Automated Clinical Data Extraction with Knowledge Conditioned LLMs

Diya Li, Asim Kadav, Aijing Gao[†], Rui Li, Richard Bourgon[†]
[†]Freenome, South San Francisco, CA, USA
[†]{aijing.gao, richard.bourgon}@freenome.com
{916lidiya, asimkadav, raylee8023}@gmail.com

Abstract

The extraction of lung lesion information from clinical and medical imaging reports is crucial for research on and clinical care of lung-related diseases. Large language models (LLMs) can be effective at interpreting unstructured text in reports, but they often hallucinate due to a lack of domain-specific knowledge, leading to reduced accuracy and posing challenges for use in clinical settings. To address this, we propose a novel framework that aligns generated internal knowledge with external knowledge through in-context learning (ICL). Our framework employs a retriever to identify relevant units of internal or external knowledge and a grader to evaluate the truthfulness and helpfulness of the retrieved internal-knowledge rules, to align and update the knowledge bases. Experiments with expert-curated test datasets demonstrate that this ICL approach can increase the F1 score for key fields (lesion size, margin and solidity) by an average of 12.9% over existing ICL methods.
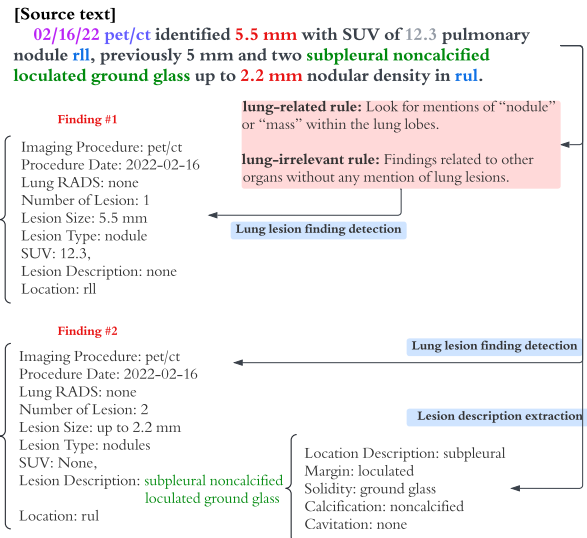
Figure 1: Example of lung lesion information extraction. Two findings (one describing a single lesion, and the other, two lesions) were identified in the source text. Example rules from the generated internal knowledge base are shown. First-stage finding detection and primary structured field parsing is followed by a second stage that further parses lesion description text.

1 Introduction

Lung lesion clinical data extraction from medical imaging and clinical reports plays a crucial role in enhancing the early detection and study of lung-related diseases (Zhang et al., 2018; Huang et al., 2024). Accurate automated extraction can reduce the manual effort required by a radiologist or physician. As illustrated in Figure 1, given a report, the task is to automatically extract information at the *finding* level, where a *finding* refers to text describing one or more closely related lesions. Since a report can have multiple findings, our task is to detect all findings and to parse each of them into a structured schema with pre-defined fields. (See Figure 1 and Tables 8 and 9).

However, interpreting unstructured text in reports presents a considerable challenge due to the complexity and variability of medical language (Wang et al., 2018). Creating specialized supervised machine learning models for specific medical terms can be effective but is often resource-intensive (Spasic et al., 2020). Recently, large language models (LLMs) have emerged as valuable assistive tools for general clinical data extraction (Singhal et al., 2023; Thirunavukarasu et al., 2023).

Nonetheless, using LLMs for clinical data extraction suffers from several challenges. First, LLMs often miss fine-grained details in clinical data extraction (Ji et al., 2023; Dagdelen et al., 2024), due to a lack of domain-specific knowledge. The extraction of lung lesion information requires an understanding of specialized fields (such as *margin* and *solidity*) that are not included in predefined schemas (Linguistic Data Consortium, 2006, 2008). Second, for extracting complex domain-specific

fields, LLMs often fail to understand nested sub-fields (Chen et al., 2024), and as a result, they may generate structurally inconsistent outputs.

To provide an automated method of clinical data extraction that addresses the above limitations, we propose a two-stage LLM framework that uses an *internal knowledge base* that is iteratively aligned with an expert-derived *external knowledge base* using in-context learning (ICL). Specifically, we first create the internal knowledge base by utilizing a manually curated medical report training corpus to generate *references*. The references that are deemed relevant to new input reports are converted into a set of higher-level *rules* that comprise the internal knowledge base. When extracting data from a report, rules from the internal knowledge base are *retrieved* and *graded* by our system to improve alignment with the external knowledge base. This process enhances the effectiveness of finding detection by leveraging relevant extraction patterns that are aligned with external knowledge. Lastly, to address the challenge of extraction of nested fields, we first extract an unstructured lesion description text field for each finding, then parse the description text into structured fields as a separate task that employs a more instructed approach (Figure 1).

We validate our approach through experiments using a curated dataset from a real-world clinical trial that includes annotations from medical experts. In addition, we define a new field schema for the lung lesion extraction task that may be useful for related lung disease studies. Our results demonstrate improvements in the accuracy of lung lesion clinical data extractions when using our framework compared to existing ICL methods.

2 Methodology

2.1 Task Definition

Our task is to extract lung lesions findings from clinical and imaging reports. Key fields include imaging procedure, lesion size, margin, solidity, lobe, and for PET/CT, standardized uptake value (SUV). [1] Extraction of the above fields is useful for oncology research and to support clinical care (American Cancer Society, 2024).

2.2 Clinical Data Extraction Framework

Given input reports $\mathcal{X}$, an internal knowledge base (KB) containing LLM-generated rules $\mathcal{D} =$

$\{d_1, ..., d_N\}$, and an expert-curated external KB $\mathcal{K} = \{k_1, ..., k_M\}$, our system aims to generate the extracted fields as $\mathcal{Y}$.

Our framework for aligning the KBs uses a retriever $\mathcal{R}$ and a grader $\mathcal{G}$. The retriever $\mathcal{R}$ retrieves the top k rules $\tilde{\mathcal{D}} = \{d_1, ..., d_k\}$ that are relevant to the input $\mathcal{X}$ from $\mathcal{D}$. The grader $\mathcal{G}$ then further selects and attempts to improve the rules $\tilde{\mathcal{D}}$ based on the input $\mathcal{X}$ and retrieved external knowledge $\tilde{\mathcal{K}} = \mathcal{R}(\mathcal{K}|\tilde{\mathcal{D}})$ from $\mathcal{K}$, resulting in $\hat{\mathcal{D}} = \mathcal{G}(\tilde{\mathcal{D}}|\mathcal{X}, \tilde{\mathcal{K}})$, where $\hat{\mathcal{D}}$ are knowledge aligned rules. By adding the improved $\hat{\mathcal{D}}$ to the default prompt, an LLM extracts the fields from reports $\mathcal{X}$ into our structured lesion field schema $\mathcal{Y}$.

2.3 Lung Lesion Knowledge Base Construction

We construct two KBs: an internal one generated by an LLM-based rule generator module using a small labeled training set, and an external one using expert knowledge resources.

Internal Knowledge Base Construction Using a small training set of annotated reports with lung lesion and non-lung lesion findings, we first ask the *rule generator* (implemented by an LLM) to create *lung-related* and *lung-irrelevant references*. A reference takes the following form:

source text "Additional soft tissue nodular density in the right upper lobe measuring 1.3 cm."

explanation "This finding is described as a 'soft tissue nodular density' measuring 1.3 cm, located in the right upper lobe. It may indicate a small mass or nodule."

To transform these references into more general, reusable *rules*, we next prompt the rule generator to identify common properties among the references and to extrapolate. For example, lung-related references often include measurements, so an extrapolated rule might be: "*Look for descriptions that include measurements (e.g., 'identified 5.5 mm', 'measures 1.8 x 1.2 cm') which often indicate lung lesions.*" The rules are generated in a multi-dialogue style, and the generation process is illustrated in Table 6. These generalized rules make up our internal knowledge base, denoted as $\mathcal{D}$, consisting of *lung-related* and *lung-irrelevant rules*. Example rules are provided in Table 3. We prioritize the *lung-irrelevant rules* since they assist
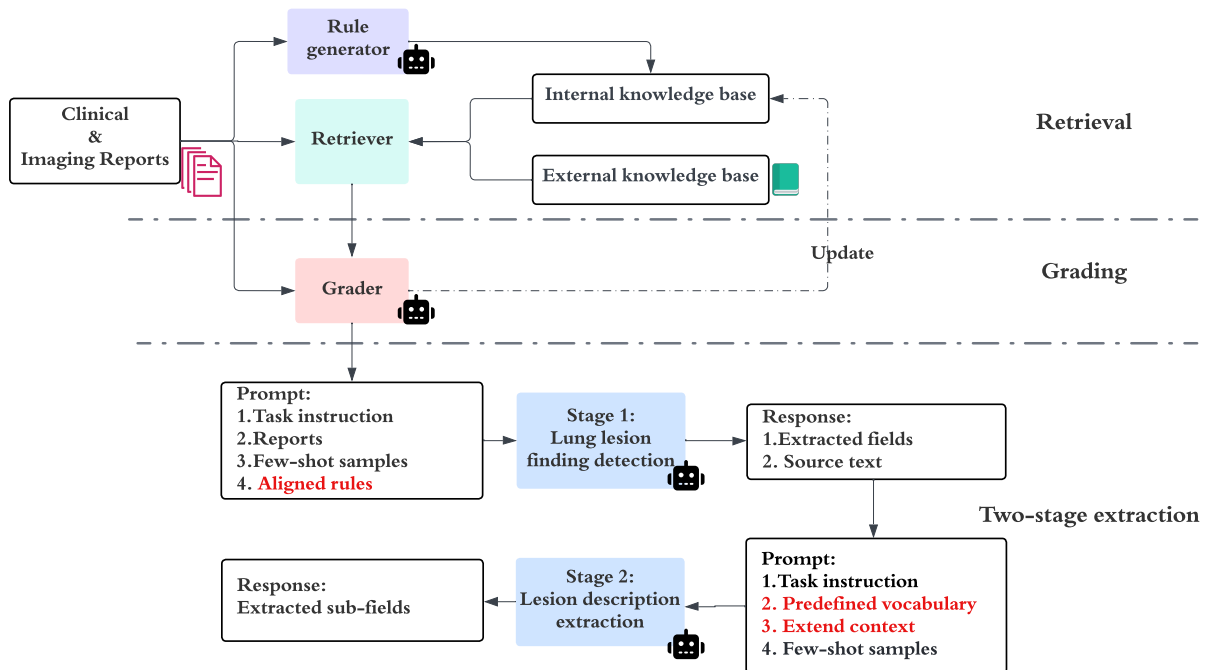

Figure 2: Framework for two-stage knowledge conditioned clinical data extraction. The 🤖 symbol indicates that the module is implemented by prompting an LLM. Rules used in prompts for lesion finding detection are derived from the internal knowledge base and aligned with external knowledge by a grader. Unstructured lesion description text is extracted in stage 1. In stage 2, this text is parsed into structured fields by providing the LLM with additional specialized inputs, including a controlled vocabulary.

LLMs in distinguishing findings that are not related to the lungs, thereby reducing false positives in inference.

External Knowledge-base Construction We manually identify external domain-specific knowledge from authoritative sources, including clinical guidelines[2,3,4] and expert medical opinions. The collected information is divided into chunks and stored in a local database for easy retrieval. The chunk size is 1000 with an overlap of 200 for indexing. The external KB, denoted as $\mathcal{K}$, encompasses a diverse range of content formats, including structured data, textual information, and procedural guidelines.

2.4 Retriever & Grader

Retriever Given reports $\mathcal{X}$, the retriever module $\mathcal{R}$ is responsible for identifying the top-k relevant *lung-related* and *lung-irrelevant rules* from the internal KB $\mathcal{D}$. This retrieval process matches the input reports with the most pertinent rules and returns these as $\tilde{\mathcal{D}} = \mathcal{R}(\mathcal{D}|\mathcal{X})$.

For each rule $d \in \tilde{\mathcal{D}}$, the retriever $\mathcal{R}$ also retrieves $\tilde{\mathcal{K}} = \mathcal{R}(d, \mathcal{K})$ from the external KB $\mathcal{K}$ for use by the grader G for knowledge alignment.

Grader To improve the quality of the retrieved rules $\tilde{\mathcal{D}}$, we introduce a grader $\mathcal{G}$, also implemented with an LLM. The grader is assigned two tasks and iterates over these tasks to refine the rules in internal KB $\mathcal{D}$.

First, $\mathcal{G}$ grades the rules in $\tilde{\mathcal{D}}$ with a *truthfulness* score, an integer ranging from 1 to 3, by comparing each d against retrieved external knowledge $\tilde{\mathcal{K}} = \mathcal{R}(d, \mathcal{K})$ and assessing its alignment with authoritative sources. If the *truthfulness* score of a rule falls below a threshold, the grader removes the rule from $\mathcal{D}$ and generates revised rules that are added back to $\mathcal{D}$. Second, the grader $\mathcal{G}$ assigns the aligned rules in $\hat{\mathcal{D}}$ a *helpfulness* score based on their relevance to the input reports $\mathcal{X}$. The *helpfulness* score is an integer ranging from 1 to 5. To assess *helpfulness*, the grader analyzes how well each rule supports the extraction and interpretation of information from $\mathcal{X}$. Rules that do not meet the helpfulness threshold are removed from $\hat{\mathcal{D}}$. This process is repeated for each rule d over I iterations, with I determined through practical experience. This iterative approach helps refine the alignment

Algorithm 1 Grading Algorithm

Input: imaging and clinical reports $\mathcal{X}$, retriever $\mathcal{R}$, grader $\mathcal{G}$, retrieved internal rules $\tilde{\mathcal{D}}$, external knowledge $\mathcal{K}$, number of iterations I, thresholds.
Output: aligned rules $\hat{\mathcal{D}}$, updated internal KB $\mathcal{D}$.

1: Initialize $\hat{\mathcal{D}} = \emptyset$;
2: **for** $i = 1$ to I **do**
3:     **for** $d \in \tilde{\mathcal{D}}$ **do**
4:         $\tilde{\mathcal{K}} = \mathcal{R}(d, \mathcal{K})$;
5:         $T = \mathcal{G}_{\text{truthfulness}}(d, \tilde{\mathcal{K}})$;
6:         **if** $T < \text{threshold}_T$ **then**
7:             $\mathcal{D} = \mathcal{D} \setminus \{d\}$;
8:             $d = \mathcal{G}_{\text{align}}(d, \tilde{\mathcal{K}})$;
9:             $\mathcal{D} = \mathcal{D} \cup \{d\}$;
10:         **end if**
11:         $\hat{\mathcal{D}} = \hat{\mathcal{D}} \cup \{d\}$;
12:     **end for**
13:     **for** $d \in \hat{\mathcal{D}}$ **do**
14:         $H = \mathcal{G}_{\text{helpfulness}}(d, \mathcal{X})$;
15:         **if** $H < \text{threshold}_H$ **then**
16:             $\hat{\mathcal{D}} = \hat{\mathcal{D}} \setminus \{d\}$;
17:         **end if**
18:     **end for**
19: **end for**

of the rules, ensuring that only the most relevant ones are retained in $\hat{\mathcal{D}}$. The prompts for assessing *truthfulness* and *helpfulness* can be found in Table 7.

The final set of high-scoring rules $\hat{\mathcal{D}}$ is used in prompts for lesion finding extractions. This iterative process is intended to increase the likelihood that rules in the updated $\hat{\mathcal{D}}$ yield outputs with the desired properties. The full grading algorithm is detailed in Algorithm 1, where $\mathcal{G}_{\text{align}}$ returns the aligned rule based on retrieved external knowledge.

2.5 Two-stage Extraction

Clinical data often contain nested information. For example, an imaging report may include two findings described in a single phrase: *"2 adjacent pulmonary nodules within the left lower lobe, the larger of the two measuring 5mm with an SUV of 2.39."* In cases like this, the LLM often fails to detect the second finding because it is not well separated from the first finding in the text.

To address this limitation, we decompose the clinical data extraction task into two stages: (i) lung lesion finding detection and primary structured field parsing, followed by (ii) further parsing of lesion description text.

For the first stage, we use $\hat{\mathcal{D}}$ as a part of the LLM prompt for the lung lesion finding detection, along with task instructions, the input reports, and few-shot samples (Table 8). The second stage aims to extract additional structured fields from the *lesion description* text. $\hat{\mathcal{D}}$ does not contribute in the second stage, as the set of valid terms to describe lesion description fields is limited. Instead, we provide the LLM with a controlled vocabulary based on the SNOMED ontology (SNOMED, 2024) (Table 9). We also note that the *lesion description* alone is often insufficient, since information missed in the first stage can lead to errors in subsequent extraction steps. To mitigate this issue, the second stage prompt combines the extracted *lesion description* text with the full *source text* from the first stage, to extend the context for extracting lesion description fields.

The two-stage extraction workflow is illustrated in Figure 2.

3 Experiments

3.1 Datasets

Our work utilizes clinical and imaging reports from Freenome's Vallania study (ClinicalTrials.gov, NCT05254834), which include lung cancer screening and other clinical results. Clinical experts manually identify all lung lesion findings and extract relevant fields based on our annotation schema (Table 5). To develop a gold standard dataset for performance evaluation, 19 subjects are randomly sampled. These subjects have a total of 31 clinical and 30 imaging reports, resulting in 189 findings. We randomly select 9 of these subjects as the training set, with the remaining 10 designated as the test set. Dataset and annotation details are discussed in Appendix A.1.

3.2 Evaluation Metrics

For a given test report, the gold standard findings and the system-detected findings may differ in number and/or ordering. The two sets of findings need to be aligned to one other. To achieve this, we perform an additional matching step and use the Hungarian algorithm[5] to match the gold-standard and system-detected findings. All extracted fields are used to construct the cost matrix for matching. We report micro precision, recall, and F1 scores for the extraction task.

---

[5] https://en.wikipedia.org/wiki/Hungarian_algorithm

Stage	Fields	Default Prompts			CoT			RAG			Ours		
		P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
# 1	Image procedure	78.2	63.6	70.1	79.5	64.2	<u>71.0</u>	75.3	61.1	67.5	86.5	72.7	**79.0**
	Lesion size †	85.9	82.1	84.0	87.3	84.2	<u>85.7</u>	83.0	78.5	80.7	92.8	85.8	**89.1**
	SUV	76.0	73.1	74.5	77.4	74.3	<u>75.8</u>	72.5	69.4	70.9	86.6	74.0	**79.7**
	Lesion type	83.7	67.3	74.6	85.1	68.7	<u>76.1</u>	80.2	63.9	71.2	88.1	73.6	**80.2**
	Lobe	72.7	60.4	66.0	74.0	61.5	<u>67.2</u>	70.0	57.5	63.0	81.9	69.6	**75.2**
#2	Margin †	68.4	65.0	66.7	68.5	67.5	67.5	75.0	63.2	<u>68.6</u>	90.0	76.3	**82.4**
	Solidity †	65.0	35.7	45.7	67.3	36.9	<u>47.7</u>	77.1	27.8	40.7	96.9	55.6	**69.2**
	Calcification	87.7	61.0	71.6	89.0	62.1	<u>73.2</u>	76.0	62.8	67.5	88.8	67.8	**75.7**
	Cavitation	50.0	100.0	66.7	60.0	100.0	<u>73.4</u>	50.0	100.0	66.7	87.5	100.0	**91.7**

Table 1: Overall precision (P), recall (R), and F1 scores evaluated on the test set. The results are averaged over 5 runs. The best results are marked in bold, and second best are underlined. Fields extracted in stage 1 vs. stage 2 are indicated. Fields marked with † are the clinically most important fields.

3.3 Module Implementation

The LLMs used for rule generation, the grader, lung lesion finding detection, and lesion description extraction are based on the official API of the Google PaLM2 model (Anil et al., 2023). All prompts used with LLMs are listed in Appendix A.3.

We use retriever $\mathcal{R}$ to obtain the top k relevant internal knowledge rules $\tilde{\mathcal{D}} = \mathcal{R}(\mathcal{D}|\mathcal{X})$ and retrieve external knowledge $\tilde{\mathcal{K}} = \mathcal{R}(\tilde{\mathcal{D}}, \mathcal{K})$ based on semantic similarity to $\tilde{\mathcal{D}}$. Specifically, we use the text embedding API (*text-embedding-004*) from Google (Google Vertex AI, 2024) to obtain the embeddings of $\mathcal{X}$, $\mathcal{D}$, and $\mathcal{K}$. Cosine similarity is used for semantic similarity scores. For hyper-parameter settings used in our system, refer to Table 10.

3.4 Comparison Baselines

As there is no prior work on lung lesion extraction using LLMs with our curated real-world dataset, we apply commonly-used ICL baseline methods and compare against the following:

Few-shot Learning Here, the LLM is provided with a small number of gold standard examples as a part of the basic prompts used for lesion finding detection and lesion description extraction (Brown et al., 2020). These prompts, referred to as *default prompts*, do not include any knowledge base content or additional guidance. We report results based on these default prompts, and other methods incrementally build upon them.

Chain of Thought (CoT) Additional instructions are added in the default prompts to guide the LLM to break down the lesion finding detection task into simpler, sequential steps by *thinking step by step* (Wei et al., 2022b). CoT is not applied at stage-two because this task is straightforward to conduct.

Retrieval Augmented Generation (RAG) RAG complements basic LLM queries, and it attempts to reduce hallucination by introducing external knowledge to improve the context (Lewis et al., 2020). We implement a RAG approach that directly retrieves information from $\mathcal{K}$ and adds the retrieved external knowledge chunks into the default prompt. This approach does not use the internal KB ($\mathcal{D}$).

3.5 Results and Analysis

Overall results The overall results are shown in Table 1. We are especially interested in the fields denoted with †, which include *lesion size*, *margin*, and *solidity*, because these are often of greatest clinical interest for cancer work (Nathan et al., 1962; Khan et al., 2011).

In our experiments, the benefit of Chain of Thought (CoT) reasoning is limited, as it appears to be more effective for traditional multi-step reasoning tasks, rather than our specialized extraction task (Wei et al., 2022a). The RAG implementation also performs poorly in the lung lesion extraction task — even worse than the default prompts. This may be due to incorrect retrieval of external knowledge based only on semantic similarity search, resulting in adding noise to the prompt. This suggests that the utility of the external knowledge ($\mathcal{K}$) may be constrained without first attempting to align it to the specific extraction task. Unlike RAG, our method first generates internal knowledge related to the specific extraction task. External knowledge is then utilized solely to align and update the internal knowledge. Results in Table 1 suggest that this improves the quality of our method's generated rules.

Our model outperforms all ICL baselines across all fields, particularly excelling in the † fields, with

Stage	Fields	w/o knowledge			w/o context			w/o grading			Ours		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
#1	Image procedure	78.2	63.6	70.1	86.5	72.7	**79.0**	84.6	61.7	71.4	86.5	72.7	**79.0**
	Lesion size †	85.9	82.1	84.0	92.8	85.8	**89.1**	92.4	85.1	88.5	92.8	85.8	**89.1**
	SUV	76.0	73.1	74.5	86.6	74.0	**79.7**	85.7	69.2	76.6	86.6	74.0	**79.7**
	Lesion type	83.7	67.3	74.6	88.1	73.6	**80.2**	91.2	69.8	78.9	88.1	73.6	**80.2**
	Lobe	72.7	60.4	66.0	81.9	69.6	**75.2**	80.8	63.2	70.8	81.9	69.6	**75.2**
#2	Margin †	68.6	67.8	67.2	85.8	75.0	80.0	84.5	66.7	74.1	90.0	76.3	**82.4**
	Solidity †	91.7	37.0	52.6	95.0	44.4	60.1	90.0	55.7	65.8	96.9	55.6	**69.2**
	Calcification	100.0	57.1	72.7	80.4	64.3	70.1	83.3	71.4	73.3	88.8	67.8	**75.7**
	Cavitation	55.6	100.0	71.5	75.0	100.0	83.4	66.7	100.0	77.8	87.5	100.0	**91.7**

Table 2: Ablation study on our two-stage knowledge conditioned model. Note that the performance of the model in the first stage without extended context ("w/o context") is the same as our full model, as the context extension is only applied during the second extraction stage.

an average of 12.9% increase in F1 score. Specifically, it achieves a 3.4% improvement in *lesion size*, over 13.8% in *margin*, and a 21.5% improvement in *solidity*.

Ablation Study To assess the contribution of each component of our method, we conduct ablation tests by removing each main module. The ablation results are listed in Table 2.

Notably, there is a significant performance decrease in the model that does not use the knowledge bases ("w/o knowledge"), indicating the importance of incorporating domain knowledge. Further, because lesion finding extraction quality degrades when the KBs are ignored, the quality of stage 2 lesion description extraction also degrades. Next, the model that omits providing extended context and the SNOMED controlled vocabulary for stage 2 ("w/o context"), performs worse for stage 2 fields. This indicates that extended context in stage 2 prompts can help prevent error propagation from stage 1, and that the controlled vocabulary standardizes the extraction of lesion description fields. Finally, we observe that the performance of the model that does not use the grader for knowledge alignment ("w/o grading") varies significantly across runs, suggesting that the grader's alignment role improves consistency and reduces noise.

3.6 Discussion

Case Study of Internal Knowledge In our knowledge conditioned model, the grader iteratively updates the internal knowledge if a rule's *truthfulness* score falls below a threshold, which is a hyper-parameter in our experiments.

To better understand the impact of the aligned rules in the internal KB, we identify the most frequently picked lung-related and lung-irrelevant rules from the test set (Table 3). Rules about *nodules* and *masses* are frequently picked, as these are two commonly used terms for lung lesion types. (See rules #2 and #3 in Table 3.) We also observe that the LLM tends be better at detecting lung lesion findings with explicit lesion sizes, using these as an anchor point to extract the full finding. *Solidity* information is sparse in clinical data, but there are many cases where the finding does not have size information yet it describes solidity. Terms like *solid*, *partly solid*, and *groundglass* often refer to the solidity field. Rule #1 in Table 3 contributes to the LLM's ability to detect lesion findings that reference lesion solidity.

For lung-irrelevant rules, the top-picked rule relates to findings in other organs, such as liver and kidney, without any mention of lung. Obviously, a rule of this type helps in distinguishing between relevant and irrelevant findings.

Effect of Retriever Top-k To determine the optimal k values for internal knowledge retrieval, we perform a grid search using the training set, evaluating the performance of *lesion size* extraction. Different values for k are considered for both lung-related and lung-irrelevant rules. As shown in Figure 3, the best extraction performance was observed when $k = 2$ for lung-related rules and $k = 1$ for lung-irrelevant rules. We use these optimal k values for extraction in the test set. An interesting finding is that using only a few rules contributes significantly to improving lesion size extraction performance.

Category	Rule #	Picked Rule
Lung-related	#1	{ "pattern": "solid | partly solid | groundglass", "rule": "Clinical notes mentioning 'solid', 'partly solid', 'groundglass' could indicate pulmonary nodule findings." }
	#2	{ "pattern": "nodule", "rule": "Look for descriptions that mention 'nodule', which often indicate lung lesions." }
	#3	{ "pattern": "mass", "rule": "Look for descriptions that mention 'mass' which often indicate lung lesions." }
Lung-irrelevant	#4	{ "pattern": "liver | kidney | other organs", "rule": "Findings related to other organs (e.g., liver, kidney) without any mention of lung lesions." }

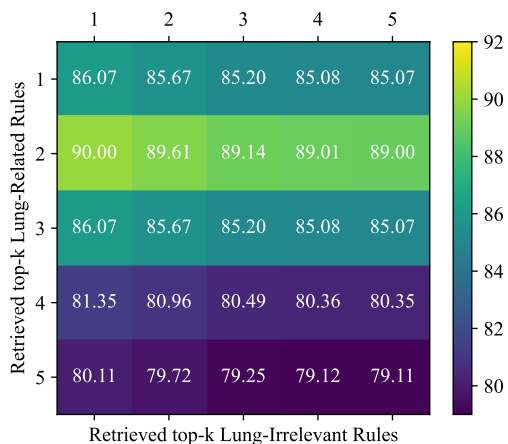Table 3: Most frequently picked lung-related and lung-irrelevant rules in test dataset.

Figure 3: Heatmap of lesion size extraction performance with varying values for the retriever's top-k hyperparameter, for both lung-related and lung-irrelevant rules.

4 Related Work

4.1 Clinical Information Extraction

Early work in clinical information extraction focused on rule-based systems and supervised machine learning techniques (Savova et al., 2010; Wang et al., 2018; Barrett et al., 2013; Denny et al., 2010; Mehrabi et al., 2015; Roberts et al., 2012; Li et al., 2015), which were labor-intensive to create and required a process that lacked scalability.

Recently, deep learning models, especially transformer-based architectures, have shown promise in clinical information extraction (Zhong et al., 2022; Spasic et al., 2020). These models reduce the need for extensive feature engineering, but they rely on high-quality annotated data.

In the past few years, LLMs have been applied to clinical information extraction (Goel et al., 2023; Wornow et al., 2024). LLMs can extract multiple fields simultaneously without requiring labeled training data for each field. While LLMs show promise in accelerating this process, high error

rates and frequent hallucinations still necessitate manual review. We propose a fully automated approach using novel techniques to improve accuracy and mitigate hallucinations.

4.2 Reference-guided Extraction

The idea of using external references or knowledge sources to guide information extraction has been explored in various domains, including clinical NLP (Demner-Fushman and Lin, 2005). Researchers have investigated the use of medical ontologies, knowledge bases, and domain-specific corpora to improve the performance of clinical information extraction systems (Goswami et al., 2019; Jin et al., 2022; Kiritchenko et al., 2010). These approaches typically involve incorporating external knowledge sources into the model architecture, or using them as auxiliary inputs during training or inference. However, existing methods may not fully leverage the evolving knowledge available in clinical references (Yan et al., 2024). In contrast, our system dynamically aligns and refines references with external knowledge, allowing for easy updates as new knowledge becomes available.

5 Conclusions

In this paper, we propose a novel framework for extracting lung lesion information from clinical and imaging reports using LLMs. Our approach aligns internal and external knowledge through in-context learning (ICL) to enhance the reliability and accuracy of extracted information. By dynamically selecting and updating internal knowledge and using external knowledge solely for internal-knowledge updates, our method outperforms commonly used ICL methods over data from real-world clinical trials. It excels in accurately detecting and extracting the most clinically relevant lesion information, such as lesion size, margin, and solidity.

Ethical Considerations

We recognize the importance of meeting all ethical and legal standard throughout our work, particularly in handling sensitive medical data and PII.

The clinical data used in this study may not be shared or distributed. All PII in the data used for this work have been fully redacted, to protect patient identities and adhere strictly to all relevant regulations, laws and guidelines. Our commitment to data security extends to our model development process, which is limited to the use of privacy friendly Google Cloud LLMs. This tool has been approved by our Data Governance Committee, ensuring that our practices align with institutional guidelines and maintain the highest standards of data security and compliance.

References

American Cancer Society. 2024. Lung cancer early detection, diagnosis, and staging. https://www.cancer.org/content/dam/CRC/PDF/Public/8705.00.pdf.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Neil Barrett, Jens H Weber-Jahnke, and Vincent Thai. 2013. Engineering natural language processing solutions for structured information from clinical text: extracting sentinel events from palliative care consult letters. In *MEDINFO 2013*, pages 594–598. IOS Press.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ruirui Chen, Chengwei Qin, Weifeng Jiang, and Dongkyu Choi. 2024. Is a large language model a good annotator for event extraction? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17772–17780.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418.

Dina Demner-Fushman and Jimmy Lin. 2005. Knowledge extraction for clinical question answering: Preliminary results. In *Proceedings of the AAAI-05 Workshop on Question Answering in Restricted Domains*, pages 9–13. AAAI Press (American Association for Artificial Intelligence) Pittsburgh, PA.

Joshua C Denny, Marylyn D Ritchie, Melissa A Basford, Jill M Pulley, Lisa Bastarache, Kristin Brown-Gentry, Deede Wang, Dan R Masys, Dan M Roden, and Dana C Crawford. 2010. Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*, 26(9):1205–1210.

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, et al. 2023. Llms accelerate annotation for medical information extraction. In *Machine Learning for Health (ML4H)*, pages 82–100. PMLR.

Google Vertex AI. 2024. Vertex ai generative ai documentation: Text embeddings. https://cloud.google.com/vertex-ai/generative-ai/docs/model-reference/text-embeddings. Accessed: 2024-06-14.

Google Vision. 2024. Cloud vision documentation. https://cloud.google.com/vision/docs. Accessed: 2024-06-14.

Raxit Goswami, Vatsal Shah, Nehal Shah, and Chetan Moradiya. 2019. Ontological approach for knowledge extraction from clinical documents. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1487–1491. IEEE.

Jingwei Huang, Donghan M Yang, Ruichen Rong, Kuroush Nezafati, Colin Treager, Zhikai Chi, Shidan Wang, Xian Cheng, Yujia Guo, Laura J Klesse, et al. 2024. A critical assessment of using chatgpt for extracting structured data from clinical notes. *npj Digital Medicine*, 7(1):106.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Song-fang Huang, Xiaozhong Liu, and Sheng Yu. 2022. Biomedical question answering: a survey of approaches and challenges. *ACM Computing Surveys (CSUR)*, 55(2):1–36.

Ali Nawaz Khan, Hamdan H Al-Jahdali, Klaus L Irion, Mohammad Arabi, and Shyam Sunder Koteyar. 2011. Solitary pulmonary nodule: A diagnostic algorithm in the light of current imaging technique. *Avicenna journal of medicine*, 1(02):39–51.

Svetlana Kiritchenko, Berry De Bruijn, Simona Carini, Joel Martin, and Ida Sim. 2010. Exact: automatic extraction of clinical trial characteristics from journal publications. *BMC medical informatics and decision making*, 10:1–17.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Qi Li, Stephen Andrew Spooner, Megan Kaiser, Nataline Lingren, Jessica Robbins, Todd Lingren, Huaxiu Tang, Imre Solti, and Yizhao Ni. 2015. An end-to-end hybrid algorithm for automated medication discrepancy detection. *BMC medical informatics and decision making*, 15:1–12.

Linguistic Data Consortium. 2006. Ace 2005 multilingual training corpus.

Linguistic Data Consortium. 2008. ACE (automatic content extraction) English annotation guidelines for relations v6.2. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-relations-guidelines-v6.2.pdf.

Saeed Mehrabi, Anand Krishnan, Alexandra M Roch, Heidi Schmidt, DingCheng Li, Joe Kesterson, Chris Beesley, Paul Dexter, Max Schmidt, Mathew Palakal, et al. 2015. Identification of patients with family history of pancreatic cancer-investigation of an nlp system portability. *Studies in health technology and informatics*, 216:604.

MH Nathan, VP Collins, and RA Adams. 1962. Differentiation of benign and malignant pulmonary nodules by growth rate. *Radiology*, 79(2):221–232.

Kirk Roberts, Bryan Rink, Sanda M Harabagiu, Richard H Scheuermann, Seth Toomay, Travis Browning, Teresa Bosler, and Ronald Peshock. 2012. A machine learning approach for identifying anatomical locations of actionable findings in radiology reports. In *AMIA Annual Symposium Proceedings*, volume 2012, page 779. American Medical Informatics Association.

Guergana K Savova, Jin Fan, Zi Ye, Sean P Murphy, Jiaping Zheng, Christopher G Chute, and Iftikhar J Kullo. 2010. Discovering peripheral arterial disease cases from radiology notes using natural language processing. In *AMIA Annual Symposium Proceedings*, volume 2010, page 722. American Medical Informatics Association.

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. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

SNOMED. 2024. Snomed ct. https://www.snomed.org/. Accessed: 2024-06-14.

Irena Spasic, Goran Nenadic, et al. 2020. Clinical text data in machine learning: systematic review. *JMIR medical informatics*, 8(3):e17984.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, et al. 2018. Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, 77:34–49.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022a. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Michael Wornow, Alejandro Lozano, Dev Dash, Jenelle Jindal, Kenneth W. Mahaffey, and Nigam H. Shah. 2024. Zero-shot clinical trial patient matching with llms. *Preprint*, arXiv:2402.05125.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.

Junjie Zhang, Yong Xia, Hengfei Cui, and Yanning Zhang. 2018. Pulmonary nodule detection in medical images: a survey. *Biomedical Signal Processing and Control*, 43:138–147.

Yizhen Zhong, Jiajie Xiao, Thomas Vetterli, Mahan Matin, Ellen Loo, Jimmy Lin, Richard Bourgon, and Ofer Shapira. 2022. Improving precancerous case characterization via transformer-based ensemble learning. *arXiv preprint arXiv:2212.05150*.

Lesion	Total	Training	Test
Subjects	19	9	10
Clinical reports	31	16	15
Imaging reports	30	14	16
Total findings	189	81	108

Table 4: Manually annotated lung lesion dataset statistics.

A   Appendix

A.1   Data Preparation and Annotation

We use a real-world dataset collected from a case-control, multicenter diagnostic study designed to gather blood samples for the development of blood-based screening tests. In the collected clinical and imaging reports, all personally identifiable information (PII) had been previously redacted. The textual information within these reports is extracted using optical character recognition (OCR) via Google's Cloud Vision API (Google Vision, 2024).

Two annotators with clinical expertise manually identify all lung lesion findings and extract relevant fields based on our annotation schema (Table 5). The inter-annotator agreement (IAA) is assessed using 10 reports reviewed by both annotators and calculated using Cohen's κ. The 10 reports include 5 clinical notes and 5 radiology reports from 2 subjects. The average Cohen's κ value for 9 lesion fields is 0.86. In cases where discrepancies are found, a third clinician participates to resolve the differences and ensure consensus. The counts of subjects, reports, and findings in the training and test splits are listed in Table 4.

A.2   Lung Lesion Annotation Schema

According to the Lung-RADS guidelines [6], the full annotation schema is described in Table 5.

A.3   Prompts

The system prompts for rule generator, grader, lung lesion finding detection, and lesion description extraction are presented in Table 6, 7, 8, and 9, respectively.

A.4   Hyper-parameters

The hyper-parameter settings for all modules are listed in Table 10.

---

[6]https://www.acr.org/-/media/ACR/Files/RADS/Lung-RADS/Lung-RADS-2022.pdf

Field	Description
Evaluator Signed On	The medical expert who signs the report, such as a physician, medical examiner, or pathologist. The expert's signature verifies the report and confirms their agreement with the findings and opinions.
Date of Report Signed	The date the medical expert signs the report.
Imaging Procedure	The imaging procedure identifying the pulmonary lesion, including documentation or comparisons of previous procedures.
Date of Imaging Procedure Performed	The date the imaging procedure is performed.
Lesion SeqNo	An auxiliary variable to help track the number of lesions described in a report, listed in chronological order if dates are available.
Number of Lesions	Indicates whether the lesions are solitary or multiple.
Lesion Size (mm)	Size can be reported in diameter, area, or all three dimensions (width, height, depth). Usually measured in millimeters; convert from centimeters if necessary.
SUV	The reported standard uptake value of the nodule, which may be provided even if lesion size is not mentioned.
Lesion Type	Terms used in medical imaging to describe small growths in the lungs, differing mainly in size. A pulmonary nodule is a rounded opacity $\leq$ 3 cm in diameter, while a pulmonary mass is > 3 cm. A pulmonary cyst is an air- or fluid-filled sac within lung tissue.
Lobe	The lobe of the lung where the nodule is located.
Lesion Description	Detailed description of the pulmonary lesion.
Margin	Describes the edge characteristics of the lesion, such as 'spiculated', 'well-defined', or 'irregular'.
Solidity (Morphology)	Refers to the shape and structure of the lesion, such as 'ground glass', 'partly-solid', or 'solid'. • For solid and part-solid nodules, the size threshold for an actionable nodule or positive screen is $\geq$ 6 mm. • For nonsolid (ground-glass) nodules, the size threshold is $\geq$ 20 mm. • On follow-up screening CT exams, the size cutoff is $\geq$ 4 mm for solid and part-solid nodules and/or an interval growth of $\geq$ 1.5 mm of preexisting nodule(s).
Calcification	Indicates if the pulmonary nodules are calcified.
Cavitation	A gas-filled space within the lung tissue.
Lung RADS Score	Lung-RADS is a classification system for findings in low-dose CT (LDCT) screening exams for lung cancer. Examples include '4A', '4B', and '4X'.

Table 5: Lung lesion annotation schema.

Role	Prompt
System	*You are a pulmonary radiologist. Your task is to extract key findings from the clinical or imaging reports.*
User	How many findings of Lung Lesions are present in the following text: {text}
System	{lesion_number}
User	Please provide detailed explanations.
System	{detailed_explanations}
User	Only {num_findings} findings should be classified as Lung Lesions, explain why they are and why the remaining findings are not. Return in JSON format of: {"lung lesion findings": ["referred text": "reason of being lung lesion finding"], "none lung lesion findings": ["referred text": "reason of not being lung lesion finding"]}
System	{references}
User	Transform the references into generalized, reusable rules by abstracting common properties. Format the output in the following JSON structure: ["pattern": "example pattern", "rule": "example rule description"]
System	{lung-relevant rules, lung-irrelevant rules}

Table 6: Multi-dialogue prompt template of rule generator.

You are a grader assessing the helpfulness and truthfulness of retrieved rules related to pulmonary (lung) lesions in the context of pulmonary lesion findings..

Given the clinical or imaging report, please evaluate the *helpfulness* of each rule on a scale from 1 to 5, where:
    1 means not helpful at all
    2 means slightly helpful
    3 means moderately helpful
    4 means very helpful
    5 means extremely helpful

Below is the clinical or imaging report:
{input_query}

Additionally, evaluate the *truthfulness* of each rule based on the retrieved knowledge on a scale from 1 to 3, where:
    1 means not truthful at all
    2 means partially truthful
    3 means completely truthful
Provide a brief explanation indicating how the rule can help in the extraction of pulmonary lesion characteristics and how the retrieved knowledge supports or refutes the rule.

Below is the retrieved external knowledge:
{external_knowledge}

Table 7: Prompt template for grader to assess helpfulness and truthfulness. Note that we chose a range score of 1-5 for truthfulness in our sample study, but extreme values of 1 and 5 are rare, so we set the range to 1-3.

You are a pulmonary radiologist. Extract key findings from the clinical or imaging report and organize them into the provided JSON structure.

Use the following JSON template as a guide:
```
[
    {
        "Imaging Procedure": "Enter imaging procedure here or 'None'",
        "Procedure Date": "Enter date in YYYY-MM-DD format here or 'None'",
        "Lung RADS": "Enter Lung RADS category here or 'None'",
        "Number of Lesion": "Enter number of lesion here or 'None'",
        "Lagest Lesion Size": "Enter lesion size here",
        "Lesion Type": "Enter lesion type here",
        "SUV": "Enter SUV here or 'None'",
        "Location": "Enter location here or 'None'",
        "Lesion Description": "Enter Lesion Description here or 'None'",
        "Text Source": "Enter text source here or 'None'",
    },
    {
        // Add additional finding as needed

    },
]       // Lung Lesion Findings
```

Below is the clinical or imaging report:
```
{input_query}
```

Below are some examples for reference:
```
{few_shot_samples}
```

Below are some lung-related rules for reference:
```
{corrected_rules}
```
Below are some lung-irrelevant rules for reference:
```
{corrected_rules}
```

Table 8: Prompt template for stage-1 lung lesion finding detection.

You are a pulmonary radiologist. Please extract location description, margin, solidity, calcification, cavitation from lesion description and organize them into the provided JSON structure.

Use the following JSON template with **preferred vocabularies** as a guide:
```
{
    "location description": "Enter location description here or 'None'",
    "margin": "Enter margin description here, preferably from the vocabulary
                ['spiculated', 'rounded', 'ill-defined', 'irregular', 'lobulated'] or 'None'"
    "solidity": "Enter solidity description only from the fixed vocabulary ['solid',
                'partly solid', 'groundglass', 'ground-glass',
                'groundglass and consolidative'] or 'None'",
    "calcification": "Enter calcification description here,
                    preferably from ['noncalcified'] or 'None'",
    "cavitation": "Enter cavitation description here,
                    preferably from ['mildly cavitary', 'cavitary'] or 'None'"
}
```

Below is the lesion description text:
```
{lesion_description_text}
```

Below is the full text of report containing the finding for reference:
```
{source_text}
```

Below are some examples for reference:
```
{few_shot_samples}
```

Table 9: Prompt template for stage-2 lesion description text structured data extraction.

Module	Hyper-parameter	Value
Rule generator	temperature	0.9
	top_p	1
Retriever	retrival threshold for external knowledge	0.9
	retrieved top-k lung-related rule	2
	retrieved top-k lung-irrelevant rule	1
Grader	number of interations I	3
	truthfulness threshold	2
	helpfulness threshold	4
Lesion finding detection	temperature	0.2
Lesion description extraction	temperature	0.2

Table 10: Hyper-parameter settings used by our clinical data extraction system.