Automating Lung-RADS Categorization And Follow-Up Recommendations Using In-Context Learning With Large Language Models

Tiancheng Zhou, M.S[1], Aokun Chen, Ph.D[1], Yu Hu, M.S[1], Xiwei Lou, M.S[1], Xing He, Ph.D[2,3], Yu Huang, Ph.D[2,3], Bruno Hochhegger, MD, Ph.D[5], Hiren Mehta, MD[5], Mattia Prosperi, Ph.D[4], Yi Guo, Ph.D[1], Jiang Bian, Ph.D[2,3]

[1]Department of Health Outcomes & Biomedical Informatics, College of Medicine, University of Florida, Gainesville, Florida, USA; [2]Department of Biostatistics and Health Data Science, School of Medicine, Indiana University, Indianapolis, Indiana, USA; [3]Regenstrief Institute, Indianapolis, Indiana, USA; [4]Department of Epidemiology, College of Public Health and Health Professions, University of Florida, Gainesville, Florida, USA; [5]Health Shands Hospital, University of Florida, Gainesville, Florida, USA

Abstract

Lung cancer remains a significant challenge in public health, ranking among the leading causes of cancer-related mortality. Low-dose computed tomography (LDCT) -based lung cancer screening has emerged as an effective tool for early detection, particularly in high-risk populations. However, interpreting lung nodule characteristics from radiology reports can often be time-consuming and labor-intensive due to the length and inherent ambiguity of the reports, even with standardized reporting requirements like Lung-RADS. Generating Lung-RADS assessments from original radiology reports is a significant task for radiologists. This study addresses these challenges by developing an in-context learning framework utilizing large language models (LLMs). In this process, we aimed to identify the best approach that accurately categorizes lung nodules and streamlines management decisions, providing robust and interpretable decision support. Overall, this research aims to reduce the time and effort of the radiologist in lung cancer screening, ultimately enhancing efficiency and accuracy and enabling timely and precise interventions.

Introduction

Lung cancer remains one of the most aggressive malignancies and is the leading cause of cancer-related deaths in the U.S.[1–3] It is the second most common cancer in men, following prostate cancer, and in women, following breast cancer.[4] In 2025 alone, approximately 226,650 new lung cancer cases and 124,730 related deaths are projected.[5] Tobacco smoking accounts for approximately 80% to 90% of lung cancer cases, placing nearly 94 million former and current smokers at elevated risk.[6,7]. Despite ongoing advancements in treatment, long-term survival rates remain low, particularly for patients diagnosed at advanced disease stages.[2]

Lung cancer screening (LCS), particularly with low-dose computed tomography (LDCT), has significantly improved early detection and thus reduced overall mortality.[8] To standardize the reporting and management of nodules found on LCS, the American College of Radiology (ACR) introduced Lung Imaging Reporting and Data System (Lung-RADS) in 2014[9], integrating findings from landmark clinical trials such as the National Lung Screening Trial (NLST)[2] and the Nederlands–Leuvens Longkanker Screenings Onderzoek (NELSON) trial.[10] In the latest Lung-RADS guideline v2022[1], two outcomes were included: Lung-RADS score and follow-up managements. In LCS reports, pulmonary nodules detected on LDCT scans are categorized into six distinct groups: 1, 2, 3, 4A, 4B, and 4X, mainly based on size (e.g., larger nodules indicating higher risk), density (e.g., solid, part-solid, or ground-glass), growth pattern (i.e., stable versus enlarging), multiplicity (i.e., single versus multiple nodules), and morphology (e.g., irregular or spiculated borders suggesting malignancy). Each Lung-RADS category determines one of four standardized management pathways, which include routine annual LDCT screening (12-month follow-up), short-term LDCT surveillance at 6 or 3 months, or immediate further evaluation with diagnostic CT, PET-CT, or biopsy for high-risk lesions and nodules**.**

Although lung-RADS has improved standardization, its clinical application remains complex. Radiologists must perform a time-consuming multistep workflow involving a comprehensive LDCT scan review, detailed documentation of nodule characteristics (e.g., size, type, and growth), precise assignment of lung-RADS categories, and clear communication of follow-up recommendations, which may introduce inter-reader variability and management discrepancies.[11] Artificial intelligence (AI) has emerged as a promising tool for enhancing radiological accuracy and workflow efficiency, aiding in tasks such as the interpretation of LDCT scans.[12] Several studies have investigated machine learning (ML) and natural language processing (NLP) techniques for automated lung-RADS

classification and lung cancer risk prediction. Tammemägi et al. proposed PLCOm2012, which incorporated the Lung-RADS score in LC risk prediction.[13] Mendoza et al. developed AI models for predicting Lung-RADS categories 3 and 4 but did not provide a comprehensive classification approach.[14] Hsu et al. developed an artificial neural network (ANN) to predict positive and negative risks for LC with lung-RADS, but ANNs are often considered black boxes and lack interpretability to explain how the predictions were made.[15] Beyer et al. applied NLP to distinguish between suspicious and positive nodules but did not assign specific Lung-RADS categories.[16] White et al. conducted a similar study with Beyer under the patients with Lung-RADS score 2, 3, 4, 4A, 4B, and 4X.[17] Additionally, early applications of large language models (LLMs), such as ChatGPT, have primarily focused on general cancer staging rather than detailed Lung-RADS classification and structured follow-up recommendations.[18]

More recently, Singh et al.[19] demonstrated that ChatGPT-4o achieved higher accuracy in assigning Lung-RADS scores from radiology reports compared to ChatGPT-3.5, Gemini, and Gemini Advanced. However, this study faced several limitations. First, it only utilized proprietary models, raising concerns regarding data privacy and potentially hindering broader clinical adoption. Second, it did not incorporate advanced prompt-engineering strategies that could significantly enhance LLM performance. Our study addresses existing gaps by investigating advanced prompt-engineering techniques—specifically, chain-of-thought prompting and prompt chaining—for directly assigning Lung-RADS scores and predicting follow-up management from radiology narratives and comparing the performance of these two approaches to gain some insights into their effectiveness in estimating malignancy risk. We systematically evaluate multiple open-source LLMs of varying sizes to assess their performance on these critical clinical tasks, employing both multiclass and hierarchical classification strategies. Additionally, we perform an error analysis using explanations generated by the LLMs to identify specific performance bottlenecks.

Methods

Study Populations

Our study cohort was derived from patients with LDCT-based LCS or LC, extracted from the Integrated Data Repository at the University of Florida Health. A total of 25,514 patients were identified based on LCS procedure codes (Current Procedural Terminology (CPT)/Healthcare Common Procedure Coding System (HCPCS): G0296, G0297, S8032, 71271) and LC diagnosis codes (International Classification of Diseases, 9th Revision (ICD-9): 1622, 1623, 1624, 1625, 1628, 1629; 10th Revision (ICD-10): C34) between January 1, 2021, and June 30, 2024.

Within this cohort, we identified 10,960 radiology report and radiology impression report pairs related to LCS. The radiology reports included only lung nodule characteristics, while Lung-RADS scores and LCS management (i.e., target) were documented in impression reports. After excluding impression notes without Lung-RADS scores, we retained 10,224 report pairs for the analysis, as shown in Figure 1-a. Figure 1-b also shows the distribution of Lung-RADS scores where the majority (77.64%) of the notes contain patients with Lung-RADS scores 1 or 2.

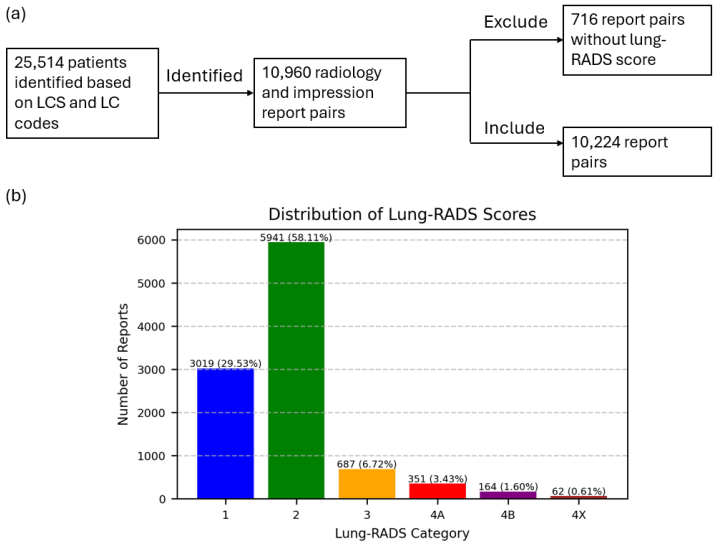

Figure 1. Study population, radiology report pairs selection, and the distribution of lung-RADS scores. (a) report pairs processing data flow. (b) distribution of Lung-RADS scores in selected reports.

System Framework

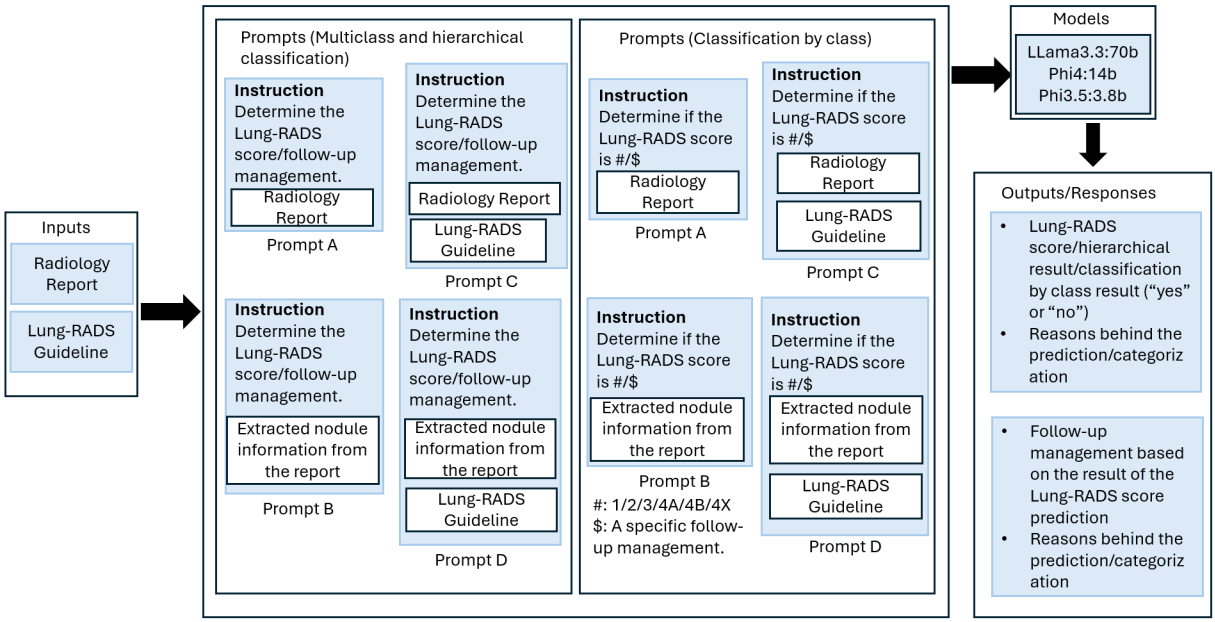

Figure 2. Lung-RADS Category and Follow-up Management Prediction with Large Language Models.

Figure 2 illustrates our proposed framework for the Lung-RADS category and follow-up management prediction with LLMs. The framework takes the prompt with two key inputs: a radiology report and the Lung-RADS guidelines for specific prompts. The radiology reports provide detailed LCS results, including nodule characteristics and other pulmonary findings for each patient. The Lung-RADS guideline provides detailed descriptions of nodule characteristics of each category that assist models by offering extra information to improve prediction performance. We insert the guideline into prompts as needed (See **Supplementary Figure 1**). This framework consists of two main components: Prompts constructed based on different prompting strategies and three LLMs of various sizes.

Prompting Strategies

We explored four prompting strategies. First, Chain-of-thought prompting (COT). A method proposed by Wei et al.[20] that introduces intermediate reasoning steps to enhance the ability of LLMs to solve complex reasoning tasks. In our study, the model processes the radiology report using instructions derived from the lung cancer screening guidelines, serving as the performance baseline. Second, Prompt-chaining (PC) is an NLP technique that structures a series of prompts to break down a complex task, guiding the model toward the desired output. In our case, the model first extracts nodule information and then utilizes this information for prediction. [21] Our third and fourth prompts are the knowledge-guided prompting[22] for COT and PC (COT-KG and PC-KG). These approaches are built upon COT and PC by incorporating Lung-RADS guidelines to enhance the model's reasoning and prediction accuracy. An example prompt, which combines the radiology report and guideline, is provided in **Supplementary Figure 1** of the supplementary materials.

Model Selection

A primary goal of our study was to identify the most accurate and efficient method for predicting LCS outcomes. Thus, we compared the state-of-the-art language models[23,24] and selected three candidates based on the following criteria: (1) deployable on a single GPU (e.g., Nvidia A100), 64 CPUs (AMD EPYC 7742 64-Core), 128GB of memory, and (2) top performed LLMs within different parameter scales based on benchmarking dashboards[23,25,26]. These include: (1) Phi-3.5 (3.8B parameters) – A small language model developed by Microsoft in August 2024, outperforming Llama 3.2 3B in two out of three reasoning tasks in benchmark comparisons.[25] (2) Phi-4 (14B parameters) – A mid-sized LLM released by Microsoft in December 2024, which outperformed Llama 3.3 8B on American Mathematics Competition (AMC) 10/12 tests that involve complex reasoning tasks, which is the core challenge of our study.[23,26] (3) Llama 3.3-70B (70B parameters) – A large-scale LLM developed by Meta and released in December 2024. It is considered one of the most advanced models available and has been shown to outperform GPT-4o.[27]

Output/Response Settings

For the framework outputs—Lung-RADS and follow-up recommendation management prediction—we conducted experiments with three classification settings: (1) Multiclass classification – Each radiology report was assigned a single Lung-RADS score; (2) Hierarchical classification – Following the design of Riel et al.[28], findings were first categorized into benign (Lung-RADS scores 1, 2, or a 12-month follow-up) or malignant (Lung-RADS scores 3, 4A, 4B, 4X, or other follow-ups, and the specific outcome was then predicted from within each category. (3) Classification by each class – This approach determined whether a report could be classified into a specific Lung-RADS category, treating the task as a binary classification problem. The model outputted "yes" or "no" for each category. We instructed the models to conduct per-class predictions for each LCS outcome, investigating potential limitations in LLMs.

For the follow-up management prediction, we consolidated and restructured specific categories based on their follow-up recommendations according to the specific version of the guideline[29] matching the date of the report: (1) Lung-RADS score 1 and 2 were merged into a single class, as both recommend a 12-month LDCT follow-up, (2) Lung-RADS score 3 was treated as an independent class, requiring a 6-month LDCT follow-up, (3) Lung-RADS score 4A was also classified separately, as it mandates a 3-month LDCT follow-up, and (4) Lung-RADS scores 4B and 4X were combined into one class, as both necessitate a chest CT for follow-up management. Following ML practice, we divided the dataset into training (70%) and testing (30%) sets, stratified by Lung-RADS score. The training data was used to refine the prompts before being applied to the testing dataset for evaluation.

Experimental Settings

All experiments were conducted on a server equipped with 64 AMD EPYC 7742 64-core CPUs, 128 GB of memory, and one Nvidia A100 80 GB GPU. The server runs an Ubuntu system, and the framework is implemented using LangChain 0.2.7, PyTorch 2.5.1+cu124, and Python 3.10.12. The models used were downloaded and hosted with Ollama v0.6.2.

Our study aims to provide a comprehensive evaluation of model performance, clinical applicability, and computational efficiency. The experiments were structured into three main components: lung-RADS score prediction, lung cancer screening follow-up management prediction, and efficiency comparison of LLMs. For lung-RADS score prediction, we evaluated the performance of three models across four prompting strategies, focusing on multi-class classification, hierarchical classification, and classification by individual class. The experimental setup for lung cancer screening follow-up management prediction mirrored the lung-RADS score prediction. Finally, we compared the computational efficiency of the models by measuring the average inference time under the worst-case scenario, i.e., the longest inputs with Chain-of-Thought + guideline prompting.

To ensure a fair assessment of the model's performance, we evaluate our framework using the following metrics: (1) Precision = TP / (TP + FP), where TP represents true positives and FP represents false positives. (2) Recall = TP / (TP + FN), where FN represents false negatives. (3) F1-score = 2 * (precision * recall) / (precision + recall), a balanced measure of precision and recall. (4) Accuracy = (TP + TN) / (TP + TN + FP + FN), where TN represents true negatives. For performance comparison, we primarily evaluate each model using F1-scores as the key performance metric.

Results

Lung-RADS Score Prediction

This section presents the performance of each model on Lung-RADS score prediction. Specifically, **Figure 3a** shows the results of multiclass classification under the four different prompt strategies, where the combination of COT or PC with the guideline in the prompt yielded the highest performance with the Llama 3.3-70B model, achieving a micro F1-score of 0.78. The complete comparison results of multiclass classification are shown in **Supplementary Figure 2** of the supplementary materials.

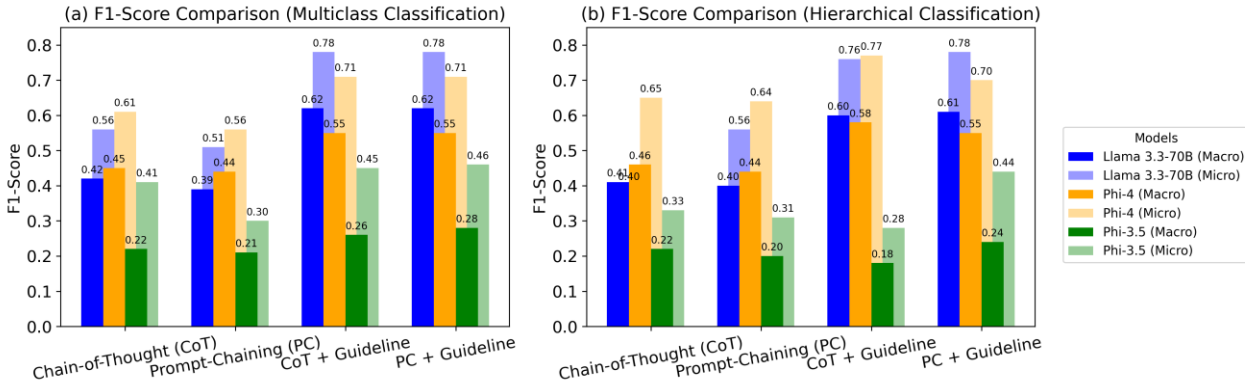

Figure 3. Model performance on multiclass and hierarchical classification. (a) Model performance on multiclass classification. (b) Model performance on hierarchical classification

Figure 3b presents the results of hierarchical classification, where Llama 3.3-70B with PC and guideline achieved the highest micro F1-score of 0.78, slightly outperforming the combination of the chain-of-thought with guideline. Meanwhile, Phi-4 received a 3% boost in F1 score with hierarchical classification, while Phi-3.5 suffered an even greater performance loss. Contrary to our assumption, the hierarchical classification setting did not improve the models on Lung-RADS score prediction. The complete comparison results of hierarchical classification are shown in **Supplementary Figure 3** of the supplementary materials.

Figure 4 illustrates the results of the Lung-RADS score prediction using Llama 3.3-70B, Phi-4, and Phi-3.5 on each Lung-RADS score independently. The highest performance across most categories was achieved with prompts incorporating both PC and Lung-RADS guideline for Llama 3.3-70B and Phi-4 models, while Phi-3.5 performed best with basic COT prompts. Llama 3.3-70B outperformed the two smaller models by a large margin, and all LLMs struggled to identify patients with a Lung-RADS score of 4X. Phi-3.5 failed to leverage the context in the prompt, making it unideal for Lung-RADS score prediction with in-context learning. The complete comparison results of each model on individual lung-RADS categories are shown in **Supplementary Figures 4-6** of the supplementary materials.

Management Classification

In this section, we present the performance of each model in predicting lung cancer screening management. **Figure 5a** illustrates multi-class classification results across all three models. Llama 3.3-70B achieved the highest performance with COT and the guideline at a micro F1-score of 0.86, with a 3% performance gap with Phi-4. Llama 3.3-70B and Phi-4 both benefit from the guideline, demonstrating its in-context learning capability. The complete comparison results of multiclass classification are shown in **Supplementary Figure 7** of the supplementary materials.

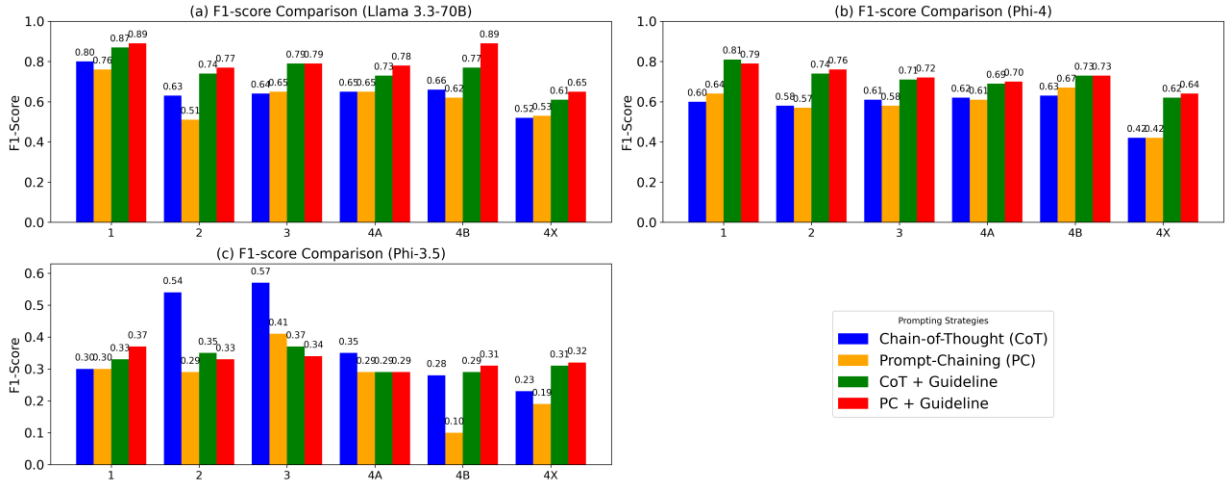

Figure 4. Model performance by Lung-RADS score. (a) performance of Llama 3.3-70B by Lung-RADS score. (b) performance of Phi-4 by lung-RADS score. (c) performance of Phi-3.5 by lung-RADS score.

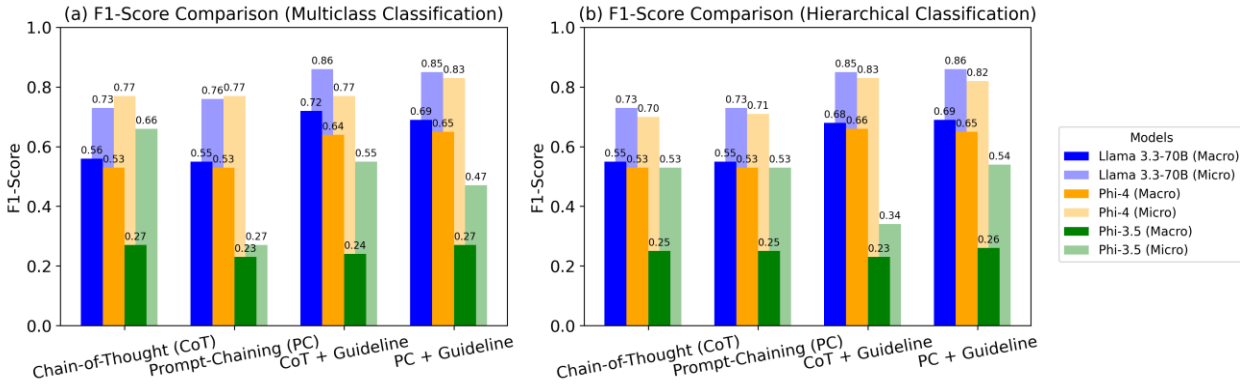

Figure 5. Model performance on multiclass classification and hierarchical classification. (a) Model performance on multiclass classification. (b) Model performance on multiclass classification

In hierarchical classification, as shown in **Figure 5b**, Llama 3.3-70B remains the leading model with a micro F1 score of 0.86, including prompts with PC and guidelines, followed by Phi-4 with a 2% performance gap in F1 score. Similar to Lung-RADS score prediction, the hierarchical classification did not improve the performance of LLM on lung cancer screening management prediction. The complete comparison results of hierarchical classification are shown in **Supplementary Figure 8** of the supplementary materials.

Figure 6 illustrates the LCS management performance of Llama 3.3-70B, Phi-4, and Phi-3.5. Similar to Lung-RADS score prediction, Llama 3.3-70B outperformed all the Phi models across all categories. Phi-4 remained the second-best model for LCS management prediction, achieving unpaired performance with Llama 3.3-70B with around 2% performance loss on the F1 score on predicting most categories of LCS management. Phi-3.5 still lacks the ability to utilize the information in prompts for LCS management prediction. The complete comparison results for each model across all follow-up categories are presented in **Supplementary Figures 9-11** of the supplementary materials.

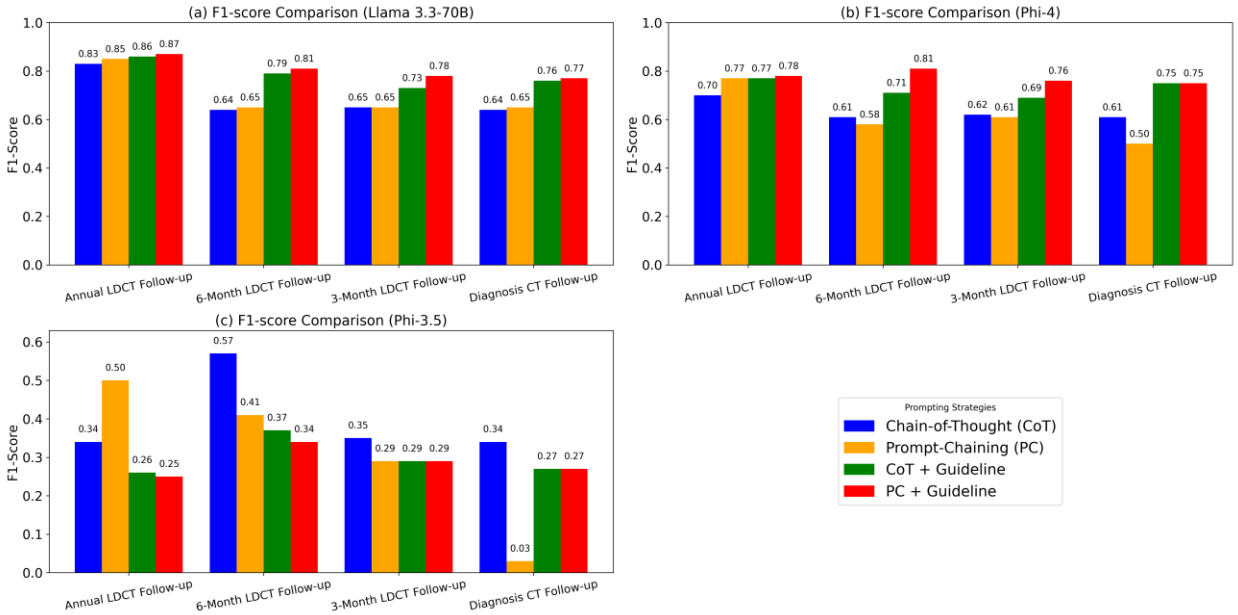

Figure 6. Model performance by lung cancer screening management. (a) Performance of Llama 3.3-70B by lung cancer screening management. (b) Performance of Phi-4 by lung cancer screening management. (c) Performance of Phi-3.5 by lung cancer screening management.

Efficiency of the Different Models

Regarding inference speed, Llama-3.3 requires significantly more time per iteration compared to the other two models

with the same computational resources. Under the worst case, where the prompt included COT and guideline, Llama-3.3 takes an average of 32.52 seconds per inference, whereas Phi-4 and Phi-3.5 used an average of 7.93 seconds and 5.04 seconds, respectively, as shown in **Table 1**.

Table 1. Computation resources and speed comparison.

Models	**Speed (worst scenario)**
Llama3.3-70b	32.52s/iteration
Phi-4	7.93s/iteration
Phi-3.5	5.04s/iteration

Error Analysis

To further understand the bottleneck in predicting the Lung-RADS score, as well as management recommendations, we conducted an error analysis with 330 misclassified cases on the best-performed setup: Llama 3.3-70B with chain-of-thought prompting and Lung-RADS score guidelines. In addition to predicting the Lung-RADS score, we instructed the LLM to output its path of reasoning, allowing us to investigate the cause of the errors. **Table 2** presents the primary causes of errors, including hallucination, failure to follow instructions, and unclear instructions. In hallucination cases, the model correctly extracted relevant information and cited the correct section of the Lung-RADS guideline but predicted the wrong Lung-RADS score, pretending to follow the guideline. In the cases of failing to follow instructions, the model disregarded the explicit instructions from the prompt, resulting in following the wrong path of reasoning. There were also cases where the instructions were unclear, especially for identifying patients with a Lung-RADS score of 4X. The Lung-RADS 4X is defined as nodules classified as category 3 or 4 with additional suspicious features, where neither the guideline nor clinicians held an exhaustive list of such features, leading to errors in identifying 4X cases. These findings suggest that while LLMs show promising results for Lung-RADS score and LCS management prediction, LLMs could still benefit from stronger reasoning abilities and more detailed prompt design.

Table 2. Primary causes of errors from error analysis (excluding potential ground truth errors).

Primary Causes	**Example Reasons**	**Explain**
Hallucinations (32%)	New hazy nodule measuring 4 mm. Given its new status and small size, it leans towards a lower suspicion category.	The model identified that the nodule is new but relied solely on its size on the prediction, disregarding nodule characteristics “new” and “hazy”.
Failed to follow instructions (23%).	It's considered a new nodule since there's no previous comparison available. The size of the nodule (5 mm) fits the criteria for Category 3 as a new nodule greater than 4 mm and less than 6 mm.	Our instruction states that if the report does not state “previous size: N/A,” then the nodule cannot be considered new, but this example ignored this.
Unclear instruction (13%)	Solid nodule less than 6 mm in diameter. Absence of specific calcifications or suspicious characteristics.	The model failed to recognize **"AP WINDOW MASS"** as a suspicious nodule feature from the report.
Others (32%)		

Discussion

In this study, we demonstrated that LLMs can achieve good performance in predicting lung-RADS categories and follow-up management recommendations with in-context learning methods. Llama 3.3-70B achieved the highest micro F1 Scores of 0.78 for lung-RADS classification and 0.87 for management classification with COT and lung-RADS guidelines, with Phi-4 following closely at 0.77 for lung-RADS classification and 0.83 for management classification. The Lung-RADS 4X category remains the hardest for all LLMs to accurately predict, while the hierarchical classification method brought little improvement to the LCS outcome prediction. While the performance is lower than the finetuned models, in-context learning techniques can be easily adapted to different note formats from different sites. Additionally, the overall results indicate that predicting follow-up management outperforms direct Lung-RADS score classification, suggesting it may be a potentially more effective approach for estimating the overall

malignancy risk of the pulmonary nodules.

Our results suggested that LLMs with 10B+ parameters with in-context learning prompts with PC and Lung-RADs guideline could be an ideal solution for LCS outcome prediction. Smaller LLMs, e.g., Phi-3.5, cannot benefit from the in-context learning prompts, and fine-tuning could be a solution to adapt such small models to the LCS task, which will be our future exploration. The hierarchical classification method failed to divide the prediction task into simpler tasks, as the benign (Lung-RADS score 1 and 2) and malignant (Lung-RADS score 3, 4A, 4B, and 4X) were well differentiated by the LLMs, resulting in minimal improvements on the macro F1 score.

Compared with other studies that seek to predict the LCS results, our system outperformed the Vancouver Risk Calculator proposed by White et al. to predict the malignancy (ours 92% v.s. White 90%), i.e., malignant vs non-malignant, among patients with lung-RADS 2, 3, 4A, and 4B[17]. Similarly, our system outperformed the ANN proposed by Hsu et al.[15] by 7% and was able to provide detailed explanations for each prediction. We cannot make a direct comparison with the system from Byers, as they focus on Lung-RADS score 4 (a combination of Lung-RADS scores 4A, 4B, and 4X), which differs from our experimental setting.[16] Overall, our system proves LLM as a better tool for LCS outcome prediction compared to statistical methods and early ANN models.

From both computational and clinical perspectives, the in-context learning approach offers practical advantages that support integration into existing radiology workflows. Unlike fine-tuning methods, in-context learning avoids the need for retraining or large annotated datasets, making it adaptable across different note formats and institutions. Models like Phi-4 achieve reasonable performance with response times under 8 seconds. Additionally, all models in our study are publicly available, can be run offline, and our prompts are openly accessible, supporting scalable and privacy-preserving deployment. These features make it feasible to use our system as a real-time assistive tool for suggesting Lung-RADS categories and follow-up plans within routine radiology practice.

Building on this feasibility, we suggest that LLM-assisted workflows could enhance radiology reporting through collaborative decision-making. In an LLM-first workflow, the model provides initial predictions and rationales, which radiologists review and refine. In a radiologist-first workflow, clinicians generate assessments first, followed by LLM-based validation or secondary review. These approaches align with previous work: Gaber et al.[30], demonstrated high clinician-LLM agreement in triage and diagnosis using dual-input designs. Similarly, Hager et al.[31] explored interactive, LLM-supported radiology reporting workflows, where clinicians retained decision-making authority while benefiting from real-time LLM assistance. Together, these studies support the validity and practicality of incorporating our system into routine lung cancer screening interpretation as a second-reader tool.

Our study has several limitations. First, the ground truth Lung-RADS scores extracted from the impression notes may not be accurate. In our error analysis, we identified approximately 12% of cases where the ground truth misaligned with the guideline, possibly due to misjudgment of the radiologist in determining Lung-RADS scores. Prior research also reported that at least 6% of Lung-RADS scores are incorrectly assigned due to the subjective assessments of nodule characteristics, such as type, size, and status.[28]. We plan to address this issue in collaboration with local radiologists in our future work. Additionally, we adopted a one-size-fits-all approach for measuring and interpreting the size of the nodules. In radiology reports, the nodule size was documented with either the largest dimension as A mm (e.g., 4 mm) or a 2-dimensional measurement as A x B mm (e.g., 6 x 4 mm). While some radiologists calculate the average size (e.g., 5 mm), we used the largest dimension (e.g., 6 mm) reported in the radiology reports, as many radiologists do, which could lead to mis-predicting the Lung-RADS score. Thirdly, our system did not address the imbalance in Lung-RADS scores, which stems from both the use of in-context learning and the inherent prevalence distribution in the population, where categories 1 and 2 cases are dominant (1: 39%, 2: 45%)[29], which will be addressed in our future study with finetuned LLMs. Moreover, our system did not include the previous radiology report, relying on history findings from the current radiology report, which resulted in our confusion with dubious descriptions, e.g., “previous size: N/A” that can be interpreted as either “No previous LCS” or “No nodule was found in previous LCS.”

Conclusion

Our study demonstrated that LLMs, particularly Llama 3.3-70B and Phi-4, can effectively predict Lung-RADS scores and follow-up management recommendations without fine-tuning. We offer a solution to automate radiology assessments, improving the efficiency in LCS result reporting and accuracy compared to existing methods. However, the inconsistency in documenting radiology findings and the limited reasoning capability of LLMs indicate the need for further improvements in both standardizing lung cancer screening reports and LLMs. To the best of our knowledge,

this is the first study to evaluate LLMs for Lung-RADS categorization and management classification.

Supplementary Materials
https://github.com/zackzhou0814/Lung-RADS-Prediction/blob/main/Supplementary_Materials.pdf

References

1. Christensen J, Prosper AE, Wu CC, Chung J, Lee E, Elicker B, et al. ACR Lung-RADS v2022: Assessment categories and management recommendations. J Am Coll Radiol. 2024 Mar;21(3):473–88.
2. National Lung Screening Trial Research Team, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med. 2011 Aug 4;365(5):395–409.
3. Jemal A, Siegel R, Xu J, Ward E. Cancer statistics, 2010. CA Cancer J Clin. 2010 Sep;60(5):277–300.
4. Lung cancer statistics [Internet]. [cited 2025 Feb 20]. Available from: https://www.cancer.org/cancer/types/lung-cancer/about/key-statistics.html
5. Cancer of the lung and Bronchus - Cancer stat facts [Internet]. SEER. [cited 2025 Feb 20]. Available from: https://seer.cancer.gov/statfacts/html/lungb.html
6. Schabath MB, Cote ML. Cancer progress and priorities: Lung cancer. Cancer Epidemiol Biomarkers Prev. 2019 Oct;28(10):1563–79.
7. Roizen MF. Prevalence of heavy smoking in California and the United States, 1965-2007. Yearb Anesthesiol Pain Manag. 2012 Jan;2012:272–3.
8. Bonney A, Malouf R, Marchal C, Manners D, Fong KM, Marshall HM, et al. Impact of low-dose computed tomography (LDCT) screening on lung cancer-related mortality. Cochrane Database Syst Rev. 2022 Aug 3;8(8):CD013829.
9. Screening for Lung Cancer with Low Dose Computed Tomography (LDCT) [Internet]. [cited 2025 Mar 16]. Available from: https://www.cms.gov/medicare-coverage-database/view/ncacal-decision-memo.aspx?proposed=Y&NCAId=274&bc=0
10. de Koning HJ, van der Aalst CM, de Jong PA, Scholten ET, Nackaerts K, Heuvelmans MA, et al. Reduced lung-cancer mortality with volume CT screening in a randomized trial. N Engl J Med. 2020 Feb 6;382(6):503–13.
11. Park S, Park H, Lee SM, Ahn Y, Kim W, Jung K, et al. Application of computer-aided diagnosis for Lung-RADS categorization in CT screening for lung cancer: effect on inter-reader agreement. Eur Radiol. 2022 Feb;32(2):1054–64.
12. Cellina M, Cacioppa LM, Cè M, Chiarpenello V, Costa M, Vincenzo Z, et al. Artificial intelligence in lung cancer screening: The future is now. Cancers (Basel). 2023 Aug 30;15(17):4344.
13. Tammemägi MC, Ten Haaf K, Toumazis I, Kong CY, Han SS, Jeon J, et al. Development and validation of a multivariable lung cancer risk prediction model that includes low-dose computed tomography screening results: A secondary analysis of data from the National Lung Screening Trial. JAMA Netw Open. 2019 Mar 1;2(3):e190204.
14. Mendoza DP, Petranovic M, Som A, Wu MY, Park EY, Zhang EW, et al. Lung-RADS category 3 and 4 nodules on lung cancer screening in clinical practice. AJR Am J Roentgenol. 2022 Jul;219(1):55–65.
15. Hsu YC, Tsai YH, Weng HH, Hsu LS, Tsai YH, Lin YC, et al. Artificial neural networks improve LDCT lung cancer screening: A comparative validation study [Internet]. Research Square. Research Square; 2020. Available from: http://dx.doi.org/10.21203/rs.3.rs-24642/v4
16. Beyer SE, McKee BJ, Regis SM, McKee AB, Flacke S, Saadawi GE, et al. Automatic Lung-RADS™ classification with a natural language processing system. J Thorac Dis. 2017 Sep;9(9):3114–22.
17. White CS, Dharaiya E, Dalal S, Chen R, Haramati LB. Vancouver risk calculator compared with ACR Lung-RADS in predicting malignancy: Analysis of the National Lung Screening Trial. Radiology. 2019 Apr;291(1):205–11.
18. Nakamura Y, Kikuchi T, Yamagishi Y, Hanaoka S, Nakao T, Miki S, et al. ChatGPT for automating lung cancer staging: feasibility study on open radiology report dataset [Internet]. medRxiv. 2023. p. 2023.12.11.23299107. Available from: http://medrxiv.org/content/early/2023/12/13/2023.12.11.23299107.abstract
19. Singh R, Hamouda M, Chamberlin JH, Tóth A, Munford J, Silbergleit M, et al. ChatGPT vs. Gemini: Comparative accuracy and efficiency in Lung-RADS score assignment from radiology reports. Clin Imaging. 2025 Mar 1;121(110455):110455.
20. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning

in large language models [Internet]. arXiv [cs.CL]. 2022. Available from: http://arxiv.org/abs/2201.11903
21. Gadesha V, Kavlakoglu E. What is prompt chaining? [Internet]. 2024 [cited 2025 Mar 5]. Available from: https://www.ibm.com/think/topics/prompt-chaining
22. Liu J, Liu A, Lu X, Welleck S, West P, Le Bras R, et al. Generated knowledge prompting for commonsense reasoning. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) [Internet]. Stroudsburg, PA, USA: Association for Computational Linguistics; 2022. Available from: http://dx.doi.org/10.18653/v1/2022.acl-long.225
23. Ecekamar. Introducing Phi-4: Microsoft's Newest Small Language Model Specializing in Complex Reasoning [Internet]. Microsoft Community Hub. 2024 [cited 2025 Mar 5]. Available from: https://techcommunity.microsoft.com/blog/aiplatformblog/introducing-phi-4-microsoft's-newest-small-language-model-specializing-in-comple/4357090
24. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models [Internet]. Meta. 2024 [cited 2025 Mar 5]. Available from: https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/
25. Microsoft/phi-3.5-mini-instruct · hugging face [Internet]. [cited 2025 Mar 22]. Available from: https://huggingface.co/microsoft/Phi-3.5-mini-instruct
26. microsoft/phi-4 · Hugging Face [Internet]. [cited 2025 Mar 22]. Available from: https://huggingface.co/microsoft/phi-4
27. Introducing Meta Llama 3: The most capable openly a [Internet]. Meta. 18AD [cited 2025 Mar 17]. Available from: https://ai.meta.com/blog/meta-llama-3/
28. van Riel SJ, Jacobs C, Scholten ET, Wittenberg R, Winkler Wille MM, de Hoop B, et al. Observer variability for Lung-RADS categorisation of lung cancer screening CTs: impact on patient management. Eur Radiol. 2019 Feb;29(2):924–31.
29. Lung-RADS® [Internet]. [cited 2025 Mar 22]. Available from: https://www.acr.org/Clinical-Resources/Clinical-Tools-and-Reference/Reporting-and-Data-Systems/Lung-RADS
30. Gaber F, Shaik M, Allega F, Bilecz AJ, Busch F, Goon K, et al. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. NPJ Digit Med. 2025 May 9;8(1):263.
31. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. 2024 Sep;30(9):2613–22.