

Large Language Models for Classifying Usual Interstitial Pneumonia from Radiology Reports: Native Reasoning Versus Structured Prompting

Ran Zhang

rzhang229@wisc.edu

University of Wisconsin School of Medicine and Public Health

Thomas M. Grist

University of Wisconsin School of Medicine and Public Health

Mark Schiebler

University of Wisconsin School of Medicine and Public Health

Yijing Wu

University of Wisconsin School of Medicine and Public Health

Nathan Sandbo

University of Wisconsin School of Medicine and Public Health

Allan R. Brasier

University of Wisconsin School of Medicine and Public Health

Guang-Hong Chen

University of Wisconsin School of Medicine and Public Health

Research Article

Keywords:

Posted Date: April 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9442328/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Extracting disease labels from radiology reports is essential for developing deep learning-based diagnostic models and enabling large-scale retrospective clinical research. Classification of usual interstitial pneumonia (UIP) patterns from high-resolution computed tomography (HRCT) reports according to Fleischner Society guidelines is a particularly demanding task, requiring synthesis of spatial distribution, fibrotic features, and exclusion criteria. As open-source large language models (LLMs) are released at an accelerating pace with steadily improving general benchmarks, a practical question arises: do these improvements translate to better performance on complex, real-world clinical classification, and does the optimal prompting strategy differ across model architectures?

While prior studies have evaluated LLMs for radiology report labeling, none have compared how prompting strategies interact with the native reasoning capabilities of newer model architectures. We evaluated 10 open-source LLMs from three architecture families (Llama, Qwen, Gemma) spanning 8 to 405 billion parameters, each tested with three prompting strategies on 270 HRCT reports classified by expert consensus of two senior thoracic radiologists. Four models with native reasoning ("thinking") capability were additionally tested in thinking mode. The best configuration achieved $\kappa = 0.70$ and 82% four-class accuracy. Structured reasoning prompting improved all Llama models but degraded all models with native reasoning capability (Qwen 3.5 and Gemma 4), revealing an architecture-dependent interaction. Thinking mode hurt performance on criteria-based prompts and never yielded the best configuration. Larger models did not consistently outperform smaller ones: Llama 3.1 405B offered no advantage over Llama 3.3 70B, and the Qwen 397B model underperformed the dense Qwen 27B.

These findings demonstrate that newer model generations with improved general benchmarks and larger parameter counts do not guarantee better performance on specialized medical classification tasks, and that prompt design must be matched to model architecture.

INTRODUCTION

Usual interstitial pneumonia (UIP) pattern represents the characteristic histopathologic and radiologic pattern of idiopathic pulmonary fibrosis (IPF), a progressive fibrosing interstitial lung disease with a median survival of 2–3 years from diagnosis [1]. Accurate classification of UIP patterns on high-resolution computed tomography (HRCT) is essential for clinical management, prognostic assessment, and therapeutic decision-making of IPF. To improve diagnostic consistency and guide clinical decision-making, the 2018 Fleischner Society and ATS/ERS/JRS/ALAT guidelines established four standardized categories for

UIP classification: typical UIP, probable UIP, indeterminate for UIP, and alternative diagnosis [2, 3].

Despite standardized criteria, consistent implementation of UIP classification in routine practice remains challenging. First, adherence to guideline-based UIP classification varies widely across radiologists, reflecting differences in training, experience, and familiarity with the guidelines. Second, even among experienced thoracic radiologists who explicitly apply the guidelines, substantial inter-reader variability persists. Earlier studies have shown only moderate inter-observer agreement, with κ values between 0.38 and 0.69 when applying the 2018 Fleischner guidelines [4, 5]. This variability reflects the inherent complexity of UIP classification, which requires integrating multiple imaging features—including spatial distribution, presence of honeycombing, reticulation patterns, traction bronchiectasis, and exclusionary findings—into a single diagnostic category.

Standardized UIP categories have value beyond terminology. Clinically, guideline-based classification supports consistent diagnostic pathways and management decisions, including escalation to multidisciplinary discussion and treatment planning. For research, standardized UIP labels define the target phenotype needed to curate large cohorts and to train and validate deep learning models that identify UIP from HRCT images. However, despite widespread imaging archives, standardized UIP categories are often not explicitly recorded in radiology reports, limiting scalable reuse of existing data. Establishing reliable methods to extract standardized UIP classifications from real-world radiology reports is therefore a critical prerequisite for both large-scale phenotyping studies and the development of AI-assisted UIP classification tools.

Large language models (LLMs) have recently demonstrated strong performance in medical natural language processing tasks, including information extraction and diagnostic categorization from clinical notes and radiology reports [6–15]. Prior work has shown that LLMs can extract tumor measurements from oncologic CT reports with high accuracy [8], label chest radiograph findings using privacy-preserving open-source models [15], and automatically categorize liver lesions according to LI-RADS criteria with moderate agreement against expert ground truth [10]. More recently, Khosravi et al. [9] reported that chain-of-thought (CoT) prompting had variable effects on binary pathology label extraction and that larger models generally performed better. Moassefi et al. [16] demonstrated consistent cross-institutional performance using optimized prompts with Llama 3.1 70B model. However, these studies focused on binary or single-feature labeling tasks. UIP classification represents a fundamentally more demanding test case: accurate classification requires synthesizing multiple imaging features, applying exclusion criteria, and resolving nuanced diagnostic language—reasoning that approaches expert-level radiologic judgment.

Meanwhile, open-source LLMs are also evolving rapidly, with recent releases such as the Qwen 3.5 and Gemma 4 families introducing reasoning-oriented training and extended thinking modes, which have improved performance on general reasoning benchmarks. Whether the architectural advances in these newer reasoning models translate to improved performance on specialized medical classification tasks, and whether their built-in reasoning interacts with external prompting strategies has not been examined. In this study, we evaluated 10 open-source LLMs from three architecture families, including both reasoning models and their predecessors for classifying UIP patterns from HRCT reports. We benchmarked agreement with consensus-derived expert labels and examined whether structured clinical reasoning prompts were associated with different performance patterns in reasoning-native versus non-reasoning model families.

MATERIALS AND METHODS

Study Design and Data Collection

This retrospective study was determined exempt by Institution's Minimal Risk Research IRB (ID 2023-0813, June 27, 2023). We identified 5,926 consecutive HRCT radiology reports from patients undergoing evaluation for suspected interstitial lung disease at a single academic medical center between 2013–2023. A total of 300 reports were randomly sampled from the full report set for manual review. After excluding 26 reports from patients with lung transplant, 274 reports remained for independent expert radiologist review. Four reports (one from each classification category) were reserved as few-shot examples, leaving 270 reports for final evaluation (Figure 1). Each report contained Findings and Impression sections, with median length of 202 words (range, 80–632). All reports were generated by board-certified radiologists.

Reference Standard

Two board-certified thoracic radiologists, each with over 30 years of experience in thoracic imaging, independently reviewed all 270 reports. Radiologists classified each report into one of four categories according to the 2018 Fleischner Society guidelines: (1) typical UIP pattern, (2) probable UIP pattern, (3) indeterminate for UIP, or (4) alternative diagnosis. Classifications were based solely on radiology report text, without access to images or additional clinical information. Radiologists were blinded to each other's assessments and to the LLM outputs.

For cases with initial disagreement (n=108 [40%]), the two radiologists conducted joint consensus sessions. During these sessions, both radiologists re-evaluated the discordant reports together, discussing the specific imaging features described and their interpretation according to Fleischner criteria. Consensus was reached through

discussion, and the final classification was recorded. This consensus-derived label served as the reference standard for model evaluation.

Large Language Model Selection

We evaluated 10 open-source LLMs spanning three architecture families and parameter counts from 8 billion to 405 billion (Table 1). Models were selected to represent a range of contemporary locally deployable open-source LLM families, model scales, and reasoning capabilities and to enable systematic comparisons across architecture families, model generations, and scale.

The Llama family (Meta AI) included Llama 3.1 8B, Llama 3.3 70B, and Llama 3.1 405B. These models were released prior to the adoption of reasoning-specific training objectives and do not support extended thinking modes. The Qwen 3.5 family (Alibaba) included 9B, 27B, and 397B variants. Qwen 3.5 models were trained with reasoning-enhanced objectives and support a native thinking mode that generates internal reasoning tokens before producing a response. The Gemma family (Google) included Gemma 3 12B, Gemma 3 27B, MedGemma 27B and Gemma 4 27B. Gemma 4 represents a newer generation with reasoning-enhanced training and native thinking capability, while the Gemma 3 models and MedGemma do not support thinking mode. MedGemma 27B shares the Gemma 3 27B base architecture but received additional pretraining on biomedical literature.

All models were deployed locally using the vLLM framework (version 0.18) on institutional GPU computing infrastructure, ensuring that all patient data remained within institutional infrastructure without requiring cloud-based APIs or external data transmission. This locally hosted approach supports compliance with healthcare privacy regulations.

Table 1. Characteristics of the 10 open-source large language models evaluated.

Family	Model	Parameters	Reasoning Mode	Release Date
Llama (Meta)	Llama 3.1 8B	8B	No	Jul 2024
	Llama 3.3 70B	70B	No	Dec 2024
	Llama 3.1 405B	405B	No	Jul 2024
Qwen 3.5 (Alibaba)	Qwen 3.5 9B	9B	Yes	Feb 2026
	Qwen 3.5 27B	27B	Yes	Feb 2026
	Qwen 3.5 397B	397B	Yes	Feb 2026
Gemma (Google)	Gemma 3 12B	12B	No	Mar 2025
	Gemma 3 27B	27B	No	Mar 2025
	MedGemma 27B	27B	No	May 2025
	Gemma 4 27B	27B	Yes	Mar 2026

Prompting Strategies

We evaluated three prompting strategies of increasing structural complexity, applied identically across all 10 models. The three prompts shared identical classification criteria (drawn from the 2018 Fleischner Society guidelines for four UIP categories) and differed only in the amount of reasoning structure and examples provided (Appendix A).

Criteria-only prompting. The model received the Fleischner classification criteria and was instructed to classify the report into one of the four UIP categories. No reasoning structure or examples were provided.

Structured reasoning prompting. The criteria-based prompt was augmented with a six-step clinical decision algorithm that guided the model through: (1) checking for explicit UIP-related terminology and clinical history, (2) semantic identification of UIP pattern language including variant phrasings, (3) assessing core fibrotic features including reticulation, traction bronchiectasis, and honeycombing, (4) evaluating spatial distribution patterns, (5) identifying exclusionary features suggesting alternative diagnoses, and (6) synthesizing findings into a classification using explicit decision rules.

Structured reasoning with few-shot examples. The structured reasoning prompt was augmented with four representative examples (one per classification category), each showing the report text and the correct consensus classification. These four examples were drawn from the 274 reviewed reports and excluded from the evaluation set.

Thinking Mode Evaluation

Four models support a native extended reasoning mode: Qwen 3.5 9B, Qwen 3.5 27B, Qwen 3.5 397B, and Gemma 4 27B. In thinking mode, the model generates internal chain-of-thought tokens before producing a final classification. We evaluated these four models under all three prompting strategies in both non-thinking and thinking mode. Non-thinking mode used greedy decoding (temperature = 0). Thinking mode used vendor-recommended sampling parameters (temperature = 0.6, top-p = 0.95, top-k = 20 for Qwen 3.5 models; temperature = 1.0, top-p = 0.95, top-k = 64 for Gemma 4 27B). This yielded 42 total experimental conditions: 30 standard (10 models × 3 prompts) plus 12 thinking mode (4 models × 3 prompts).

Report Characterization

To characterize the inferential complexity of the classification task, we analyzed whether reports contained explicit UIP-related terminology. Reports were categorized as "UIP mentioned" if they contained any of the following terms: "UIP," "usual interstitial pneumonia," "IPF," or "idiopathic pulmonary fibrosis." Reports without any of these terms were categorized as "no UIP mentioned," indicating that classification required inference from described imaging features alone. This analysis was used to quantify the proportion of cases where classification requires integration of multiple imaging findings and application of diagnostic criteria versus simple keyword extraction.

Statistical Analysis

Sample size was determined based on expected κ of 0.60 with a 95% confidence interval half-width of ± 0.10 , requiring 250 cases. We calculated Cohen's kappa for inter-rater agreement. Model performance metrics included overall accuracy, category-specific precision, recall, and F1 scores. Bootstrap resampling ($n=1,000$) provided 95% confidence intervals for all metrics. For simplified classification schemes (binary and 3-class), we combined categories as follows: binary (typical UIP + probable UIP vs. others) and 3-class (typical UIP vs. probable UIP vs. others). Analyses were conducted using Python 3.10 with scikit-learn 1.3.0 and statsmodels 0.14.0.

RESULTS

Dataset Characteristics and Inter-Reader Agreement

The 270 HRCT reports represented 259 unique patients with a median age of 69 years (IQR 60–75 years) and 58% male patients. Reports spanned the full study period with a relatively uniform distribution, capturing changes in reporting practices over the decade. Patient demographics are shown in Table 2.

Table 2. Patient Demographics and Clinical Characteristics (n = 259)

Characteristic	Value
Age, years	
Mean $\pm$ SD	66.7 $\pm$ 13.1
Median (IQR)	69 (60–75)
Range	23–91
Sex, n (%)	
Male	151 (58.3)
Female	108 (41.7)
Race, n (%)	
White	246 (95.0)
Black or African American	8 (3.1)
Other	5 (1.9)
Ethnicity, n (%)	
Not Hispanic or Latino	246 (95.0)
Hispanic or Latino	11 (4.2)
Other/Unknown	2 (0.8)

Abbreviations: IQR = interquartile range; SD = standard deviation.

Initial independent classification by the two expert radiologists showed moderate agreement ($\kappa = 0.43$, 95% CI: 0.36–0.49) with a 60% concordance (162/270 cases). Disagreement occurred most frequently between indeterminate and alternative diagnosis categories (68/108, 63% of discordant cases), followed by probable UIP and indeterminate categories (15/108, 14% of discordant cases). Category-specific positive agreement, defined as twice the number of concordant cases divided by the sum of each reader's category totals, was highest for typical UIP (80%) and alternative diagnosis (70%), and lowest for indeterminate (30%) and probable UIP (58%).

Following consensus adjudication of 108 initially discordant cases, final reference classifications comprised typical UIP pattern (40 cases, 14.8%), probable UIP (27 cases, 10.0%), indeterminate for UIP (39 cases, 14.4%), and alternative diagnosis (164 cases, 60.7%).

UIP Terminology in Reports

Analysis of terminology revealed that only 74 reports (27%) contained UIP- or IPF-related terminology, while the majority (196/270, 73%) contained no UIP-related terms. The proportion of reports mentioning UIP/IPF varied substantially by consensus classification: 90% (36/40 reports) of typical UIP cases and 67% (18/27) of probable UIP cases contained UIP/IPF terminology, compared to only 31% (12/39) of indeterminate cases and 5% (8/164) of alternative diagnosis cases. These findings suggest that accurate classification of most reports requires integration of described imaging features and application of Fleischner criteria rather than simple keyword matching.

LLM Classification Performance

Overall, the large language models evaluated demonstrated strong potential for report-based UIP classification, with the best model (Llama 3.3 70B) achieving 81.9% accuracy and a Cohen's kappa of 0.70 when using a structured reasoning prompt with few-shot examples (Table 3). Five of the evaluated models achieved a kappa of 0.65 or higher under their optimal configurations. However, the most successful prompting strategy varied significantly based on the model's underlying architecture.

Table 3. Four-class UIP classification performance across 10 open-source LLMs.

All models were evaluated in standard mode with three prompting strategies. Four reasoning-capable models were additionally evaluated in thinking mode; the last two columns show the best displayed thinking-mode kappa and prompt. Bold values indicate the highest displayed standard-mode performance for each model.

Model	Standard mode			Thinking mode	
	CO κ / Acc (%)	SR κ / Acc (%)	SR+FS κ / Acc (%)	Best κ	Best prompt
Llama 3.1 8B	0.19 / 31.1	0.24 / 36.7	0.28 / 45.9	-	-
Llama 3.3 70B	0.61 / 76.3	0.67 / 80.0	0.70 / 81.9	-	-
Llama 3.1 405B	0.60 / 75.6	0.69 / 81.1	0.69 / 81.5	-	-
Qwen 3.5 9B	0.66 / 79.3	0.61 / 77.0	0.61 / 77.0	0.61	SR
Qwen 3.5 27B	0.69 / 81.9	0.57 / 73.3	0.60 / 75.9	0.64	CO
Qwen 3.5 397B	0.64 / 78.9	0.61 / 75.9	0.62 / 77.8	0.64	SR+FS
Gemma 3 12B	0.58 / 73.3	0.52 / 69.6	0.52 / 69.6	-	-
Gemma 3 27B	0.62 / 75.9	0.63 / 77.0	0.65 / 78.1	-	-
MedGemma 27B	0.68 / 81.5	0.57 / 73.0	0.63 / 78.1	-	-
Gemma 4 27B	0.68 / 81.9	0.58 / 74.4	0.66 / 80.0	0.65	CO

Abbreviations: CO = criteria-only; SR = structured reasoning; SR+FS = structured reasoning with few-shot examples; Acc = accuracy.

Impact of Structured Prompting

The impact of adding structured clinical reasoning to a prompt depended heavily on whether the model was built with native reasoning capabilities. For models without native reasoning (Llama family), providing a structured clinical reasoning algorithm consistently improved classification performance. In contrast, among reasoning-native models (Qwen 3.5 and Gemma 4), structured reasoning prompts consistently reduced performance relative to criteria-only prompting. They performed best when given a simpler, criteria-only prompt. For example, the Qwen 3.5 27B model achieved a kappa of 0.69 under criteria-based prompting but dropped to 0.57 when forced to use a structured reasoning prompt.

Impact of Extended Thinking Mode

Four reasoning-capable models (Qwen 3.5 9B, 27B, 397B, and Gemma 4 27B) were additionally evaluated with thinking mode enabled across all three prompting strategies. As shown in Table 3, when extended thinking is enabled, the best κ for each model was consistently lower than or equal to its best standard-mode κ. These results suggest that extended thinking mode did not provide benefit for this structured classification task in the reasoning-capable models evaluated.

Effect of Model Scale and Medical Pretraining

Within the Llama family, scaling from 8B to 70B parameters yielded a large performance gain ($\kappa = 0.28$ vs. 0.70 at best prompt), but further scaling from 70B to 405B provided no meaningful improvement ($\kappa = 0.70$ vs. 0.69). In the Qwen family, the 27B model ($\kappa = 0.69$) outperformed the 397B model ($\kappa = 0.64$), indicating that larger parameter counts do not guarantee better performance. Meanwhile, the compact Qwen 3.5 9B achieved $\kappa = 0.66$ with criteria-based prompting alone, substantially outperforming older models of comparable size (Llama 3.1 8B, $\kappa = 0.28$; Gemma 3 12B, $\kappa = 0.58$), suggesting that model generation and training strategy may matter more than parameter count alone for this task.

MedGemma 27B, which received additional biomedical pretraining on the Gemma 3 27B architecture, outperformed its base model on criteria-based prompting ($\kappa = 0.68$ vs. 0.62). However, MedGemma showed a larger performance degradation with structured reasoning than its base model, suggesting that medical pretraining may have changed the model's sensitivity to prompt structure.

Error Analysis of Best-Performing Configuration

Among the 49 misclassifications by the best-performing configuration (Llama 3.3 70B, structured reasoning with few-shot examples, $\kappa = 0.70$), the most common error involved confusion between indeterminate and alternative diagnosis categories (29/49, 59%): 20 consensus alternative diagnosis cases were classified as indeterminate, and 9 consensus indeterminate cases were classified as alternative diagnosis (Figure 2). This pattern mirrors the predominant source of disagreement between the two expert radiologists, in which 63% of discordant cases (68/108) also involved these two categories.

The model performed best for typical UIP ($F1 = 0.87$) and alternative diagnosis ($F1 = 0.90$), which represent the most distinctive radiologic patterns. Performance was lower for probable UIP ($F1 = 0.64$) and indeterminate ($F1 = 0.60$), reflecting the inherent challenges in these intermediate classifications.

Performance on Simplified Classification Tasks

For binary classification (typical UIP + probable UIP vs. all other categories), the best model achieved 95.2% accuracy with $\kappa = 0.87$. For three-class classification (typical UIP vs. probable UIP vs. others), accuracy was 92.6% with $\kappa = 0.81$. These simplified schemes showed substantially higher agreement, suggesting robust performance for clinical triage and downstream decision support.

DISCUSSION

This study shows that open-source large language models can achieve substantial agreement with consensus-derived expert labels for report-based classification of usual interstitial pneumonia (UIP) patterns. The best-performing configuration achieved $\kappa = 0.70$ and 81.9% accuracy for four-class classification, despite the intrinsic difficulty of the task and only moderate initial agreement between the two thoracic radiologists ($\kappa = 0.43$). These results suggest that LLM-based extraction can be effective for complex guideline-based report classification, while also highlighting that residual errors likely reflect both the ambiguity of report-based UIP categorization and limitations of current LLM-based approaches.

A central finding of this study is that the effect of prompting strategy was architecture-dependent. In the Llama family, structured reasoning prompts consistently improved performance, and the addition of few-shot examples provided further gains. In contrast, among reasoning-native models (Qwen 3.5 and Gemma 4), structured reasoning prompts consistently reduced performance relative to criteria-only prompting. Extended thinking mode also did not improve performance and never yielded the best overall configuration. Together, these findings suggest that prompt/model combinations should be benchmarked empirically rather than selected based on a general assumption that more elaborate prompting or reasoning will improve clinical NLP performance. For newer reasoning-oriented architectures, simpler prompts may be more effective for structured classification tasks.

Our results suggest two primary applications for LLM-based UIP classification. First, automated classification may serve as a triage and notification tool within clinical workflows. The high binary classification accuracy (95% for UIP/probable UIP versus other categories) indicates that the model could reliably identify cases warranting expedited pulmonology or interstitial lung disease specialist review, as well as identifying patients for clinical trial enrollment. Importantly, such deployment would be most appropriate within a human-in-the-loop framework, in which radiologists retain final interpretive authority.

Second, automated extraction of UIP classifications from free-text radiology reports may substantially accelerate AI-driven research in interstitial lung disease. Manual phenotyping of large ILD cohorts is resource-intensive and often infeasible at scale. By enabling reliable extraction of standardized labels from existing report archives, LLM-based tools could facilitate the creation of large, well-labeled datasets for training and validating computer vision models, studying disease progression, and conducting retrospective outcome analyses.

This study has several limitations. First, this study analyzed final chest CT radiology reports generated in routine clinical practice by multiple readers with heterogeneous

training backgrounds and reporting styles. Accordingly, the evaluation was conducted at the level of report documentation rather than standardized image-based adjudication. We did not incorporate multidisciplinary conference adjudication or perform case-by-case correlation with independent expert thoracic radiologist image review; such analyses would provide an important complementary benchmark for future studies but were beyond the scope of the present report-focused investigation. The single-institution design may limit generalizability, as reporting conventions, terminology usage, and disease prevalence can vary across institutions. Performance was assessed retrospectively using expert consensus on report-derived labels rather than prospective clinical outcomes. For triage-oriented applications, future work should assess whether automated flagging improves time to diagnosis, rates of multidisciplinary discussion, or downstream management decisions. For research use, it remains to be determined whether LLM-extracted labels support downstream model training and evaluation comparably to manually curated datasets. Although prior work has demonstrated that image classifiers are resilient to LLM labeling noise for binary tasks [9], this has not been established for complex multi-class classification schemes such as UIP categorization.

CONCLUSIONS

Large language models can classify UIP patterns from radiology reports with substantial agreement to consensus-derived expert labels. However, the optimal prompting strategy is architecture-dependent: structured reasoning improves models without built-in reasoning capabilities but degrades reasoning-native models. These findings suggest that prompt/model combinations for clinical report classification should be empirically benchmarked rather than selected on the assumption that greater prompt complexity or explicit reasoning will be universally beneficial.

References

1. Ley B, Collard HR, King TE, Jr. Clinical course and prediction of survival in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med*. 2011;183(4):431–40. doi: 10.1164/rccm.201006-0894CI.
2. Lynch DA, Sverzellati N, Travis WD, Brown KK, Colby TV, Galvin JR, et al. Diagnostic criteria for idiopathic pulmonary fibrosis: a Fleischner Society White Paper. *Lancet Respir Med*. 2018;6(2):138–53. doi: 10.1016/S2213-2600(17)30433-2.
3. Raghu G, Remy-Jardin M, Myers JL, Richeldi L, Ryerson CJ, Lederer DJ, et al. Diagnosis of Idiopathic Pulmonary Fibrosis. An Official ATS/ERS/JRS/ALAT Clinical Practice Guideline. *Am J Respir Crit Care Med*. 2018;198(5):e44–e68. doi: 10.1164/rccm.201807-1255ST.

4. Nathan SD, Pastre J, Ksovreli I, Barnett S, King C, Aryal S, et al. HRCT evaluation of patients with interstitial lung disease: comparison of the 2018 and 2011 diagnostic guidelines. *Ther Adv Respir Dis*. 2020;14:1753466620968496. doi: 10.1177/1753466620968496.
5. Widell J, Liden M. Interobserver variability in high-resolution CT of the lungs. *Eur J Radiol Open*. 2020;7:100228. doi: 10.1016/j.ejro.2020.100228.
6. Adams LC, Truhn D, Busch F, Kader A, Niehues SM, Makowski MR, et al. Leveraging GPT-4 for Post Hoc Transformation of Free-text Radiology Reports into Structured Reporting: A Multilingual Feasibility Study. *Radiology*. 2023;307(4):e230725. doi: 10.1148/radiol.230725.
7. Cozzi A, Pinker K, Hidber A, Zhang T, Bonomo L, Lo Gullo R, et al. BI-RADS Category Assignments by GPT-3.5, GPT-4, and Google Bard: A Multilanguage Study. *Radiology*. 2024;311(1):e232133. doi: 10.1148/radiol.232133.
8. Fink MA, Bischoff A, Fink CA, Moll M, Kroschke J, Dulz L, et al. Potential of ChatGPT and GPT-4 for Data Mining of Free-Text CT Reports on Lung Cancer. *Radiology*. 2023;308(3):e231362. doi: 10.1148/radiol.231362.
9. Khosravi B, Dapamede T, Li F, Chisango Z, Bikmal A, Garg S, et al. Role of Model Size and Prompting Strategies in Extracting Labels from Free-Text Radiology Reports with Open-Source Large Language Models. *J Imaging Inform Med*. 2026;39(1):995–1004. doi: 10.1007/s10278-025-01505-7.
10. Matute-Gonzalez M, Darnell A, Comas-Cufi M, Pazo J, Soler A, Saborido B, et al. Utilizing a domain-specific large language model for LI-RADS v2018 categorization of free-text MRI reports: a feasibility study. *Insights Imaging*. 2024;15(1):280. doi: 10.1186/s13244-024-01850-1.
11. Santomartino SM, Zech JR, Hall K, Jeudy J, Parekh V, Yi PH. Evaluating the Performance and Bias of Natural Language Processing Tools in Labeling Chest Radiograph Reports. *Radiology*. 2024;313(1):e232746. doi: 10.1148/radiol.232746.
12. Savage CH, Park H, Kwak K, Smith AD, Rothenberg SA, Parekh VS, et al. General-Purpose Large Language Models Versus a Domain-Specific Natural Language Processing Tool for Label Extraction From Chest Radiograph Reports. *AJR Am J Roentgenol*. 2024;222(4):e2330573. doi: 10.2214/AJR.23.30573.
13. Wiest IC, Ferber D, Zhu J, van Treeck M, Meyer SK, Juglan R, et al. Privacy-preserving large language models for structured medical information retrieval. *NPJ Digit Med*. 2024;7(1):257. doi: 10.1038/s41746-024-01233-2.
14. Wihl J, Rosenkranz E, Schramm S, Berberich C, Griessmair M, Woznicki P, et al. Data extraction from free-text stroke CT reports using GPT-4o and Llama-3.3-70B: the impact of annotation guidelines. *Eur Radiol Exp*. 2025;9(1):61. doi: 10.1186/s41747-025-00600-2.

15. Dorfner FJ, Jurgensen L, Donle L, Al Mohamad F, Bodenmann TR, Cleveland MC, et al. Comparing Commercial and Open-Source Large Language Models for Labeling Chest Radiograph Reports. *Radiology*. 2024;313(1):e241139. doi: 10.1148/radiol.241139.
16. Moassefi M, Houshmand S, Faghani S, Chang PD, Sun SH, Khosravi B, et al. Cross-Institutional Evaluation of Large Language Models for Radiology Diagnosis Extraction: A Prompt-Engineering Perspective. *J Imaging Inform Med*. 2026;39(1):989–94. doi: 10.1007/s10278-025-01523-5.

Figures

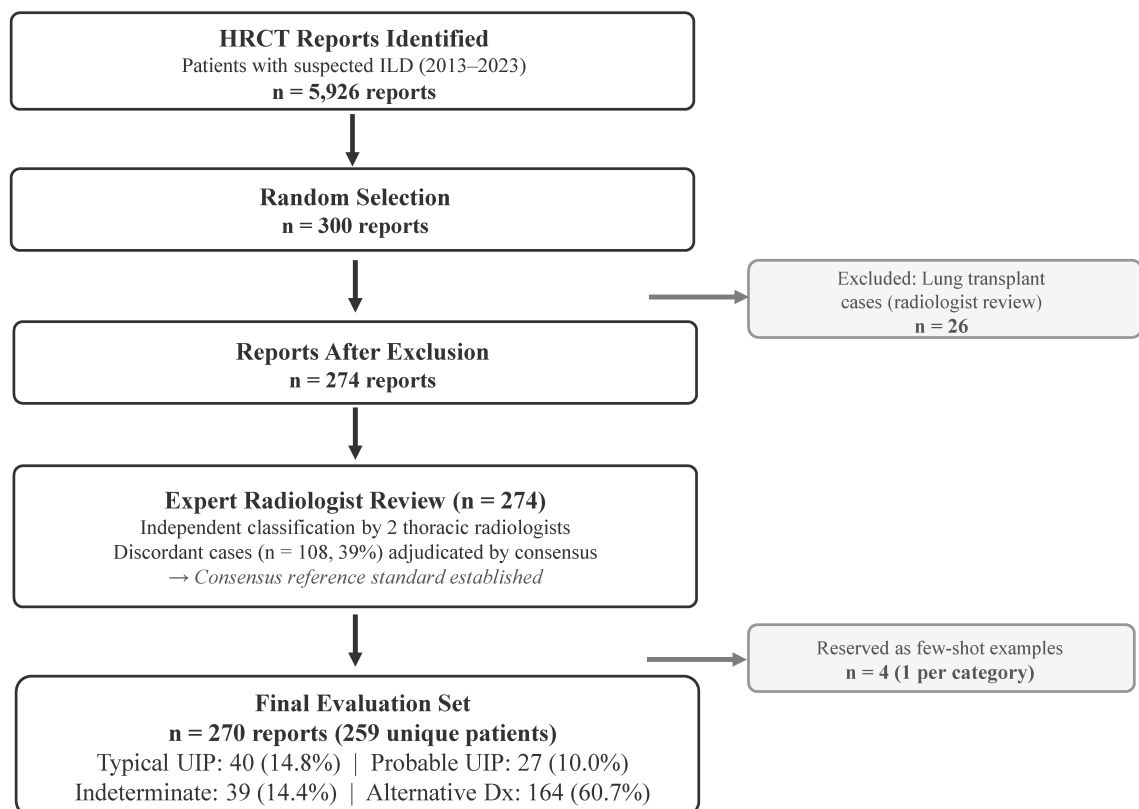


Figure 1

Flowchart of study design.

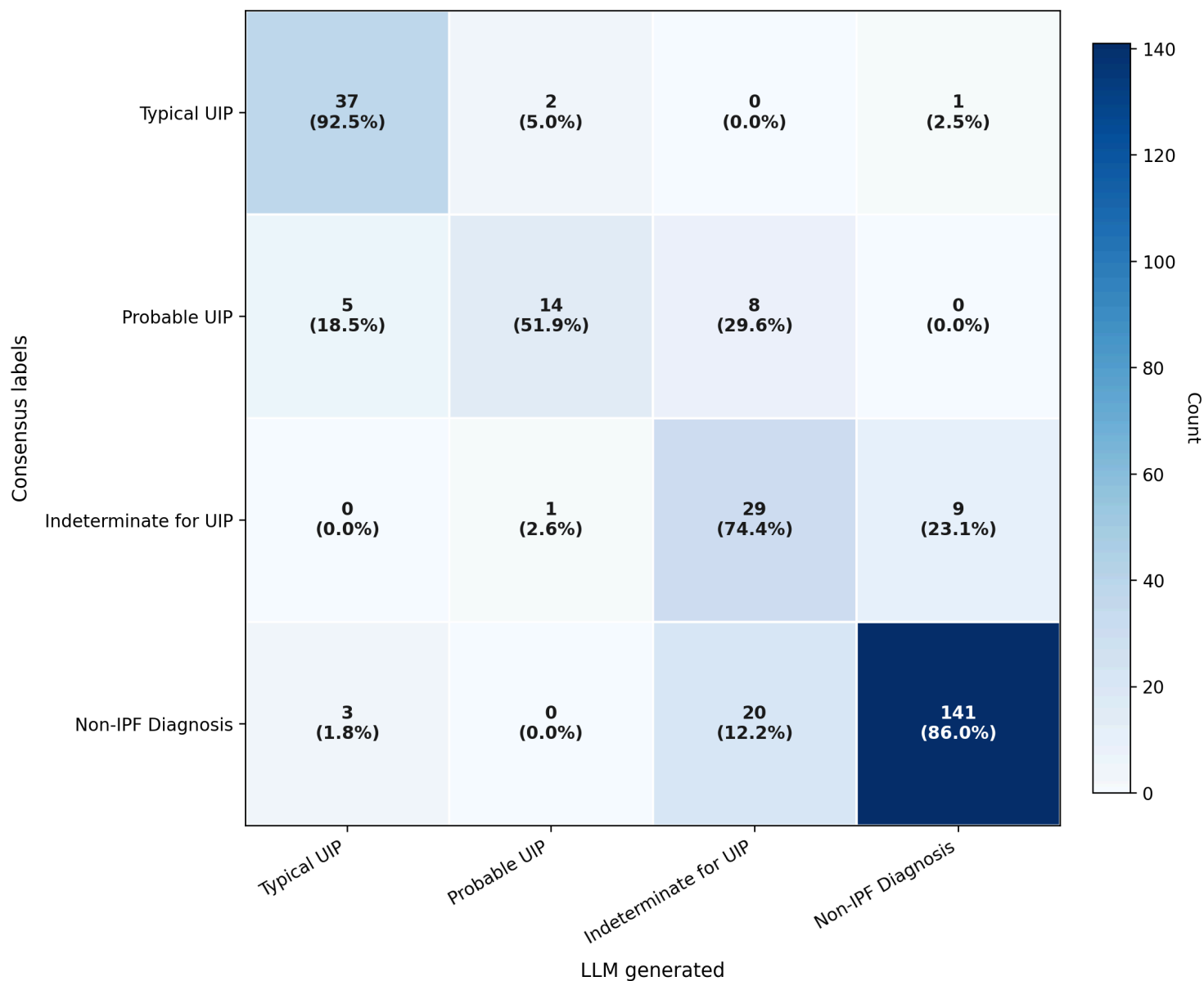


Figure 2

Confusion matrix of the Llama 3.3 70B model with structured prompting and few-shot examples.