Enhancing Pulmonary Disease Prediction Using Large Language Models with Feature Summarization and Hybrid Retrieval-Augmented Generation: Multicenter Methodological Study based on Radiology Report

Ronghao Li, Shuai Mao, Congmin Zhu, Yingliang Yang, Chunting Tan, Li Li, Xiangdong Mu, Honglei Liu, Yuqing Yang

Submitted to: Journal of Medical Internet Research
on: February 13, 2025

Table of Contents

Enhancing Pulmonary Disease Prediction Using Large Language Models with Feature Summarization and Hybrid Retrieval-Augmented Generation: Multicenter Methodological Study based on Radiology Report

Ronghao Li[1*]; Shuai Mao[2*]; Congmin Zhu[1*]; Yingliang Yang[1*]; Chunting Tan[3*]; Li Li[4*]; Xiangdong Mu[4*]; Honglei Liu[1*] DR (AM); Yuqing Yang[2*]

[1]No.10, Xitoutiao, You An Men, Fengtai District, Beijing, P.R.China, 100069 Capital Medical University Beijing CN
[2] Beijing University of Posts and Telecommunications Beijing CN
[3] Beijing Friendship Hospital Beijing CN
[4] Tsinghua Changgung Hospital Beijing CN
[*]these authors contributed equally

Corresponding Author:
Honglei Liu DR (AM)
No.10, Xitoutiao, You An Men, Fengtai District, Beijing, P.R.China, 100069
Capital Medical University
Fengtai District, Capital Medical University, Beijing, China
Beijing
CN

Abstract

Background: The rapid advancements in natural language processing (NLP), particularly the development of large language models (LLMs), have opened new avenues for managing complex clinical text data. However, the inherent complexity and specificity of medical texts present significant challenges for the practical application of prompt engineering in diagnostic tasks.

Objective: In this study, LLMs with new prompt engineering technology are explored to enhance model interpretability and improve the prediction performance of pulmonary disease based on traditional deep learning model.

Methods: A retrospective dataset including 2965 chest CT radiology reports was constructed. The reports were from four cohorts, healthy individuals, patients with pulmonary tuberculosis, lung cancer, and pneumonia. Then a novel prompt engineering strategy that integrates feature summarization(F-Sum), chain of thought (CoT) reasoning, and a hybrid retrieval-augmented generation (RAG) framework was proposed. A feature summarization approach, leveraging TF-IDF and K-means clustering, was employed to extract and distill key radiological findings related to three diseases. Simultaneously, the hybrid RAG framework combined dense and sparse vector representations to enhance LLMs' comprehension of disease-related text. Three state-of-the-art LLMs, GLM-4-Plus, GLM-4-air, and GPT-4o, were integrated with the prompt strategy to evaluate the efficiency in recognizing pneumonia, tuberculosis, and lung cancer. The traditional deep learning model, BERT, was also compared to assess the superiority of LLMs. Finally, the proposed method was tested on an external validation dataset consisted of 343 chest CT report from another hospital.

Results: Comparing with BERT-based prediction model and various other prompt engineering techniques, our method with GLM-4-Plus achieved the best performance on test dataset, attaining an F1 score of 0.8887 and accuracy of 0.8947. On external validation dataset, F1 score (0.8631) and accuracy (0.9167) of the proposed method with GPT-4o were the highest. Compared to the popular strategy with manually selected typical samples (Few-shot) and CoT designed by doctor (F1:0.8294, Accuracy: 0.8342), the proposed method that summarized disease characteristics (F-Sum) based on LLM and automatically generated COT performed better (F1: 0.8870, Accuracy: 0.8974). Although the BERT-based model got similar results on test dataset (F1: 0.8514, Accuracy: 0.8763), its predictive performance significantly decreases on the external validation set (F1: 0.4836, Accuracy: 0.7833).

Conclusions: These findings highlight the potential of LLMs to revolutionize pulmonary disease prediction, particularly in resource-constrained settings, by surpassing traditional models in both accuracy and flexibility. The proposed prompt engineering strategy not only improves predictive performance but also enhances the adaptability of LLMs in complex medical contexts, offering a promising tool for advancing disease diagnosis and clinical decision making.

(JMIR Preprints 13/02/2025:72638)
DOI: https://doi.org/10.2196/preprints.72638

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Original Manuscript

Enhancing Pulmonary Disease Prediction Using Large Language Models with Feature Summarization and Hybrid Retrieval-Augmented Generation: Multicenter Methodological Study based on Radiology Report

Abstract:

Background: The rapid advancements in natural language processing (NLP), particularly the development of large language models (LLMs), have opened new avenues for managing complex clinical text data. However, the inherent complexity and specificity of medical texts present significant challenges for the practical application of prompt engineering in diagnostic tasks.

Objective: In this study, LLMs with new prompt engineering technology are explored to enhance model interpretability and improve the prediction performance of pulmonary disease based on traditional deep learning model.

Methods: A retrospective dataset including 2965 chest CT radiology reports was constructed. The reports were from four cohorts, healthy individuals, patients with pulmonary tuberculosis, lung cancer, and pneumonia. Then a novel prompt engineering strategy that integrates feature summarization(F-Sum), chain of thought (CoT) reasoning, and a hybrid retrieval-augmented generation (RAG) framework was proposed. A feature summarization approach, leveraging TF-IDF and K-means clustering, was employed to extract and distill key radiological findings related to three diseases. Simultaneously, the hybrid RAG framework combined dense and sparse vector representations to enhance LLMs' comprehension of disease-related text. Three state-of-the-art LLMs, GLM-4-Plus, GLM-4-air, and GPT-4o, were integrated with the prompt strategy to evaluate the efficiency in recognizing pneumonia, tuberculosis, and lung cancer. The traditional deep learning model, BERT, was also compared to assess the superiority of LLMs. Finally, the proposed method was tested on an external validation dataset consisted of 343 chest CT report from another hospital.

Results: Comparing with BERT-based prediction model and various other prompt engineering techniques, our method with GLM-4-Plus achieved the best performance on test dataset, attaining an F1 score of 0.8887 and accuracy of 0.8947. On external validation dataset, F1 score (0.8631) and accuracy (0.9167) of the proposed method with GPT-4o were the highest. Compared to the popular strategy with manually selected typical samples (Few-shot) and CoT designed by doctor (F1:0.8294, Accuracy: 0.8342), the proposed method that summarized disease characteristics (F-Sum) based on LLM and automatically generated COT performed better (F1: 0.8870, Accuracy: 0.8974). Although the BERT-based model got similar results on test dataset (F1: 0.8514, Accuracy: 0.8763), its predictive performance significantly decreases on the external validation set (F1: 0.4836, Accuracy:

0.7833).

Conclusion: These findings highlight the potential of LLMs to revolutionize pulmonary disease prediction, particularly in resource-constrained settings, by surpassing traditional models in both accuracy and flexibility. The proposed prompt engineering strategy not only improves predictive performance but also enhances the adaptability of LLMs in complex medical contexts, offering a promising tool for advancing disease diagnosis and clinical decision making.

Key Words: Retrieval-Augmented Generation; Large Language Models; Prompt Engineering; Pulmonary Disease Prediction

Introduction

The rapid advancement of artificial intelligence (AI) has significantly driven the development of medical auxiliary diagnosis systems. Algorithms based on Electronic Health Records (EHRs), utilizing machine learning (ML) and deep learning (DL) techniques, have demonstrated considerable potential in predicting various clinical conditions and outcomes, such as lung cancer [1,2], breast cancer [3], and type 1 diabetes [4].

The Transformer architecture [5] has effectively addressed the challenge of capturing long-range dependencies within sentences, significantly enhancing models' abilities to understand contextual relationships. Consequently, Transformer-based models, such as Bidirectional Encoder Representations from Transformers (BERT) [6], have become the dominant approach for text representation learning. More recently, the advent of large language models (LLMs), including ChatGPT, BARD, and LLaMA, has provided researchers with novel tools for analyzing complex medical data [7-10], particularly clinical text. LLMs have shown excellent performance in tasks such as information extraction [11-14], text summarization [15,16], and text generation [17]. For example, BiomedGPT, the first open-source and lightweight visual LLM, was designed to perform various biomedical tasks [18], such as multimodal-based disease examination, auxiliary diagnosis, and information extraction. However, while benchmarks for general natural language processing (NLP) tasks exist, their efficacy in auxiliary diagnosis tasks remains underexplored [19-21]. These findings underscore the need for targeted design and optimization of LLMs to address the unique demands of specific medical tasks. One critical issue is that most general-purpose LLMs are trained on open-domain text and lack sufficient domain-specific adaptation, making it difficult for them to interpret the highly specialized, ambiguous, and context-dependent language used in clinical texts. Furthermore, commonly used prompt strategies, such as few-shot learning or manually designed reasoning chains (e.g., chain of thought (CoT)), often rely on handcrafted examples and static templates, which are difficult to scale and generalize across diverse disease types and institutions. As a result, such models often suffer from limited robustness, reduced interpretability, and suboptimal diagnostic performance in real-world clinical settings.

Pulmonary diseases, including lung cancer, pneumonia, and tuberculosis, remain among the leading causes of morbidity and mortality worldwide. These three diseases were selected due to their high clinical prevalence, significant public health burden, and distinct imaging and pathological characteristics, which provide a robust foundation for evaluating the performance of AI-based diagnostic tools. Additionally, their associated radiology reports often contain intricate medical terminology, ambiguous symptom descriptions, and heterogeneous clinical information, posing

challenges for accurate analysis and decision-making. Therefore, developing high-quality NLP methods targeting these conditions is crucial to improve disease prediction, facilitate early diagnosis, and enhance prognosis modeling. In recent years, AI-based approaches, particularly Transformer-based architectures, have shown significant potential in pulmonary disease diagnosis by extracting informative features from diverse types of clinical data, such as radiology reports, pathology results, and electronic health records (EHRs) [2, 22, 23]. For instance, Jun Shao et al. combined BERT and other deep learning models with the Swin Transformer to distinguish between bacterial, fungal, and viral pneumonia, as well as tuberculosis, demonstrating the effectiveness of hybrid architectures in multi-class classification tasks [24]. Hyeongmin Cho et al. used pathology reports and gold standard to generate prompt-response pairs for training and then applied GLMs to extract information for staging from pathology reports [25].

Despite these advancements, existing methods often struggle to capture the nuanced clinical context embedded in free-text narratives while maintaining interpretability in decision-making processes. In medical-assisted diagnosis, the predictive performance of traditional fine-tuning-based deep learning models is often constrained by the limited availability of single-center data, imbalanced data distributions, and variations in reporting standards across multiple institutions. Additionally, fine-tuning and optimizing these models require extensive expertise from AI specialists, which to some extent hinders their widespread clinical adoption.

In this study, we aimed to address the practical limitations of existing large language models (LLMs) in real-world diagnostic scenarios, particularly their insufficient domain-specific adaptation, challenges in interpreting complex clinical language, and reliance on rigid and handcrafted prompting strategies. To overcome these issues and facilitate accurate pulmonary disease diagnosis, we developed a novel prompt engineering strategy based on LLMs, combined with unsupervised learning technique, to explore new effective diagnostic assistance method. Inspired by the clinical reasoning process, our approach integrates automatic feature summarization, structured CoT reasoning, and a hybrid retrieval-augmented generation (RAG) mechanism. Together, these components enhance the ability of LLMs to understand and reason over radiology report texts, enabling accurate prediction of diseases such as pneumonia, tuberculosis, and lung cancer. We systematically compared our method with multiple prompting strategies and validated its performance using a multi-center dataset. Furthermore, comparing to the traditional fine-tuned BERT model, we demonstrated the superior predictive performance and generalizability of our approach. The proposed new method can achieve favorable prediction results without the need for complex fine-tuning process, making it convenient for integration into existing clinical workflows to assist

physicians in diagnosing diseases.

Methods

Dataset Description

This study utilized 2,965 chest CT radiology reports collected from Beijing Friendship Hospital in Beijing, China, spanning the years 2012 to 2018. The dataset comprised 367 reports from healthy individuals, 534 from patients with pulmonary tuberculosis, 553 from patients with lung cancer, and 1,511 from patients with pneumonia. The original texts of reports were extracted from medical information system and labels for the image findings were annotated by two respiratory physicians based on the report texts. Personal privacy information in the reports and reports with vague descriptions, diagnostic discrepancies, and text extraction errors were removed. Any discrepancies between their annotations were resolved through consensus with a senior physician to ensure accuracy. Each radiology report contained two key sections: image findings and labels. The dataset was then used to develop a multi-classification model for pulmonary disease prediction. Reports were further divided into a training set and a test set in an 8:2 ratio.

Additionally, an external validation dataset of 343 chest CT radiology reports was collected from Beijing Tsinghua Changgung Hospital. This dataset included 31 reports from healthy individuals, 23 reports from patients with pulmonary tuberculosis, 93 reports from patients with lung cancer, and 196 reports from patients with pneumonia.

The study was approved by the Human Research Ethics Committees of Beijing Friendship Hospital and Capital Medical University, Beijing, China. To safeguard patient privacy, all identifying information was removed prior to analysis.

LLM-based Classification Model

We improved the prompt engineering technology and developed a novel LLM-based classification framework by integrating three key components: clustering-based feature summarization, CoT, and hybrid RAG. Initially, TF-IDF and K-means clustering techniques are employed to extract representative radiological reports, which are subsequently summarized by LLMs. Based on summarized diseases features, LLM is used to generate CoT-based diagnostic inquiries to enhance structured reasoning. Then a hybrid RAG framework is implemented with both dense and sparse vector representations to retrieve similar reports. Finally, combined information from feature summarization, CoT reasoning, and hybrid retrieval was included in the prompt for LLMs to perform

pulmonary disease prediction. This framework is designed to emulate a physician's cognitive process in learning and diagnosing diseases, thereby enhancing predictive accuracy by identifying characteristic patterns in diseased populations and analyzing similarities with comparable cases. **Figure 1** illustrates the workflow of the proposed prompt engineering strategy.

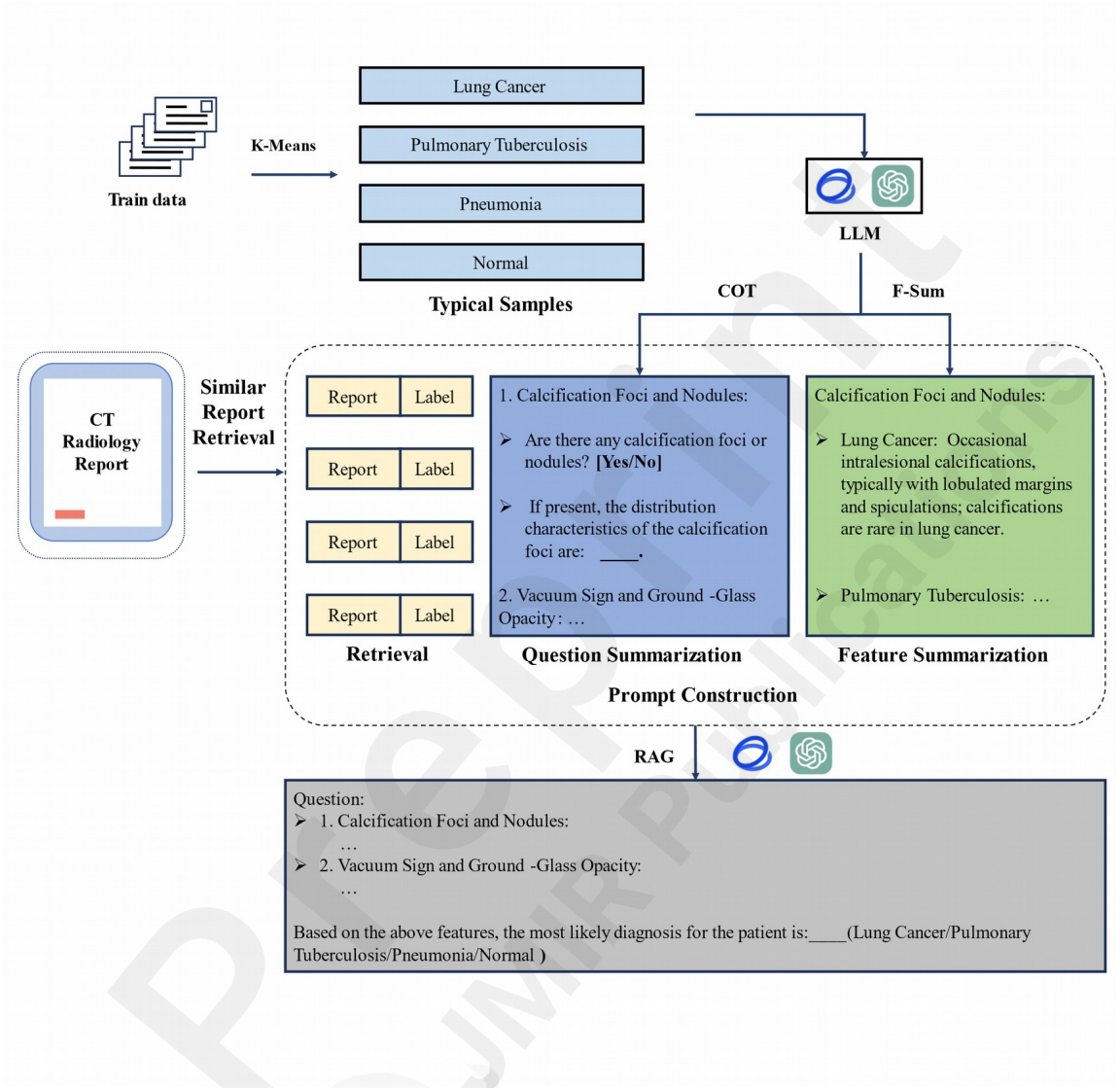


Figure 1. Workflow of pulmonary disease prediction using LLMs with feature summarization and hybrid RAG.

To enhance the understanding of complex medical terminology and radiological features, the workflow begins with feature summarization, where representative imaging findings for each disease are obtained using the TF-IDF and K-means clustering method. These findings are further analyzed by the LLM to extract disease-specific features and rank their importance. Then the summarized features are used by the LLM to generates diagnostic questions to construct a logical reasoning pathway. During the prediction phase, the workflow retrieves similar imaging reports via a hybrid RAG framework to refine the LLM's understanding of disease patterns, ultimately generating comprehensive and precise results for disease prediction.

Feature Summarization (F-Sum)

Feature summarization serves as the foundational step for learning disease characteristics by analyzing radiology text and summarizing key features. To facilitate interpretable and efficient feature extraction from radiology report texts, we adopted the TF-IDF method, which assigns importance scores to medical terms based on their frequency and specificity. This enables the selection of clinically relevant descriptors without requiring extensive labeled data. For clustering, we employed K-means due to its simplicity and effectiveness in grouping similar reports, which supports coherent feature summarization.

First radiology report text is transformed into a numerical feature matrix using TF-IDF and subsequently normalized. The top 5,000 features with the highest TF-IDF weights are retained. For each disease category, the K-means algorithm is applied, with the maximum number of clusters set to 20 or the total sample size of that category. Representative samples are proportionally selected from each cluster based on cluster size to ensure diversity. The samples closest to the cluster centroid are chosen to maintain category representativeness and balance. This step is pre-developed before analyzing the text with LLMs.

Then the representative samples derived from clustering are analyzed by the LLM to summarize key disease-specific features (**Figure 2A**). The LLM ranks these features by importance, allowing the model to prioritize critical characteristics in the diagnostic decision-making process. The tasks of feature summarization and Chain-of-Thought (COT) analysis were executed utilizing the GLM4-Plus model.

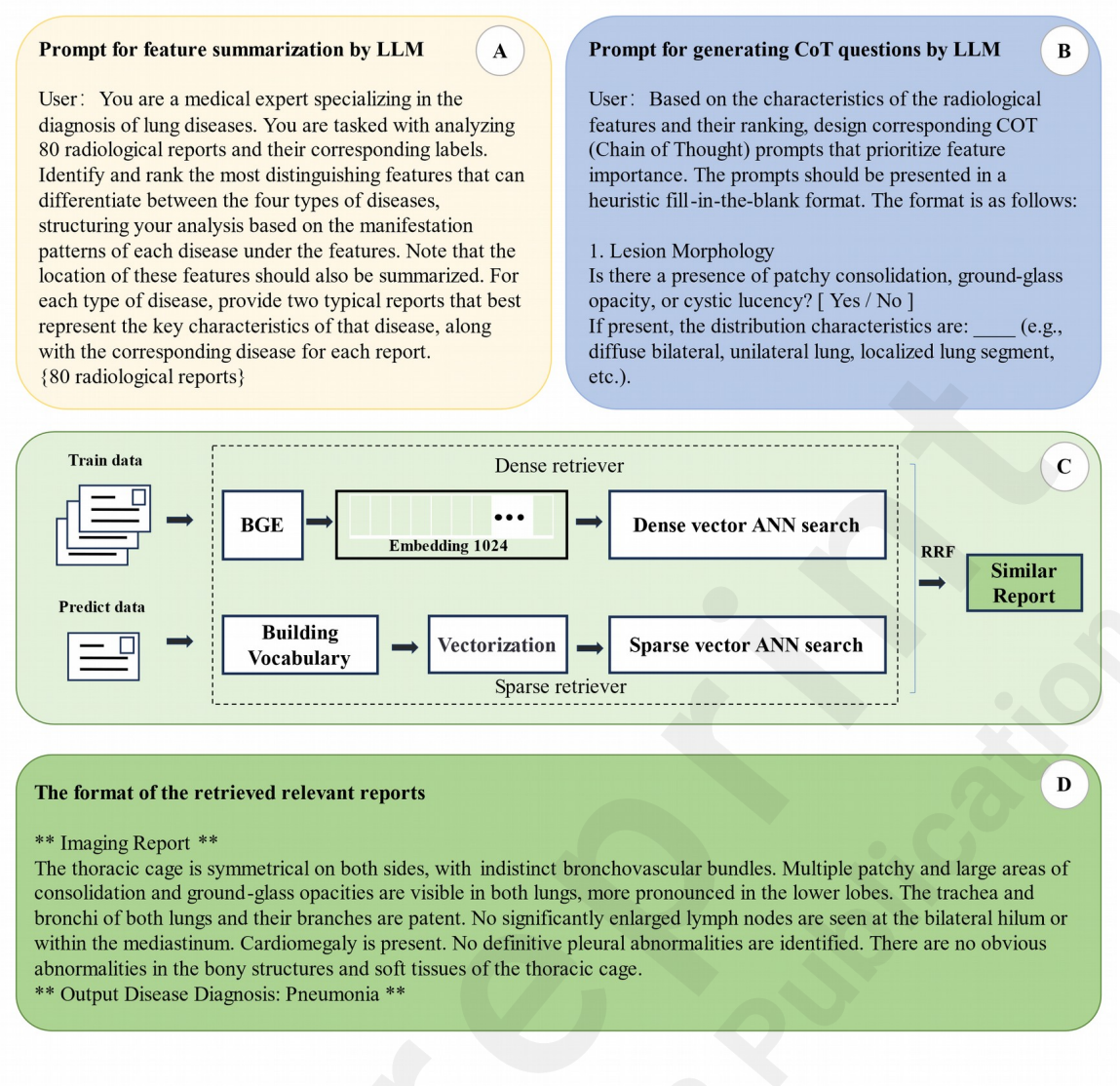


Figure 2. Detailed Strategy Diagrams for Each Section.
A. The prompt for feature summarization using the LLM. **B**. The prompt for generating Chains of Thought (CoT) questions with LLM. **C**. The hybrid Retrieval-Augmented Generation (RAG) process for retrieving similar reports based on both dense and sparse vector representations. **D**. The format of similar reports retrieved by RAG in the final prompt for LLMs.

CoT Reasoning

The CoT reasoning component guides the model through a step-by-step diagnostic process, enhancing interpretability and replicating clinical reasoning. Based on the representative imaging findings from clustering, the LLM generates diagnostic questions designed to emulate a physician's thought process (Figure 2B). These questions support the model in systematically deriving disease predictions by simulating a logical diagnostic workflow.

Hybrid RAG

The RAG technique enhances the LLM's generative capabilities by incorporating external knowledge through retrieval. In the disease classification task, RAG enables the model to extract

insights from similar cases, improving diagnostic precision.

We implemented a hybrid retrieval system that integrates dense and sparse vector representations (**Figure 2C**), utilizing the Milvus database. Dense vectors were generated using the pre-trained language model BGE-large-zh-v1.5. Average pooling was applied to create fixed-length embeddings. A custom vocabulary was built based on the imaging reports. Texts were converted into a term frequency matrix to generate sparse vector representation.

During retrieval, similarity queries were performed using L2 Euclidean distance, and the top 50 results were identified through nearest neighbor search. The results from both retrieval methods were then re-ranked using the Relevance-Weighted Rank Fusion (RRF) Ranker, merging them into a single retrieval result table to balance the contributions of different retrieval approaches,

$$RRF(d)=\sum_{r\in R}\frac{1}{k+r(d)}$$

where k =60, and r represents the similarity score in a specific retriever method. The top four most relevant imaging reports and disease labels were retained for subsequent predictions. Retrieved reports were formatted into prompts as shown in Figure 2D, enabling the LLM to leverage similar case knowledge for more accurate predictions.

Mode Training and Evaluation

LLM Selection

We employed novel prompt engineering strategies tailored for three prominent LLMs: GLM4-plus, GLM4-air, and GPT-4o. GLM4-plus and GLM4-air, developed by Zhipu AI, are general-purpose Transformer-based models optimized for Chinese NLP tasks, excelling in text comprehension and generation. GPT-4o, launched by OpenAI, inherits the advanced generative capabilities of the GPT series and introduces significant enhancements in multimodal processing and algorithm efficiency (https://openai.com/index/hello-gpt-4o).

To compare our approch with traditional deep learning approaches, we selected Bidirectional Encoder Representations from Transformers (BERT) as the baseline. BERT is a pre-trained Transformer-based language representation model that has demonstrated exceptional performance in a range of NLP tasks, including text classification, sentiment analysis, and named entity recognition.

Comparisons among Various LLM Prompt Engineering Strategies

As shown in Table 1, we evaluated multiple different prompt engineering strategies by combining few-shot, F-Sum, CoT, and RAG, across the three LLMs: GLM4-air, GLM4-plus, and GPT-4o. Few-

shot learning improves LLM performance by exposing the model to a limited number of examples relevant to the target task. For the few-shot, fixed samples for each disease category were provided to guide the model's learning process before performing prediction tasks. Additionally, we compared these strategies with manually designed prompt, denotated as the ‘Few-shot +CoT+ RAG（manual prompt）.’In this approach, both the few-shot learning samples and the CoT questions were designed through an expert-driven process, involving direct communication with medical specialists. This methodology emulates expert diagnostic procedures by incorporating critical features for distinguishing between diseases and formulating CoT questions that promote a deeper understanding and differentiation of medical conditions. The comparative analysis of predictive performance across these strategies provides a more comprehensive evaluation of the F-Sum + CoT + RAG approach, particularly highlighting its advantages in handling complex medical tasks.

Table 1. Performance Comparison of Different Prompt Engineering Strategies based on LLMs.

	GLM4-air	GLM4-Plus	GPT4o
Few-shot+CoT		+	
F-Sum+Few-shot	+	+	+
F-Sum +RAG	+		
F-Sum+Few-shot+CoT	+	+	+
F-Sum +CoT+RAG	+	+	+
F-Sum +CoT+RAG(BERT embed Cluster)	+		
Few-shot +CoT+ RAG(manual prompt)	+		

Comparison with BERT-based Model

To ensure a fair comparison, the BERT-based model was trained and tested using the same datasets as the LLMs. Text data was encoded into vector representations using bert-base-chinese as the backbone. The model consisted of 12 stacked BERT layers, each with 768 filters, and produced an embedding dimension of 768 after the BertPooler layer. A linear classification layer was added on top of the word vectors, and the Softmax function was used for predicting the categories of pulmonary diseases (Figure S1). During training, the AdamW optimizer was employed with a learning rate of 1e-5. The model was trained over 30 epochs with a batch size of 16. The maximum input length was set to 510 tokens to accommodate longer sequences.

The experiments were conducted on a single NVIDIA A6000 GPU (48 GB memory) with Python 3.8.5, PyTorch 2.0.0+cu117, CUDA 11.7, and Ubuntu 22.04.1.

Evaluation Metrics

We employed multiple evaluation metrics, including accuracy (Acc), precision (Pre), recall, and F1-score, to comprehensively assess model performance. For multi-classification tasks, we used the macro-average approach, which calculates the metrics (precision, recall, and F1-score) for each class individually and then computes their arithmetic mean across all classes.

Results

The performance of lung disease prediction using different prompt engineering strategies with LLMs and BERT is summarized in **Table 2**. On the test dataset, the F-Sum + CoT + RAG strategy achieved the highest results among all LLM-based methods. GLM-4-air achieved an F1-score of 0.887 and accuracy of 0.8974, while GLM-4-plus reached an F1-score of 0.8887 and accuracy of 0.8947. Notably, the F-Sum + CoT + RAG strategy demonstrated a balanced performance, achieving the highest recall (0.9210) with GLM-4-air and the highest precision (0.8837) with GPT-4o. These results indicate that the F-Sum + CoT + RAG strategy enhances predictive accuracy and achieves a good trade-off between recall and precision.

On the external validation dataset, GPT-4o with F-Sum + CoT + RAG outperformed other models, achieving the highest F1-score (0.8631) and accuracy (0.9167). It also maintained high recall (0.9131) and precision (0.8611), showcasing its strong generalization capabilities across diverse data distributions. In contrast, BERT exhibited significantly poorer performance, with its F1-score dropping from 0.8514 on the test set to 0.4836 on the external validation set.

Partial results of the feature summarization have been listed at **Table 3** and further analysis of confusion matrices (**Figure 3**) revealed nuanced insights into the model performance for specific disease categories. While LLM-based strategies performed well on lung cancer, they were less accurate for pneumonia, likely due to the larger sample size in the training data favoring BERT's fine-tuned recognition. Using the F-Sum+CoT+RAG method, GPT-4o exhibited low recall (65.31%, 32/49) with 15 cases misclassified as pneumonia (PN) in tuberculosis (TB) prediction, which is similar with the BERT (67.35%, 33/49), while GLM-4-air (recall=85.71%, 42/49) and GLM-4-Plus (91.84%, 45/49) demonstrated significantly superior performance. For PN, GPT-4o (recall=93.56%, 189/202) and BERT (99.01%, 200/202) were the highest, whereas GLM-4-air (85.64%, 173/202) and GLM-4-Plus (85.15%, 172/202) showed notable confusion between PN and TB (14 and 25 misclassifications, respectively). In lung cancer (LC) detection, GPT-4o (91.04%, 61/67), GLM-4-air (92.54%, 62/67), and GLM-4-Plus (94.03%, 63/67) surpassed BERT's recall (64.18%, 43/67), which misclassified 23 LC cases as PN. All models detected healthy individuals well: GPT-4o (95.24%, 60/63), GLM-4-air (98.39%, 61/62), GLM-4-Plus (96.77%, 60/62), and BERT (91.94%, 57/62).

Table 2. Performance of LLM-based and BERT-based lung disease prediction models.

Test Dataset	LLM	F1	Accuracy	Recall	Precision
F-Sum +CoT+RAG		**0.8870**	**0.8974**	**0.9210**	**0.8660**
F-Sum +RAG	GLM-4-air	0.8627	0.8711	0.9039	0.8427
F-Sum+Few-shot+CoT		0.8304	0.8342	0.888	0.8111

F-Sum+Few-shot		0.85	0.8579	0.9027	0.8324
Few-shot +CoT+ RAG(manual prompt)		0.8294	0.8342	0.8614	0.8419
F-Sum+CoT+RAG	GLM-4-PLUS	**0.8887**	**0.8947**	**0.9195**	**0.8728**
Few-shot+CoT		0.8044	0.7895	0.8838	0.8014
F-Sum+Few-shot		0.8121	0.7974	0.8917	0.8043
F-Sum+Few-shot+CoT		0.8103	0.7974	0.8931	0.7992
F-Sum+CoT+RAG	GPT4o	**0.8725**	**0.8974**	**0.8627**	**0.8837**
F-Sum+Few-shot		0.8128	0.8026	0.8797	0.8041
F-Sum+Few-shot+CoT		0.8512	0.8553	0.9006	0.8316
BERT	BERT	0.8514	0.8763	0.8062	0.9260
External Validation Dataset					
F-Sum+CoT+RAG	GPT4o	**0.8631**	**0.9167**	**0.9131**	**0.8611**
F-Sum+CoT+RAG	GLM-4-PLUS	0.7227	0.8167	0.7860	0.7417
F-Sum+CoT+RAG	GLM-4-air	0.7864	0.8500	0.9293	0.7377
CoT+ RAG(manual prompt)	GLM-4-air	0.7725	0.8500	0.8329	0.7956
BERT	BERT	0.4836	0.7833	0.4833	0.5051

Note: F1, recall, and precision were all Macro averages. The ‘Few-shot +CoT+ RAG (manual prompt)’ meant that the samples for few-shot and the questions for CoT were selected and written by expert.

Table 3. Partial results of the feature summarization.

Feature	Tuberculosis	Lung Cancer	Pneumonia	No Disease
Calcifications and Nodular Shadows ([illegible])	Spotty, patchy, or nodular calcifications, predominantly in the apical and posterior segments of the upper lobes, often accompanied by fibrous streaks and pleural traction. [illegible]	Occasional intralesional calcifications, typically with lobulated margins and spiculations; calcifications are rare in lung cancer. [illegible]	Calcifications are uncommon; if present, they are scattered and not associated with pleural traction. [illegible]	No calcifications or abnormal nodules. [illegible]
Cavity Formation ([illegible])	Thick-walled cavities with smooth inner walls, relatively large in diameter, often accompanied by patchy or nodular shadows and surrounding fibrosis. [illegible]	Rare cavitation; when present, it has irregular wall thickness and surrounding infiltration. [illegible]	Cavities are rare, observed only in prolonged cases. [illegible]	No cavitation. [illegible]

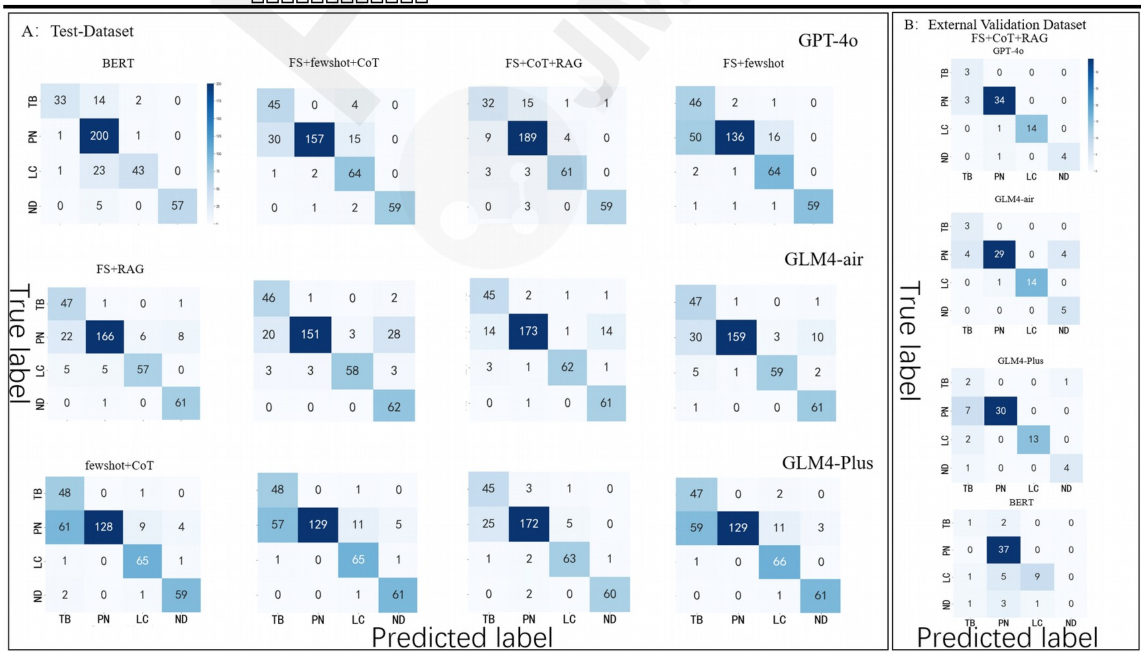


Figure 3. Confusion matrixes of LLM-based and BERT-based models. TB stands for Tuberculosis, LC for Lung Cancer, and PN for Pneumonia, while ND indicates No Disease.

Discussion

LLMs have shown exceptional potential in clinical applications, including disease diagnosis, treatment recommendation, and prognosis prediction. However, challenges such as suboptimal diagnostic accuracy and adherence to clinical guidelines persist. This study addresses these issues by integrating feature summarization, CoT reasoning, and hybrid RAG into a prompt engineering framework, resulting in significant performance improvements in pulmonary disease prediction.

The proposed strategy (F-Sum + CoT + RAG) demonstrated superior performance on both the test dataset (GLM-4-air: F1 = 0.887, Accuracy = 0.8974; GLM-4-plus: F1 = 0.8887, Accuracy = 0.8947) and the external validation dataset (GPT-4o: F1 = 0.8631, Accuracy = 0.9167). These results highlight the effectiveness of combining feature learning, CoT, and RAG to enhance LLM reasoning and representation capabilities across varying data distributions. Across different LLMs, F-Sum + CoT + RAG consistently outperformed F-Sum + few-shot + CoT and F-Sum + few-shot. Notably, GLM-4-Plus achieved the highest F1 score on the test dataset, while GPT-4o outperformed others on the external validation dataset, achieving the highest F1 score. The observed performance divergence may arise from the different distributions of diseases in the two datasets. On test set, the method with GLM-4-Plus recognizes the TB (45/49=91.8%) more accurately than the GPT-4o (32/49=65.31%). However, the accuracy (189/202=93.6%) of GPT-4o in identifying the PN is higher than the GLM-4-Plus (172/202=85.15%). For the external dataset, the number of samples classified as the PN (N=37) is far more than the TB (N=3), therefore, the superiority of the GLM-4-Plus in identifying the TB may not be fully reflected in its predictive performance. In general, GPT- or GLM-series LLMs outperform BERT in overall predictive accuracy. However, they exhibit varying predictive performance across different disease categories. Therefore, selecting an appropriate LLM for clinical application may require careful consideration of the primary diagnostic and therapeutic priorities.

Except GLM-series and GPT4-o LLMs, the DeepSeek-V3 is further tested to validated the generalizability of our method. With the same strategy (F-Sum+CoT+RAG), it achieved 93.16% accuracy on the test set and 86.67% on external validation, which is comparable to GLM/GPT-4o variants. Notably, the result with DeepSeek demonstrated particularly advantage in detecting LC, with an F1 score of 0.9177, which may be attributed to its training focus on oncology literature. Good predictive performance across multiple LLMs (GLM-4, GPT-4o, DeepSeek, etc.) confirms the benefits of the proposed method again.

To further evaluate the generalizability of the proposed prompt engineering strategies, we applied the optimal strategy (F-Sum + CoT + RAG) to the external validation dataset. LLMs all demonstrated strong performances, confirming the generalizability of the strategies. In contrast, the BERT model

showed a marked decline in generalization capability, with its F1 score dropping significantly from the test dataset (0.8514) to the external validation dataset (0.4836). These results underscore the superiority of LLM-based prompt engineering strategies in maintaining performance across multi-center datasets, compared to traditional fine-tuning models.

Feature summarization, a key component of this approach, enables the model to prioritize critical disease characteristics, enhancing its domain-specific knowledge. Compared to the Few-shot + CoT + RAG strategy, feature summarization consistently achieved better results, with an F1-score of 0.8870 versus 0.8294 on the test set. On the external validation dataset, Few-shot + CoT + RAG achieved an F1 score of 0.7725, while the feature summarization strategy achieved 0.7864. Before K-means clustering, TF-IDF was employed to extract important medical words to encode report texts. For comparative analysis, instead of TF-IDF, reports texts were represented with BERT-based embedding vectors. With the GLM4-air, the proposed method (F-Sum with BERT-based embeddings +CoT+RAG) achieved inferior performance (F1 = 0.8596, Accuracy = 0.8632, and Precision = 0.8447) to the method with TF-IDF (F1=0.8870, Accuracy = 0.8974, and Precision = 0.8660), which verify the efficiency of the combination of TF-IDF and K-means.

CoT reasoning further improved performance by simulating a clinician's step-by-step diagnostic process. For example, the inclusion of CoT reasoning raised the F1-score from 0.8627 to 0.8870 for GLM-4-air. RAG enhanced the generative ability of the model by retrieving contextually relevant information, enabling more accurate predictions. Experimental results revealed that multi-path retrieval outperformed single-path retrieval, as it provided more comprehensive background knowledge, enabling the model to make more accurate judgments.

As shown in Figure 4, compared to RAGs with either dense or sparse vector representations, the hybrid RAG configuration (dense + sparse) demonstrated obvious superiority, achieving 93.16% accuracy vs. 91.32% (sparse) and 88.42% (dense). Using the shapley value decomposition to quantified two representations’ contributions, the dense retrieval is +2.33% and sparse retrieval is +0.88%. The dense vectors capture semantic similarity (e.g., "nodular opacities"), while sparse representations ensure keyword matching (e.g., "caseating granulomas" for TB).

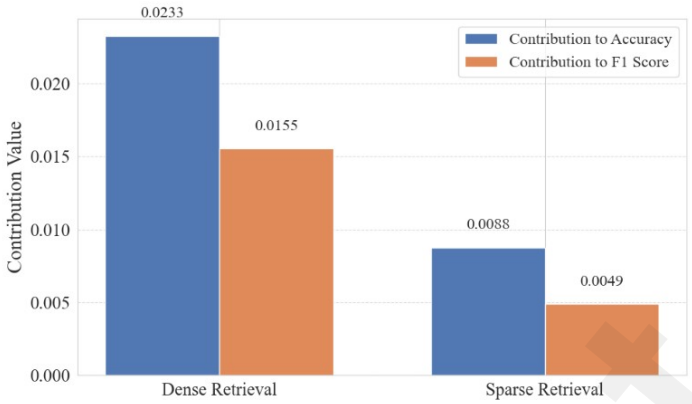


Figure 4. Analysis of contributions of sparse and dense representations of hybrid RAG to disease prediction

Despite these advances, certain limitations remain. This method struggled to distinguish between tuberculosis and pneumonia due to overlapping radiological features. For instance, acute miliary tuberculosis may present as diffuse micronodules in both lungs, resembling the findings in viral or Mycoplasma pneumonia. While LLMs sacrifice some pneumonia recall in favor of predicting other subcategories, their overall performance remains superior. In contrast, BERT, despite its seemingly good performance in distinguishing between tuberculosis and pneumonia, is highly influenced by factors such as sample imbalance and class size. As a result, BERT tends to overpredict pneumonia due to the higher prevalence of pneumonia cases in the dataset. These challenges illustrate the difficulty of differentiating pneumonia and tuberculosis in certain clinical scenarios. Future studies should integrate richer clinical expertise and multimodal data to address these challenges. Besides, when deploying models in the hospital, to ensure the data security, the model should automatically de-identify patient information. Replacing the existing online-called APIs of LLMs with locally deployed models will be necessary to reduce the risk of data leakage.

Based on our newly proposed method, the system can be practically deployed in real-world hospital settings by integrating it into existing hospital information systems (HIS) and picture archiving and communication systems (PACS). Radiology reports of various pulmonary diseases can be continuously added to the RAG-based knowledge base, enabling automatic summarization of disease features and real-time prediction for new reports. In clinical workflows, the system can be triggered after a radiology report is generated, providing diagnostic suggestions to physicians as decision support. The required infrastructure includes connection to EHR/PACS, a secure backend for model inference, and an interface for physicians to review and interact with AI-generated outputs, thereby enhancing diagnostic accuracy and efficiency.

Conclusions

In this study, we developed an LLM-based classification model for pulmonary disease diagnosis, leveraging typical disease patterns, simulating diagnostic reasoning processes, and systematically analyzing similar cases. The proposed model achieved significant improvements in diagnostic performance for pulmonary diseases, as validated across multi-center datasets, demonstrating robust generalization capabilities.

By integrating advanced prompt engineering techniques, including feature summarization, CoT reasoning, and RAG, this approach not only enhances predictive accuracy but also lays a foundation for extending auxiliary diagnostic and treatment tasks to other diseases. This work offers valuable insights and a theoretical framework to support the broader application of LLMs in clinical decision-making and medical research.

Data Availability

Data is available upon reasonable request to the corresponding author. The code related to model building is available on GitHub (https://github.com/zhegerushigui666/Enhancing-Pulmonary-Disease-Prediction-Using-Rrompting-and-RAG).

Author Contributions

Ronghao Li: Data curation, Methodology, Formal analysis, Software
Shuai Mao: Data curation, Methodology, Formal analysis, Software
Congmin Zhu: Data curation
Yingliang Yang: Data curation
Chunting Tan: Data curation
Li Li: Writing-review & editing
Xiangdong Mu: Writing-review & editing
Honglei Liu: Conceptualization, Writing-original draft, Funding acquisition, Supervision
Yuqing Yang: Conceptualization, Writing-original draft, Resources, Supervision

Funding

This work was supported by the Beijing Natural Science Foundation (No. 7242264, L246059), the National Natural Science Foundation of China (No. 62203060, 62403492, 82202299) and the R&D Program of Beijing Municipal Education Commission (No. KM202310025020)

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Huang S, Yang J, Shen N, Xu Q, Zhao Q. Artificial intelligence in lung cancer diagnosis and prognosis: Current application and future perspective. *Semin Cancer Biol.* 2023;89:30-37.
2. Feng X, Goodley P, Alcala K, et al. Evaluation of risk prediction models to select lung cancer screening participants in Europe: a prospective cohort consortium analysis. *The Lancet Digital Health.* 2024;6(9):e614-e624.
3. Uwimana A, Gnecco G, Riccaboni M. Artificial intelligence for breast cancer detection and its health technology assessment: A scoping review. *Comput Biol Med.* 2025;184.
4. Daniel R, Jones H, Gregory JW. Predicting type 1 diabetes in children using electronic health records in primary care in the UK_ development and validation of a machine-learning algorithm. *Lancet Digit Health.* 2024;6.
5. A V. Attention is all you need. *Advances in Neural Information Processing Systems.* 2017.
6. J D. Bert: Pre-training of deep bidirectional transformers for language understanding. *arxiv preprint.* 2018;arxiv:1810.04805.
7. Wu C, Lin W, Zhang X, Zhang Y, Xie W, Wang Y. PMC-LLaMA: toward building open-source language models for medicine. *J Am Med Inform Assoc.* 2024;31(9):1833-1843.
8. Kasthuri VS, Glueck J, Pham H, et al. Assessing the Accuracy and Reliability of AI-Generated Responses to Patient Questions Regarding Spine Surgery. *J Bone Joint Surg Am.* 2024;106(12):1136-1142.
9. Zhang J, Sun K, Jagadeesh A, et al. The potential and pitfalls of using a large language model such as ChatGPT, GPT-4, or LLaMA as a clinical assistant. *J Am Med Inform Assoc.* 2024;31(9):1884-1891.
10. Wang H, Gao C, Dantona C, Hull B, Sun J. DRG-LLaMA : tuning LLaMA model to predict diagnosis-related group for hospitalized patients. *NPJ Digit Med.* 2024;7(1):16.
11. Huang J, Yang DM, Rong R, et al. A critical assessment of using ChatGPT for extracting structured data from clinical notes. *npj Digital Medicine.* 2024;7(1).
12. Li T, Shetty S, Kamath A, et al. CancerGPT for few shot drug pair synergy prediction using large pretrained language models. *npj Digital Medicine.* 2024;7(1).
13. Oh Y, Park S, Byun HK, et al. LLM-driven multimodal target volume contouring in radiation oncology. *Nature Communications.* 2024;15(1).
14. Bhayana R, Nanda B, Dehkharghanian T, et al. Large Language Models for Automated Synoptic Reports and Resectability Categorization in Pancreatic Cancer. *Radiology.* 2024;311(3).
15. Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat Med.* 2024;30(4):1134-1142.
16. Guevara M, Chen S, Thomas S, et al. Large language models to identify social determinants of health in electronic health records. *NPJ Digit Med.* 2024;7(1):6.
17. Yalamanchili A, Sengupta B, Song J, et al. Quality of Large Language Model Responses to Radiation Oncology Patient Care Questions. *JAMA Network Open.* 2024;7(4).
18. Zhang K, Zhou R, Adhikarla E, et al. A generalist vision-language foundation model for diverse biomedical tasks. *Nat Med.* 2024;30(11):3129-3141.
19. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine.* 2024.
20. Shah-Mohammadi F, Finkelstein J. Accuracy Evaluation of GPT-Assisted Differential Diagnosis in Emergency Department. *Diagnostics.* 2024;14(16).
21. Longwell JB, Hirsch I, Binder F, et al. Performance of Large Language Models on Medical Oncology Examination Questions. *Jama Network Open.* 2024;7(6).
22. Kukhareva PV, Li H, Caverly TJ, et al. Lung Cancer Screening Before and After a Multifaceted Electronic Health Record Intervention. *JAMA Network Open.* 2024;7(6).
23. Saad MB, Hong L, Aminu M, et al. Predicting benefit from immune checkpoint inhibitors in patients with non-small-cell lung cancer by CT-based ensemble deep learning: a retrospective study. *The Lancet Digital Health.* 2023;5(7):e404-e420.
24. Shao J, Ma J, Yu Y, et al. A multimodal integration pipeline for accurate diagnosis, pathogen identification, and prognosis prediction of pulmonary infections. *Innovation (Camb).* 2024;5(4):100648.

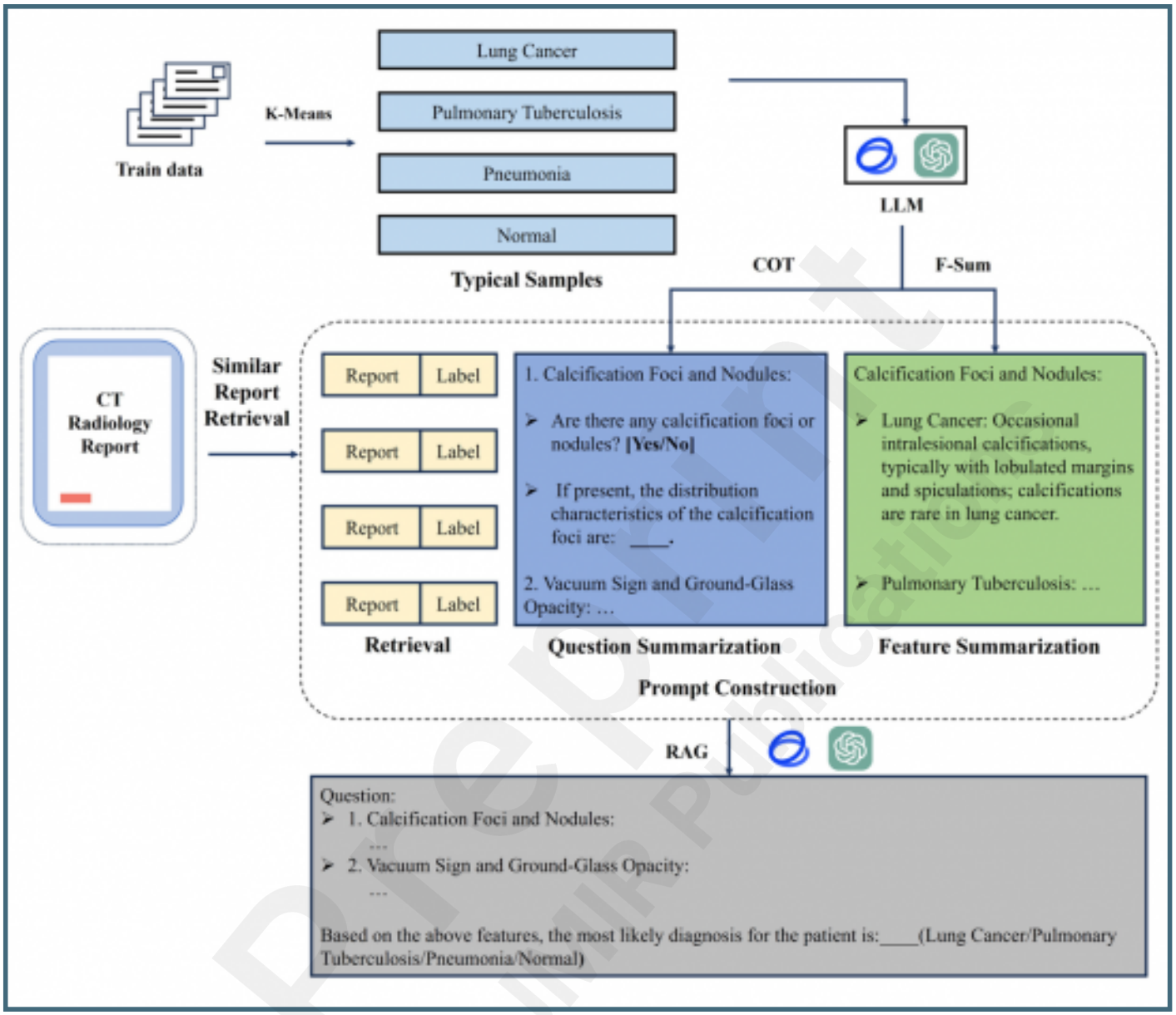
25 Cho H, Yoo S, Kim B, et al. Extracting lung cancer staging descriptors from pathology reports: A generative language model approach[J]. Journal of Biomedical Informatics, 2024, 157: 104720.

Supplementary Files

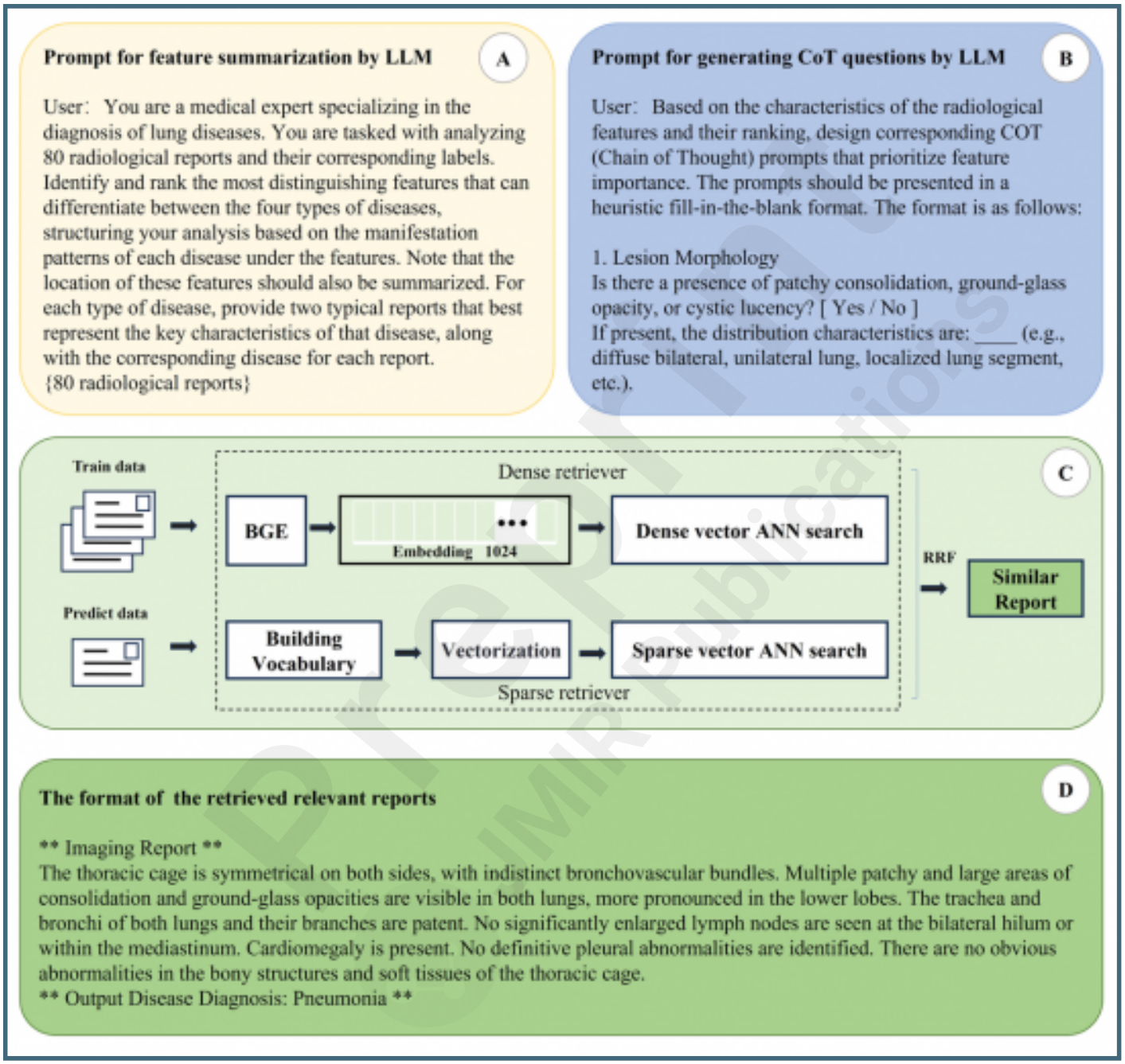
Supplementary files.
URL: http://asset.jmir.pub/assets/a5c2b2139e47fdcda231a596cbc6b414.docx

Figures

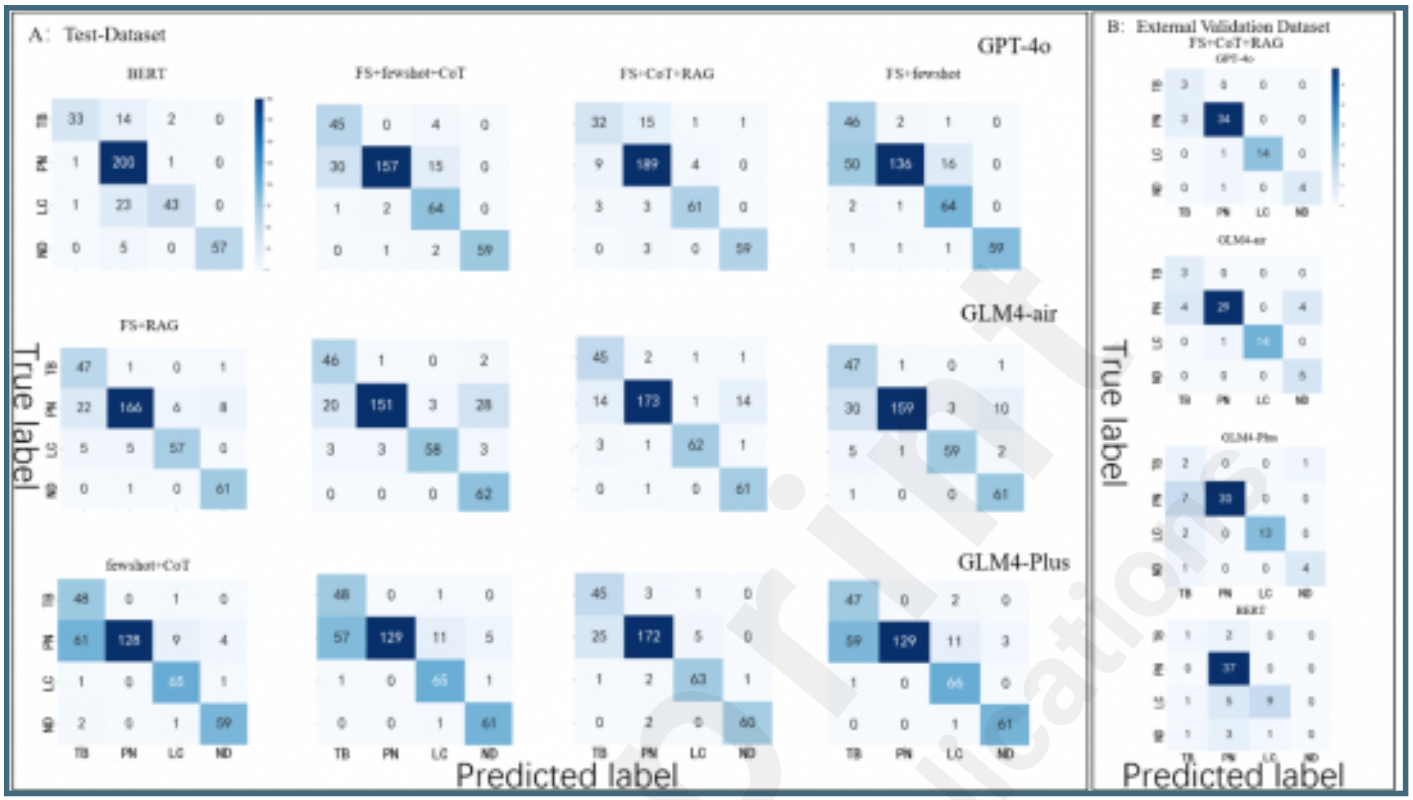
Workflow of pulmonary disease prediction using LLMs with feature summarization and hybrid RAG.



Detailed Strategy Diagrams for Each Section. A. The prompt for feature summarization using the LLM. B. The prompt for generating Chains of Thought (CoT) questions with LLM. C. The hybrid Retrieval-Augmented Generation (RAG) process for retrieving similar reports based on both dense and sparse vector representations. D. The format of similar reports retrieved by RAG in the final prompt for LLMs.



Confusion matrixes of LLM-based and BERT-based models. TB stands for Tuberculosis, LC for Lung Cancer, and PN for Pneumonia, while ND indicates No Disease.



Analysis of contributions of sparse and dense representations of hybrid RAG to disease prediction.