Automated Identification of Incidentalomas Requiring Follow-Up: A Multi-Anatomy Evaluation of LLM-Based and Supervised Approaches

Namu Park[1], Farzad Ahmed[2], Zhaoyi Sun[1], Kevin Lybarger[2], Ethan Breinhorst[3], Julie Hu[3], Özlem Uzuner[2], Martin Gunn[4], Meliha Yetisgen[1]

[1]*Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, USA*
[2]*Department of Information Sciences and Technology, George Mason University, Fairfax, VA, USA*
[3]*Department of Radiology, Te Whatu Ora Health New Zealand, Te Toka Tumai Auckland, Auckland, New Zealand*
[4]*Department of Radiology, School of Medicine, University of Washington, Seattle, WA, USA*

Abstract

Objective: To evaluate large language models (LLMs) against supervised baselines for fine-grained, lesion-level detection of incidentalomas requiring follow-up, addressing the limitations of current document-level classification systems.

Methods: We utilized a dataset of 400 annotated radiology reports containing 1,623 verified lesion findings. We compared three supervised transformer-based encoders (BioClinicalModernBERT, ModernBERT, Clinical Longformer) against four generative LLM configurations (Llama 3.1-8B, GPT-4o, GPT-OSS-20b). We introduced a novel inference strategy using lesion-tagged inputs and anatomy-aware prompting to ground model reasoning. Performance was evaluated using class-specific F1-scores.

Results: The anatomy-informed GPT-OSS-20b model achieved the highest performance, yielding an incidentaloma-positive macro-F1 of 0.79. This surpassed all supervised baselines (maximum macro-F1: 0.70) and closely matched the inter-annotator agreement of 0.76. Explicit anatomical grounding yielded statistically significant performance gains across GPT-based models ($p < 0.05$), while a majority-vote ensemble of the top systems further improved the macro-F1 to 0.90. Error analysis revealed that anatomy-aware LLMs demonstrated superior contextual reasoning in distinguishing actionable findings from benign lesions.

Conclusion: Generative LLMs, when enhanced with structured lesion tagging and anatomical context, significantly outperform traditional supervised encoders and achieve performance comparable to human experts. This approach offers a reliable, interpretable pathway for automated incidental finding surveillance in radiology workflows.

Keywords: Incidental Findings; Large Language Models; Natural Language Processing; Radiology; Clinical Decision Support

1. Introduction

Incidental findings, or *incidentalomas*, refer to unexpected abnormalities discovered during imaging studies performed for unrelated reasons [1]. Their detection has increased as imaging utilization has grown across healthcare. A systematic review estimated that incidental findings appear in up

to one-third of computed tomography (CT) and positron emission tomography CT (PET-CT) examinations [2]. These findings create a clinical dilemma, since most are benign while some represent early-stage disease that requires intervention. Balancing overdiagnosis and missed pathology has become an important concern in medical practice.

To support clinicians, the American College of Radiology (ACR) and other professional groups provide standardized decision rules for managing incidental findings [1]. These guidelines outline follow-up pathways based on organ type, lesion size, and imaging features. Ensuring adherence remains difficult in real-world settings. Radiology reports are unstructured narratives that often include multiple findings and subtle recommendations, so significant incidental findings may be overlooked or receive no follow-up, which contributes to patient risk and healthcare inefficiency. Decision rules also contain little clinical context about comorbidities, although this information is relevant for determining the appropriate timing and type of follow-up.

In large hospital systems, millions of radiology reports are generated each year, so manual review is infeasible. Automated systems that identify incidental findings and follow-up recommendations could improve care coordination. Early natural language processing (NLP) work demonstrated that extracting such information is possible [3, 4, 5], but earlier systems were limited by linguistic variability, contextual ambiguity, and the lack of lesion-level supervision. Distinguishing an incidental lesion from the primary diagnostic concern often requires nuanced understanding that rule-based or shallow learning approaches cannot provide.

Recent advances in clinical NLP offer new opportunities. Domain-specific encoders such as BioClinicalBERT [6] and ClinicalBERT [7] introduced contextual representations tailored to clinical text. Generative models such as GPT-4 [8] and Med-PaLM [9] show strong reasoning abilities and can interpret subtle contextual cues, recognize implicit recommendations, and integrate cross-sentence information that is essential for identifying incidentalomas accurately.

This study uses improved reasoning and text-understanding capabilities of modern NLP systems to develop LLM-based methods for anatomy-aware identification of incidentalomas. By integrating lesion-tagged inputs and prompting strategies grounded in anatomical structure, our approach aligns computational predictions with how radiologists interpret imaging findings. We compare LLM-based systems with strong supervised baselines to evaluate their accuracy and clinical consistency. We find that anatomical grounding improves performance and interpretability. The best LLM configuration achieves an incidentaloma-positive macro-F1 of 0.79, which closely matches the inter-annotator agreement macro-F1 of 0.76.

2. Related Work

2.1. Clinical NLP using Supervised Models

Transformer-based language models [10] have transformed clinical NLP by leveraging contextual embeddings from large-scale biomedical and clinical corpora. BioClinicalBERT [6], trained on MIMIC-III [11] and PubMed texts, has demonstrated strong performance on concept extraction and relation detection tasks. Variants such as ClinicalBERT [7] and other transfer learning approaches have further improved robustness across multiple clinical domains [12]. To handle longer clinical documents, architectures such as Clinical Longformer and Clinical BigBird extend transformers with sparse attention, enabling efficient modeling of radiology reports and other long sequences [13]. More recently, domain-extended models like ModernBERT [14], BioClinicalModernBERT [15] incorporate long-context adaptations specifically designed for biomedical applications.

Statement of significance

Problem	Radiology reports frequently contain incidental findings, yet many are under-recognized or lack appropriate follow-up, leading to missed or delayed diagnoses. Manual review is infeasible at scale, and existing automated systems often fail to distinguish incidentalomas from clinically indicated findings.
What is Already Known	Previous NLP and machine learning methods have demonstrated the feasibility of extracting incidental findings from radiology text but mostly rely on document-level classification. These systems struggle with contextual ambiguity and lack lesion-level interpretability.
What this Paper Adds	This paper presents a lesion-tagged, anatomy-aware framework for automated identification of incidentalomas requiring follow-up across six anatomies. By integrating structured lesion information with LLMs and supervised encoders, the study shows that anatomy-informed prompting substantially improves detection accuracy, interpretability, and clinical reliability. It also introduces the first lesion-level benchmark and error analysis for incidentaloma classification.
Who Would Benefit from the New Knowledge in this Paper	Radiologists, clinical informaticians, and health systems developing scalable AI tools for incidental finding surveillance and data-driven quality improvement in imaging workflows.

In radiology, these models have been applied to tasks such as report classification, lesion entity extraction, and malignancy detection [16, 17]. Comparative evaluations highlight their strengths for document-level prediction but also reveal challenges in multi-class classification tasks where false negatives are especially costly [18, 19]. Our work contributes to this body of research by systematically benchmarking BioClinicalModernBERT, ModernBERT, and Clinical Longformer for incidentaloma detection, providing evaluation across multiple anatomies.

2.2. Clinical NLP using Generative Large Language Models

Generative LLMs have recently emerged as powerful tools for clinical NLP, capable of zero-shot and few-shot generalization across diverse medical tasks. GPT-4 [20] and GPT-4o [21] both have shown considerable performance on medical reasoning benchmarks, demonstrating competency in both multiple-choice exams and free-text reasoning tasks [8, 21, 22]. Similarly, Med-PaLM [9], an instruction-tuned variant of PaLM [23], has exhibited strong performance on clinically-oriented question answering and reasoning tasks. Open-source foundation models such as LLaMA [24] have also been widely adopted in clinical NLP research as the backbone for lightweight adaptations. Methods like low-rank adaptation (LoRA) [25] have enabled efficient fine-tuning of LLMs for domain-specific tasks without retraining full models.

Within radiology domain, LLMs have been explored for report summarization [26], clinical reasoning [27], and structured entity extraction from imaging reports [28]. Instruction-tuned medical LLMs such as MedAlpaca [29] have further demonstrated the feasibility of adapting general-purpose models to healthcare through domain-specific instruction datasets. Despite these advances, few

studies have explored lesion-level classification tasks that explicitly leverage lesion and anatomy information. Our work addresses this gap by 1) marking candidate lesions with structured tags in the report text and 2) providing LLMs with explicit lesion-anatomy associations during inference. These structured inputs enhance classification and interpretability by ensuring models attend to specific findings within their correct anatomical context.

2.3. Identification of Incidentalomas in Radiology Reports

Large-scale analyses have highlighted the frequency of incidental findings in abdominal and pelvic CT exams and their downstream clinical implications [2]. Organ-specific investigations have characterized incidentalomas of the adrenal glands, thyroid, and lungs, providing prevalence estimates and evidence-based follow-up recommendations [30, 31, 19]. Traditional approaches have relied on structured radiology databases and manual chart reviews, which ensure accuracy but limit scalability across large health systems [32, 18].

Earlier NLP systems established the foundation for automated detection of incidental findings in radiology reports. Dutta *et al.* [33] and Trivedi *et al.* [34] implemented rule-based and interactive NLP frameworks for identifying incidental observations in narrative imaging text, while Kang *et al.* [35] focused on incidental pulmonary nodules. These early studies demonstrated the feasibility of text mining for large-scale retrospective identification of incidental findings but were limited by hand-engineered features and domain-specific lexicons. Building on these foundations, more recent NLP pipelines have leveraged machine learning and transformer-based architectures to automatically extract incidental findings from unstructured radiology reports [3, 28]. However, most prior NLP efforts have emphasized document-level classification of reports containing incidental findings rather than fine-grained lesion-level annotation. Closely related to incidentaloma identification, recent studies have advanced the extraction of follow-up imaging recommendations from radiology reports using NLP approaches [4, 36, 5].

Subsequent advances in deep learning enabled more robust classification and extraction pipelines. Canton *et al.* [37] proposed a multi-modal system that combined imaging metadata and report text to automatically detect adrenal and thyroid incidentalomas. Schumm *et al.* [38] automated adrenal nodule extraction from electronic health records, and Bala *et al.* [39] developed a web-based application for incidentaloma tracking and management, illustrating the integration of AI systems into clinical follow-up workflows.

Recent work has highlighted the growing role of LLMs in incidental finding interpretation and communication. Woo *et al.* [40] evaluated a HIPAA-compliant instance of GPT-4 [20] for identifying actionable incidental findings and generating patient-facing summaries from radiology reports, achieving high factual alignment with expert reviews. Bhayana *et al.* [41] similarly demonstrated that single-shot prompting with GPT-4 could identify incidental findings across multiple report types with minimal supervision. Aksu *et al.* [42] compared ChatGPT [20], Gemini [43], and Claude [44] for adrenal incidentaloma management, showing that GPT-based reasoning systems can approximate endocrinologist-level decision-making. These developments highlight the potential of multimodal and text-based LLM systems to enhance incidentaloma triage and follow-up workflows. Our study extends this line of work by introducing lesion-specific annotations across six anatomies and explicitly modeling both incidentaloma presence and follow-up requirements, while exploring how LLM-based prompting can improve lesion-level interpretability and clinical decision support.

Table 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

3. Methods

3.1. Dataset

3.1.1. Preprocessing

We utilized an existing clinical database comprising 6,668,323 radiology reports from 2007 to 2020, representing the general patient population across four hospitals within the UW Medicine system. All reports were automatically de-identified and processed with *PL-Marker++*, a radiology-specific BERT-based model for lesion-level entity and relation extraction [45, 28]. PL-Marker++ was trained on a manually annotated corpus of 609 radiology reports labeled by medical residents and board-certified radiologists, achieving inter-annotator agreement F1 = 0.762 and macro-F1 = 0.88 for lesion finding extraction.

PL-Marker++ extracts three main categories of information: clinical indications, lesion findings, and medical problems, each accompanied by finding-specific arguments such as anatomy and size trend attributes. 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 represent mass-forming pathologic structures (e.g., *"hepatic lesion", "adrenal nodule"*), whereas medical problems represent non mass-forming pathologic processes. Anatomical and descriptive attributes capture details such as size, count, and temporal size trend with five possible values: *increasing*, *decreasing*, *no change*, *disappeared*, and *new*. Anatomy information associated with each lesion, medical problem, and clinical indication is extracted using entity–relation extraction and mapped to predefined anatomy categories such as *lung*, *liver*, and *kidney*.

Table 1 presents token-level lesion identification performance on the annotated dataset described by Park et al. [28]. On the test set, PL-Marker++, a BERT-based model trained on annotated radiology reports, outperformed n-shot approaches using Llama 3.1-8B and GPT-4o. This is consistent with recent work [46, 47] showing that LLMs often struggle with token-level clinical named entity recognition tasks, likely due to limitations of decoder-based architectures.

These structured outputs, particularly the integration of categorized clinical indications and lesion-level findings with anatomical and size trend attributes, formed the basis for this study. They enabled the identification of reports describing new or potentially actionable lesions and provided clinical context to improve LLM inference performance described in the following sections. The study protocol was reviewed and approved by the University of Washington Institutional Review Board (IRB).

3.1.2. Sampling Process

The low prevalence of incidentaloma-containing reports (e.g. 9.9% for adrenal incidentalomas [39] and less than 5% for chest computed tomography for incidental pulmonary embolism [48]) and the substantial linguistic variability with which radiologists describe these findings make identifying such cases challenging. Directly searching for terms such as "incidental" or "incidentaloma" would introduce bias and miss many true incidentalomas, since radiologists often describe them without using these terms. To avoid this limitation, we used a broader sampling strategy that leveraged PL-Marker++ outputs and SVM-based recommendation sentence identification, aiming for more comprehensive coverage of incidentaloma cases. After discussions with a board-certified radiologist, we developed a sampling strategy designed to achieve a high positivity rate for reports likely to contain incidentalomas. Although this approach may introduce some sampling bias, it increased the proportion of incidentaloma-positive reports, which was important for constructing a robust gold standard. Our sampling strategy applied four sequential filters:

(1) **Target relevant anatomies:** Using structured clinical findings extracted with PL-Marker++, we analyzed all radiology reports and identified findings mapped to six anatomical regions: kidney, liver, lung, pancreas, adrenal gland, and thyroid. These regions were selected because they are more likely to contain incidentalomas. Among all reports, 24.7 % (n = 1,767,623) contained at least one finding (clinical indication, medical problem, or lesion) from these anatomies, yielding 7,519,138 findings. We then selected reports across all imaging modalities that included lesion findings in these regions with assertion values labeled as *"present"* or *"possible"*. In total, 6.3 % of all reports (n = 112,100) satisfied these criteria and were retained.

(2) **Exclude previously identified lesions:** Using size trend data, we excluded reports with size trend values of "*increasing*", "*decreasing*", "*disappeared*", or "*no change*", which indicate comparison with prior imaging and therefore previously identified lesions. Accordingly, 5.7% of reports identified in step (1) were excluded, leaving 105,729 reports.

(3) **Filter surveillance cases:** PL-Marker++ was used to extract clinical indication types (*trauma*, *symptom*, *neoplastic diagnosis*, and *non-neoplastic diagnosis*). Reports with a neoplastic diagnosis in the clinical indication were excluded to minimize inclusion of expected findings rather than incidental ones. After this step, 39.6% of reports (n = 41,833) were retained.

(4) **Select reports with follow-up recommendations:** Among the 41,833 reports retained, we further selected those containing follow-up recommendation sentences, as incidentalomas are frequently accompanied by such recommendations. Recommendation sentences were identified using a support vector machine (SVM)–based classifier [36]. This step yielded 19,690 reports with the highest probability of containing a clinically important incidentaloma.

To ensure accurate anatomical assignment for each lesion and facilitate lesion-level and document-level analyses, we implemented a double-verification process for anatomy extraction. PL-Marker++ initially provided sentence-level lesion–anatomy mappings, but its inference was sometimes limited when relevant anatomy mentions appeared outside the immediate sentence. To address this, we used an additional LLM (Llama 3.1-8B-Instruct) [49] trained to infer anatomical associations from full report context, which achieved a micro-F1 score of 0.895 on the annotated dataset from the PL-Marker++ study. Discrepancies between systems were manually reviewed, and a final verified

anatomy label was assigned to each annotated lesion. This dual-level verification ensured robust and consistent anatomical categorization across six target anatomies: kidney, liver, lung, pancreas, adrenal gland, and thyroid.

3.1.3. Data Annotation

Of the 19,690 radiology reports that met the selection criteria, 400 reports were randomly selected and annotated. Annotation guidelines were refined through multiple iterations by two board-certified radiologists to ensure both precision and clinical relevance. Rather than relying solely on document-level labels, the guidelines were designed to leverage lesion-level findings. Each lesion was annotated as *No Incidentaloma*, *Incidentaloma–No Risk*, or *Incidentaloma–Follow-up Required*, with the distinction based on clinical severity (e.g. *"Benign simple cysts"* were considered as not requiring follow-up). To expedite annotation and minimize missed lesion mentions, lesion findings extracted using PL-Marker++ [28] were pre-annotated in the BRAT annotation tool [50]. Differential diagnoses were excluded.

Two medical residents participated in the annotation process. Initially, we conducted double-annotation training rounds in which both annotators independently labeled the same reports, followed by weekly feedback sessions supervised by a board-certified radiologist. Discrepancies were reviewed and resolved through consensus, and remaining disagreements were adjudicated by the radiologist who developed the guidelines. Once a document-level inter-annotator agreement (IAA) of 0.896 F1 for incidentaloma status was achieved, we transitioned to single-annotation rounds, where each annotator was assigned distinct reports.

As a result, 160 double-annotated reports contained 1,623 pre-identified lesion findings, averaging 10.15 lesions per report. Of these reports, 117 (73.1%) contained at least one incidentaloma, demonstrating an enriched sample. Annotators reached agreement on 1,498 incidentaloma labels, resulting in an agreement rate of 92.3% (0.838 F1). Common incidentaloma terms included "lesion", "nodule", "cyst", and "mass". Annotator 1 identified an average of 2.538 incidentalomas per report, slightly higher than Annotator 2's average of 2.294. We then moved to single annotation for the remaining reports. The overall dataset distribution at the document level is summarized in Table 2, while Table 3 presents the anatomy-specific distribution of all annotated lesion findings.

Table 2: Distribution of annotated reports with and without incidentalomas.

	Total (n)	W/ Incidentaloma	W/O Incidentaloma
Double-annotated	160	117 (73.1%)	43 (26.9%)
Single-annotated	240	158 (65.8%)	82 (34.2%)
Total	400	275 (69.25%)	125 (30.75%)

3.2. Incidentaloma Classification

Using the annotated reports, we defined the incidentaloma classification task to support both lesion-level and document-level analyses. Leveraging anatomy information from Section 3.1.2 together with the annotated lesion findings, we framed incidentaloma identification as seven anatomy-specific, three-way classifications per report. For each anatomy (*Lung, Liver, Kidney, Adrenal, Pancreas, Thyroid, Other*), the model generated one label from {0 = *No Incidentaloma*, 1 = *Incidentaloma–No Risk*, 2 = *Incidentaloma–Follow-up Required*}.

Table 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

During inference, all models first generated lesion-level predictions for each anatomy and then aggregated these predictions 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 $L_{\text{lung}} = 2$. These class indices (0, 1, and 2) are used throughout the paper to refer to their corresponding labels.

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

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

Inter-annotator agreement (IAA) for the document-level annotations, calculated without considering anatomy-specific labels, was 0.93 F1 for *No Incidentaloma*, 0.81 F1 for *Incidentaloma–No Risk*, and 0.70 F1 for *Incidentaloma–Follow-up Required*, resulting in a macro-average F1 of 0.81. This schema provided consistent, anatomy-specific labeling and served as the foundation for subsequent model training and evaluation. Summary statistics are presented in Table 3.

3.2.1. Supervised Learning Approach

We fine-tuned three transformer-based encoders on the annotated dataset: BioClinicalModernBERT, ModernBERT, and Clinical Longformer. BioClinicalModernBERT [15] is a recent BERT-based model for bioclinical applications pre-trained on large-scale biomedical and clinical corpora, optimized for domain-specific terminology and extended context modeling. ModernBERT [14] combines general-domain and clinical text during pre-training, providing a balanced representation that enables comparison with a less domain-specialized encoder. Clinical Longformer [13] uses a sparse attention mechanism to efficiently process lengthy radiology reports that often exceed the input capacity of standard BERT-like models. All models were trained according to the classification task described in Section 3.2, to extract lesion-level incidentaloma labels.

Model hyperparameters were tuned on the validation set. We explored learning rates in $\{1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}, 5 \times 10^{-5}\}$, batch sizes of $\{8, 16, 32\}$, and up to 15 training epochs, with weight decay $\{0.01, 0.05\}$ and dropout $\{0.1, 0.2\}$. The final configuration, selected according to the highest validation macro-F1 for incidentaloma classes, used a batch size of 16, learning rate 2×10^{-5}, linear warmup over 10% of total steps, weight decay 0.01, and dropout 0.1. The maximum sequence length was 512 tokens for BioClinicalModernBERT and ModernBERT, and 2,048 tokens for Clinical Longformer. Models were trained for up to 10 epochs with early stopping (patience = 3), using the AdamW optimizer [51], gradient clipping [52] (maximum norm = 1.0), and mixed-precision (FP16) training on NVIDIA A100 GPUs.

The input features comprised a structured textual representation derived from the preprocessed radiology reports rather than the full report narrative. For each report, we generated seven anatomy-specific input strings, one for each anatomy category. Each input took the form `Anatomy: <organ> | Lesion: <lesion description> | Context: <neighboring text>`, allowing the model to reason separately about potential incidental findings in each anatomical region. Anatomical information was obtained from the verified anatomy labels generated during preprocessing (Section 3.1.1) and inserted directly into the input as plain text. Lesion descriptions were obtained from the lesion spans identified during the same stage, and a maximum of 100 characters of surrounding report text were included to capture relevant local context. This window size was selected through validation-set analysis to balance lesion specificity with contextual information.

Each sequence was tokenized and padded to the maximum input length supported by the encoder model, using the model-specific tokenizer. This yielded a unified text representation capturing three components, the anatomy of interest, the lesion description, and the immediate local context, while ensuring consistent formatting across models with different sequence capacities. During inference, the supervised models first generated lesion-level predictions for each anatomy and then aggregated these into final document-level labels, as described in Section 3.2.

Since the gold standard labels were substantially imbalanced (Table 3), with relatively few instances labeled *Incidentaloma–Follow-up Required*, we implemented cost-sensitive learning strategies to enhance performance on underrepresented yet clinically important classes. Three approaches were evaluated. First, class-weighted cross-entropy was applied, assigning weights inversely proportional to class frequencies. Second, focal loss was employed with $\gamma = 2$ and class-specific weighting $\alpha = w_c$, emphasizing hard-to-classify and minority examples. Finally, an expected-cost objective was tested using a 3×3 asymmetric cost matrix C designed to penalize clinically consequential misclassifications:

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

where p_θ denotes the softmax output probabilities. At inference time, we also used cost-aware decision making by selecting the label that minimized expected misclassification cost under the predicted distribution.

3.2.2. LLM-based Approach

Similar to supervised learning methods, the LLM-based approach first extracts lesion-level incidentaloma labels and then aggregates these predictions into document-level labels, as described in Section 3.2. For LLMs, prior to inference, candidate lesion entity spans in the original report were enclosed with XML-format tags using lesion information extracted by PL-Marker++, while the surrounding text remained intact. Numbered lesion tags support detailed error analysis and

help reduce noise. Without tags, it can be difficult to identify the lesion of interest, particularly when terms like "nodule" appear in multiple contexts within the same report. By highlighting lesions directly, tags disambiguate such cases, expedite manual review by clinicians, and minimize the influence of irrelevant or distracting text, which improves interpretability and clinical reliability.

Our ablation study using the Llama 3.1-8B Instruct model showed that the inclusion of lesion tags substantially improved performance compared with untagged inputs. In the 0-shot setting, inputs with lesion tags achieved higher macro F1-score than all 0-shot, 1-shot and 5-shot settings without tags, indicating that structural cues can be more beneficial than additional examples. The improvement was primarily driven by higher precision with comparable recall, suggesting that the tags helped models focus on the correct lesion spans. Accordingly, we used lesion-tagged inputs throughout all LLM experiments. Generation was performed with settings intended to minimize stochastic variation (temperature = 0, top-p = 1).

Using these lesion-tagged radiology reports, we evaluated two prompt settings:

1. **Base Prompt:** The lesion-tagged report is passed as-is.
2. **With-anatomy Prompt:** The lesion-tagged report is supplemented with a concise mapping that pins each tag to previously extracted anatomy information described in 3.1.1; the mapping is a single line such as `LESION1=Thyroid; LESION2=Pancreas; LESION3=Adrenal; ...`. This pinning removes residual ambiguity about anatomy assignment while keeping all evidence in situ.

An example of the With-anatomy prompt is shown in Figure 2. Both prompting approaches use the same instruction described in Figure 1 and the same report content, where lesions are tagged using numbered XML-format tags. In the With-anatomy approach, additional anatomical information is included. As illustrated in the LLM reasoning trace in Figure 2, the model correctly attends to the target lesions by leveraging the numbered tags during inference.

We evaluated four generative LLM-based approaches under identical settings using the two prompts described above. The first was a LoRA (Low-Rank Adaptation) [25] fine-tuned Llama 3.1-8B model, adapted on our annotated radiology reports to generate incidentaloma labels in a domain-specific manner while keeping parameter updates lightweight. The second approach was a prompt-engineered version of Llama 3.1-8B, applied in a zero-shot fashion without gradient updates and evaluated in both prompt settings. The third approach employed GPT-4o, a proprietary state-of-the-art model, to benchmark performance against open-source alternatives. Finally, we evaluated GPT-OSS [53], an open-source GPT-style model that has demonstrated competitive performance across multiple public benchmarks. For computational feasibility, we used GPT-OSS-20B rather than GPT-OSS-120B, balancing inference time and resource requirements while maintaining strong performance. All experiments were conducted on our HIPAA-compliant secure server environment. Fine-tuning GPT-OSS-20B was not feasible because the model relies on generating explicit reasoning traces, which were not available for our clinical dataset. Fine-tuning GPT-4o was not performed to maintain consistency across proprietary models and to ensure a fair comparison within a prompt-only evaluation framework.

3.3. Evaluation

The held-out test set contains 80 double-annotated reports, yielding 560 anatomy-specific labels in total (seven anatomies per report). Of these labels, 50 are *Incidentaloma–No Risk* (1) and 29 are *Incidentaloma–Follow-up Required* (2); the remainder are *No Incidentaloma* (0). The training set

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

(n = 280) and validation set (n = 40) were used exclusively for prompt refinement in the LLM-based approaches and for model optimization in the supervised learning experiments.

Model performance was primarily evaluated using the F1-score, which balances sensitivity and precision. F1-scores were calculated for all three classes. To better capture performance on clinically meaningful cases, we additionally report an *incidentaloma macro-F1*, defined as the mean of the F1 values for the two incidentaloma classes (1 and 2). Because the *No Incidentaloma* class comprised the majority of samples and consistently yielded high performance across models, the incidentaloma macro-F1 offers a more balanced assessment focused on cases where incidentalomas are present.

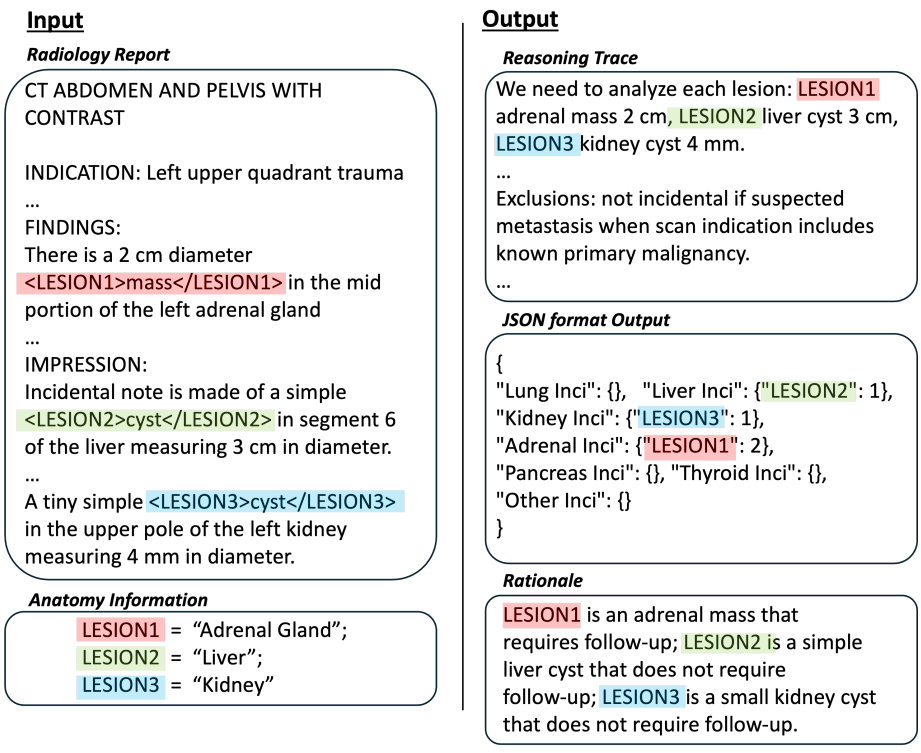

Figure 2: Example of input and output used for LLM-based incidentaloma identification using GPT-OSS-20B. Lesions that are not returned in the JSON output are treated as No Incidentaloma (Class 0). Reasoning traces are available only in GPT-OSS-20B inferences, as GPT-4o does not provide reasoning trace outputs.

4. Results

4.1. Performance Comparison

As shown in Table 4, the best overall performance was achieved by the GPT-OSS-20b (With Anatomy) model, which obtained the highest F1-scores across all labels and an incidentaloma-positive macro-F1 of 0.79. GPT-4o (With Anatomy) was the second-best performer, with F1 scores of 0.82 and 0.71 for the incidentaloma-positive classes. These two anatomy-informed methods consistently outperformed all other systems, indicating that explicit anatomy context substantially improves incidentaloma classification.

Among supervised encoders, BioClinicalModernBERT (w/o CS) and ModernBERT (CS) both reached an incidentaloma macro-F1 of 0.70, with BioClinicalModernBERT performing better for Class 1 (0.79 F1) and ModernBERT slightly stronger for Class 2 (0.63 F1). Cost-sensitive (CS) learning produced only modest gains and mainly increased recall for minority classes.

Comparing model families, LLMs showed clear gains over supervised encoders. Only Llama 3.1-8B was fine-tuned using LoRA; all other LLMs, including GPT-4o and GPT-OSS-20B, were evaluated in a prompt-only configuration. GPT-4o (Base) already matched the strongest non-LLM baselines, and adding anatomical context further improved performance, yielding up to a Δ+0.08 increase in macro-F1 (classes 1–2) and clearly surpassing all non-LLM methods. The consistent

benefit of anatomy-aware prompting across both Llama and GPT-based architectures highlights the value of anatomy-aware contextualization of lesion findings in improving incidentaloma detection.

Table 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). Overall Accuracy and Incidentaloma Macro-F1 (computed on Class 1 and 2 only) are also reported to show each model's overall correctness and ability to detect and correctly classify incidentaloma-positive cases. 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**

4.2. Pairwise Statistical Significance Testing of Models on Incidentaloma-Positive Lesions

To evaluate statistical significance in model performance, we conducted a lesion-level bootstrap analysis restricted to lesions annotated as incidentalomas. This approach quantifies variability at the lesion level and focuses on clinically meaningful cases. Figure 3 summarizes these comparisons, with each horizontal line representing the estimated CI for Δ Macro-F1. Llama-based models were excluded because of their consistently lower performance across all metrics.

Overall, both GPT-OSS (Base) and GPT-OSS (Anatomy) achieved higher Macro-F1 values than the supervised baselines (ModernBERT, BioclinicalModernBERT). GPT-OSS (Anatomy) provided the best overall performance, with confidence intervals lying entirely above zero when contrasted with all supervised encoders ($p < 0.01$). All GPT-based models outperformed the supervised methods by a statistically significant margin, underscoring the advantage of LLMs for lesion-level classification.

Consistent with the previous section, both GPT-OSS and GPT-4o showed significant gains when anatomical context was incorporated through anatomy-aware prompting. GPT-4o (Anatomy) significantly outperformed its base version ($p = 0.012$), and GPT-OSS (Anatomy) also improved over GPT-OSS (Base). These results demonstrate that anatomy-aware prompting improves robustness and clinical reliability of LLM-based systems for incidentaloma classification and is a generalizable strategy for radiology-specific information extraction.

4.3. Error Analysis

4.3.1. Errors in Supervised Models

We examined error patterns for BioClinicalModernBERT and ModernBERT, both with and without cost-sensitive (CS) training, to understand the limitations of supervised encoders. Clinical

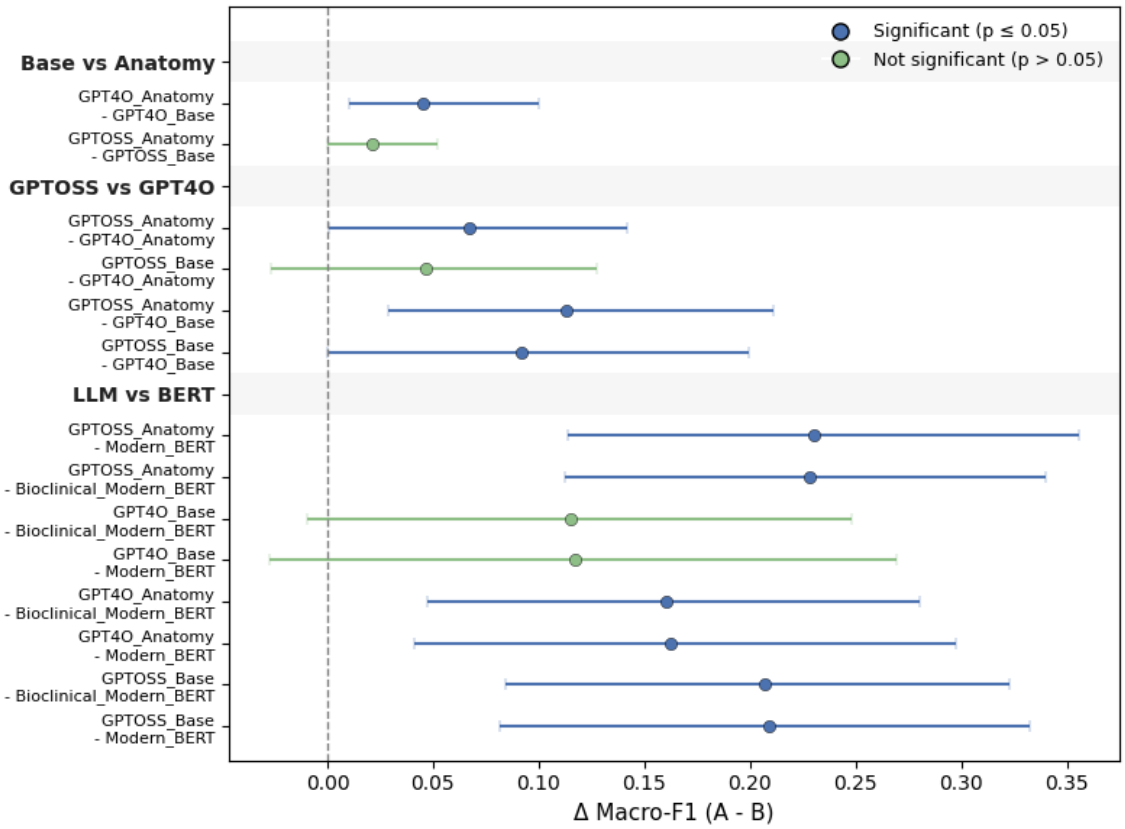

Figure 3: Pairwise non-parametric bootstrap comparison of model performance on incidentaloma-positive lesions. Each point represents the mean difference in Macro-F1 (A–B) across 1,000 lesion-level bootstrap samples, with horizontal bars showing 95% confidence intervals. Comparisons to the right of zero indicate that Model A outperformed Model B.

Longformer was excluded from detailed analysis because of its poor recall for follow-up-required cases and inconsistent error behavior.

BioClinicalModernBERT showed the most balanced performance among the supervised models but still underestimated lesion severity, frequently labeling follow-up-required incidentalomas as low-risk. Cost-sensitive training reduced some of these misses but increased false positives for benign lesions. ModernBERT showed a similar trend with generally lower recall and precision, often missing follow-up recommendations that appeared only in impression sections or in abbreviated form (e.g., "f/u CT advised").

Qualitative inspection indicated that both models struggled with hedged phrasing (e.g., "likely benign," "probably cystic"), complex size descriptions, and multi-lesion narratives mixing benign and higher-risk lesions. ModernBERT particularly struggled when follow-up recommendations were expressed indirectly or outside the main lesion description. LIME-based interpretability analyses [54] supported these findings, showing that token-level attributions were dominated by explicit lesion descriptors and common modifiers (e.g., "subcentimeter," "stable," "mass"), while contextual

indicators of clinical intent were underweighted.

4.3.2. Errors in Generative LLMs

The most critical errors involved missed incidentalomas, which represent potentially high-risk findings. GPT-4o (Base) missed 10 of 29 (34.5%) such cases, GPT-4o (Anatomy) reduced this to 7 of 29 (24.1%), GPT-OSS (Base) to 6 of 29 (20.7%), and GPT-OSS (Anatomy) to 5 of 29 (17.2%). These trends indicate that anatomy-aware prompting improves focus on lesion-specific information and reduces clinically important misses.

False positives, where normal lesions were misclassified as incidentalomas (Class $0 \rightarrow$ Class 1 or 2), were uncommon (3–6%). Anatomy inclusion slightly reduced these rates for GPT-4o ($4.8\% \rightarrow 3.5\%$) and GPT-OSS ($5.6\% \rightarrow 5.4\%$). None of the models predicted no-risk incidentalomas as requiring follow-up. Across all models, a recurring pattern was underestimation of severity, where follow-up-required lesions were misclassified as no-risk (Class $2 \rightarrow$ Class 1). GPT-4o showed such underestimation in 13.8% of error cases, while GPT-OSS reduced this to 10.3%. These errors often appeared in reports with equivocal language or indirectly expressed follow-up intent. GPT-OSS (Anatomy) achieved the best balance, minimizing hazardous misses while keeping false positives low.

To further characterize GPT-OSS (Anatomy), we reviewed its 33 errors among 560 evaluated lesions. Errors occurred across multiple anatomical sites (lung, liver, kidney, thyroid), indicating broader reasoning challenges rather than anatomy-specific effects. Three dominant mismatch types were identified:

Follow-Up Required vs. No-Risk (Gold = Class 2 → Model = Class 1, n=3)
GPT-OSS (Anatomy) correctly identified the lesion but underestimated its clinical significance. These cases involved conditional language such as *"tiny sub-6 mm lung nodule, follow-up if high risk"* or *"indeterminate hypodensity on unenhanced liver CT"*, where reassuring phrases were overemphasized and qualifiers related to higher-risk populations or incomplete characterization were underweighted.

Incidentaloma Missed (Gold = Class 1 or 2 → Model = Class 0, n = 4)
Four lesions labeled as incidentalomas were classified as non-incidentalomas because annotators had overlooked contextual cues such as "stable" or "compared to prior exam," indicating prior evaluation. These were annotation errors rather than model failures. GPT-OSS (Anatomy) correctly identified all four as non-incidental, suggesting sensitivity to subtle contextual indicators that human annotators sometimes miss.

False Incidentaloma Detection (Gold = 0 → Model = 1 or 2, n = 26)
False positives were the largest error category. GPT-OSS (Anatomy) misclassified 26 non-incidental cases as incidentalomas, split between no-risk (13) and follow-up-required (13). Most arose from differences in interpreting the relationship between the clinical indication and the lesion. For example, with clinical indication *"previous CXR abnormal"* and *"small opacity in the left lower lung"*, annotators linked the opacity to the prior abnormal CXR and labeled it non-incidental, whereas the model treated it as a separate finding requiring follow-up. In another case with indication *"restaging"*, human annotators considered a lesion in a different anatomical region incidental, while the model interpreted the context as evidence of existing malignancy and classified it as non-incidental. These examples highlight how divergent interpretations of clinical intent and anatomical context can lead to disagreement between human experts and LLMs.

5. Discussion

5.1. 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. Most of the GPT-correct and BERT-incorrect cases contained nodule management guideline templates that are automatically inserted into reports by the radiology information system (RIS) macros (e.g., *"follow-up per Fleischner Society guidelines"*). These standardized 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 semantic context. In contrast, GPT models demonstrated stronger contextual reasoning, correctly linking each recommendation to its corresponding lesion finding. While these results highlight the contextual adaptability of LLMs, the BERT-based models showed a more conservative bias that often missed positive incidentaloma cases but produced fewer false positives, reflecting a precision-oriented decision boundary.

To integrate the complementary strengths of different modeling approaches, a simple majority-vote ensemble approach was constructed across six strong systems: 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 across models. In case of ties during majority voting, the lowest numerical label was chosen (e.g., $0 < 1 < 2$), providing a conservative bias toward non-incidentaloma predictions. This parameter-free ensemble achieved the highest overall performance, yielding a Macro-F1 of 0.902 and surpassing all individual models. The resulting 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 gains over the best individual model, GPT-OSS (With Anatomy), which achieved 0.968, 0.842, and 0.727 for the same categories, respectively. The ensemble therefore improved each class by a notable margin, particularly for identifying incidentalomas requiring follow-up, where the F1 value increased by more than 0.10. This indicates that combining the contextual reasoning of LLMs with the structured precision of supervised encoders may enhance reliability in clinical information extraction from radiology reports, particularly in cases involving ambiguous or uncertain phrasing.

While the ensemble achieved clear gains in robustness and overall accuracy, this approach also introduces practical trade-offs. Deploying multiple large-scale models in parallel increases computational overhead, inference latency, and system complexity, which may limit real-world scalability. Furthermore, ensembling can obscure the interpretability of individual model decisions, complicating clinical validation and downstream error analysis. Future implementations should therefore balance the benefits of ensemble stability against operational efficiency and transparency, particularly when integrating LLM-based systems into clinical workflows.

5.2. Clinical Implications and Potential Applications

The proposed incidentaloma identification approach has several potential applications for clinical decision support (CDS), workflow optimization, and follow-up tracking in radiology practice. It could be integrated at the point of order entry within CDS systems to automatically identify patients with a prior incidentaloma and recommend the most appropriate follow-up imaging study. When incorporated into radiology reporting software, the model could provide real-time guidance based on lesion size, imaging characteristics, and patient demographics, suggesting the most relevant clinical practice guideline for follow-up. In addition, the structured lesion-level outputs could be used to populate follow-up tracking systems, enabling longitudinal monitoring of incidental findings

and improving adherence to recommended follow-up intervals. Moreover, it forms a foundation for evaluation about the downstream clinical and economic impact of imaging incidentalomas in the setting of patient co-morbidities and other lesions.

Our study also offers opportunities for improved explainability and transparency in clinical AI systems. Lesion-level analyses and LLM-generated explanations could be integrated into interactive dashboards, enabling radiologists to visualize model reasoning and validate follow-up recommendations. Such interpretability features would enhance clinician trust and support safe integration of AI-assisted tools into routine radiology workflows.

Collectively, these applications illustrate how automated incidentaloma detection can extend beyond research evaluation to support safer, more consistent, and guideline-concordant imaging care.

5.3. Limitations

Despite these findings, several limitations warrant consideration. First, the annotated dataset, while carefully curated and reviewed by radiologists, remains limited in size due to the intensive nature of manual lesion-level annotation. This constraint may limit the generalizability of results to less common anatomies or atypical phrasing styles. Second, radiology reporting conventions and imaging modalities vary widely across institutions, which could reduce the transferability of model performance to other health systems without additional adaptation or domain-specific tuning. Third, the evaluation primarily focused on text-based single-report analysis and did not incorporate temporal context from prior studies or longitudinal patient histories that could refine incidentaloma characterization and follow-up reasoning.

In addition, annotation subjectivity may have introduced minor inconsistencies, particularly in borderline cases where incidentaloma status or follow-up need was open to interpretation. While annotation review and consensus procedures were employed, subtle differences in annotator judgment could influence ground-truth labels. Furthermore, although model explainability was partially examined through token-level attribution and qualitative inspection, deeper interpretability analyses involving radiology experts are needed to fully understand the decision-making behavior of llms in complex radiology narratives.

Future work should therefore explore the integration of longitudinal patient data, cross-institutional validation, and more scalable annotation strategies, alongside improved interpretability frameworks, to enhance both the reliability and clinical trustworthiness of AI-assisted incidentaloma detection systems.

6. Conclusions

This study presents a comprehensive evaluation of supervised transformer-based encoders and generative LLMs for automated identification and classification of incidentalomas in radiology reports. By introducing lesion-tagged inputs and anatomy-aware prompting, we demonstrate that generative LLMs, particularly GPT-OSS (With Anatomy), achieve substantially stronger and more balanced performance than conventional supervised encoders. These gains were most pronounced in incidentaloma-positive cases, where explicit lesion and anatomical context enabled more precise reasoning about clinical relevance and follow-up necessity.

Beyond quantitative performance, this study emphasizes a detailed lesion-level error analysis. Incorporating structured lesion cues, such as lesion tags and anatomical information, promoted more consistent lesion-level inference in LLMs, while transformer-based supervised models produced fewer

false positives, reflecting a precision-oriented bias. Ensemble analysis further demonstrated that combining generative and encoder-based systems through majority voting stabilized predictions across clinically meaningful categories.

Although this work was conducted on data from a single institution and limited to individual reports without longitudinal context, the results provide clear evidence that structured lesion context and anatomical grounding are key to reliable and transparent LLM-based clinical reasoning in radiology. Future research should extend these findings through multi-institutional validation, temporal modeling, and incorporation of human-in-the-loop frameworks to ensure clinical robustness and generalizability.

Overall, this work advances the development of clinically aligned, anatomy-aware LLM frameworks for radiology report understanding, bridging the gap between automated text classification and clinical decision support in real-world radiology workflows.

Acknowledgments

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

References

[1] Berland LL, Silverman SG, Gore RM, Mayo-Smith WW, Megibow AJ, Yee J, et al. Managing Incidental Findings on Abdominal CT: White Paper of the ACR Incidental Findings Committee. Journal of the American College of Radiology. 2010;7(10):754-73.

[2] Lumbreras B, Donat L, Hernández-Aguado I. Incidental Findings in Imaging Diagnostic Tests: A Systematic Review. The British Journal of Radiology. 2010;83(988):276-89.

[3] Cai T, Giannopoulos AA, Yu S, Kelil T, Ripley B, Kumamaru KK, et al. Natural Language Processing Technologies in Radiology Research and Clinical Applications. RadioGraphics. 2016;36(3):738-53.

[4] Mabotuwana T, Hall CS, Tieder J, Gunn ML. Improving Quality of Follow-Up Imaging Recommendations in Radiology. In: AMIA Annual Symposium Proceedings. vol. 2017; 2018. p. 1196-204.

[5] Dalal S, Hombal V, Weng WH, Mankovich G, Mabotuwana T, Hall CS, et al. Determining Follow-Up Imaging Study Using Radiology Reports. Journal of Digital Imaging. 2020 Feb;33(1):121-30.

[6] Alsentzer E, Murphy J, Boag W, Weng WH, Jindi D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. Minneapolis, Minnesota, USA: ACL; 2019. p. 72-8.

[7] Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. arXiv Preprint arXiv:190405342. 2019.

[8] Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems. arXiv Preprint arXiv:230313375. 2023.

[9] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large Language Models Encode Clinical Knowledge. Nature. 2023;620(7972):172-80.

[10] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All You Need. Advances in Neural Information Processing Systems. 2017;30.

[11] Johnson AE, Pollard TJ, Shen L, Lehman LwH, Feng M, Ghassemi M, et al. MIMIC-III, A Freely Accessible Critical Care Database. Sci Data. 2016;3(1):1-9.

[12] Peng Y, Yan S, Lu Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In: Proceedings of the 18th BioNLP Workshop and Shared Task; 2019. p. 58-65.

[13] Li Y, Wehbe RM, Ahmad FS, Wang H, Luo Y. Clinical-Longformer and Clinical-BigBird: Transformers for Long Clinical Sequences. arXiv Preprint arXiv:220111838. 2022.

[14] Warner B, Chaffin A, Clavié B, Weller O, Hallström O, Taghadouini S, et al. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. arXiv Preprint arXiv:241213663. 2024.

[15] Sounack T, Davis J, Durieux B, Chaffin A, Pollard TJ, Lehman E, et al. BioClinical ModernBERT: A State-of-the-Art Long-Context Encoder for Biomedical and Clinical NLP. arXiv Preprint arXiv:250610896. 2025.

[16] Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33; 2019. p. 590-7.

[17] Gerevini AE, Lavelli A, Maffi A, Maroldi R, Minard AL, Serina I, et al. Automatic Classification of Radiological Reports for Clinical Care. Artificial Intelligence in Medicine. 2018;91:72-81.

[18] Welch HG, Black WC. Overdiagnosis in Cancer. Journal of the National Cancer Institute. 2010;102(9):605-13.

[19] MacMahon H, Naidich DP, Goo JM, Lee KS, Leung AN, Mayo JR, et al. Guidelines for Management of Incidental Pulmonary Nodules Detected on CT Images: From the Fleischner Society 2017. Radiology. 2017;284(1):228-43.

[20] Achiam, et al. GPT-4 Technical Report. arXiv Preprint arXiv:230308774. 2023.

[21] Hurst A, Lerer A, Goucher AP, Perelman A, Ramesh A, Clark A, et al. GPT-4o System Card. arXiv Preprint arXiv:241021276. 2024.

[22] Park N, Ramachandran GK, Lybarger K, Xia F, Uzuner O, Yetisgen M, et al. Identifying Imaging Follow-Up in Radiology Reports: A Comparative Analysis of Traditional ML and LLM Approaches. arXiv preprint arXiv:251111867. 2025.

[23] Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. Palm: Scaling Language Modeling with Pathways. Journal of Machine Learning Research. 2023;24(240):1-113.

[24] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:230213971. 2023.

[25] Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: Low-Rank Adaptation of Large Language Models. ICLR. 2022;1(2):3.

[26] Liu Z, Li Y, Shu P, Zhong A, Jiang H, Pan Y, et al. Radiology-GPT: A Large Language Model for Radiology. Meta-Radiology. 2025:100153.

[27] Wang G, Yang G, Du Z, Fan L, Li X. ClinicalGPT: large language models finetuned with diverse medical data and comprehensive evaluation. arXiv preprint arXiv:230609968. 2023.

[28] Park N, Lybarger K, Ramachandran GK, Lewis S, Damani A, Uzuner O, et al. A novel corpus of annotated medical imaging reports and information extraction results using BERT-based language models. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024); 2024. p. 1280-92.

[29] Han T, Adams LC, Papaioannou JM, Grundmann P, Oberhauser T, Löser A, et al. MedAlpaca—An Open-Source Collection of Medical Conversational AI Models and Training Data. arXiv Preprint arXiv:230408247. 2023.

[30] Maher DI, Williams E, Grodski S, Serpell JW, Lee JC. Adrenal Incidentaloma Follow-Up is Influenced by Patient, Radiologic, and Medical Provider Factors: A Review of 804 Cases. Surgery. 2018;164(6):1360-5.

[31] Song Z, Wu C, Kasmirski J, Gillis A, Fazendin J, Lindeman B, et al. Incidental Thyroid Nodules on Computed Tomography: A Systematic Review and Meta-Analysis Examining Prevalence, Follow-Up, and Risk of Malignancy. Thyroid. 2024;34(11):1389-400.

[32] Reiner BI. Radiology Reporting Past, Present, and Future: The Radiologist's Perspective. Journal of the American College of Radiology. 2007;4(5):313-9.

[33] Dutta S, Long WJ, Brown DF, Reisner AT. Automated Detection Using Natural Language Processing of Radiologists Recommendations for Additional Imaging of Incidental Findings. Annals of Emergency Medicine. 2013;62(2):162-9.

[34] Trivedi G, Dadashzadeh ER, Handzel RM, Chapman WW, Visweswaran S, Hochheiser H. Interactive NLP in Clinical Care: Identifying Incidental Findings in Radiology Reports. Applied Clinical Informatics. 2019;10(04):655-69.

[35] Kang SK, Garry K, Chung R, Moore WH, Iturrate E, Swartz JL, et al. Natural Language Processing for Identification of Incidental Pulmonary Nodules in Radiology Reports. Journal of the American College of Radiology. 2019;16(11):1587-94.

[36] Lau W, Payne TH, Uzuner O, Yetisgen M. Extraction and Analysis of Clinically Important Follow-Up Recommendations in a Large Radiology Dataset. AMIA Summits on Translational Science Proceedings. 2020;2020:335.

[37] Canton SP, Dadashzadeh E, Yip L, Forsythe R, Handzel R. Automatic Detection of Thyroid and Adrenal Incidentals Using Radiology Reports and Deep Learning. Journal of Surgical Research. 2021;266:192-200.

[38] Schumm M, Hu MY, Sant V, Kim J, Tseng CH, Sanz J, et al. Automated Extraction of Incidental Adrenal Nodules from Electronic Health Records. Surgery. 2023;173(1):52-8.

[39] Bala W, Steinkamp J, Feeney T, Gupta A, Sharma A, Kantrowitz J, et al. A web application for adrenal incidentaloma identification, tracking, and management using machine learning. Applied Clinical Informatics. 2020;11(04):606-16.

[40] Woo KmC, Simon GW, Akindutire O, Aphinyanaphongs Y, Austrian JS, Kim JG, 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. 2024;31(9):1983-93.

[41] Bhayana R, Elias G, Datta D, Bhambra N, Deng Y, Krishna S. Use of GPT-4 with Single-Shot Learning to Identify Incidental Findings in Radiology Reports. American Journal of Roentgenology. 2024;222(2):e2330651.

[42] Baş Aksu Ö, Aydın RF, Gökçay Canpolat A, Demir Ö, Şahin M, Emral R, et al. Artificial intelligence in endocrine practice: comparing ChatGPT, Gemini, and Claude for adrenal incidentaloma care. Journal of Endocrinological Investigation. 2025:1-11.

[43] Team G, Anil R, Borgeaud S, Alayrac JB, Yu J, Soricut R, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:231211805. 2023.

[44] Anthropic. Claude; 2024. Accessed: 08/2025. Available from: `https://www.anthropic.com/claude/sonnet`.

[45] Lee K, Dobbins NJ, McInnes B, Yetisgen M, Uzuner Ö. Transferability of Neural Network Clinical Deidentification Systems. Journal of the American Medical Informatics Association. 2021 09;28(12):2661-9. Available from: `https://doi.org/10.1093/jamia/ocab207`.

[46] Hu Y, Chen Q, Du J, Peng X, Keloth VK, Zuo X, et al. Improving Large Language Models for Clinical Named Entity Recognition via Prompt Engineering. Journal of the American Medical Informatics Association. 2024;31(9):1812-20.

[47] Lu Q, Li R, Wen A, Wang J, Wang L, Liu H. Large Language Models Struggle in Token-Level Clinical Named Entity Recognition. In: AMIA Annual Symposium Proceedings. vol. 2024; 2025. p. 748.

[48] O'Sullivan JW, Muntinga T, Grigg S, Ioannidis JP. Prevalence and Outcomes of Incidental Imaging Findings: Umbrella Review. BMJ. 2018;361.

[49] Grattafiori A, Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, et al. The Llama 3 Herd of Models. arXiv Preprint arXiv:240721783. 2024.

[50] Stenetorp P, Pyysalo S, Topić G, et al. BRAT: A Web-Based Tool for NLP-Assisted Text Annotation. In: Proceedings of the Demonstrations at the Conference of the European Chapter of the Association for Computational Linguistics. Avignon, France; 2012. p. 102-7. Available from: `https://aclanthology.org/E12-2021`.

[51] Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. arXiv Preprint arXiv:171105101. 2017.

[52] Zhang J, He T, Sra S, Jadbabaie A. Why Gradient Clipping Accelerates Training: A Theoretical Justification for Adaptivity. arXiv Preprint arXiv:190511881. 2019.

[53] Agarwal S, Ahmad L, Ai J, Altman S, Applebaum A, Arbus E, et al. gpt-oss-120b & gpt-oss-20b model card. arXiv Preprint arXiv:250810925. 2025.

[54] Ribeiro MT, Singh S, Guestrin C. " Why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016. p. 1135-44.