. We specifically targeted newly identified lesions across six anatomies commonly associated with incidental findings: lung, kidney, liver, adrenal gland, pancreas, and thyroid. This filtering step yielded 2,095 CT reports and 464 MRI reports, each containing at least one new lesion mention and one recommendation sentence.

From this refined pool, we randomly selected 720 index reports for manual review. Two trained annotators, supervised by a radiologist, reviewed these reports to remove false positives resulting from automated extraction errors. Reports containing vague or ambiguous recommendations (e.g., "attention on follow-up is recommended" or "clinical correlation is advised") or non-imaging follow-up suggestions were excluded. After this manual refinement, 586 index reports were retained, representing a diverse and well-balanced sample of clinically meaningful recommendation contexts. For each index report, all subsequent radiology reports from the same patient were retrieved to construct patient-level timelines. These subsequent studies were defined as *candidate reports*. Each index report was paired with all its candidate reports, forming index–candidate pairs that served as the primary units for model training and evaluation. The final corpus consisted of 586 index reports and 5,807 candidate reports, totaling 6,393 annotated radiology documents.

4.1.3 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 × 2.4 cm, previously 3.2 × 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 using the BERT tokenizer [34], which converts text into subword tokens.

4.1.4 Modeling Approaches

Prior work in follow-up imaging identification applied discrete machine learning models that combined textual representations and engineered linguistic features to predict whether a candidate report fulfilled a follow-up recommendation [84]. Building on this foundation, we formulated the task as a binary classification problem by constructing index–candidate pairs and labeling each pair as positive when the candidate report represented the correct follow-up for the corresponding index report. Each index–candidate pair therefore served as a unique sample for model training, validation, and testing.

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 [125]. 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 4.3 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.

Supervised Learning using Transformer-based Models

We also investigated supervised learning with Transformer-based models capable of extended inputs, including Longformer [11], BioClinicalModernBERT [123], Llama3-8B-Instruct [1], and Llama3.1-8B-Instruct [49], 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 4.4). 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.

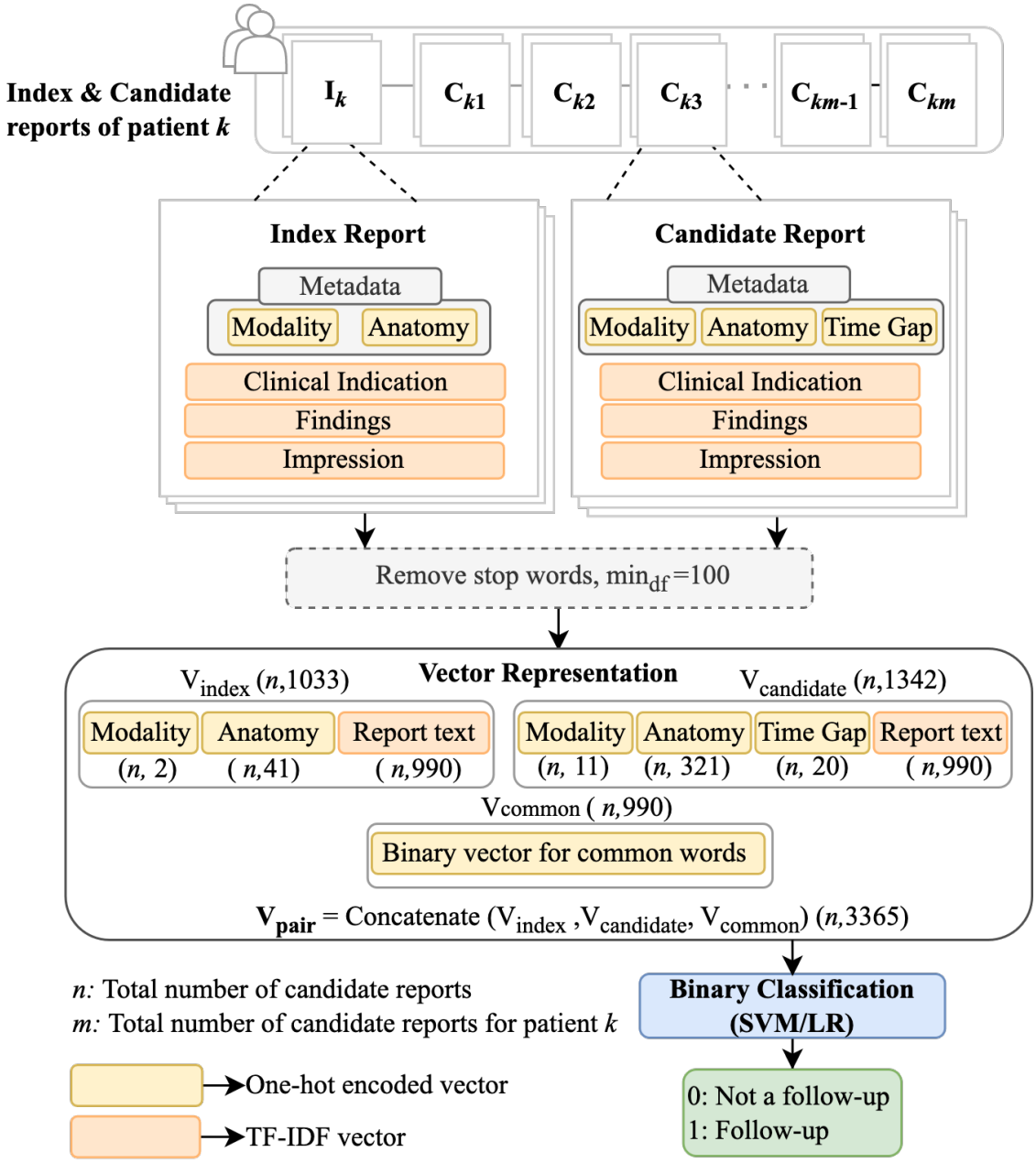

Figure 4.3: Follow-up identification using feature-based traditional models (SVM and LR).

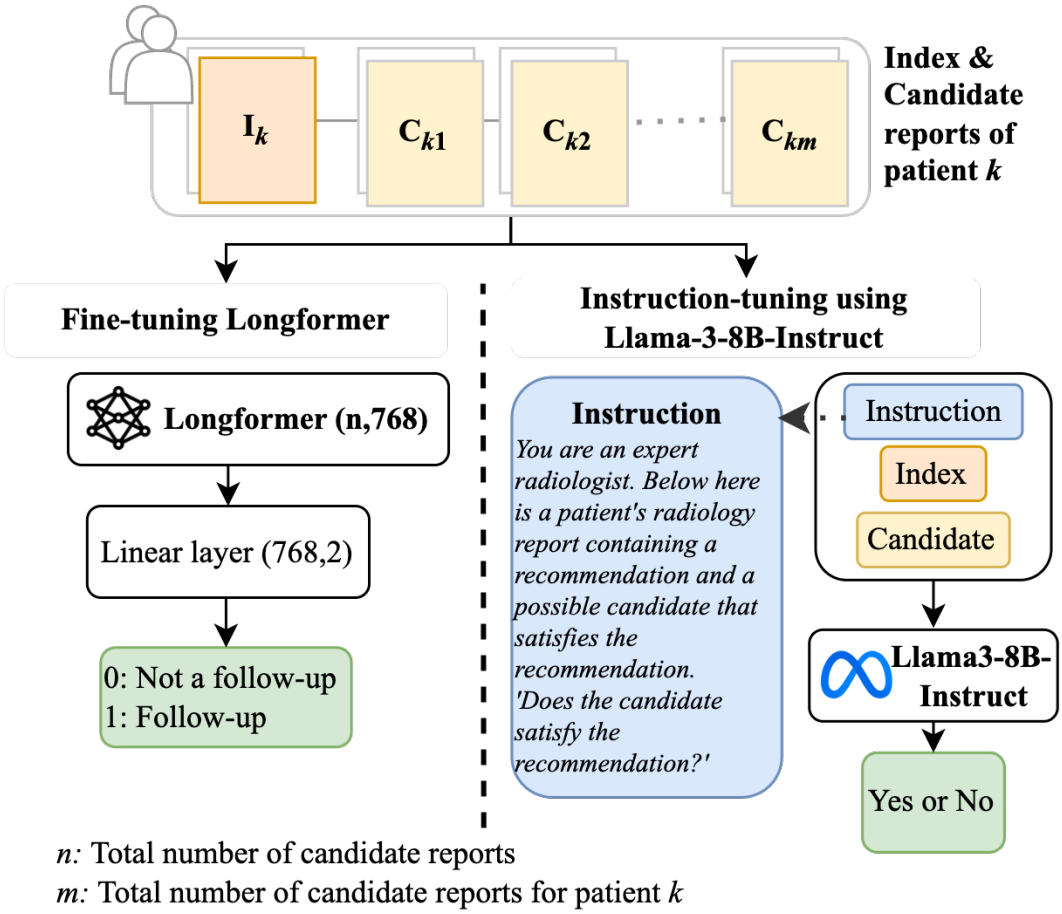

Figure 4.4: Follow-up identification using Transformer-based models – Longformer and Llama3-8B-Instruct

Generative Large Language Model

To evaluate generative LLMs for identifying follow-up imaging studies, we assessed GPT-4o [59] (version 2024-05-13) and GPT-OSS-20B [4] in our HIPAA-compliant environment using two strategies: a baseline setting and an advanced, task-optimized setting (Figure 4.5). 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.

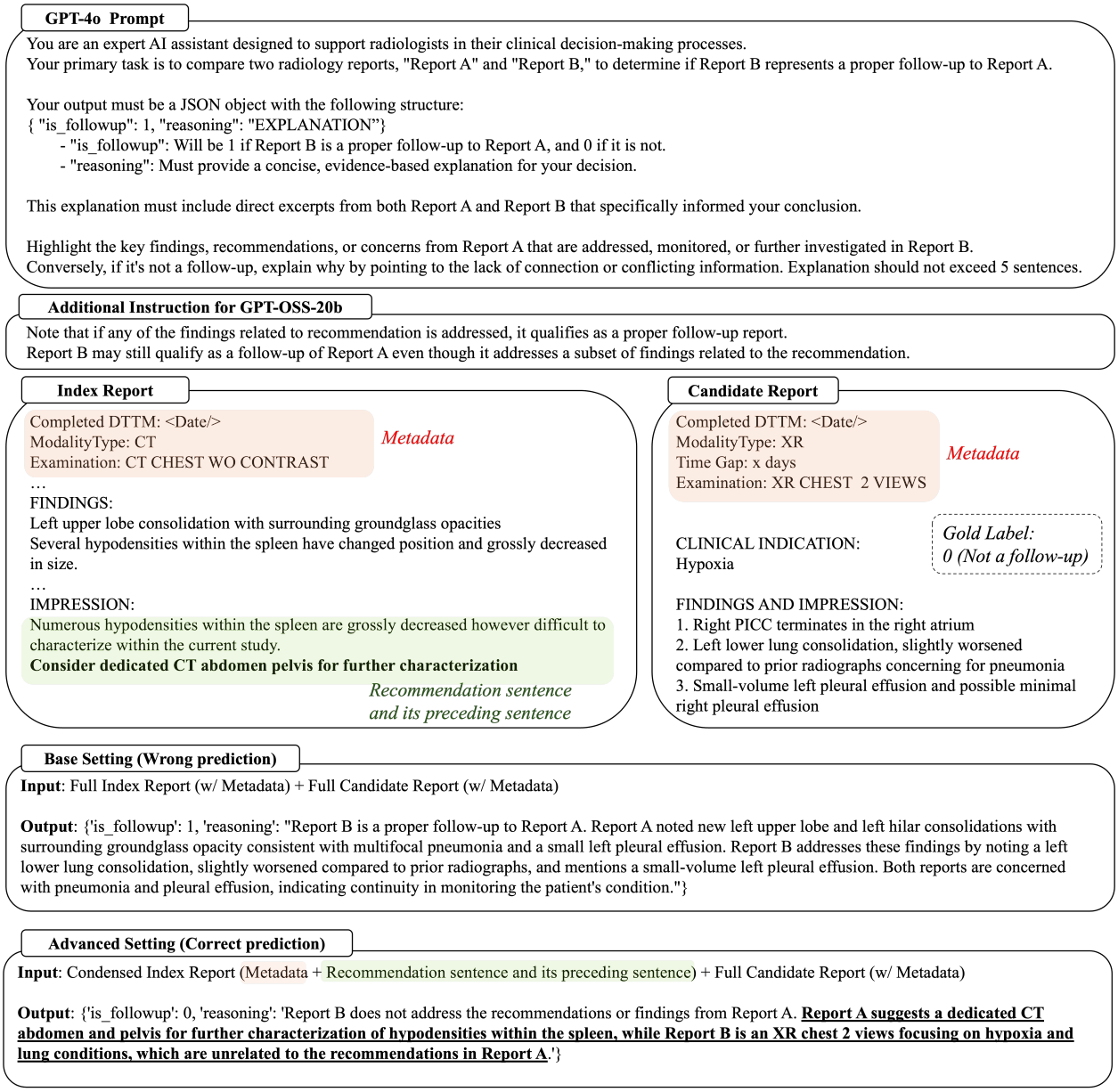

Figure 4.5: Follow-up identification using generative LLMs. GPT-OSS-20B prompt is created by adding 2-sentence instructions on top of GPT-4o prompt. Outputs for base and advanced settings are actual predictions from GPT-4o.

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 4.5), 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 4.5).

4.1.5 Evaluation

We evaluated performance at the index report-level chronologically (Figure 4.6) 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

	Time → Candidate 1	Candidate 2	Candidate 3	Candidate 4	...	Candidate *n*
True Positive (TP) : Correctly predicted the earliest follow-up report						
Case 1. Model identified a follow-up report, which is the correct follow-up report						
Truth	X	X	✓	X	…	X
Prediction	X	X	✓	X	…	X
Case 2. Model identified multiple follow-up reports, the earliest of which is the correct follow-up report						
Truth	X	X	✓	X	…	X
Prediction	X	X	✓	✓	…	X
True Negative (TN) : Correctly predicted that no follow-up reports exist for the target source report						
Truth	X	X	X	X	…	X
Prediction	X	X	X	X	…	X
False Positive (FP) : Incorrectly predicted the existence of one or more follow-up reports when none actually existed						
Truth	X	X	X	X	…	X
Prediction	✓	X	✓	X	…	X
False Negative (FN) : Failed to identify the earliest follow-up report, despite its existence						
Case 1. Model did not identify any follow-up reports						
Truth	X	X	✓	X	…	X
Prediction	X	X	X	X	…	X
Case 2. Model identified one or more follow-up reports. Failed to choose the correct follow-up report						
Truth	X	X	✓	X	…	X
Prediction	X	X	X	✓	…	✓
Case 3. Model identified one or more follow-up reports, including the correct follow-up report. However, the earliest candidate report selected by the model is not the true follow-up report.						
Truth	X	X	✓	X	…	X
Prediction	X	✓	✓	X	…	X

Figure 4.6: 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

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.

4.2 Results

4.2.1 Model Performance and Significance Testing

Table 4.1 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 4.2) 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 achieved the highest precision among all systems (0.907, 95% CI: 0.872–0.943), yet its recall remained substantially

Table 4.1: 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)

Table 4.2: 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

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.

4.2.2 Error Analysis

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 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 4.5 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.

4.3 Conclusion

In this chapter, 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.

4.4 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 [113], MedVLM-R1 [100]), represent promising avenues for advancing performance and clinical relevance.

Finally, 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.

Chapter 5

AUTOMATIC IDENTIFICATION OF INCIDENTALOMAS IN RADIOLOGY REPORTS

This chapter presents the development and evaluation of anatomy-aware LLM frameworks for identifying incidental findings, or *incidentalomas*, within radiology reports. Incidentalomas are unexpected lesions discovered during imaging performed for reasons unrelated to the detected abnormality [13]. Their increasing frequency with modern imaging utilization has created new challenges in ensuring appropriate follow-up [80]. Traditional NLP methods demonstrated the feasibility of automating incidental finding extraction [18, 84, 27], yet these systems were limited by handcrafted rules and poor contextual generalization.

This chapter integrates three core advances. First, we construct a lesion-level annotated dataset covering six key anatomical regions associated with incidental findings: kidney, liver, lung, pancreas, adrenal gland, and thyroid. Second, we design a prompting framework that introduces lesion-tagged and anatomy-informed contextual cues, enabling generative LLMs to reason in ways that align with radiologic interpretation. Third, we compare these LLM systems against strong supervised transformer baselines to evaluate accuracy, interpretability, and generalizability. The overall goal is to advance scalable, clinically consistent methods for incidentaloma identification, supporting automated surveillance and improving continuity of care.

5.1 Method

5.1.1 Dataset

Preprocessing

We utilized an existing clinical database comprising 6,668,323 radiology reports collected between 2007 and 2020, representing the general patient population across four hospitals within the University of Washington (UW) Medicine system. All reports were automatically de-identified using a neural de-identification system [74], ensuring full compliance with institutional and HIPAA standards. Structured information extraction was then performed using *PL-Marker++* [101], a radiology-specific BERT-based model optimized for lesion-level entity and relation extraction described in Chapter 3.

PL-Marker++ extracts three main categories of information: clinical indications, lesion findings, and medical problems. Each category is accompanied by argument-level attributes such as anatomy, size, and temporal size trend. Clinical indications describe the reason for the imaging study (e.g., *"evaluation of lung nodule," "follow-up for prior mass,"* or *"abdominal pain"*) and are automatically categorized into four subtypes: *neoplastic diagnosis*, *non-neoplastic diagnosis*, *symptom*, and *trauma*. Lesion findings and medical problems represent observed abnormalities (e.g., *"hepatic lesion"* or *"adrenal nodule"*), while descriptive attributes capture site-specific details such as size, count, and trend.

The size trend attribute includes five standardized values: *increasing*, *decreasing*, *no change*, *disappeared*, and *new*. Anatomical information associated with each lesion, medical problem, and clinical indication is extracted using entity–relation linking and mapped to a controlled vocabulary of predefined anatomy categories such as *lung*, *liver*, *kidney*, *pancreas*, *adrenal gland*, and *thyroid*. Table 5.1 summarizes the token-level lesion identification performance on the annotated dataset described in Chapter 3. Results show that *PL-Marker++* significantly outperforms few-shot LLM approaches using Llama 3.1-8B and GPT-4o. These findings are consistent with recent studies [57, 79] indicating that decoder-based LLMs underperform on

token-level clinical named entity recognition tasks due to their generative architecture.

The structured outputs produced by *PL-Marker++*, particularly the integration of categorized clinical indications with lesion-level findings and anatomical context, formed the foundation for this study. These outputs enabled systematic identification of reports describing new or potentially actionable lesions and provided rich contextual information for downstream LLM inference. The full study protocol was reviewed and approved by the University of Washington Institutional Review Board (IRB).

Table 5.1: Lesion identification performance using multiple approaches on annotated radiology reports (Park et al., 2024).

Method	Overlap Match			Exact Match		
	Precision	Recall	F1-score	Precision	Recall	F1-score
PL-Marker++	0.880	0.888	**0.884**	0.740	0.712	**0.726**
Llama 3.1-8B (0-shot)	0.377	0.322	0.348	0.014	0.012	0.013
Llama 3.1-8B (1-shot)	0.449	0.387	0.415	0.064	0.055	0.059
GPT-4o (0-shot)	0.599	0.374	0.460	0.170	0.106	0.131
GPT-4o (1-shot)	0.576	0.489	0.529	0.192	0.163	0.177

Sampling Strategy

Given the low prevalence of reports containing incidentalomas within the clinical corpus (for example, 9.9% for adrenal incidentalomas [7] and less than 5% for chest computed tomography identifying incidental pulmonary embolisms [98]), direct identification of such cases was challenging. Searching for explicit terms such as *"incidental"* or *"incidentaloma"* would have introduced lexical bias and missed many true incidentalomas, since radiologists often describe these findings without using those words. To overcome this limitation, we designed a broader, data-driven sampling strategy that combined structured information

extracted from *PL-Marker++* with a support vector machine (SVM)–based classifier for recommendation sentence detection. This approach ensured a more comprehensive and representative coverage of incidentaloma cases across anatomies and modalities.

Following iterative consultations with a board-certified radiologist, we established a multi-stage sampling pipeline optimized for high precision in identifying reports likely to contain incidental findings. The sampling process comprised four sequential steps:

(1) Using structured clinical findings extracted with *PL-Marker++*, we analyzed all radiology reports in the database and identified findings mapped to six anatomical regions: kidney, liver, lung, pancreas, adrenal gland, and thyroid. These organs were selected because they have the highest likelihood of harboring incidentalomas based on prior literature and clinical experience. Among all reports, 24.7% (n = 1,767,623) contained at least one finding (clinical indication, medical problem, or lesion) located in these regions, yielding a total of 7,519,138 findings. Reports across all imaging modalities were then selected if they included lesion findings with assertion values labeled as *"present"* or *"possible"*. A total of 6.3% of all reports (n = 112,100) met these inclusion criteria and were retained for downstream processing.

(2) Using extracted size-trend information, we excluded reports that contained findings with trend values of *"increasing"*, *"decreasing"*, *"disappeared"*, or *"no change"*. These attributes typically indicate a prior imaging comparison and therefore represent lesions already under surveillance rather than newly detected incidental findings. This exclusion removed 5.7% of the candidate reports, resulting in a subset of 105,729 reports for the next phase.

(3) The *Clinical Indication* section of each report provides insight into the rationale for imaging and the patient's underlying condition. As described in Section 5.1.1, *PL-Marker++* categorizes clinical indications into four subtypes: *trauma*, *symptom*, *neoplastic diagnosis*, and *non-neoplastic diagnosis*. Reports containing a neoplastic diagnosis were

excluded to minimize inclusion of findings that were likely intentional rather than incidental. After this filtering, 39.6% of reports (n = 41,833) remained for further review.

(4) Finally, among the 41,833 reports retained, we identified those that contained follow-up recommendation sentences, since incidental findings are frequently associated with such recommendations. Recommendation sentences were detected using an SVM-based classifier developed by Lau et al. [71]. This step yielded 19,690 reports with the highest estimated probability of containing an incidentaloma.

To ensure accurate anatomical labeling for each lesion and to support lesion-level and document-level analyses described in later sections, we implemented a dual verification process for anatomy extraction. *PL-Marker++* initially provided sentence-level lesion–anatomy mappings, but its performance was occasionally limited when relevant anatomical mentions appeared outside the immediate sentence context, such as in section headers or prior narrative segments. To address this limitation, we introduced a large language model, Llama 3.1-8B Instruct [49], trained to infer anatomical associations from the entire report context.

When evaluated on the CAMIR dataset described in Chapter 3, the Llama model achieved a micro-F1 score of 0.895. Discrepancies between the two systems were manually reviewed, and a final verified anatomy label was assigned to each lesion. This verified anatomical information was subsequently used in all downstream analyses, ensuring consistent lesion-to-organ mapping across the six target anatomies (kidney, liver, lung, pancreas, adrenal gland, and thyroid). The dual-level verification process provided a robust foundation for both model evaluation and cross-anatomy performance comparisons presented in subsequent sections.

Data Annotation

A total of 19,690 radiology reports met the selection criteria described in the previous section, from which 400 reports were randomly selected for manual annotation. The annotation guidelines were developed through multiple iterations by two board-certified radiologists to

ensure both precision and clinical relevance. Rather than relying solely on document-level labeling for incidentaloma presence, the annotation protocol was structured around detailed lesion-level findings that better reflect clinical interpretation. Each lesion was evaluated to determine whether it represented a previously unidentified incidental finding and was assigned one of three labels: *No Incidentaloma*, *Incidentaloma–No Risk*, or *Incidentaloma–Follow-up Required*. The distinction between the latter two categories was based on clinical severity and management implications; for example, clearly benign findings such as *"simple renal cyst"* were classified as not requiring follow-up.

To facilitate efficient and consistent annotation, lesion mentions automatically extracted using *PL-Marker++* [101] were preloaded into the BRAT annotation tool [127]. This pre-annotation step reduced the likelihood of missed lesion mentions and allowed annotators to focus on semantic and contextual labeling rather than text span identification. Differential diagnoses and speculative statements were excluded from the scope of annotation to avoid ambiguity.

Two medical residents participated as primary annotators. Initial training was conducted through several double-annotation rounds, during which both annotators independently labeled the same reports to gain familiarity with the annotation schema and the BRAT interface. Each week, annotators completed a shared batch of reports, followed by feedback sessions led by a board-certified radiologist. During these sessions, disagreements were systematically reviewed, and guideline refinements were made through consensus. Once stable inter-annotator agreement was achieved (document-level F1 = 0.896 for incidentaloma status), the process transitioned to single-annotation rounds, with each annotator assigned distinct report sets.

To ensure gold-standard quality, all remaining disagreements and ambiguous cases were reviewed by the supervising radiologist who authored the guidelines. This final adjudication ensured consistency and eliminated residual discrepancies prior to dataset release.

In total, 160 double-annotated reports contained 1,623 pre-identified lesion findings, averaging 10.15 lesions per report. Among these, 117 reports (73.1%) included at least one

incidentaloma, confirming that the sampling approach successfully enriched the dataset for incidental findings. Annotators reached agreement on 1,498 of 1,623 lesion labels (92.3% agreement; 0.838 F1). Commonly annotated incidentaloma terms included *lesion*, *nodule*, *cyst*, and *mass*. Annotator 1 identified an average of 2.54 incidentalomas per report, slightly higher than Annotator 2's average of 2.29.

After completion of the double-annotation phase, both annotators proceeded with single-annotation rounds for the remaining 240 reports. Table 5.2 summarizes the document-level distribution of reports with and without incidentalomas, while Table 5.3 presents the anatomy-specific distribution of all annotated lesion findings across the corpus.

Table 5.2: Distribution of annotated reports with and without incidentalomas.

	Total (n)	With Incidentaloma	Without Incidentaloma
Double-annotated	160	117 (73.1%)	43 (26.9%)
Single-annotated	240	158 (65.8%)	82 (34.2%)
Total	400	275 (69.3%)	125 (30.8%)

5.1.2 Incidentaloma Classification Task

Using the annotated dataset described in the previous section, we formulated the incidentaloma classification task to support both lesion-level and document-level analyses. This framework established a unified and clinically grounded foundation for large-scale incidentaloma identification across radiology reports. Building upon the verified anatomical mappings described in Section 5.1.1, each lesion was associated with its corresponding anatomical site, enabling multi-anatomy classification within a single report.

We defined the task as a series of seven anatomy-specific, three-way classification problems, corresponding to the following categories: *Lung, Liver, Kidney, Adrenal, Pancreas,*

Table 5.3: Number of incidentalomas across six target anatomies in our dataset of 400 radiology reports. A single report may include multiple incidentalomas in different anatomical sites. Percentages in parentheses indicate the proportion of low-risk incidentalomas and those requiring follow-up within each anatomy.

	No Incidentaloma	Incidentaloma	
		No Risk	Follow-up Required
Lung	324	18 (4.5%)	58 (14.5%)
Liver	302	78 (19.5%)	20 (5.0%)
Kidney	238	128 (32.0%)	34 (8.5%)
Adrenal	379	6 (1.5%)	15 (3.8%)
Pancreas	384	5 (1.2%)	11 (2.8%)
Thyroid	378	14 (3.5%)	8 (2.0%)
Others	341	25 (6.3%)	34 (8.5%)
Total	2346	274	180

Thyroid, and Other. For each anatomy, the model generated one label from the set $\{0 =$ *No Incidentaloma*$, 1 =$ *Incidentaloma–No Risk*$, 2 =$ *Incidentaloma–Follow-up Required*$\}$. This formulation provided structured, interpretable outputs while enabling both fine-grained lesion-level evaluation and holistic report-level aggregation.

During inference, each model first produced lesion-level predictions for every anatomical site. These lesion-level predictions improved interpretability and supported post-hoc error analysis by highlighting lesion-specific contexts. The lesion-level outputs were then aggregated into a single anatomy-level label using a severity precedence rule. Formally, for a given anatomy a with lesion labels $\{l_1, l_2, \ldots, l_n\}$, the aggregated label L_a was defined as:

$$L_a = \max\{l_1, l_2, \ldots, l_n\}, \quad l_i \in \{0, 1, 2\},$$

where $0 =$ *No Incidentaloma*, $1 =$ *Incidentaloma–No Risk*, and $2 =$ *Incidentaloma–Follow-*

up Required. For example, if three lung lesions were labeled $\{0, 0, 2\}$, the aggregated anatomy label for lung would be $L_{\text{lung}} = 2$.

Each radiology report was thus represented as a structured vector of seven anatomy-specific incidentaloma labels:

$$R = (L_{\text{lung}}, L_{\text{liver}}, L_{\text{kidney}}, L_{\text{adrenal}}, L_{\text{pancreas}}, L_{\text{thyroid}}, L_{\text{other}}), \quad L_a \in \{0, 1, 2\}.$$

This structured schema enabled simultaneous lesion-level and anatomy-level interpretation, facilitating both fine-grained clinical analysis and report-level classification.

Inter-annotator agreement (IAA) for the document-level labels, computed without anatomy stratification, demonstrated strong consistency: 0.93 F1 for *No Incidentaloma*, 0.81 F1 for *Incidentaloma–No Risk*, and 0.70 F1 for *Incidentaloma–Follow-up Required*, yielding a macro-average F1 of 0.81. These agreement scores confirmed the reliability of the schema and annotation process described earlier. The resulting dataset served as the foundation for all subsequent model training, evaluation, and analysis.

Summary statistics for the annotation schema are presented in Table 5.3. As shown, the distribution of incidentalomas varies substantially across anatomical regions, reflecting the heterogeneous prevalence of incidental findings in clinical imaging. For example, the kidney and liver exhibited higher proportions of incidentalomas, whereas the pancreas and adrenal glands were less frequently involved. This variability underscores the clinical diversity of incidentaloma detection and motivates the need for anatomy-specific modeling approaches.

Supervised Learning Approach

We implemented a supervised learning framework for the classification of incidentalomas using three transformer-based encoder models: BioClinicalModernBERT [123], ModernBERT [139], and Clinical Longformer [75]. Each model was fine-tuned on the annotated dataset, in which lesion-level and anatomy-specific labels were assigned to radiology reports. This approach allowed us to compare model architectures that differ in domain specialization,

pre-training scope, and maximum sequence length capacity, providing insight into how these factors influence incidentaloma classification performance.

BioClinicalModernBERT [123] represents the most recent BERT-based model tailored for biomedical and clinical applications. It integrates an extended vocabulary and long-context optimization, enabling accurate modeling of specialized clinical terminology. ModernBERT [139], which is pre-trained on both general-domain and clinical text, serves as a balanced baseline for evaluating the effects of domain adaptation. Clinical Longformer [75] extends the traditional transformer architecture with a sparse attention mechanism, allowing efficient processing of long radiology reports that often exceed the 512-token limit of standard BERT models. Together, these models capture a diverse range of linguistic and contextual modeling strategies, providing complementary perspectives on supervised learning for radiology NLP.

Radiology reports were preprocessed following the standardized pipeline described in the previous section. The input text preserved the original narrative structure while retaining lesion-specific details essential for classification. Key lesion mentions, including terms such as *"nodule"*, *"cyst"*, and *"mass"*, were preserved, and their anatomical context (*Kidney, Liver, Lung, Pancreas, Adrenal, Thyroid, or Other*) was appended as categorical tokens to provide localized contextual information. The classification labels followed the schema introduced earlier (*No Incidentaloma*, *Incidentaloma–No Risk*, *Incidentaloma–Follow-up Required*). This input representation enabled the models to capture lesion-level semantics while maintaining the broader clinical narrative of each report.

Each model was trained with its respective maximum token capacity: 512 tokens for BioClinicalModernBERT and ModernBERT, and 2,048 tokens for Clinical Longformer. Tokenized sequences were padded to uniform lengths, and the resulting embeddings represented three main components: anatomical tokens, lesion mentions, and surrounding contextual text. These combined features provided both local and global cues for accurate incidentaloma detection.

Hyperparameters were optimized through systematic grid search on the validation set. We tested learning rates of $\{1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}, 5 \times 10^{-5}\}$, batch sizes of $\{8, 16, 32\}$,

and training epochs of $\{5, 10, 15\}$. Weight decay values of $\{0.01, 0.05\}$ and dropout rates of $\{0.1, 0.2\}$ were also explored. The final configuration, selected for the highest validation macro-F1 across incidentaloma-positive classes, used a learning rate of 2×10^{-5}, batch size of 16, weight decay of 0.01, and dropout of 0.1. Training was conducted for up to 10 epochs with early stopping (patience = 3) based on validation performance. Optimization employed the AdamW algorithm [78] with gradient clipping (maximum norm = 1.0) [153], linear learning rate warmup over 10% of total steps, and mixed-precision (FP16) training on NVIDIA A100 GPUs to enhance computational efficiency and numerical stability.

Given the inherent class imbalance in the dataset (Table 5.3), where the *Incidentaloma–Follow-up Required* class was substantially underrepresented, we applied cost-sensitive learning strategies to prioritize clinically meaningful predictions. Three complementary approaches were implemented. First, we applied class-weighted cross-entropy, assigning weights inversely proportional to class frequency to improve recall on minority classes. Second, we employed focal loss with $\gamma = 2$ and class-specific weighting $\alpha = w_c$, which emphasized hard-to-classify and underrepresented examples. Third, we introduced an expected-cost (EC) objective based on a 3×3 asymmetric cost matrix C, constructed to penalize clinically critical misclassifications such as failure to identify lesions that required follow-up. The model minimized the probability-weighted expected misclassification cost using the following objective:

$$\mathcal{L}_{\text{EC}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{2} C_{y_i,k}\, p_\theta(k \mid x_i),$$

where $p_\theta(k \mid x_i)$ denotes the softmax probability of class k given the input x_i. During inference, a cost-aware decision rule was applied by selecting the label that minimized expected misclassification cost under the predicted probability distribution. This ensured that predictions reflected the relative clinical importance of errors rather than uniform statistical weighting.

Through these modeling and training strategies, the supervised encoders were optimized to capture nuanced contextual information about incidentalomas, balancing interpretability and clinical reliability. This framework served as the baseline for comparison against generative

LLMs described in the subsequent section.

LLM-based Approach

Similar to the supervised learning framework, the LLM–based approach also generated lesion-level predictions that were subsequently aggregated into document-level labels. Each model received structured lesion-tagged radiology reports as input, allowing it to explicitly identify and reason about individual lesions. Prior to inference, candidate lesion entity spans in the original report were enclosed in XML-style tags (`<LESION1>...</LESION1>`) using lesion information extracted by *PL-Marker++*, while the surrounding text remained unchanged.

The inclusion of numbered lesion tags was found to be critical for accurate incidentaloma classification. Lesion tagging supports interpretability, enables detailed error analysis, and helps minimize the influence of irrelevant findings. Without these structural cues, the model can easily misattribute terms such as *"nodule"* or *"mass"* to unrelated findings, especially in reports containing multiple anatomical regions. Tagging highlights the true lesion spans, providing unambiguous contextual references and facilitating efficient clinical review. This approach also prevents the model from overinterpreting unrelated portions of the report, thereby enhancing reliability and clinical interpretability.

An ablation study using the LLaMA 3.1–8B Instruct model confirmed the utility of lesion tags. Input sequences containing lesion tags achieved a higher macro F1 score than all untagged configurations, including 0-shot, 1-shot, and 5-shot settings. This improvement was primarily driven by increased precision while maintaining comparable recall, indicating that the tags helped the model focus on the correct lesion spans. Structural cues thus provided greater benefit than few-shot demonstrations. As a result, all LLM experiments in this study used lesion-tagged inputs. To minimize stochastic variation during inference, generation parameters were fixed at temperature = 0 and top-p = 1.

Two primary prompting configurations were evaluated.

1. **Base Prompt:** The lesion-tagged report was provided directly to the model without

additional metadata.

2. **With-Anatomy Prompt:** The lesion-tagged report was supplemented with a concise mapping that linked each lesion tag to its corresponding anatomical site, extracted by *PL-Marker++*. The mapping was appended as a single line of text (e.g., `LESION1=Thyroid; LESION2=Pancreas; LESION3=Adrenal; ...`), reducing ambiguity about anatomy assignment while preserving the full contextual narrative.

Both prompts used the same task instruction shown in Figure 5.1 and an example of the With-Anatomy prompt and model output is illustrated in Figure 5.2. The figure demonstrates how numbered lesion tags and anatomical pinning guide the model's reasoning during inference. As shown in the reasoning trace, the model correctly attends to the target lesion descriptions while ignoring irrelevant findings, highlighting the interpretability benefits of structured input design.

We evaluated four generative LLM frameworks under identical inference and prompting conditions. The first was a LoRA-fine-tuned [56] version of LLaMA 3.1-8B [49], adapted using our annotated radiology reports to improve domain alignment while maintaining efficient parameter updates. The second approach applied the same LLaMA 3.1-8B model in a zero-shot setting, using prompt engineering alone without any gradient updates, and was evaluated under both Base and With-Anatomy prompts. The third model, GPT-4o, served as a closed-source benchmark representing state-of-the-art proprietary LLM performance. Finally, we evaluated GPT-OSS [4], an open-source GPT-style architecture comparable to GPT-4 in scale and reasoning ability. To balance inference efficiency and computational cost, we used GPT-OSS-20B rather than the larger 120B variant. All experiments were conducted within a HIPAA-compliant secure computing environment.

This experimental framework enabled systematic comparison of generative models across two dimensions: the effect of optimal prompting (Base vs. With-Anatomy) and the influence of architectural design (open-source versus proprietary systems).

Prompt for Incidentaloma Identification

```
Role: You are a board-certified radiologist.
Task: Analyze each report to verify incidentaloma status in target anatomies.
The list of lesions to consider is provided in the input using
<LESION> </LESION> tags.

Exclusions:
Do not classify a lesion as an incidentaloma if:
- It is suspected or potential metastasis when the scan indication includes a known
  primary malignancy.
- It is found on surveillance scans (e.g., cirrhosis patients undergoing
repeated liver imaging for HCC or patients with known malignancy undergoing
  routine scans for metastases).
- It has been previously identified in a prior study.
- Its size change is mentioned ("stable", "decreased", "increased", "unchanged", etc.)
- Its clinical indication is related to the target lesion.

Example output:
{ 'Lung Inci': {},
 'Liver Inci': {"LESION2":1},
 'Kidney Inci': {},
 'Adrenal Inci': {},
 'Pancreas Inci': {},
 'Thyroid Inci': {"LESION4":2},
 'Other Inci': {} }

Empty dict: No incidentalomas
Category 1: Incidentalomas not requiring follow-up.
Category 2: Incidentalomas requiring follow-up.

Provide brief reasoning (<5 sentences) after JSON output.
```

Figure 5.1: Prompt used for verifying incidentaloma status across target anatomies. Exclusion criteria and examples were derived from the annotation guidelines.

5.1.3 Evaluation

The held-out test set comprised 80 double-annotated reports, corresponding to 560 anatomy-specific labels (seven anatomies per report). Among these labels, 50 were categorized as *Incidentaloma–No Risk* (1) and 29 as *Incidentaloma–Follow-up Required* (2), while the remaining labels represented *No Incidentaloma* (0). The training set ($n = 280$) and validation set ($n = 40$) were used exclusively for model optimization in the supervised learning experiments and for iterative prompt refinement in the LLM-based approaches.

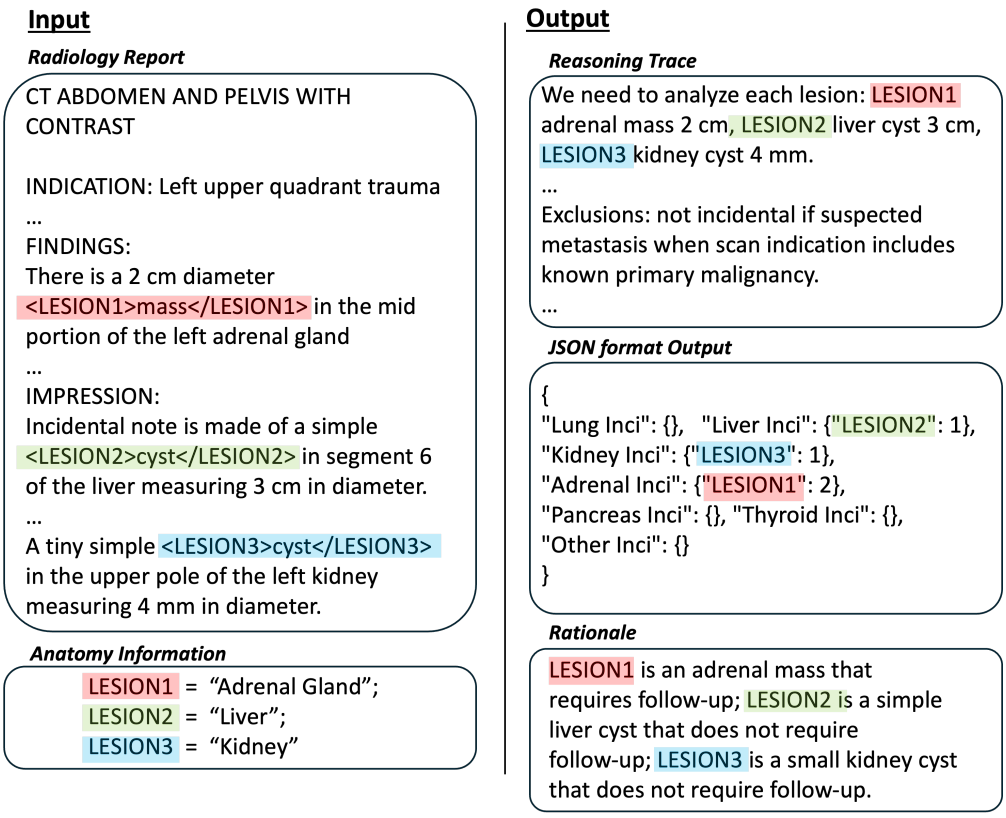

Figure 5.2: Example of input and output used for LLM-based incidentaloma identification using GPT-OSS-20B. Lesions that are not returned in the JSON output are treated as No Incidentaloma (Class 0). Reasoning traces are available only in GPT-OSS-20B inferences, as GPT-4o does not expose internal reasoning outputs.

Model performance was evaluated using the F1-score, which balances sensitivity and precision and provides a more reliable metric than overall accuracy for imbalanced clinical datasets. F1-scores were computed for all three classes in this study. To more effectively assess performance on clinically meaningful cases, we additionally report an *incidentaloma macro-F1*, defined as the mean of the F1 values for the two incidentaloma-positive classes (1 and 2). Since the *No Incidentaloma* class constituted the majority of samples and consistently achieved high performance across models, the incidentaloma macro-F1 offers a more balanced and clinically informative measure of classification quality in reports containing true incidental findings. Accordingly, incidentaloma macro-F1 was selected as the main metric for evaluation.

5.2 Results

5.2.1 Performance Comparison

As summarized in Table 5.4, the GPT-OSS-20B (With Anatomy) model achieved the highest overall performance among all evaluated systems. It attained the strongest incidentaloma-positive macro-F1 of 0.79, representing the best performance in this study. The second-best model, GPT-4o (With Anatomy), achieved F1-scores of 0.82 and 0.71 for the *Incidentaloma–No Risk* and *Incidentaloma–Follow-up Required* classes, respectively. Both anatomy-informed approaches consistently outperformed all other models, confirming that incorporating explicit anatomical context markedly enhances incidentaloma classification accuracy.

Among supervised transformer encoders, BioClinicalModernBERT (without cost-sensitive learning) and ModernBERT (cost-sensitive) achieved the highest incidentaloma macro-F1 (0.70). BioClinicalModernBERT demonstrated stronger performance on the *No Risk* class (0.79 F1), whereas ModernBERT showed a slight advantage on the *Follow-up Required* class (0.63 F1). The application of cost-sensitive (CS) training had only a modest effect overall, though it marginally improved recall for the minority follow-up class, suggesting limited benefit beyond class weighting in transformer optimization.

When comparing across model families, generative LLMs consistently outperformed all

Table 5.4: Performance comparison of supervised encoders (with and without cost-sensitive learning (CS)) and LLM-based approaches on incidentaloma classification. F1 values are reported for each class (0: No Incidentaloma, 1: Incidentaloma–No Risk, 2: Incidentaloma–Follow-up Required). Best values for each category are in bold.

Model	**Class 0**	**Class 1**	**Class 2**	**Accuracy**	**Incidentaloma Macro-F1**
Inter-annotator Agreement (IAA)	0.93	0.81	0.70	0.89	0.76
Supervised Encoder-based Models					
BioClinicalModernBERT (w/o CS)	0.99	**0.79**	0.61	0.95	0.70
BioClinicalModernBERT (CS)	0.99	0.72	0.60	0.95	0.66
ModernBERT (w/o CS)	0.99	0.76	0.60	0.95	0.68
ModernBERT (CS)	**0.99**	0.77	**0.63**	**0.95**	**0.70**
Large Language Models					
Fine-tuned Llama 3.1-8B	0.96	0.62	0.46	0.91	0.54
Llama 3.1-8B (Base)	0.89	0.59	0.46	0.81	0.52
Llama 3.1-8B (With Anatomy)	0.90	0.61	0.64	0.82	0.63
GPT-4o (Base)	0.96	0.76	0.62	0.92	0.69
GPT-4o (With Anatomy)	0.97	0.82	0.71	0.94	0.77
GPT-OSS-20b (Base)	0.96	0.81	0.71	0.93	0.76
GPT-OSS-20b (With Anatomy)	**0.97**	**0.84**	**0.73**	**0.94**	**0.79**

supervised baselines. Even the GPT-4o (Base) configuration achieved performance comparable to the best non-LLM systems, and the inclusion of anatomical context produced an additional macro-F1 improvement of up to Δ+0.08 on incidentaloma-positive classes. This pattern was consistent across both GPT- and Llama-based architectures, reinforcing that structured, anatomy-aware prompting enhances the model's ability to accurately interpret lesion-level findings and identify clinically actionable incidentalomas.

Overall, the results highlight that the integration of anatomical grounding and lesion tagging substantially improves both precision and recall, particularly for the more clinically consequential *Follow-up Required* category. These findings underscore the clinical potential

of anatomy-aware LLM frameworks for scalable and interpretable incidentaloma detection across diverse radiology report corpora.

5.2.2 Pairwise Significance Analysis on Incidentaloma-Positive Lesions

To assess the statistical significance of performance differences among models for incidentaloma identification, we performed a lesion-level bootstrap analysis. This approach quantifies the variability of model performance across individual lesion predictions, providing a more granular and clinically meaningful assessment than report-level analysis. To emphasize the subset of greatest clinical relevance, bootstrap resampling was restricted to lesions annotated as incidentalomas. Each resample was used to estimate the difference in Macro-F1 between model pairs, and the resulting confidence intervals were used to evaluate statistical significance. Figure 5.3 presents these pairwise comparisons, where each horizontal line represents the 95% confidence interval (CI) for the mean difference (Δ Macro-F1). Llama-based models were excluded from this analysis because of their lower and less stable performance across metrics.

Across all comparisons, the GPT-OSS family achieved the most consistent and statistically robust improvements. Both GPT-OSS (Base) and GPT-OSS (With Anatomy) produced higher Macro-F1 scores than the supervised baselines (ModernBERT and BioClinicalModernBERT). Among these, GPT-OSS (With Anatomy) demonstrated the strongest overall performance, with all confidence intervals positioned entirely above zero when compared to encoder-based models ($p < 0.01$). These results confirm the statistical superiority of GPT-based models in lesion-level incidentaloma classification.

Furthermore, the incorporation of anatomical context significantly improved performance for both GPT-OSS and GPT-4o models. GPT-4o (With Anatomy) achieved a statistically significant improvement over its base configuration ($p = 0.012$), and a similar trend was observed for GPT-OSS (With Anatomy), confirming the effectiveness of anatomy-aware prompting across architectures. These improvements reflect the ability of explicit anatomical grounding to reduce model uncertainty and improve classification consistency across diverse lesion types.

Taken together, these findings demonstrate that anatomy-informed prompting substantially enhances the robustness and clinical reliability of LLM-based systems for incidentaloma detection. The consistent and statistically significant gains across different architectures highlight the potential of structured, anatomy-aware contextualization as a generalizable strategy for improving performance in radiology-specific information extraction tasks.

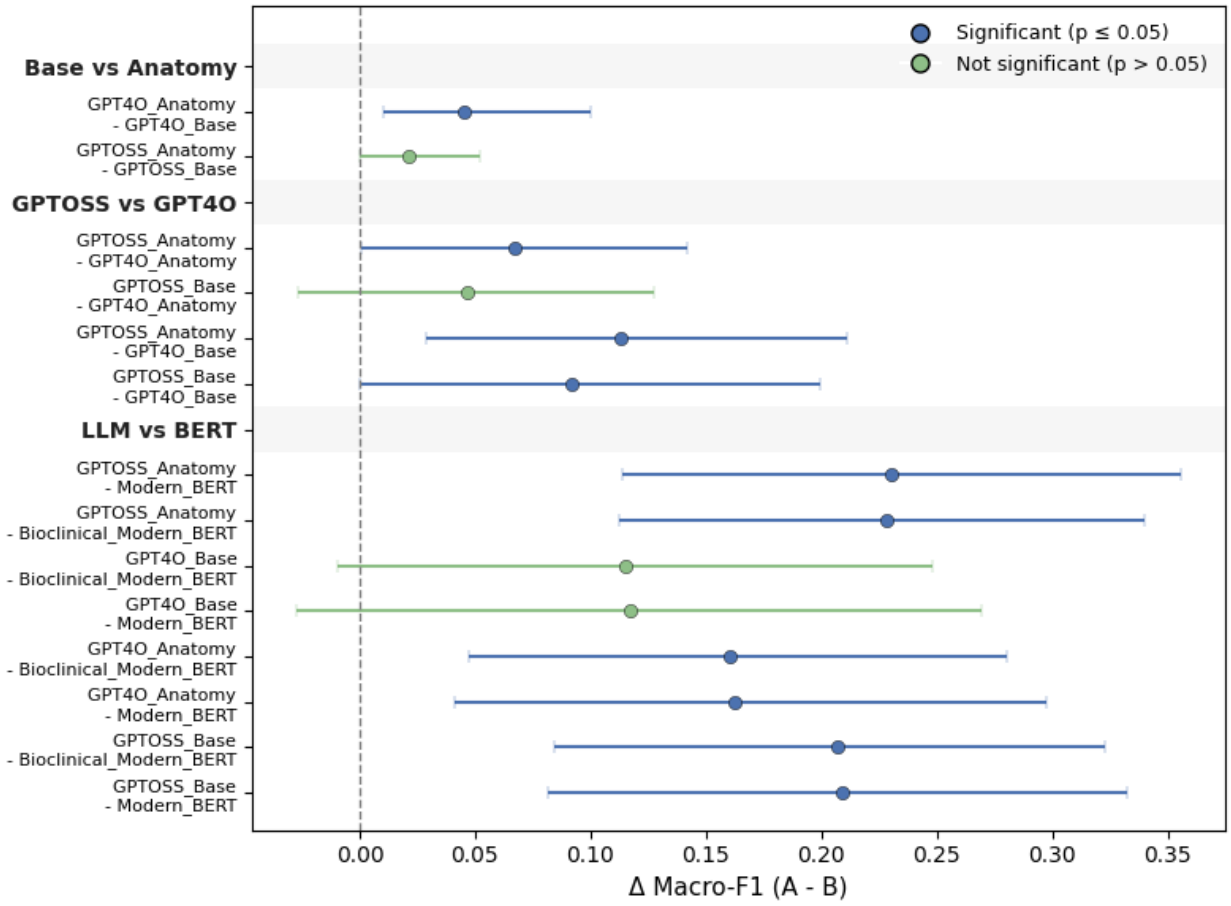

Figure 5.3: Pairwise non-parametric bootstrap comparison of model performance on incidentaloma-positive lesions. Each point represents the mean difference in Macro-F1 (Model A vs. Model B) across 1,000 lesion-level bootstrap samples, with horizontal bars indicating 95% confidence intervals. Points to the right of zero indicate that Model A outperformed Model B.

5.2.3 Error Analysis

Errors in Supervised Models

We examined error patterns for BioClinicalModernBERT and ModernBERT, both with and without cost-sensitive (CS) training, to better understand the limitations of supervised encoders. Clinical Longformer was excluded from detailed analysis because of its low recall for follow-up-required cases and its inconsistent error behavior, suggesting a misalignment between its architecture and the radiology report structure.

BioClinicalModernBERT demonstrated the most balanced performance among the supervised encoders but still showed consistent weaknesses. The most frequent error involved underestimating lesion severity, where lesions requiring follow-up were predicted as low-risk incidentalomas. Applying cost-sensitive learning reduced these misses but slightly increased false positives, particularly in benign findings. ModernBERT exhibited a similar trend with generally lower precision and recall. It frequently failed to capture abbreviated follow-up references appearing only in impression sections (e.g., "f/u CT advised") or separated from the primary lesion description.

Qualitative review identified linguistic characteristics that contributed to these errors. Both encoders struggled with hedged expressions such as "likely benign" or "probably cystic," size-related modifiers, and reports containing multiple lesions of mixed clinical significance. ModernBERT, in particular, often failed to associate implicit follow-up recommendations appearing outside the main descriptive context, reflecting limited discourse-level reasoning. These errors suggest that supervised encoders depend heavily on lexical and structural regularities, which restrict their adaptability to the broader contextual nuances common in radiology reports.

Model interpretability analyses using LIME further supported these findings. Token-level attribution weights were dominated by explicit lesion terms (e.g., "mass," "nodule," "cyst") and quantitative descriptors (e.g., "subcentimeter," "stable"), while contextual phrases conveying diagnostic uncertainty or clinical judgment were underrepresented. This behavior indicates

that, although supervised encoders capture surface-level radiologic terminology effectively, they lack the deeper contextual inference required to determine when a lesion warrants follow-up.

Errors in Generative LLMs

The most clinically significant errors were missed incidentalomas, as these represent high-risk findings that may require additional evaluation. GPT-4o (Base) failed to identify 10 of 29 (34.5%) such cases, while GPT-4o (With Anatomy) reduced this to 7 of 29 (24.1%). GPT-OSS (Base) further improved with 6 of 29 (20.7%) misses, and GPT-OSS (With Anatomy) achieved the lowest miss rate of 5 of 29 (17.2%). These results indicate that incorporating anatomical information enables the model to focus more effectively on relevant lesion-level context, reducing clinically important omissions.

False positives, where normal or non-incidental findings were incorrectly classified as incidentalomas (Class 0 $\rightarrow$ Class 1 or 2), were relatively rare, occurring in approximately 3–6% of cases. The inclusion of anatomical context modestly reduced these errors for both GPT-4o (4.8% to 3.5%) and GPT-OSS (5.6% to 5.4%). None of the LLMs misclassified low-risk incidentalomas as follow-up-required, suggesting that false positive cases primarily stemmed from boundary ambiguity rather than overestimation of clinical risk.

A consistent trend across models was the underestimation of incidentaloma severity, where follow-up-required lesions were labeled as no-risk (Class 2 $\rightarrow$ Class 1). GPT-4o showed this pattern in 13.8% of errors, while GPT-OSS reduced it to 10.3%. These cases often contained equivocal or conditional language (e.g., "likely benign," "probably cystic lesion, follow-up recommended") or phrasing where follow-up intent was implied rather than directly stated (e.g., "repeat MRI in six months to ensure stability"). Although the models identified the lesion correctly, they underestimated its clinical urgency. Among all systems, GPT-OSS (With Anatomy) demonstrated the best balance between sensitivity and specificity, minimizing missed follow-up-required findings without introducing excessive false positives.

To further characterize these behaviors, we analyzed errors made by GPT-OSS (With

Anatomy), the best-performing approach. Across 560 lesion instances, the model produced 33 errors, which were distributed across multiple anatomies including the lung, liver, kidney, and thyroid. This distribution suggests that errors arose from reasoning complexity rather than anatomy-specific bias. Three primary mismatch types were observed:

Follow-up Required vs. No-Risk (Gold = Class 2 → Model = Class 1, n = 3)

In these cases, the model correctly detected the lesion but underestimated its significance. Common examples included conditional phrasing such as "tiny sub-6 mm lung nodule, follow-up if high risk" or "indeterminate hypodensity on unenhanced liver CT." The model tended to prioritize the reassuring portion of the statement (e.g., "no follow-up needed for low-risk patients") and overlooked the qualifying clause referencing higher-risk contexts. This behavior indicates insufficient weighting of linguistic cues expressing diagnostic uncertainty or conditional recommendations, leading to under-classification of borderline lesions.

Incidentaloma Missed (Gold = Class 1 or 2 → Model = Class 0, n = 4)

Four lesions originally labeled as incidentalomas were predicted as non-incidentalomas. Manual review revealed that these were annotation inconsistencies rather than true model errors. Each report contained contextual clues such as "stable" or "compared to prior exam," suggesting that the findings had been previously evaluated and therefore should not have been labeled as incidentalomas. Interestingly, GPT-OSS (With Anatomy) correctly identified these cases as non-incidental, demonstrating sensitivity to temporal and comparative context that sometimes surpassed human annotation accuracy.

False Incidentaloma Detection (Gold = 0 → Model = 1 or 2, n = 26)

False positive classifications accounted for the majority of residual errors. GPT-OSS (With Anatomy) mislabeled 26 non-incidental lesions as incidentalomas, evenly split between no-risk (13) and follow-up-required (13) categories. These discrepancies were largely related to differences in how the model and human annotators interpreted the relationship between the clinical indication and the described lesion. For example, in a report with the indication "previous CXR abnormal" and a "small opacity in the left lower lung," annotators considered the opacity related to the prior abnormal chest X-ray and thus not incidental, whereas the

model treated it as an unrelated finding warranting follow-up. Similarly, in another case with the indication "restaging," the model inferred that the lesion reflected ongoing malignancy and thus labeled it as non-incidental, while annotators judged it incidental because it appeared in a separate anatomical site with no explicit linkage to the known primary tumor. These cases highlight how nuanced interpretation of anatomical and clinical context can lead to divergent conclusions between human experts and language models.

Overall, these analyses reveal that even the best-performing LLM occasionally misjudges clinical intent or contextual relationships between lesions and imaging indications. However, the inclusion of anatomical grounding and lesion tagging substantially improved model reliability, particularly for identifying clinically meaningful follow-up-required cases. The remaining errors primarily reflect interpretive ambiguities that are also challenging for human experts, underscoring the importance of ongoing refinement of annotation guidelines and model interpretability tools in future work.

5.3 Discussion

5.3.1 Ensemble Effects and Majority-Vote Performance

Twelve lesions were correctly classified by all GPT-based models but misclassified by both BERT-based models, whereas the opposite occurred only five times. Many of the GPT-correct and BERT-incorrect cases involved nodule management guideline templates that were automatically inserted into reports by the radiology information system (RIS) macros (e.g., *"follow-up per Fleischner Society guidelines"*). These template insertions often appear without explicit contextual linkage to an active finding, which likely caused confusion for the supervised models that rely heavily on local lexical cues rather than broader contextual reasoning. In contrast, GPT-based models demonstrated stronger semantic understanding, accurately linking follow-up recommendations to the appropriate lesion findings. While these results highlight the contextual adaptability of LLMs, BERT-based models displayed a more conservative bias that frequently missed positive incidentaloma cases but produced fewer

false positives, suggesting a precision-oriented decision threshold.

To combine the complementary strengths of different modeling approaches, a majority-vote ensemble was constructed using six models: GPT-4o (Base and Anatomy), GPT-OSS (Base and Anatomy), ModernBERT with cost-sensitive (CS) training, and BioClinicalModernBERT without CS. Each lesion's final label was determined by the most frequent prediction among models. In cases where a tie occurred, the lowest numerical label was selected ($0 < 1 < 2$), creating a conservative bias toward non-incidentaloma classification. This simple, parameter-free ensemble achieved the highest overall performance, yielding a Macro-F1 of 0.902, surpassing all individual systems. Lesion-level F1 scores were well balanced across all classes (No Incidentaloma = 0.988, No Risk = 0.889, Follow-up Required = 0.830), representing consistent improvements over the best single model, GPT-OSS (With Anatomy), which achieved 0.968, 0.842, and 0.727 for the same categories. The ensemble improved each class's performance, particularly for identifying follow-up-required incidentalomas, where the F1 increased by more than 0.10. These results suggest that combining the contextual reasoning strength of LLMs with the structured precision of supervised encoders enhances reliability in clinical information extraction from radiology reports, especially when interpreting ambiguous or partially specified findings.

Despite these gains, ensembling introduces practical challenges for deployment. Running multiple large models in parallel increases computational cost, inference time, and system complexity, limiting real-world scalability. In addition, ensemble predictions can obscure the interpretability of individual model contributions, complicating clinical validation and error attribution. Future implementations should therefore weigh the benefits of ensemble stability against operational efficiency and transparency, especially when integrating multimodel systems into production-grade clinical pipelines.

5.3.2 Clinical Implications and Potential Applications

The proposed incidentaloma identification framework offers several opportunities for advancing clinical decision support (CDS) and workflow optimization in radiology. Integrated at the point

of order entry, the system could automatically identify patients with prior incidentalomas and recommend appropriate follow-up imaging studies based on the detected anatomy, lesion type, and historical findings. When incorporated into radiology reporting software, the model could provide real-time guidance by aligning imaging observations with relevant clinical practice guidelines, considering lesion size, imaging characteristics, and patient risk factors. These capabilities would enable more consistent follow-up recommendations and reduce missed or delayed evaluations.

The structured lesion-level outputs generated by the model can also serve as a foundation for automated follow-up tracking. By linking individual lesions to temporal data from prior studies, the system could facilitate longitudinal monitoring of incidental findings and identify cases with incomplete or overdue follow-up. This capability would support quality improvement initiatives within radiology departments and strengthen adherence to guideline-based care.

Beyond automation, the model design promotes interpretability and clinical trust. The lesion-tagged framework and LLM-generated rationales can be integrated into interactive dashboards that allow radiologists to visualize model reasoning, inspect decision paths, and verify follow-up recommendations. Such explainability features would help clinicians validate AI outputs and maintain oversight in diagnostic decision-making.

Together, these applications demonstrate the potential of automated incidentaloma detection systems to move beyond research settings toward practical deployment in clinical environments. By supporting consistent, transparent, and guideline-concordant care, this framework has the potential to improve follow-up adherence, reduce diagnostic delays, and enhance the overall quality of imaging-based patient management.

5.4 Conclusion

In this chapter, we introduced a comprehensive evaluation of supervised transformer-based encoders and generative LLMs for the automated identification and classification of incidentalomas in radiology reports. By introducing lesion-tagged inputs and anatomy-aware

prompting, the study demonstrated that generative LLMs, particularly GPT-OSS (With Anatomy), achieved superior and more balanced performance compared with traditional supervised encoders. These improvements were most evident in incidentaloma-positive cases, where explicit lesion and anatomical context enhanced model reasoning regarding clinical significance and follow-up requirements.

In addition to quantitative performance gains, the analyses highlighted the importance of structured lesion representation for improving interpretability and consistency. The integration of lesion tags and anatomical mapping enabled LLMs to focus on clinically relevant findings while minimizing confusion from unrelated text segments. In contrast, supervised transformer models exhibited stronger precision but lower recall, reflecting a conservative bias that limited sensitivity to subtle or ambiguous findings. Ensemble analyses further demonstrated that combining encoder-based and generative systems through majority voting improved stability and performance across clinically meaningful categories.

Collectively, these findings establish the value of anatomy-aware and lesion-structured prompting as a generalizable framework for radiology report understanding. By aligning model inference with the clinical reasoning patterns of radiologists, this approach bridges the gap between automated text classification and actionable clinical decision support. The results underscore the potential of structured LLM frameworks to enhance the accuracy, interpretability, and reliability of information extraction in real-world radiology workflows.false positives in the sequence classification tasks.

5.5 Limitations

Despite the promising findings presented in this study, several limitations warrant careful consideration. First, the annotated dataset, although curated and reviewed by board-certified radiologists, remains relatively small due to the labor-intensive nature of manual lesion-level annotation. This limited dataset size may constrain model generalizability to less common anatomies, atypical lesion descriptions, or reporting styles not represented in the current sample. Second, radiology reporting conventions and imaging protocols vary substantially

across institutions and modalities, which may limit direct transferability of model performance to other healthcare systems without additional fine-tuning or contextual adaptation. Third, the present evaluation focused primarily on single-report text analysis and did not incorporate longitudinal or temporal information from prior studies, which could provide valuable context for distinguishing stable from newly detected lesions and improve reasoning about follow-up necessity.

Annotation subjectivity also presents a potential source of variability. Although consensus reviews were conducted to standardize labeling, subtle differences in interpretation may persist in borderline cases where incidentaloma status or follow-up need is ambiguous. These inconsistencies could introduce minor noise into the gold-standard dataset, particularly in complex reports with multiple overlapping findings. Furthermore, while preliminary interpretability analyses were performed using token-level attributions and qualitative case inspection, more extensive radiologist-guided explainability studies are necessary to elucidate how large language models weigh contextual and semantic cues in decision making. Understanding these mechanisms is essential for validating model reliability in clinical deployment settings.

Future work should therefore pursue three main directions: expansion of dataset scale and diversity through semi-automated or active-learning annotation pipelines; cross-institutional validation to assess model robustness in heterogeneous clinical environments; and deeper integration of temporal patient information to support longitudinal reasoning. Parallel efforts in interpretability research and clinician-in-the-loop evaluation will be critical for ensuring that AI-assisted incidentaloma detection systems achieve both technical accuracy and clinical trustworthiness.

Chapter 6

CASE STUDY: ADRENAL INCIDENTALOMA

Adrenal incidentalomas represent one of the most frequent unexpected findings on cross-sectional imaging. Their clinical relevance, variable reporting patterns, and evolving diagnostic guidelines make them an ideal exemplar for demonstrating the end-to-end integration of lesion-level extraction, longitudinal follow-up identification, and incidentaloma classification. This chapter applies the full methodology developed in Chapters 3–5 to a focused use case: large-scale adrenal incidentaloma analysis in our institutional radiology corpus.

6.1 Motivation

Adrenal lesions are detected in approximately 4–5% of abdominal CT examinations [116], and most are benign. However, determining which lesions require further evaluation remains challenging, particularly when findings are documented inconsistently across free-text radiology reports. Traditional adrenal imaging practices have relied on two long-standing CT-based principles: (1) an upper threshold of 10 Hounsfield Unit (HU) on non-contrast CT to infer benignity, and (2) the use of adrenal washout CT (AWCT) for lesions above this threshold. Recent evidence, however, has challenged both assumptions, concluding that AWCT was never validated in true incidentaloma populations and proposing a revised 20 HU threshold for lesions under 4 cm [116].

Given these shifts, there is significant value in using real-world clinical data to evaluate adrenal incidentalomas at scale. Automated lesion extraction, incidentaloma classification, and follow-up identification provide an opportunity to generate empirical evidence that contextualizes evolving imaging recommendations and highlights practice patterns across more than a decade of clinical imaging.

6.2 Methods

6.2.1 Overview of the Integrated Pipeline

This case study brings together the three primary components developed in prior chapters. Lesion-level extraction using CAMIR and PL-Marker++ provides anatomically grounded detection of adrenal lesions and captures attributes such as size and descriptive qualifiers. The longitudinal follow-up model introduced in Chapter 4 enables temporal reasoning across patients' imaging histories, allowing us to determine whether appropriate follow-up occurred. The incidentaloma classification framework described in Chapter 5 then assigns each adrenal lesion to clinically meaningful categories based on both textual context and anatomical cues.

Together, these components create a unified end-to-end pipeline for identifying adrenal lesions, determining their incidentaloma status, and evaluating downstream follow-up imaging across more than a decade of institutional reports.

6.2.2 Sampling Process

Sampling of adrenal lesions was conducted using a staged filtering process applied to the full PL-Marker++ extraction output. We first selected all radiology reports containing at least one adrenal lesion mention, which resulted in 86,236 reports across all modalities. To ensure that the extracted lesions reflected clinically meaningful observations rather than hypothetical or negated statements, we restricted the set to reports in which the lesion assertion status was marked as either "possible" or "present," yielding a total of 84,643 reports. Because this case study focuses on newly identified adrenal findings, we next excluded lesions whose size-trend annotations indicated stability, decrease, or other changes inconsistent with new discovery. This reduced the cohort to 44,389 reports.

To isolate true adrenal incidentalomas, we further excluded reports from patients who had documented neoplasm-related indications (identical to the approach used in Chapter 5) combined with target imaging modalities such as CT, MRI, ultrasound, or PET. This step removed cases in which adrenal lesions may have been detected in the context of known or

suspected malignancy rather than incidentally. After applying this final criterion, 8,454 reports remained and constituted the adrenal lesion cohort used as index reports for downstream incidentaloma classification and follow-up identification.

6.2.3 Follow-Up Imaging Identification for Adrenal Incidentalomas

Follow-up imaging was identified using the longitudinal model described in Chapter 4, with adrenal-specific criteria refined through consultation with a board-certified radiologist. For each of the 8,454 adrenal index reports, we examined subsequent imaging studies in the patient's timeline to determine whether any qualified as potential follow-up examinations. The radiologist advised that follow-up assessments for adrenal incidentalomas should rely only on CT or MR studies, since these modalities provide adequate characterization of adrenal morphology and are standard for interval evaluation. Based on this guidance, all ultrasound and PET-CT examinations were excluded from consideration as candidate follow-up studies.

The radiologist also recommended using a maximum timeframe of 16 months to capture clinically meaningful follow-up opportunities. This window encompasses typical adrenal follow-up intervals used in practice and ensures that delayed but still relevant assessments are included. Using this 16-month upper bound and restricting modality to CT or MR, we identified 3,475 index reports that had at least one subsequent study meeting these criteria. Across these reports, the search yielded a total of 12,324 candidate examinations, corresponding to an average of 3.54 qualifying studies per index report. The resulting distribution was heavily skewed toward cases with only one or two candidate follow-up studies, although a smaller number accumulated substantially more.

This approach provides a scalable mechanism for evaluating follow-up behavior across thousands of adrenal incidentaloma cases while grounding the methodology in radiologist-informed operational definitions of appropriate imaging. It enables systematic assessment of whether patients received timely follow-up within a clinically reasonable interval and whether follow-up practices varied across different types of adrenal lesions.

6.3 Results

6.3.1 Prevalence and Classification of Adrenal Incidentalomas

Using the best performing configuration from Chapter 5 (GPT OSS-20B with anatomy aware prompting and lesion tags), we processed 8,454 reports that mentioned the adrenal lesions. End-to-end incidentaloma identification required 92.5 hours of compute time.

Although the sampling was centered on adrenal lesions, incidentalomas were present across other anatomies as well. Table 6.1 summarizes class distributions by anatomy type, with counts and within-row percentages. *Adrenal Incidentaloma* had the largest proportion requiring follow-up (29.8%) and the lowest share of "No Incidentaloma" (34.4%), which reflects the adrenal-focused sampling.

Across anatomies, the combined proportion of incidentalomas (classes 1 and 2) varies widely: Adrenal 65.6%, Other 34.1%, Kidney 32.9%, Lung 22.2%, Liver 20.9%, Thyroid 5.3%, and Pancreas 3.0%. Adrenal shows the highest burden, and nearly half of adrenal incidentalomas require follow-up (29.8% out of 65.6% $\approx$ 45%). In contrast, kidney and liver have moderate combined prevalence but relatively small follow-up fractions among their incidentalomas (10.6% and 12.4% respectively), while pancreas and thyroid show low overall prevalence yet higher follow-up share within their incidentalomas (56.7% and 35.8% respectively), although absolute counts for these two groups are small.

6.3.2 Adherence to Follow-Up Recommendation

Among the 2,520 reports predicted as having at least one adrenal incidentaloma requiring follow-up, 1,638 had at least one candidate follow-up examination (CT or MR within 16 months), which is 65.0%. Within those 1,638 index–candidate sets, 513 had a qualifying follow-up exam, which is 31.3% among reports with at least one candidate report and 20.4% relative to all 2,520 predicted as requiring follow-up. Processing all index–candidate pairs took 10.2 hours using GPT-OSS 20b with the best-performing configuration described in Chapter 4.

Table 6.1: Report-level incidentaloma class prediction results among reports with at least one adrenal lesion (n=8,454).

Anatomy Type	**No Incidentaloma (0)**	**Incidentaloma, No-Risk (1)**	**Incidentaloma, Follow-up Required (2)**
Lung Incidentaloma	6571 (77.7%)	1355 (16.0%)	528 (6.2%)
Liver Incidentaloma	6689 (79.1%)	1548 (18.3%)	217 (2.6%)
Kidney Incidentaloma	5668 (67.0%)	2487 (29.4%)	299 (3.5%)
Adrenal Incidentaloma	2908 (34.4%)	3026 (35.8%)	2520 (29.8%)
Pancreas Incidentaloma	8200 (97.0%)	110 (1.3%)	144 (1.7%)
Thyroid Incidentaloma	8012 (94.8%)	284 (3.4%)	158 (1.9%)
Other Incidentaloma	5575 (65.9%)	2297 (27.2%)	582 (6.9%)

These adherence levels are consistent with prior evidence. Prior adrenal incidentaloma studies show similarly low rates of recommended follow-up. Feeney et al. [42] reported that only a minority of patients with incidental adrenal masses complete recommended radiographic evaluation, with a typical adherence near one third across institutions. A separate adrenal incidentaloma cohort of 121 patients found that only 27% received any follow-up imaging within the recommended interval [111]. Another adrenal incidentaloma study reported that only 23% of patients with an in-network primary care provider underwent an appropriate cross-sectional imaging examination [41]. Taken together, these findings show that the integrated pipeline yields empirical adherence estimates that closely reflect actual clinical behavior.

Figure 6.1 summarizes the distribution of follow-up intervals among the 513 reports with adrenal incidentalomas that had qualifying follow-up exam. The overall pattern shows a strong right-skewed distribution, with most follow-up examinations occurring relatively early. The median interval was 42 days, and the first quartile was 9 days, which indicates that a substantial proportion of follow-up studies were completed within the first few weeks after

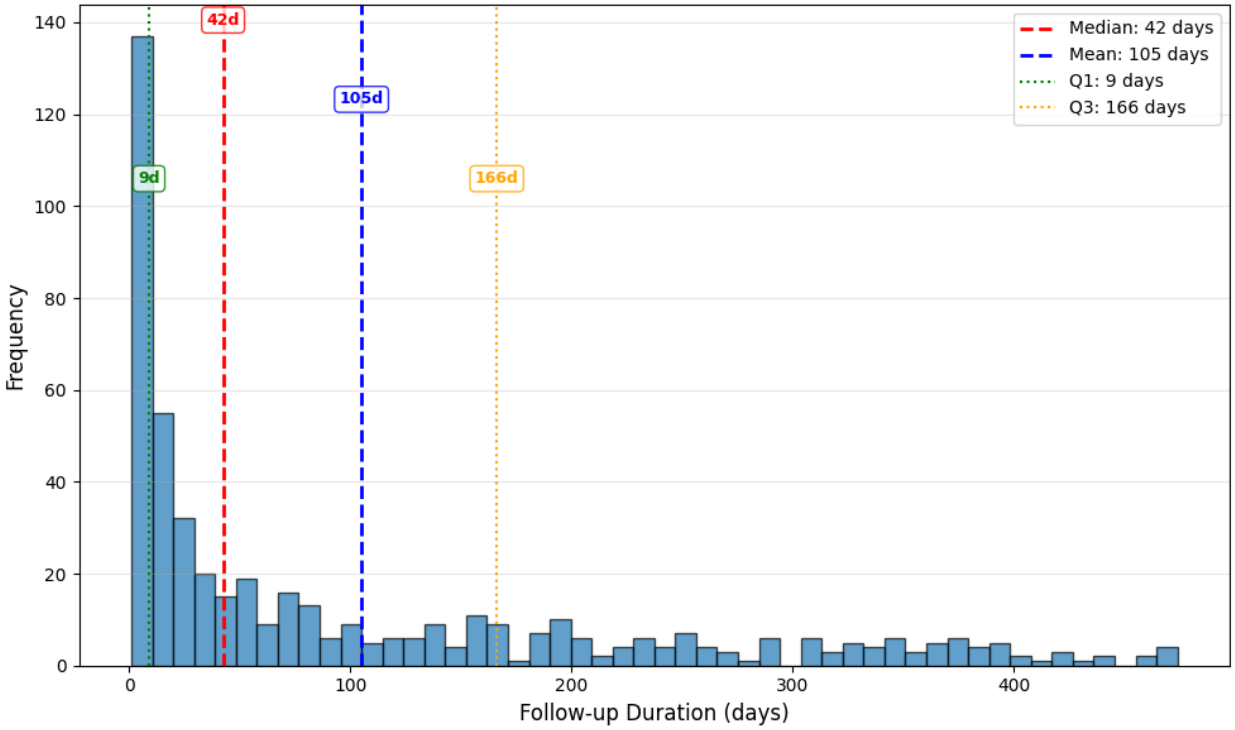

Figure 6.1: Distribution of follow-up intervals (days) for qualifying adrenal follow-up studies.

the index report. These early follow-up examinations likely represent patients for whom the imaging concern prompted timely clinical action, and they illustrate that when follow-up does occur, it is often completed promptly. Beyond this early cluster, a long tail of delayed follow-up is evident. The mean interval was 105 days, and the third quartile extended to 166 days, with scattered examinations occurring more than one year after the index report. The combined pattern shows that even among patients who eventually receive follow-up imaging, the timing is highly variable.

6.4 Discussion

Using the reports with extracted adrenal incidentalomas, we further evaluate recent study [116] suggesting that current management guidelines require revision. Although traditional practice relies on a 10 HU threshold on non-contrast CTs, contemporary large cohort studies

show that these criteria are overly restrictive and provide limited diagnostic value in true incidentaloma populations. Benign lesions frequently exceed 10 HU, and a 20 HU cutoff for lesions smaller than 4 cm retains a very high predictive value for benignity. These findings indicate that many lesions previously labeled indeterminate can be safely reclassified without additional imaging. Therefore, among 5,546 reports with at least one adrenal incidentaloma predicted, we limited our analysis to 2,150 reports that were actually non-contrast CTs. In this pool, we conducted LLM-based inference to extract HU values, lesion size, and whether a follow-up was recommended. HU values were often expressed qualitatively, such as *"less than 10 HU"*, so we applied a rule-based quantification system to convert these descriptions into numeric values. From this process, we identified 908 adrenal incidentalomas with explicit HU mentions. A total of 893 lesions contained both HU and size information, and after excluding measurements greater than 100 HU, we retained 749 adrenal incidentalomas for further analysis (Figure 6.2). HU values above 100 are uncommon for benign adrenal pathology and typically indicate measurement error, technical artifact, or a distinct diagnostic entity, which justifies their exclusion. All clinical attributes used in this analysis, including attenuation, lesion size, and follow-up recommendation, were extracted using GPT-OSS-20B.

As shown in Table 6.2, applying the proposed guideline leads to a substantial reclassification of adrenal incidentalomas. Under the previous system, 234 lesions were labeled indeterminate and typically required additional imaging. When the new criteria are applied, only 120 lesions fall into categories that recommend or seriously consider follow-up (Categories 2 and 3). In other words, approximately 50% of lesions previously considered indeterminate can be reclassified as definitively benign under the updated criteria. This shift highlights the potential for meaningful reductions in unnecessary imaging and patient burden.

Breaking down the 629 benign adrenal incidentalomas under the proposed guideline, the LLM identified that 104 of these lesions received a follow-up recommendation in the original report, representing 16.53% of the newly classified benign group. For example, a lesion with 16 HU and a size of 1.9 cm was recommended for follow-up, even though such a lesion is considered benign under the updated criteria and does not require additional

imaging. Among these 629 benign findings, our pipeline identified that 18 lesions had an actual follow-up examination performed. All of these were confirmed to be adenomas, which provides additional evidence that many follow-up examinations performed in practice may not have been clinically necessary.

Through this analysis, our integrated extraction and follow-up identification framework provides a detailed view of how guideline updates would affect real-world clinical decision-making. The results show that modernized criteria would reduce the number of patients undergoing avoidable imaging while maintaining appropriate surveillance for higher-risk lesions. They also demonstrate that LLM-based extraction can reliably characterize lesion features at scale, allowing healthcare systems to evaluate the downstream consequences of guideline changes using routinely generated radiology reports.

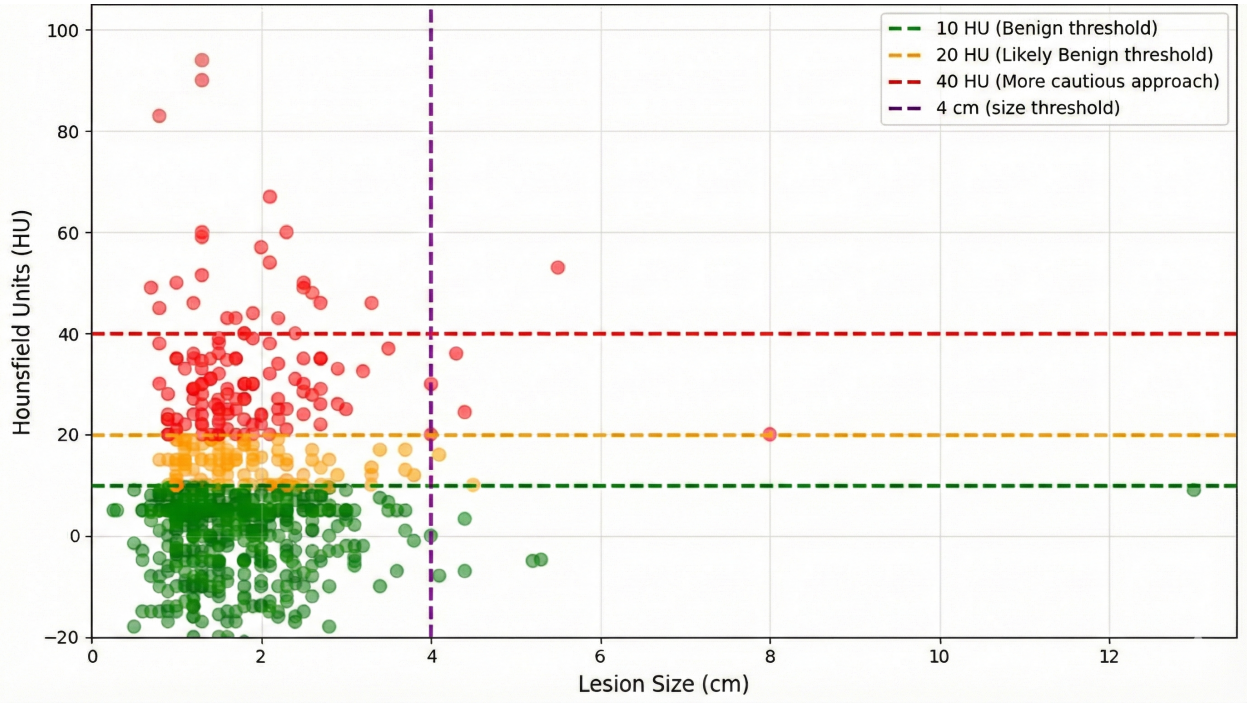

Figure 6.2: Distribution of Hounsfield Units and lesion sizes among 749 predicted adrenal incidentalomas identified in non-contrast CT reports

Category	Count (%)	Definition
Previous Guidelines		
Benign	515 (68.76%)	Non-contrast HU < 10, interpreted as lipid-rich adenoma.
Indeterminate	234 (31.24%)	HU ≥ 10 or size-based concern, traditionally required adrenal washout CT.
Proposed Guidelines *(Seow et al., 2025)*		
Category 1: Benign	629 (83.98%)	No further imaging required. $<$10 HU | 10–20 HU with size < 4 cm.
Category 2: Likely Benign	117 (15.62%)	Recommend 6–12 month non-contrast CT. 10–20 HU and size > 4 cm | HU > 20 and 1-4 cm
Category 3: Higher Risk	3 (0.40%)	Multidisciplinary or surgical referral. Lesions > 20 HU and/or $>$ 4cm. A more cautious approach suggested when HU > 40 or size $>$ 6 cm.

Table 6.2: Lesion categorization under previous and proposed adrenal incidentaloma guidelines.

6.5 Conclusion

This case study shows that an LLM-based information extraction pipeline can characterize adrenal incidentalomas at scale using only free-text radiology reports. The system identified many adrenal incidentalomas across the dataset, extracted key attributes, and quantified downstream adherence. Among those predicted to need follow-up, 65.0% had at least one candidate CT or MR within sixteen months, and 20.4% received qualifying follow-up that satisfied modality and timing criteria.

Beyond adherence estimation, the framework enabled a data-driven evaluation of emerging adrenal incidentaloma guidelines. Using automatically extracted attenuation values and lesion

sizes, we found that applying the proposed criteria would substantially reduce the number of lesions labeled indeterminate. Approximately one half of lesions that previously required follow-up could be reclassified as benign using the revised HU and size thresholds. Among the 104 benign lesions for which a follow-up was recommended in the original report, 18 had an actual follow-up examination, and all were ultimately confirmed to be adenomas. These findings illustrate that many follow-up studies performed in routine practice may not have been clinically necessary when evaluated under contemporary evidence-based criteria.

These results underscore the potential of NLP-based systems to support quality monitoring in radiology by identifying incidental findings, characterizing clinically actionable subgroups, and evaluating whether recommended imaging occurs within the expected timeframe. The ability to perform these assessments across thousands of reports using only narrative text highlights the scalability and real-world applicability of the pipeline. This work also demonstrates that LLM-based extraction can enable retrospective simulation of guideline changes, which provides an empirical foundation for estimating the impact of revised imaging recommendations on patient management.

Several limitations should be acknowledged. The analysis relies exclusively on radiology report text from a single institution and may miss imaging performed outside the institution or documented in nonstandard ways. Lesion-level descriptors extracted from narrative text may lack the precision required to fully stratify adrenal lesions. In addition, follow-up imaging appropriateness does not incorporate biochemical evaluations or other clinical factors that may influence follow-up decisions.

Future work will focus on expanding the framework to include structured endocrine laboratory values and incorporating linkages to additional clinical documentation. These enhancements may allow for more comprehensive risk stratification and more accurate determination of whether lesions receive appropriate surveillance in real-world practice.

Overall, this case study highlights how large-scale, report-driven analysis can reveal important patterns in adrenal incidentaloma management and demonstrates how NLP-based approaches can contribute to improving guideline-consistent imaging care.

Chapter 7
CONCLUSION AND FUTURE WORK

This chapter concludes the dissertation by summarizing its key contributions, discussing major limitations, and outlining future research directions. It also offers final reflections on the implications of this work for radiology NLP and the broader landscape of LLMs in clinical applications.

7.1 Key Contributions

This dissertation introduced an integrated framework for extracting clinically meaningful information from radiology reports, emphasizing lesion-level structure, anatomy-aware reasoning, and scalable application to real-world clinical data. The work advances both the technical and translational aspects of radiology-focused natural language processing. The primary contributions are summarized below.

- **Creation of a high-quality radiology corpus for information extraction.** A richly annotated dataset was developed to capture lesion findings, anatomical attributes, clinical indications, and relations among these elements. This corpus supported the development and evaluation of PL-Marker++, a radiology-specific information extraction model that achieved strong performance and provided structured outputs for large-scale deployment across more than six million radiology reports.

- **Systematic evaluation of supervised and generative models for identifying follow-up imaging.** The dissertation examined multiple modeling strategies for detecting follow-up imaging in longitudinal report sequences. Generative LLMs, particularly GPT-based architectures, demonstrated strong performance through structured prompt-

ing, while supervised encoders provided conservative, precision-oriented behavior. The findings showed that structured contextual cues such as lesion attributes and anatomical information enhance the reliability of automated follow-up identification.

- **Development of an NLP framework for incidentaloma detection.** A new method for incidentaloma classification was introduced using lesion-tagged inputs and anatomy-aware prompts. This approach enabled fine-grained identification of incidentalomas across several anatomical regions and demonstrated improved sensitivity compared to supervised transformer models. The explicit representation of lesions using XML-style tags proved essential for recovering subtle or ambiguously stated findings.

- **Clinical case study on adrenal incidentalomas.** A focused case study extended the incidentaloma detection framework to adrenal incidentalomas, demonstrating an end-to-end pipeline for lesion identification, incidentaloma classification, and follow-up adherence assessment. The findings were consistent with published evidence on follow-up rates and illustrated how this framework can be used to characterize incidentaloma patterns across a broader clinical population.

Together, these contributions establish a scalable and clinically aligned approach for radiology NLP, with demonstrated utility across several downstream applications involving lesion-level reasoning.

7.2 Limitations

- **Dataset scale and diversity.** The annotated corpora developed in this work, such as CAMIR and the incidentaloma dataset, are relatively small due to the labor-intensive nature of manual lesion-level annotation. For instance, CAMIR includes 609 reports, and the incidentaloma evaluation set comprises 400 reports. This limited scale may constrain the ability of the models to generalize to atypical lesion descriptions, rare pathologies, or diverse reporting styles not represented in the samples.

- **Single-institutional bias.** The study relies exclusively on data from a single hospital system. Radiology reporting conventions, templates, and imaging protocols vary substantially across healthcare institutions, which limits the direct transferability of model performance to other systems without additional fine-tuning or domain adaptation.

- **Unimodal text analysis.** The current framework relies solely on the narrative text of radiology reports and does not incorporate pixel-level information from the images. This limits the system's ability to cross-verify textual findings against the ground-truth imaging appearance or use visual cues to resolve ambiguities in the text.

- **Evaluation metrics.** The evaluation focused primarily on standard NLP performance metrics such as precision, recall, and F1-score. While necessary for benchmarking, these metrics do not fully capture the clinical reasoning quality, factual consistency, or downstream impact on patient outcomes required for safe clinical deployment.

7.3 Future Work

Building on the foundations laid in this dissertation, I plan to pursue several specific research directions to bridge the gap between technical NLP performance and clinical utility.

- **Multimodal integration.** I plan to extend the current text-based frameworks by integrating imaging data. Combining pixel-level data with text has strong potential to improve lesion-level characterization, reduce ambiguity, and support joint reasoning across modalities.

- **Cross-institutional validation.** To address the single-institution limitation, I intend to validate the proposed framework on multi-institutional datasets. This will allow me to assess generalizability and identify settings that require fine-tuning to accommodate heterogeneous reporting styles.

- **Scalable annotation methods.** To overcome the bottleneck of manual data curation, I aim to implement semi-supervised or active-learning annotation pipelines. These methods will help expand the scale of lesion-level annotations and reduce the expert time required for high-quality dataset development.

- **Temporal and longitudinal modeling.** I will move beyond single-report analysis to model sequences of reports across patient timelines. This work will focus on enhancing the identification of disease progression and improving reasoning regarding the stability, growth, or resolution of lesions over time.

- **Clinician-centered interpretability.** I plan to design and evaluate interactive visual tools that highlight lesion-level evidence and explain model reasoning. These tools are essential for improving clinician trust and supporting the safe integration of LLM-assisted systems into radiology workflows.

7.4 Final Remarks

This dissertation demonstrates that large-scale annotated corpora and modern language models can meaningfully improve the automated understanding of radiology reports. Across three major aims—dataset construction, longitudinal follow-up detection, and incidentaloma identification—the work provides a unified framework for transforming unstructured radiology text into structured clinical information that is actionable, interpretable, and scalable. The results highlight the strong potential of LLMs to support radiologists in routine reporting tasks and to strengthen population-level surveillance of incidental findings and follow-up adherence.

Integrating these models into clinical workflows offers several opportunities. Automated extraction systems can assist radiologists by surfacing key findings, tracking follow-up recommendations across time, and standardizing documentation for downstream quality improvement and research. However, deployment also requires careful consideration of real-world

challenges, including model drift, domain shifts across hospitals, hallucination risks in generative models, and the need for transparent decision-making that clinicians can trust. Ensuring seamless workflow integration will require collaboration with radiologists, informatics teams, and clinical operations groups to design interfaces that augment, rather than disrupt, clinical practice.

Despite these challenges, the methods and resources developed in this dissertation provide a strong foundation for future innovation in radiology NLP. By continuing to refine dataset quality, integrate multimodal information, and enhance model reliability, this line of research is well positioned to contribute to safer, more efficient, and more intelligent clinical imaging workflows.

Appendix A

CAMIR

A.1 Labeled Entities Statistics

Table. A.1 demonstrates the dataset statistics groupbed by entities.

Table A.1: Distribution of Labeled Clinical Concepts

Type	Subtype	Count
Abdomen		
		5
	Abdominal_Wall	50
	Adrenal_Gland	71
	Mesentery	52
	Peritoneal_Sac	64
	Retroperitoneal	162
	Spleen	108
	Undetermined	488
Assertion		
	absent	2767
	possible	1052
	present	8513
Body_Regions		
	Entire_Body	7

Continued on next page

Type	Subtype	Count
	Lower_Limb	172
	Pelvis	400
	Undetermined	106
	Upper_Limb	202
Cardiovascular		
		1
	Arterial	257
	Coronary_Artery	86
	Heart	224
	Pericardial_Sac	66
	Pulmonary_Artery	21
	Undetermined	40
	Venous	75
Characteristic		
		2524
Count		
		363
Digestive		
	Esophagus	70
	Intestine	79
	Large_Intestine	150
	Small_Intestine	54
	Stomach	72
F_Reproductive_Obstetric		

Continued on next page

Type	Subtype	Count
	Adnexal	9
	Breast	175
	Female_Genital_Structure	26
	Ovary	28
	Uterus	31
	Undetermined	3
Head_Neck		
		2
	Ear	4
	Eye	9
	Laryngeal	4
	Mouth	58
	Neck	476
	Nasal_Sinus	241
	Pharynx	19
	Undetermined	202
Hepato-Biliary		
	Bile_Duct	75
	Gallblader	34
	Liver	425
	Pancreas	73
	Undetermined	2
Indication		
		1494

Continued on next page

Type	Subtype	Count
Indication_Type		
		1
	neoplastic_dx	558
	nonneoplastic_dx	305
	symptom	433
	trauma	76
Lesion		
		5709
Lymphatic		
	Undetermined	559
M_Reproductive		
	Prostate	40
	Testis	3
	Undetermined	6
Medical_Problem		
		6254
Miscellaneous		
	Adipose_Tissue	1
	Biomedical_Device	20
	Connective_Tissue	33
	Undetermined	5
Musculo-Skeletal		
	Bone_and_or_Joint	1559
	Skeletal_and_or_Smooth_Muscle	242

Continued on next page

Type	Subtype	Count
	Undetermined	10
Neurological		
	Brain	765
	Cerebrospinal_Fluid_Pathway	156
	Cerebrovascular_System	124
	Extraaxial	148
	Nerve	52
	Pituitary	10
	Spine_Cervical	382
	Spine_Lumbar	860
	Spine_Sacral	52
	Spine_Thoracic	246
	Spine_Unspecified	107
	Spine_Cord	210
	Undetermined	123
Respiratory		
	Lung	946
	Pleural_Membrane	198
	Tracheobronchial	51
	Undetermined	5
Size		
		6
	current	1016
	past	145

Continued on next page

Type	Subtype	Count
Size_Trend		
	decreasing	99
	disappear	66
	increasing	129
	new	168
	no_change	381
Skin		
	Skin_and_or_Mucous_Membrane	46
	Subcutaneous	11
	Undetermined	1
Thoracic		
	Mediastinal	345
	Undetermined	427
Urinary		
	Kidney	298
	Ureter	36
	Urinary_Bladder	36
	Undetermined	8

A.2 Labeled Events Statistics

Table. A.2 demonstrates the dataset statistics grouped by event types.

Table A.2: Distribution of Labeled Clinical Concepts by Event Type

Event type	Type	Subtype	Count
Indication			
Indication	Abdomen		1
Indication	Abdomen	Adrenal_Gland	3
Indication	Abdomen	Mesentery	1
Indication	Abdomen	Peritoneal_Sac	3
Indication	Abdomen	Retroperitoneal	3
Indication	Abdomen	Spleen	5
Indication	Abdomen	Undetermined	51
Indication	Assertion	absent	17
Indication	Assertion	possible	157
Indication	Assertion	present	1320
Indication	Body_Regions	Lower_Limb	47
Indication	Body_Regions	Pelvis	14
Indication	Body_Regions	Undetermined	48
Indication	Body_Regions	Upper_Limb	28
Indication	Cardiovascular	Arterial	5
Indication	Cardiovascular	Pulmonary_Artery	1
Indication	Cardiovascular	Undetermined	2
Indication	Cardiovascular	Venous	1
Indication	Digestive	Esophagus	10
Indication	Digestive	Intestine	7
Indication	Digestive	Large_Intestine	23
Indication	Digestive	Small_Intestine	3
Indication	Digestive	Stomach	11
Indication	F_Reproductive_Obstetric	Adnexal	2
Indication	F_Reproductive_Obstetric	Breast	72
Indication	F_Reproductive_Obstetric	Female_Genital_Structure	7
Indication	F_Reproductive_Obstetric	Ovary	11

Continued on next page

Event type	Type	Subtype	Count
Indication	F_Reproductive_Obstetric	Undetermined	1
Indication	F_Reproductive_Obstetric	Uterus	7
Indication	Head_Neck	Ear	2
Indication	Head_Neck	Mouth	8
Indication	Head_Neck	Nasal_Sinus	3
Indication	Head_Neck	Neck	18
Indication	Head_Neck	Pharynx	3
Indication	Head_Neck	Thyroid	2
Indication	Head_Neck	Undetermined	28
Indication	Hepato-Biliary	Gallblader	1
Indication	Hepato-Biliary	Liver	29
Indication	Hepato-Biliary	Pancreas	7
Indication	Hepato-Biliary	Undetermined	1
Indication	Indication		1494
Indication	Indication_Type		1
Indication	Indication_Type	neoplastic_dx	558
Indication	Indication_Type	nonneoplastic_dx	305
Indication	Indication_Type	symptom	433
Indication	Indication_Type	trauma	76
Indication	Lymphatic	Undetermined	11
Indication	M_Reproductive	Prostate	10
Indication	M_Reproductive	Testis	3
Indication	Miscellaneous	Biomedical_Device	2
Indication	Miscellaneous	Connective_Tissue	2
Indication	Miscellaneous	Undetermined	2
Indication	Musculo-Skeletal	Bone_and_or_Joint	83
Indication	Musculo-Skeletal	Skeletal_and_or_Smooth_Muscle	4
Indication	Neurological	Brain	37
Indication	Neurological	Cerebrospinal_Fluid_Pathway	2
Indication	Neurological	Cerebrovascular_System	7
Indication	Neurological	Extraaxial	13

Continued on next page

Event type	Type	Subtype	Count
Indication	Neurological	Nerve	9
Indication	Neurological	Pituitary	2
Indication	Neurological	Spine_Cervical	14
Indication	Neurological	Spine_Cord	1
Indication	Neurological	Spine_Lumbar	19
Indication	Neurological	Spine_Thoracic	4
Indication	Neurological	Spine_Unspecified	2
Indication	Neurological	Undetermined	6
Indication	Respiratory	Lung	62
Indication	Respiratory	Pleural_Membrane	3
Indication	Skin	Skin_and_or_Mucous_Membrane	2
Indication	Thoracic	Mediastinal	8
Indication	Thoracic	Undetermined	19
Indication	Urinary	Kidney	10
Indication	Urinary	Urinary_Bladder	7
Lesion			
Lesion	Abdomen		4
Lesion	Abdomen	Abdominal_Wall	22
Lesion	Abdomen	Adrenal_Gland	53
Lesion	Abdomen	Mesentery	47
Lesion	Abdomen	Peritoneal_Sac	39
Lesion	Abdomen	Retroperitoneal	160
Lesion	Abdomen	Spleen	74
Lesion	Abdomen	Undetermined	376
Lesion	Assertion	absent	1617
Lesion	Assertion	possible	378
Lesion	Assertion	present	3714
Lesion	Body_Regions	Entire_Body	5
Lesion	Body_Regions	Lower_Limb	79
Lesion	Body_Regions	Pelvis	354
Lesion	Body_Regions	Undetermined	42

Continued on next page

Event type	Type	Subtype	Count
Lesion	Body_Regions	Upper_Limb	162
Lesion	Cardiovascular	Arterial	27
Lesion	Cardiovascular	Heart	33
Lesion	Cardiovascular	Pericardial_Sac	12
Lesion	Cardiovascular	Pulmonary_Artery	2
Lesion	Cardiovascular	Undetermined	1
Lesion	Cardiovascular	Venous	10
Lesion	Characteristic		2563
Lesion	Count		381
Lesion	Digestive	Esophagus	42
Lesion	Digestive	Intestine	29
Lesion	Digestive	Large_Intestine	39
Lesion	Digestive	Small_Intestine	25
Lesion	Digestive	Stomach	32
Lesion	F_Reproductive_Obstetric	Adnexal	6
Lesion	F_Reproductive_Obstetric	Breast	91
Lesion	F_Reproductive_Obstetric	Female_Genital_Structure	19
Lesion	F_Reproductive_Obstetric	Ovary	18
Lesion	F_Reproductive_Obstetric	Uterus	12
Lesion	Head_Neck		1
Lesion	Head_Neck	Ear	4
Lesion	Head_Neck	Eye	1
Lesion	Head_Neck	Laryngeal	5
Lesion	Head_Neck	Mouth	38
Lesion	Head_Neck	Nasal_Sinus	91
Lesion	Head_Neck	Neck	457
Lesion	Head_Neck	Pharynx	12
Lesion	Head_Neck	Thyroid	65
Lesion	Head_Neck	Undetermined	104
Lesion	Hepato-Biliary	Bile_Duct	9
Lesion	Hepato-Biliary	Gallblader	5

Continued on next page

Event type	Type	Subtype	Count
Lesion	Hepato-Biliary	Liver	341
Lesion	Hepato-Biliary	Pancreas	39
Lesion	Lesion		5709
Lesion	Lymphatic	Undetermined	587
Lesion	M_Reproductive	Prostate	13
Lesion	M_Reproductive	Undetermined	3
Lesion	Miscellaneous	Biomedical_Device	11
Lesion	Miscellaneous	Connective_Tissue	10
Lesion	Miscellaneous	Undetermined	3
Lesion	Musculo-Skeletal	Bone_and_or_Joint	852
Lesion	Musculo-Skeletal	Skeletal_and_or_Smooth_Muscle	104
Lesion	Musculo-Skeletal	Undetermined	6
Lesion	Neurological	Brain	430
Lesion	Neurological	Cerebrospinal_Fluid_Pathway	19
Lesion	Neurological	Cerebrovascular_System	20
Lesion	Neurological	Extraaxial	99
Lesion	Neurological	Nerve	14
Lesion	Neurological	Pituitary	3
Lesion	Neurological	Spine_Cervical	67
Lesion	Neurological	Spine_Cord	39
Lesion	Neurological	Spine_Lumbar	81
Lesion	Neurological	Spine_Sacral	34
Lesion	Neurological	Spine_Thoracic	111
Lesion	Neurological	Spine_Unspecified	18
Lesion	Neurological	Undetermined	36
Lesion	Respiratory	Lung	658
Lesion	Respiratory	Pleural_Membrane	40
Lesion	Respiratory	Tracheobronchial	27
Lesion	Respiratory	Undetermined	4
Lesion	Size		6
Lesion	Size	current	1071

Continued on next page

Event type	Type	Subtype	Count
Lesion	Size	past	150
Lesion	Size_Trend	decreasing	107
Lesion	Size_Trend	disappear	70
Lesion	Size_Trend	increasing	130
Lesion	Size_Trend	new	171
Lesion	Size_Trend	no_change	392
Lesion	Skin	Skin_and_or_Mucous_Membrane	25
Lesion	Skin	Subcutaneous	8
Lesion	Skin	Undetermined	1
Lesion	Thoracic	Mediastinal	325
Lesion	Thoracic	Undetermined	379
Lesion	Urinary	Kidney	187
Lesion	Urinary	Undetermined	6
Lesion	Urinary	Ureter	12
Lesion	Urinary	Urinary_Bladder	11
Medical_Problem			
Medical_Problem	Abdomen		1
Medical_Problem	Abdomen	Abdominal_Wall	31
Medical_Problem	Abdomen	Adrenal_Gland	21
Medical_Problem	Abdomen	Mesentery	9
Medical_Problem	Abdomen	Peritoneal_Sac	38
Medical_Problem	Abdomen	Retroperitoneal	14
Medical_Problem	Abdomen	Spleen	37
Medical_Problem	Abdomen	Undetermined	140
Medical_Problem	Assertion	absent	1830
Medical_Problem	Assertion	possible	724
Medical_Problem	Assertion	present	3700
Medical_Problem	Body_Regions	Entire_Body	3
Medical_Problem	Body_Regions	Lower_Limb	64
Medical_Problem	Body_Regions	Pelvis	111
Medical_Problem	Body_Regions	Undetermined	23

Continued on next page

Event type	Type	Subtype	Count
Medical_Problem	Body_Regions	Upper_Limb	25
Medical_Problem	Cardiovascular		2
Medical_Problem	Cardiovascular	Arterial	263
Medical_Problem	Cardiovascular	Coronary_Artery	100
Medical_Problem	Cardiovascular	Heart	199
Medical_Problem	Cardiovascular	Pericardial_Sac	57
Medical_Problem	Cardiovascular	Pulmonary_Artery	18
Medical_Problem	Cardiovascular	Undetermined	40
Medical_Problem	Cardiovascular	Venous	66
Medical_Problem	Digestive	Esophagus	25
Medical_Problem	Digestive	Intestine	53
Medical_Problem	Digestive	Large_Intestine	104
Medical_Problem	Digestive	Small_Intestine	30
Medical_Problem	Digestive	Stomach	35
Medical_Problem	F_Reproductive_Obstetric	Adnexal	2
Medical_Problem	F_Reproductive_Obstetric	Breast	25
Medical_Problem	F_Reproductive_Obstetric	Female_Genital_Structure	4
Medical_Problem	F_Reproductive_Obstetric	Ovary	2
Medical_Problem	F_Reproductive_Obstetric	Uterus	14
Medical_Problem	F_Reproductive_Obstetric	Undetermined	2
Medical_Problem	Hepato-Biliary	Bile_Duct	70
Medical_Problem	Hepato-Biliary	Gallblader	36
Medical_Problem	Hepato-Biliary	Liver	95
Medical_Problem	Hepato-Biliary	Pancreas	37
Medical_Problem	Hepato-Biliary	Undetermined	1
Medical_Problem	Head_Neck		4
Medical_Problem	Head_Neck	Eye	10
Medical_Problem	Head_Neck	Mouth	17
Medical_Problem	Head_Neck	Nasal_Sinus	167
Medical_Problem	Head_Neck	Neck	55
Medical_Problem	Head_Neck	Pharynx	7

Continued on next page

Event type	Type	Subtype	Count
Medical_Problem	Head_Neck	Thyroid	20
Medical_Problem	Head_Neck	Undetermined	98
Medical_Problem	Lymphatic	Undetermined	9
Medical_Problem	M_Reproductive	Prostate	19
Medical_Problem	M_Reproductive	Undetermined	5
Medical_Problem	Medical_Problem		6254
Medical_Problem	Miscellaneous	Adipose_Tissue	1
Medical_Problem	Miscellaneous	Biomedical_Device	8
Medical_Problem	Miscellaneous	Connective_Tissue	26
Medical_Problem	Musculo-Skeletal	Bone_and_or_Joint	788
Medical_Problem	Musculo-Skeletal	Skeletal_and_or_Smooth_Muscle	168
Medical_Problem	Musculo-Skeletal	Undetermined	5
Medical_Problem	Neurological	Brain	378
Medical_Problem	Neurological	Cerebrospinal_Fluid_Pathway	140
Medical_Problem	Neurological	Cerebrovascular_System	109
Medical_Problem	Neurological	Extraaxial	49
Medical_Problem	Neurological	Nerve	37
Medical_Problem	Neurological	Pituitary	6
Medical_Problem	Neurological	Spine_Cervical	363
Medical_Problem	Neurological	Spine_Cord	175
Medical_Problem	Neurological	Spine_Lumbar	860
Medical_Problem	Neurological	Spine_Sacral	26
Medical_Problem	Neurological	Spine_Thoracic	166
Medical_Problem	Neurological	Spine_Unspecified	99
Medical_Problem	Neurological	Undetermined	113
Medical_Problem	Respiratory	Lung	342
Medical_Problem	Respiratory	Pleural_Membrane	180
Medical_Problem	Respiratory	Tracheobronchial	27
Medical_Problem	Respiratory	Undetermined	2
Medical_Problem	Skin	Skin_and_or_Mucous_Membrane	22
Medical_Problem	Skin	Subcutaneous	5

Continued on next page

Event type	Type	Subtype	Count
Medical_Problem	Thoracic	Mediastinal	35
Medical_Problem	Thoracic	Undetermined	81
Medical_Problem	Urinary	Kidney	122
Medical_Problem	Urinary	Undetermined	2
Medical_Problem	Urinary	Ureter	29
Medical_Problem	Urinary	Urinary_Bladder	20

A.3 Annotation Guidelines for CAMIR

The goal of this annotation project is to label four types of events in radiology reports: Indication, Medical Problem, Lesion, and attributes related to each of these findings. Each event is represented by a trigger span and a set of attributes that capture detailed contextual information. Some attributes are span-with-value types such as assertion or anatomy, while others are span-only such as size.

Event Overview

- **Indication**: The clinical reason for obtaining the imaging study.
- **Medical Problem**: Abnormalities identified by imaging that do not represent lesions.
- **Lesion**: Focal abnormalities requiring detailed annotation, including assertion, anatomy, size, size trend, count, and characteristic.

Trigger

Table A.3 describes examples of medical problem and lesion triggers.

Table A.3: Medical Problem and Lesion triggers.

Medical Problem	**Lesion**
air trapping, aneurysmal, ascites, atelectasis, atrophic, calcifications, cirrhosis, dilation, distention, ectatic, effusions, emboli, embolisms, embolus, fat stranding, fractures, hernias, honeycombing, hypertension, injury, intussusceptions, mosaic attenuation, necrosis, pneumothorax, reticulation, scarring, stones, stricture, thickening, tortuous	abscesses, adenocarcinoma, consolidations, cysts, fluid collections, focus, granulomas, hamartomas, hemangioma, lesions, lipomas, lymphadenopathy, masses, metastases, neoplasm, nodes (enlarged or abnormal), nodularity, nodules, sarcomas, foci

Assertion

Assertion indicates whether an event is present, absent, or possible. Absent and possible assertions usually contain explicit cues such as "denies" or "likely", while present assertions often do not include an explicit cue.

Table A.4: Assertion examples for absent and possible.

Assertion: Absent	**Assertion: Possible**
absent, denies, no, no evidence	cannot be excluded, concern, consistent, could be, could represent, likely, likely indicates, may be, possible, presumed, question, suggestive, suggests

(1) Indication

Indication consists of a concise statement of the reason for the imaging study. These statements appear near the beginning of the report under headings such as "Indication," "Clinical Indication," or "History." Indication events include a trigger, type, assertion, and anatomy.

1.1 Indication – Trigger (required) The trigger describes the reason for the imaging test. These spans commonly reference trauma events, diagnoses, or symptoms.

Table A.5: Indication triggers.

Example	**Notes**
INDICATION: High speed motor vehicle collision.	Events that cause a need for imaging, such as accidents or falls, are labeled as indication.
INDICATION: MVC	Shortened forms of trauma mechanisms remain indication triggers.
Clinical Indication: question of pulmonary sarcoid	Captures diagnostic concerns prompting imaging.
INDICATION: Found down	Represents symptom-based clinical presentations.
INDICATION: abdominal pain.	The symptom is the trigger. Anatomy such as "abdominal" is annotated separately.
INDICATION: Abnormal — CT will be repeated at [LOCATION]	Indication may reference findings from a previous exam.

1.2 Indication – Type Type categorizes the nature of the clinical reason for imaging: Trauma, Symptom, Neoplastic Diagnosis, or Non-neoplastic Diagnosis. The trigger and type generally share the same span.

1.3 Indication – Assertion (required) Assertions indicate whether the concern prompting imaging is present, absent, or possible. All indication events require an assertion.

Table A.6: Indication types.

Type	Example Sentence
Trauma	INDICATION: High speed motor vehicle accident. INDICATION: MVC.
Symptom	INDICATION: found down.
Neoplastic Diagnosis	Clinical History: x year old man with seminoma.
Non-neoplastic Diagnosis	Clinical History: concern for amiodarone toxicity. Clinical Indication: question of pulmonary sarcoid.

Table A.7: Indication assertion categories.

Value	Example and Notes
present	abd pain, bloody diarrhea; follow-up of primary CNS lymphoma. These cases describe diagnoses or symptoms.
absent	Negation is uncommon in indication, but may occur.
possible	r/o abscess; question of pulmonary sarcoid; concern for PE. Possible cases capture uncertainty or rule-out reasoning.

1.4 Indication – Anatomy Anatomy identifies the body part associated with the indication. Each anatomical region referenced in the indication should be annotated separately.

Table A.8: Indication anatomy examples.

Example Sentence	Notes
Indication: Suspected pulmonary embolus.	"pulmonary" identifies the region of concern.
INDICATION: R-sided chest pain.	Side-specific anatomy is annotated separately.
Invasive lobular carcinoma of the right breast.	The anatomical region belongs to the indication event.

(2) Medical Problem Finding

Medical Problem findings include abnormalities discovered by imaging that do not qualify as lesions. These events include trigger, assertion, and anatomy.

2.1 Medical Problem – Trigger (required) The trigger identifies the medical problem described in the text. Spans may be single or multiword expressions, such as fracture or osteophyte formation. Descriptors such as cavitation or air-fluid level should not be annotated as triggers.

Table A.9: Medical Problem triggers.

Examples: There is a nondisplaced linear fracture of the lateral epicondyle; Heart: no pericardial fluid; The liver is nodular and atrophic; Increased amount of ascites; No evidence of renal stones or scarring; The thoracic aorta is tortuous and ectatic; Aneurysmal thoracic aorta; Sarcoma with central hypodense focus; Aortic valve calcification; Increased pulmonary artery diameter; Cysts related to ductal plate anomalies.

2.2 Medical Problem – Assertion (required) Assertions indicate whether the medical problem is present, absent, or possible.

2.3 Medical Problem – Anatomy Anatomy identifies the specific body part associated with the medical problem. The most specific anatomical description should be annotated.

(3) Lesion Finding

Lesion findings describe focal abnormalities and include trigger, assertion, anatomy, size, size trend, count, and characteristic.

Table A.10: Medical Problem assertion examples.

Value	**Example Sentence**
present	Patient is status post distal esophagectomy; lucency consistent with osteophyte formation.
absent	No acute osseous abnormality.
possible	Possible nondisplaced fracture; mildly nodular liver possibly representing cirrhosis; consider postobstructive pneumonia; consider carcinomatous process.

Table A.11: Medical Problem anatomy examples.

Example Sentence	**Notes**
Stable subcentimeter low-attenuation in the left thyroid gland.	Use the most specific anatomical span.
Right kidney: No injury.	Anatomy in section headers can be annotated if linked to a problem.
Small bowel intussusceptions in the left abdomen.	Multiple anatomy spans may apply.
The liver appears increased in density.	Organ-specific.
Tree-in-bud opacities in the right middle lobe.	Use entire note if needed to determine anatomical region.

3.1 Lesion – Trigger The trigger is the descriptive term for the lesion. Common triggers include mass, node, nodule, opacity, or tumor.

Additional notes: (1) A phrase may include both trigger and anatomy, which should be annotated separately. (2) Lymph nodes qualify as lesions only if enlarged or cancer-related. (3) Normal or reactive lymph nodes are not annotated.

Table A.12: Lesion trigger examples.

prominent nodes in the right lower quadrant; partially calcified peripheral pulmonary nodule possibly representing hamartoma; nonenlarged mesenteric lymph nodes; hepatic hypodensity too small to characterize; lobular cyst; mediastinal lymphadenopathy; soft tissue density mass; numerous small nodular densities; increased number and size of nodular opacities; no abnormal enhancement.

3.2 Lesion – Assertion Assertions capture whether the lesion is present, absent, or possible.

Table A.13: Lesion assertion examples.

Value	**Example Sentence**
present	Intense uptake within a small nodule consistent with invasive adenocarcinoma.
absent	No suspicious enhancing nodule; no intracystic septations; no lymphadenopathy; primary lesion not visualized.
possible	Hypovascular round lesions likely representing satellite nodules; likely ovarian cystic neoplasm; lung nodule with possible metastases.

3.3 Lesion – Anatomy Anatomy captures the body part associated with the lesion.

3.4 Lesion Size Size captures dimensions or descriptive size modifiers.

Size is not annotated when the assertion is absent.

3.5 Lesion Size Trend Size trend identifies change in size, classified as new, disappear, increasing, decreasing, or no-change.

Table A.14: Lesion anatomy examples.

Example Sentence	**Notes**
Prominent nodes in the right lower quadrant.	Location linked to lesion.
Left lower lobe pulmonary nodule.	Multi-region specification.
Right hepatic lobe hypodensity.	Organ-based.
Innumerable hepatic cysts.	Multiple anatomy spans possible.
No mediastinal, hilar, or axillary lymph node enlargement.	Separate spans for each region.

Table A.15: Lesion size examples.

Value	**Example**
Present	The lesion measures 39.8 × 23.4 cm.
Past	Multiple new small cysts noted previously.
Present	Upper lobe nodules up to 5 mm.
Present	Pretracheal lymph node measuring 9 mm.

Table A.16: Lesion size trend examples.

Trend	**Examples**
new	Newly identified hypodense mass.
disappear	Previously seen mass no longer present.
increasing	Nodule increased from 0.5 cm to 0.6 cm; increased pleural effusions.
decreasing	Hepatic lesion decreasing from 1.5 cm to 0.6 cm.
no-change	Adrenal nodule unchanged since previous exam.

3.6 Lesion Count Count identifies the number of lesions or uses numeric modifiers such as multiple, several, or numerous.

3.7 Lesion Characteristic Characteristic describes specific lesion qualities such as peripheral, destructive, indeterminate, or enlarged.

Table A.17: Lesion characteristic examples.

Characteristic	**Example Sentence**
peripheral	Two peripheral nodules in the right upper lobe.
destructive	No destructive osseous lesion.
indeterminate	Indeterminate 3 mm nodules in the right upper lobe.
enlarged	No enlarged cervical lymph nodes seen.

A.4 Annotation Agreement per Rounds

Figure. A.1 demonstrates the trend in inter-annotator agreement per rounds.

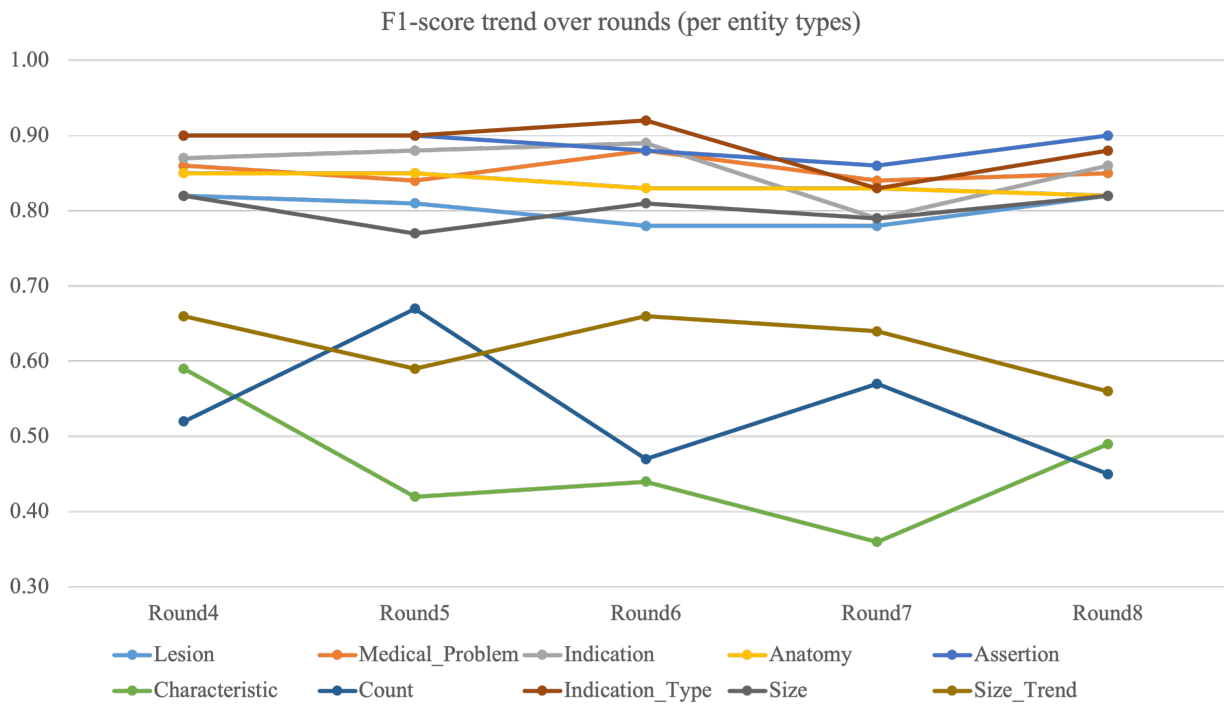

Figure A.1: F1-score trends across annotation rounds for each entity type. Includes only double-annotated rounds.

Appendix B

FOLLOW-UP IDENTIFICATION IN RADIOLOGY REPORTS

B.1 Implementation Details for Supervised Models

Table B.1 describes configurations used for supervised models for follow-up identification in radiology reports.

Table B.1: Model configurations and training hyperparameters.

Model	**Details**
TF-IDF Vectorizer	min_df = 100; stop_words = True
SVM	random_state = 51; class_weight = balanced; kernel = sigmoid; C = 0.1
Logistic Regression (LR)	class_weight = balanced; solver = lbfgs; penalty = l2; C = 0.1
Longformer (allenai/longformer-base-4096)	max_seq_length = 2048; train_batch_size = 16; learning rate = 5e-5; gradient_acc_steps = 4; epochs = 20; warmup_steps = 500; weight_decay = 0.01
LLaMA3-8B (LLaMA3-8B-instruct)	gradient_accumulation_steps = 1; learning_rate = 5e-05; epochs = 10; max_seq_length = 2048; bf16 = True; per_device_train_batch_size = 4; seed = 42; warmup_ratio = 0.1

B.2 Follow-Up Exam Annotation Guidelines

This appendix describes the annotation process for identifying follow-up imaging examinations. Definitions, clinical context, and annotation instructions are included to ensure consistent labeling across annotators. All content in this appendix is derived from the Follow-Up Exam Annotation Guidelines.

B.2.1 Definitions

Index / Source Report: A radiology report containing a follow-up recommendation. This is the reference point from which subsequent imaging is evaluated.

Candidate Report: Any radiology report that occurs chronologically after the index/source report. All imaging modalities may serve as candidate reports.

Each index/source report is presented to annotators through OneDrive, while candidate reports are presented using BRAT. Both the source and candidate reports include metadata such as modality, exam description, date, and time. Candidate reports additionally display the number of days since the index exam.

B.2.2 Annotation Task

For each index/source report, annotators review all subsequent radiology reports in chronological order and label the **first report that meets criteria** as a follow-up exam.

Only one follow-up report is annotated per index/source exam.

B.2.3 Clinical Context and Annotation Principles

Although many radiology studies contain incidental findings, this task focuses specifically on determining whether a subsequent exam qualifies as follow-up for the findings or recommendations described in the index/source report.

Anatomy Criterion

A candidate exam must include imaging of the **same anatomical region** as the finding or lesion referenced in the source report. If the anatomy does not match, the exam should not be labeled as follow-up.

Incidentalomas and Routine Surveillance

Most reports in the cohort do not contain incidentalomas, and many patients undergo routine disease surveillance unrelated to incidental findings. Examples include:

- serial CT scans for known lung cancer
- imaging related to ongoing therapy monitoring

Radiologists often do not issue explicit follow-up recommendations in surveillance cases, because imaging intervals follow standard practice or treatment-driven schedules.

Follow-Up When the Index Exam Is Already Part of a Follow-Up Chain

The index exam may itself be a follow-up for a previously identified lesion. Even in these cases, the annotator must:

- treat the given index report as the reference point, and
- label the first reasonable candidate exam after the index as follow-up

For example, if an incidental lung nodule was identified months earlier and the index report is already follow-up at 3 months, then a 6- or 12-month candidate exam still qualifies as follow-up.

Opportunistic Follow-Up

A candidate exam qualifies as follow-up even when the imaging was performed for an unrelated reason, as long as:

- the anatomy matches the lesion or finding in the index exam, and
- the exam timing is reasonable based on clinical norms

For example, a CT chest obtained after a motor vehicle collision should still be labeled as follow-up if the index exam contained an incidental lung nodule.

Caution: Very early opportunistic imaging may not represent meaningful follow-up if it would not normally allow characterization of the lesion.

Opportunistic Exams That Resolve the Need for Further Follow-Up

Occasionally, an opportunistic exam may conclusively characterize the lesion such that further follow-up is unnecessary. Annotators should still label that exam as follow-up, since it fulfills the clinical purpose of the prior recommendation.

Multiple Lesions or Multiple Organs

Some imaging studies contain findings across several organs. Annotators should:

- determine whether the candidate exam appropriately evaluates the lesion described in the index report, and
- consult a radiologist when uncertain

Follow-Up Using Different Imaging Modalities

A follow-up exam may be performed using a different modality if it is clinically appropriate for lesion characterization. Examples include:

- thyroid nodule on CT → ultrasound
- adrenal lesion on CT → MRI or dedicated CT for adenoma evaluation
- liver nodule on CT → MRI or multiphase CT
- lung opacity on X-ray → CT to assess reality or malignancy
- lung nodule on CT → PET-CT
- pancreatic or renal lesion → CT or MRI characterization

Thus, modality changes do not disqualify a candidate exam from being labeled as follow-up.

B.2.4 Typical Follow-Up Intervals

In most cases, same-modality follow-up exams occur at least **six weeks** after the index exam. However, some exceptions apply:

- A less-specific modality (e.g., chest X-ray) may be rapidly followed by CT to confirm or characterize a finding.
- A radiologist may anticipate quick resolution of a finding (e.g., infection-like lung nodules) and suggest short-interval follow-up.

B.2.5 Standard Follow-Up Timing for Common Incidentalomas

Typical radiology practice guidelines include:

- **Incidental lung nodules:** Follow-up CT typically at 3, 6, 12, or 24 months depending on size and risk.
- **Incidental thyroid nodules:** Follow-up typically occurs months to years (1–24 months), often using ultrasound.

Appendix C

IDENTIFICATION OF INCIDENTALOMAS

C.1 Definition of an Incidentaloma

Table C.1: Condensed Incidentaloma Annotation Guidelines

Category	Definition and Examples
Yes (Incidentaloma)	**Definition:** A lesion incidentally discovered and unrelated to the study indication. Not part of routine surveillance imaging. **Important exclusions:** suspected or potential metastases when the study indication involves a known primary malignancy; surveillance for chronic diseases (e.g., cirrhosis with HCC screening); surveillance for known primary cancers. **Examples:** enhancing renal mass on CT for trauma; incidental liver lesion on CT for abdominal pain; second primary tumor (e.g., incidental RCC in a patient with unrelated cancer).
Indeterminate	**Definition:** Lesion may or may not be related to the study indication; relationship to clinical concern is unclear. **Examples:** lung nodule on colon cancer restaging CT that could represent metastasis, primary lung cancer, or benign granuloma.

C.2 Cohen's Kappa between Models

Table C.2 summarizes the pairwise agreement of lesion-level predictions across all evaluated models using Cohen's kappa. The results show that agreement is highest within model families, such as between GPT-4o Base and GPT-4o Anatomy, and between the GPT-OSS

variants. Encoder models (ModernBERT and BioClinical-ModernBERT) also show very strong mutual agreement, reflecting highly similar decision behavior. Cross-family agreement between generative LLMs and supervised encoders is lower, which indicates that LLM-based systems and encoder-based systems capture overlapping but not identical lesion-level patterns. These kappa values provide a quantitative view of how similar the different models' lesion-level outputs are.

Table C.2: Pairwise Cohen's kappa agreement between lesion-level predictions across all models. Values are truncated to two decimals.

	GPT4O_Base	GPT4O_Anatomy	GPTOSS_Base	GPTOSS_Anatomy	ModernBERT	BioClinModernBERT
GPT4O_Base	1.00	0.90	0.71	0.70	0.61	0.63
GPT4O_Anatomy	0.90	1.00	0.71	0.75	0.68	0.70
GPTOSS_Base	0.71	0.71	1.00	0.84	0.63	0.66
GPTOSS_Anatomy	0.70	0.75	0.84	1.00	0.64	0.66
ModernBERT	0.61	0.68	0.63	0.64	1.00	0.93
BioClinModernBERT	0.63	0.70	0.66	0.66	0.93	1.00

BIBLIOGRAPHY

[1] Introducing meta llama 3: The most capable openly available llm to date, 4 2024. URL https://ai.meta.com/blog/meta-llama-3/.

[2] Achiam et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Lisa C Adams, Daniel Truhn, Felix Busch, Avan Kader, Stefan M Niehues, Marcus R Makowski, and Keno K Bressem. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. *Radiology*, 307(4):e230725, 2023. doi: 10.1148/radiol.230725.

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

[5] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[6] American College of Radiology. Practice parameters and technical standards. https://www.acr.org/Clinical-Resources/Clinical-Tools-and-Reference/Practice-Parameters-and-Technical-Standards, 2025. Accessed: 2025-11-03.

[7] Wasif Bala, Jackson Steinkamp, Timothy Feeney, Avneesh Gupta, Abhinav Sharma, Jake Kantrowitz, Nicholas Cordella, James Moses, and Frederick Thurston Drake. A web application for adrenal incidentaloma identification, tracking, and management using machine learning. *Applied Clinical Informatics*, 11(04):606–616, 2020.

[8] Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, et al. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.

[9] Luisa Barzon and Marco Boscaro. Diagnosis and management of adrenal incidentalomas. *The Journal of urology*, 163(2):398–407, 2000.

[10] Luisa Barzon, Nicoletta Sonino, Francesco Fallo, Giorgio Palu, and Marco Boscaro. Prevalence and natural history of adrenal incidentalomas. *European journal of endocrinology*, 149(4):273–285, 2003.

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

[12] Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. An empirical investigation of statistical significance in NLP. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL `https://aclanthology.org/D12-1091`.

[13] Lincoln L Berland, Stuart G Silverman, Richard M Gore, William W Mayo-Smith, Andrew J Megibow, Judy Yee, and James A Brink. Managing incidental findings on abdominal ct: white paper of the acr incidental findings committee. *Journal of the American College of Radiology*, 7(10):754–773, 2010. doi: 10.1016/j.jacr.2010.06.013.

[14] Francesco Bertagna, Giorgio Treglia, Emanuela Orlando, Lodovica Dognini, Luca Giovanella, Ramin Sadeghi, and Raffaele Giubbini. Prevalence and clinical significance of incidental f18-fdg breast uptake: a systematic review and meta-analysis. *Japanese journal of radiology*, 32:59–68, 2014.

[15] Jérôme Bertherat, Helen Mosnier-Pudar, and Xavier Bertagna. Adrenal incidentalomas. *Current opinion in oncology*, 14(1):58–63, 2002.

[16] Rajesh Bhayana, Gavin Elias, Daksh Datta, Nishaant Bhambra, Yangqing Deng, and Satheesh Krishna. Use of gpt-4 with single-shot learning to identify incidental findings in radiology reports. *American Journal of Roentgenology*, 222(2):e2330651, 2024.

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

[18] Tommaso Cai, Athanasios A Giannopoulos, Sean Yu, Tarek Kelil, Beth Ripley, Kanako K Kumamaru, Frank J Rybicki, and Dimitrios Mitsouras. Natural language processing technologies in radiology research and clinical applications. *RadioGraphics*, 36(3): 738–753, 2016. doi: 10.1148/rg.2016150162.

[19] Matías F Callejas, Juan I Errázuriz, Felipe Castillo, Claudia Otárola, Carlos Riquelme, Claudia Ortega, Álvaro Huete, and Pablo Bächler. Incidental venous thromboembolism detected by pet-ct in patients with cancer: prevalence and impact on survival rate. *Thrombosis Research*, 133(5):750–755, 2014.

[20] Joanne L. Callen et al. Failure to follow-up test results for ambulatory patients: a systematic review. *J Gen Intern Med*, 27(10):1334–1348, October 2012. ISSN 1525-1497. doi: 10.1007/s11606-011-1949-5.

[21] Emmanuel Carrodeguas, Ronilda Lacson, Whitney Swanson, and Ramin Khorasani. Use of machine learning to identify follow-up recommendations in radiology reports. *Journal of the American College of Radiology*, 16(3):336–343, 2019. doi: 10.1016/j.jacr.2018.10.020.

[22] Arlene Casey, Emma Davidson, Michael Poon, Hang Dong, Daniel Duma, Andreas 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, June 2021. ISSN 1472-6947. doi: 10.1186/s12911-021-01533-7. URL https://doi.org/10.1186/s12911-021-01533-7.

[23] Rita Maria Chidiac and David C Aron. Incidentalomas: a disease of modern technology. *Endocrinology and Metabolism Clinics*, 26(1):233–253, 1997.

[24] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[25] Laila R Cochon, Neena Kapoor, Emmanuel Carrodeguas, Ivan K Ip, Ronilda Lacson, Giles Boland, and Ramin Khorasani. Variation in follow-up imaging recommendations in radiology reports: patient, modality, and radiologist predictors. *Radiology*, 291(3): 700–707, 2019.

[26] Tessa S. Cook, Darco Lalevic, Caroline Sloan, Seetharam C. Chadalavada, Curtis P. Langlotz, Mitchell D. Schnall, and Hanna M. Zafar. Implementation of an Automated Radiology Recommendation-Tracking Engine for Abdominal Imaging Findings of Possible Cancer. *J Am Coll Radiol*, 14(5):629–636, May 2017. ISSN 1558-349X. doi: 10.1016/j.jacr.2017.01.024.

[27] Sandeep Dalal, Vadiraj Hombal, Wei-Hung Weng, Gabe Mankovich, Thusitha Mabotuwana, Christopher S. Hall, Joseph Fuller, Bruce E. Lehnert, and Martin L. Gunn. Determining Follow-Up Imaging Study Using Radiology Reports. *J Digit Imaging*, 33 (1):121–130, February 2020. ISSN 1618-727X. doi: 10.1007/s10278-019-00260-w.

[28] Surabhi Datta and Kirk Roberts. Fine-grained spatial information extraction in radiology as two-turn question answering. *International Journal of Medical Informatics*, 158:104628, 2022. doi: 10.1016/j.ijmedinf.2021.104628.

[29] Surabhi Datta, Jordan Godfrey-Stovall, and Kirk Roberts. Radlex normalization in radiology reports. In *AMIA Annual Symposium Proceedings*, volume 2020, page 338. American Medical Informatics Association, 2020.

[30] Surabhi Datta, Yuqi Si, Laritza Rodriguez, Sonya E Shooshan, Dina Demner-Fushman, and Kirk Roberts. Understanding spatial language in radiology: Representation framework, annotation, and spatial relation extraction from chest x-ray reports using deep learning. *Journal of Biomedical Informatics*, 108:103473, 2020. doi: 10.1016/j.jbi.2020.103473.

[31] Marco Del Chiaro, Robert J Torphy, and Richard D Schulick. Pancreatic incidentalomas: Investigation and management. *Journal of Internal Medicine*, 290(5):969–979, 2021.

[32] Dina Demner-Fushman, Wendy W Chapman, and Clement J McDonald. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772, 2009. doi: 10.1016/j.jbi.2009.08.007.

[33] Francesco Dentali, Walter Ageno, Cecilia Becattini, Luca Galli, Monica Gianni, Nicoletta Riva, Davide Imberti, Alessandro Squizzato, Achille Venco, and Giancarlo Agnelli. Prevalence and clinical history of incidental, asymptomatic pulmonary embolism: a meta-analysis. *Thrombosis research*, 125(6):518–522, 2010.

[34] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. doi: 10.18653/v1/N19-1423.

[35] Richard K. G. Do, Kaelan Lupton, Pamela I. Causa Andrieu, Anisha Luthra, Michio Taya, Karen Batch, Huy Nguyen, Prachi Rahurkar, Lior Gazit, Kevin Nicholas, Christopher J. Fong, Natalie Gangai, Nikolaus Schultz, Farhana Zulkernine, Varadan

Sevilimedu, Krishna Juluru, Amber Simpson, and Hedvig Hricak. Patterns of metastatic disease in patients with cancer derived from natural language processing of structured CT radiology reports over a 10-year period. *Radiology*, 301(1): 115–122, October 2021. ISSN 0033-8419. doi: 10.1148/radiol.2021210043. URL https://pubs.rsna.org/doi/full/10.1148/radiol.2021210043.

[36] Kevin Donnelly et al. Snomed-ct: The advanced terminology and coding system for ehealth. *Stud Health Technol Inform*, 121:279, 2006. PMID: 17095826.

[37] Lane F. Donnelly, Robert Grzeszczuk, and Carolina V. Guimaraes. Use of natural language processing (NLP) in evaluation of radiology reports: An update on applications and technology advances. *Seminars in Ultrasound, CT and MRI*, 43(2):176–181, April 2022. doi: 10.1053/j.sult.2022.02.007. URL https://doi.org/10.1053/j.sult.2022.02.007.

[38] Sayon Dutta, William J Long, David FM Brown, and Andrew T Reisner. Automated detection using natural language processing of radiologists recommendations for additional imaging of incidental findings. *Annals of emergency medicine*, 62(2):162–169, 2013.

[39] Markus Eberts and Adrian Ulges. Span-based joint entity and relation extraction with transformer pre-training. In *24th ECAI*, 2020. URL https://ebooks.iospress.nl/volumearticle/55116.

[40] Christopher S Evans, Hugh D Dorris, Michael T Kane, Benjamin Mervak, Jane H Brice, Benjamin Gray, and Carlton Moore. A natural language processing and machine learning approach to identification of incidental radiology findings in trauma patients discharged from the emergency department. *Annals of Emergency Medicine*, 81(3): 262–269, 2023.

[41] Timothy Feeney, Stephanie Talutis, Megan Janeway, Praveen Sridhar, Avneesh Gupta,

Philip E Knapp, James Moses, David McAneny, and Frederick Thurston Drake. Evaluation of incidental adrenal masses at a tertiary referral and trauma center. *Surgery*, 167(5):868–875, 2020.

[42] Timothy Feeney, Andrea Madiedo, Philip E Knapp, Avneesh Gupta, David McAneny, and Frederick Thurston Drake. Incidental adrenal masses: adherence to guidelines and methods to improve initial follow-up: a systematic review. *Journal of Surgical Research*, 269:18–27, 2022.

[43] Matthias A Fink, Arved Bischoff, Christoph A Fink, Martin Moll, Jonas Kroschke, Luca Dulz, Claus Peter Heußel, Hans-Ulrich Kauczor, and Tim F Weber. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. *Radiology*, 308(3):e231362, 2023. doi: 10.1148/radiol.231362.

[44] Andrea Franconeri, Jieming Fang, Benjamin Carney, Almamoon Justaniah, Laura Miller, Hye-Chun Hur, Louise P King, Roa Alammari, Salomao Faintuch, Koenraad J Mortele, et al. Structured vs narrative reporting of pelvic mri for fibroids: clarity and impact on treatment planning. *European radiology*, 28(7):3009–3017, 2018.

[45] Alfonso Emilio Gerevini, Alberto Lavelli, Alessandro Maffi, Roberto Maroldi, Anne-Lyse Minard, Ivan Serina, and Guido Squassina. Automatic classification of radiological reports for clinical care. *Artificial intelligence in medicine*, 91:72–81, 2018.

[46] Axel Gerstmair, Philipp Daumke, Kai Simon, Mathias Langer, and Elmar Kotter. Intelligent image retrieval based on radiology reports. *European Radiology*, 22(12): 2750–2758, 2012. doi: 10.1007/s00330-012-2608-x.

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

[48] AT Grady, JA Sosa, TP Tanpitukpongse, KR Choudhury, RT Gupta, and JK Hoang.

Radiology reports for incidental thyroid nodules on ct and mri: high variability across subspecialties. *American Journal of Neuroradiology*, 36(2):397–402, 2015.

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

[50] Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023.

[51] Saeed Hassanpour and Curtis P. Langlotz. Information extraction from multi-institutional radiology reports. *Artificial Intelligence in Medicine*, 66:29–39, January 2016. ISSN 0933-3657. doi: 10.1016/j.artmed.2015.09.007. URL `https://www.sciencedirect.com/science/article/pii/S0933365715001244`.

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

[53] Nathan Hitzeman and Erin Cotton. Incidentalomas: initial management. *American Family Physician*, 90(11):784–789, 2014.

[54] George Hripcsak and Adam S Rothschild. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3): 296–298, 2005.

[55] 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, 2024.

[56] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[57] Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, et al. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31(9):1812–1820, 2024.

[58] Kevin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

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

[60] Stephanie L Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, et al. Maira-1: A specialised large multimodal model for radiology report generation. *arXiv preprint arXiv:2311.13668*, 2023.

[61] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jenna Seekins, David A Mong, Safwan Halabi, Jacob Sandberg, Russell 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 AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019. doi: 10.1609/aaai.v33i01.3301590.

[62] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng,

Curtis P. Langlotz, and Pranav Rajpurkar. RadGraph: Extracting clinical entities and relations from radiology reports. In *Neural Information Processing Systems*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/c8ffe9a587b126f152ed3d89a146b445-Paper-round1.pdf.

[63] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Sci. Data*, 3(1):1–9, 2016. doi: 10.1038/sdata.2016.35.

[64] Stella K Kang, Kira Garry, Ryan Chung, William H Moore, Eduardo Iturrate, Jordan L Swartz, Danny C Kim, Leora I Horwitz, and Saul Blecker. Natural language processing for identification of incidental pulmonary nodules in radiology reports. *Journal of the American College of Radiology*, 16(11):1587–1594, 2019.

[65] Gregory L Katzman, Azar P Dagher, and Nicholas J Patronas. Incidental findings on brain magnetic resonance imaging from 1000 asymptomatic volunteers. *Jama*, 282(1): 36–39, 1999.

[66] Sabrina Just Kousgaard, Michael Gade, Lars Jelstrup Petersen, and Ole Thorlacius-Ussing. Incidental detection of colorectal lesions on 18f-fdg-pet/ct is associated with high proportion of malignancy: A study in 549 patients. *Endoscopy International Open*, 8(12):E1725–E1731, 2020.

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

[68] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[69] Curtis P Langlotz. Radlex: a new method for indexing online educational materials, 2006.

[70] Paul A. Larson et al. Actionable findings and the role of IT support: report of the ACR Actionable Reporting Work Group. *J Am Coll Radiol*, 11(6):552–558, June 2014. ISSN 1558-349X. doi: 10.1016/j.jacr.2013.12.016.

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

[72] Wilson Lau, Kevin Lybarger, Martin L Gunn, and Meliha Yetisgen. Event-based clinical finding extraction from radiology reports with pre-trained language model. *Journal of Digital Imaging*, pages 1–14, 2022. doi: 10.1007/s10278-022-00717-5.

[73] Bastien Le Guellec, Alexandre Lefèvre, Charlotte Geay, Lucas Shorten, Cyril Bruge, Lotfi Hacein-Bey, Philippe Amouyel, Jean-Pierre Pruvo, Gregory Kuchcinski, and Aghiles Hamroun. Performance of an open-source large language model in extracting information from free-text radiology reports. *Radiology: Artificial Intelligence*, 6(4): e230364, 2024.

[74] Kahyun Lee, Nicholas J Dobbins, Bridget McInnes, Meliha Yetisgen, and Özlem Uzuner. Transferability of neural network clinical deidentification systems. *Journal of the American Medical Informatics Association*, 28(12):2661–2669, 09 2021. ISSN 1527-974X. doi: 10.1093/jamia/ocab207. URL https://doi.org/10.1093/jamia/ocab207.

[75] Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*, 2022.

[76] Qianchu Liu, Stephanie Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Maria Teodora Wetscherek, Robert Tinn, Harshita Sharma, Fernando Pérez-García, Anton Schwaighofer, et al. Exploring the boundaries of gpt-4 in radiology. *arXiv preprint arXiv:2310.14573*, 2023.

[77] Zhengliang Liu, Yiwei Li, Peng Shu, Aoxiao Zhong, Hanqi Jiang, Yi Pan, Longtao Yang, Chao Ju, Zihao Wu, Chong Ma, et al. Radiology-gpt: a large language model for radiology. *Meta-Radiology*, page 100153, 2025.

[78] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[79] Qiuhao Lu, Rui Li, Andrew Wen, Jinlian Wang, Liwei Wang, and Hongfang Liu. Large language models struggle in token-level clinical named entity recognition. In *AMIA Annual Symposium Proceedings*, volume 2024, page 748, 2025.

[80] Blanca Lumbreras, Lucas Donat, and Ildefonso Hernández-Aguado. Incidental findings in imaging diagnostic tests: a systematic review. *The British journal of radiology*, 83 (988):276–289, 2010.

[81] Kevin Lybarger, Aashka Damani, Martin Gunn, Özlem Uzuner, and Meliha Yetisgen. Extracting radiological findings with normalized anatomical information using a span-based BERT relation extraction model. In *AMIA Informatics Summit*, 2022. URL `https://pubmed.ncbi.nlm.nih.gov/35854739`.

[82] Kevin Lybarger, Nicholas J Dobbins, Ritche Long, Angad Singh, Patrick Wedgeworth, Özlem Ozuner, and Meliha Yetisgen. Leveraging natural language processing to augment structured social determinants of health data in the electronic health record. *Journal of the American Medical Informatics Association*, 2023. doi: 10.1093/jamia/ocad073.

[83] Chong Ma, Zihao Wu, Jiaqi Wang, Shaochen Xu, Yaonai Wei, Zhengliang Liu, Lei Guo, Xiaoyan Cai, Shu Zhang, Tuo Zhang, et al. Impressiongpt: an iterative optimizing framework for radiology report summarization with chatgpt. *arXiv preprint arXiv:2304.08448*, 2023.

[84] Thusitha Mabotuwana, Christopher S Hall, Joel Tieder, and Martin L. Gunn. Improving Quality of Follow-Up Imaging Recommendations in Radiology. *AMIA Annu Symp*

Proc, 2017:1196–1204, April 2018. ISSN 1942-597X. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977608/.

[85] Thusitha Mabotuwana, Christopher S Hall, Vadiraj Hombal, et al. Automated tracking of follow-up imaging recommendations. *American Journal of Roentgenology*, 212(6): 1287–1294, 2019. doi: 10.2214/AJR.18.20586.

[86] Heber MacMahon, David P Naidich, Jin Mo Goo, Kyung Soo Lee, Ann NC Leung, John R Mayo, Atul C Mehta, Yoshiharu Ohno, Charles A Powell, Mathias Prokop, et al. Guidelines for management of incidental pulmonary nodules detected on ct images: from the fleischner society 2017. *Radiology*, 284(1):228–243, 2017.

[87] Dominic I Maher, Evan Williams, Simon Grodski, Jonathan W Serpell, and James C Lee. Adrenal incidentaloma follow-up is influenced by patient, radiologic, and medical provider factors: a review of 804 cases. *Surgery*, 164(6):1360–1365, 2018.

[88] Hans-Jonas Meyer, Andreas Wienke, and Alexey Surov. Incidental pulmonary embolism in oncologic patients—a systematic review and meta-analysis. *Supportive Care in Cancer*, 29:1293–1302, 2021.

[89] Pritam Mukherjee, Benjamin Hou, Ricardo B Lanfredi, and Ronald M Summers. Feasibility of using the privacy-preserving large language model vicuna for labeling radiology reports. *Radiology*, 309(1):e231147, 2023. doi: 10.1148/radiol.231147.

[90] Keelin Murphy, Henk Smits, Arnoud JG Knoops, Michael BJM Korst, Tijs Samson, Ernst T Scholten, Steven Schalekamp, Cornelia M Schaefer-Prokop, Rick HHM Philipsen, Annet Meijers, et al. Covid-19 on chest radiographs: a multireader evaluation of an artificial intelligence system. *Radiology*, 296(3):E166–E172, 2020.

[91] Smriti Nayan, Jayant Ramakrishna, and Michael K Gupta. The proportion of malignancy in incidental thyroid lesions on 18-fdg pet study: a systematic review and meta-analysis. *Otolaryngology–Head and Neck Surgery*, 151(2):190–200, 2014.

[92] Daiki Nishigaki, Yuki Suzuki, Tomohiro Wataya, Kosuke Kita, Kazuki Yamagata, Junya Sato, Shoji Kido, and Noriyuki Tomiyama. Bert-based transfer learning in sentence-level anatomic classification of free-text radiology reports. *Radiology: Artificial Intelligence*, 5(2):e220097, 2023. doi: 10.1148/ryai.220097.

[93] J Martijn Nobel, Koos van Geel, and Simon GF Robben. Structured reporting in radiology: a systematic review to explore its potential. *European radiology*, 32(4): 2837–2854, 2022.

[94] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

[95] European Society of Radiology (ESR) communications@ myesr. org. Esr paper on structured reporting in radiology. *Insights into imaging*, 9(1):1–7, 2018.

[96] OpenAI. Gpt-4 technical report, 2023.

[97] K O'Shea. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[98] Jack W O'Sullivan, Tim Muntinga, Sam Grigg, and John PA Ioannidis. Prevalence and outcomes of incidental imaging findings: umbrella review. *bmj*, 361, 2018.

[99] Miguel Hernandez Pampaloni and Aung Z Win. Prevalence and characteristics of incidentalomas discovered by whole body fdg petct. *International Journal of Molecular Imaging*, 2012(1):476763, 2012.

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

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

[102] Namu Park, Farzad Ahmed, Zhaoyi Sun, Kevin Lybarger, Ethan Breinhorst, Julie Hu, Ozlem Uzuner, Martin Gunn, and Meliha Yetisgen. Automated identification of incidentalomas requiring follow-up: A multi-anatomy evaluation of llm-based and supervised approaches. *arXiv preprint arXiv:2512.05537*, 2025.

[103] Namu Park, Giridhar Kaushik Ramachandran, Kevin Lybarger, Fei Xia, Ozlem Uzuner, Meliha Yetisgen, and Martin Gunn. Identifying imaging follow-up in radiology reports: A comparative analysis of traditional ml and llm approaches. *arXiv preprint arXiv:2511.11867*, 2025.

[104] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*, 2019.

[105] Fernando Pérez-García, Sam Bond-Taylor, Pedro P Sanchez, Boris van Breugel, Daniel C Castro, Harshita Sharma, Valentina Salvatelli, Maria TA Wetscherek, Hannah Richardson, Matthew P Lungren, et al. Radedit: stress-testing biomedical vision models via diffusion image editing. *arXiv preprint arXiv:2312.12865*, 2023.

[106] Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Matthew P Lungren, et al. Rad-dino: Exploring scalable medical image encoders beyond text supervision. *arXiv preprint arXiv:2401.10815*, 2024.

[107] Filippo Pesapane, Priyan Tantrige, Paolo De Marco, Serena Carriero, Fabio Zugni, Luca Nicosia, Anna Carla Bozzini, Anna Rotili, Antuono Latronico, Francesca Abbate,

et al. Advancements in standardizing radiological reports: a comprehensive review. *Medicina*, 59(9):1679, 2023.

[108] Anne-Dominique Pham, Aurélie Névéol, Thomas Lavergne, Daisuke Yasunaga, Olivier Clément, Guy Meyer, Rémy Morello, and Anita Burgun. Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings. *BMC bioinformatics*, 15:1–10, 2014.

[109] Bryan Rink, Kirk Roberts, Sanda Harabagiu, Richard H Scheuermann, Seth Toomay, Travis Browning, Teresa Bosler, and Ronald Peshock. Extracting actionable findings of appendicitis from radiology reports using natural language processing. *AMIA Summits on Translational Science*, 2013:221, 2013. URL `https://pubmed.ncbi.nlm.nih.gov/24303268`.

[110] Daniel L Rubin and Charles E Kahn Jr. Common data elements in radiology. *Radiology*, 283(3):837–844, 2017. doi: 10.1148/radiol.2016161553.

[111] Max Schumm, Ming-Yeah Hu, Vivek Sant, Jiyoon Kim, Chi-Hong Tseng, Javier Sanz, Steven Raman, Run Yu, and Masha Livhits. Automated extraction of incidental adrenal nodules from electronic health records. *Surgery*, 173(1):52–58, 2023.

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

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

[114] Hassan Semaan, Tawfik Obri, Jacob Bieszczad, Paul Aldinger, Mohammed Al-Natour, Mohamad Bazerbashi, Nicholas Peters, and Hossein K Elgafy. Incidental extra-spinal findings in lumbar spine mri: Incidence and clinical significance. *The Spine Journal*, 15 (10):S197–S198, 2015.

[115] Joeky T Senders, Aditya V Karhade, David J Cote, Alireza Mehrtash, Nayan Lamba, Aislyn DiRisio, Ivo S Muskens, William B Gormley, Timothy R Smith, Marike LD Broekman, et al. Natural language processing for automated quantification of brain metastases reported in free-text radiology reports. *JCO clinical cancer informatics*, 3: 1–9, 2019.

[116] James H Seow, Damien L Stella, Christopher J Welman, Arjuna J Somasundaram, and Jan F Gerstenmaier. Washed up: the end of an era for adrenal incidentaloma ct. *Insights into Imaging*, 16:136, 2025.

[117] Jerry Sheehan, Steven Hirschfeld, Erin Foster, Udi Ghitza, Kerry Goetz, Joanna Karpinski, Lisa Lang, Richard P Moser, Joanne Odenkirchen, Dianne Reeves, et al. Improving the value of clinical research through the use of common data elements. *Clinical Trials*, 13(6):671–676, 2016.

[118] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[119] Christopher L. Sistrom, Keith J. Dreyer, Pragya P. Dang, Jeffrey B. Weilburg, Giles W. Boland, Daniel I. Rosenthal, and James H. Thrall. Recommendations for additional imaging in radiology reports: multifactorial analysis of 5.9 million examinations. *Radiology*, 253(2):453–461, November 2009. ISSN 1527-1315. doi: 10.1148/radiol.2532090200.

[120] Caroline E. Sloan et al. Assessment of follow-up completeness and notification preferences for imaging findings of possible cancer: what happens after radiologists submit their reports? *Acad Radiol*, 21(12):1579–1586, December 2014. ISSN 1878-4046. doi: 10.1016/j.acra.2014.07.006.

[121] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. Chexbert: combining automatic labelers and expert annota-

tions for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*, 2020.

[122] Rebecca Smith-Bindman, Diana L Miglioretti, and Eric B Larson. Rising use of diagnostic medical imaging in a large integrated health system. *Health affairs*, 27(6): 1491–1502, 2008.

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

[124] David Allen Spak, JS Plaxco, L Santiago, MJ Dryden, and BE Dogan. Bi-rads® fifth edition: A summary of changes. *Diagnostic and interventional imaging*, 98(3):179–190, 2017.

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

[126] Jackson Steinkamp, Charles Chambers, Darco Lalevic, and Tessa Cook. Automatic fullycontextualized recommendation extraction from radiology reports. *Journal of Digital Imaging*, 34:374–384, 2021. doi: 10.1007/s10278-021-00423-8.

[127] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, 2012.

[128] 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, April 2021. ISSN

15320464. doi: 10.1016/j.jbi.2021.103729. URL https://linkinghub.elsevier.com/retrieve/pii/S1532046421000587.

[129] Zhaoyi Sun, Hanley Ong, Patrick Kennedy, Liyan Tang, Shirley Chen, Jonathan Elias, Eugene Lucas, George Shih, and Yifan Peng. Evaluating gpt-4 on impressions generation in radiology reports. *Radiology*, 307(5):e231259, 2023.

[130] Charles Sutton, Andrew McCallum, et al. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373, 2012.

[131] Gerry H Tan and Hossein Gharib. Thyroid incidentalomas: management approaches to nonpalpable nodules discovered incidentally on thyroid imaging. *Annals of internal medicine*, 126(3):226–231, 1997.

[132] Janneke JC Tersteeg, Paul D Gobardhan, Rogier MPH Crolla, Peter AM Kint, Ilse Niers-Stobbe, Leandra JM Boonman-de Winter, and Jennifer MJ Schreinemakers. Improving the quality of mri reports of preoperative patients with rectal cancer: effect of national guidelines and structured reporting. *American Journal of Roentgenology*, pages 1240–1244, 2018.

[133] Hugo Touvron, Louis Martin, et al. Llama 2: Open foundation and fine-tuned chat models, 2023.

[134] Gaurav Trivedi, Esmaeel R. Dadashzadeh, Robert M. Handzel, Wendy W. Chapman, Shyam Visweswaran, and Harry Hochheiser. Interactive NLP in Clinical Care: Identifying Incidental Findings in Radiology Reports. *Applied Clinical Informatics*, 10(4): 655–669, August 2019. ISSN 1869-0327. doi: 10.1055/s-0039-1695791.

[135] Ozum Tuncyurek, Alejandro Garces-Descovich, Adrian Jaramillo-Cardoso, Elena Esteban Durán, Thomas E Cataldo, Vitaliy Y Poylin, Said Fettane Gómez, Atenea Morcillo Cabrera, Tarek Hegazi, Kevin Beker, et al. Structured versus narrative reporting of

pelvic mri in perianal fistulizing disease: impact on clarity, completeness, and surgical planning. *Abdominal Radiology*, 44(3):811–820, 2019.

[136] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[137] Guangyu Wang, Guoxing Yang, Zongxin Du, Longjun Fan, and Xiaohu Li. Clinicalgpt: large language models finetuned with diverse medical data and comprehensive evaluation. *arXiv preprint arXiv:2306.09968*, 2023.

[138] Yanshan Wang, Saeed Mehrabi, Sunghwan Sohn, Elizabeth J Atkinson, Shreyasee Amin, and Hongfang Liu. Natural language processing of radiology reports for identification of skeletal site-specific fractures. *BMC Medical Informatics and Decision Making*, 19: 23–29, 2019. doi: 10.1186/s12911-019-0780-5.

[139] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*, 2024.

[140] H Gilbert Welch and William C Black. Overdiagnosis in cancer. *Journal of the National Cancer Institute*, 102(9):605–613, 2010. doi: 10.1093/jnci/djq099.

[141] Tiantian White, Mark D Aronson, Scot B Sternberg, Umber Shafiq, Seth J Berkowitz, James Benneyan, Russell S Phillips, and Gordon D Schiff. Analysis of radiology report recommendation characteristics and rate of recommended action performance. *JAMA Network Open*, 5(7):e2222549–e2222549, 2022.

[142] Walter F Wiggins, Felipe Kitamura, Igor Santos, and Luciano M Prevedello. Natural language processing of radiology text reports: Interactive text classification. *Radiology: Artificial Intelligence*, page e210035, 2021. doi: 10.1148/ryai.2021210035.

[143] Martin J Willemink, Wojciech A Koszek, Cailin Hardell, et al. Preparing medical imaging data for machine learning. *Radiology*, 295(1):4–15, 2020. doi: 10.1148/radiol.2020192224.

[144] Kar-mun C Woo, Gregory W Simon, Olumide Akindutire, Yindalon Aphinyanaphongs, Jonathan S Austrian, Jung G Kim, Nicholas Genes, Jacob A Goldenring, Vincent J Major, Chloé S Pariente, et al. Evaluation of gpt-4 ability to identify and generate patient instructions for actionable incidental radiology findings. *Journal of the American Medical Informatics Association*, page ocae117, 2024.

[145] Yan Xu, Junichi Tsujii, and Eric I-Chao Chang. 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, 2012.

[146] An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y Chang, Amilcare Gentili, and Chun-Nan Hsu. Radbert: Adapting transformer-based language models to radiology. *Radiol Artif Intell*, 4(4):e210258, 2022. doi: 10.1148/ryai.210258.

[147] Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of medical imaging*, 5(3):036501–036501, 2018.

[148] Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. Packed levitated marker for entity and relation extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 4904–4917, Dublin, Ireland, May 2022. doi: 10.18653/v1/2022.acl-long.337.

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

[150] Wen-wai Yim, Tyler Denman, Sharon W Kwan, and Meliha Yetisgen. Tumor information extraction in radiology reports for hepatocellular carcinoma patients. *AMIA Summits on Translational Science*, 2016:455, 2016. URL `https://pubmed.ncbi.nlm.nih.gov/27570686/`.

[151] William F Young Jr. Management approaches to adrenal incidentalomas: a view from rochester, minnesota. *Endocrinology and metabolism clinics of North America*, 29(1): 159–185, 2000.

[152] John Zech, Margaret Pain, Joseph Titano, et al. Natural language–based machine learning models for the annotation of clinical radiology reports. *Radiology*, 287(2): 570–580, 2018. doi: 10.1148/radiol.2018171093.

[153] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019.

[154] Henghui Zhu, Ioannis Ch Paschalidis, and Amir Tahmasebi. Clinical concept extraction with contextual word embedding. *arXiv preprint arXiv:1810.10566*, 2018.