

# Diagnostic Reasoning with Large Language Models for a Rare Disease: Case Study of Primary Ciliary Dyskinesia

Swati Rajwal<sup>1</sup>, Mary Ellen Fain<sup>2,3</sup>, Lokesh Guglani<sup>2,3</sup>, Abeer Sarker<sup>1</sup>

<sup>1</sup>Dept. of Biomedical Informatics, Emory University, <sup>2</sup>Dept. of Pediatrics, Emory University

<sup>3</sup> Children's Healthcare of Atlanta

{swati.rajwal, mary.ellen.fain, lokesh.guglani, abeed.sarker}@emory.edu

## Abstract

Primary ciliary dyskinesia (PCD) is a rare pediatric lung disease that is frequently underdiagnosed due to nonspecific early symptoms and limited clinical exposure. We investigate whether large language models (LLMs) can support early diagnostic reasoning using real-world pediatric pulmonology notes written before the final diagnosis. We curated 58 de-identified first-visit notes (28 confirmed PCD, 30 controls) and evaluated five open-source LLMs using a standardized zero-shot prompt to produce structured outputs, including PCD evaluation recommendations, justifications, and suggested tests. Quantitative performance was assessed against expert-validated labels using sensitivity, specificity, and accuracy, and a clinician qualitatively reviewed all explanations and testing recommendations for clinical soundness. Sensitivity ranged from 0.48 to 1.00 and specificity from 0.10 to 0.48 (excluding uncertain outputs), with a best accuracy of 0.75. A majority-vote ensemble of five open-source LLMs achieved perfect sensitivity (1.00) with accuracy of 0.73. While models often identified clinically relevant signals in unstructured notes, explanations and testing recommendations were frequently only partially sound. These findings suggest LLMs may serve as cautious early screening aids for rare disease suspicion, but not as standalone diagnostic tools. This work further highlights the need for larger, multi-site evaluation on longitudinal clinical text.

**Keywords:** Large Language Models, Primary Ciliary Dyskinesia, Rare Disease, Natural Language Processing

## 1. Introduction

Primary ciliary dyskinesia (PCD) is a rare inherited condition that affects the normal movement of cilia in the airways and other organs. Children with PCD often experience chronic cough, recurrent respiratory infections, ear and sinus disease, and in some cases organ laterality differences. Despite these characteristic features, PCD is frequently diagnosed late or underdiagnosed (Salinas et al., 2025). There is no single definitive test for PCD (Behan et al., 2016; Knowles et al., 2016), and diagnosis depends on combining clinical history with several specialized investigations (Zariwala et al., 1993–2025). Many children are evaluated multiple times before PCD is seriously considered, leading to delays in care and preventable lung damage.

These diagnostic challenges occur in the context of limited clinical exposure. Most clinicians encounter very few PCD cases during training or practice, and symptoms overlap with much more common childhood respiratory conditions. As a result, PCD may not be included early in the diagnostic reasoning process, even when classic features are present, a challenge that has prompted the development of several international registries and multi-center cohorts in the field (PCD Foundation, 2026; PCD Support UK, 2026). Prior work has shown that expert review of large collections of electronic medical records (EMRs) can help identify patients with previously missed PCD diagnoses,

highlighting how early clinical clues are often documented but not acted upon (Fecho et al., 2022; Zhou et al., 2025).

Large language models (LLMs) are now widely accessible to clinicians, patients, and families. These systems can read and summarize long clinical narratives, recognize symptom patterns, and generate structured responses to medical inquiries (Wei et al., 2024; Das et al., 2025) or errors (Rajwal et al., 2024). Parents report using such systems to ask whether their child's symptoms could indicate a rare condition,<sup>1</sup> and clinicians increasingly report LLM use as decision support tools (Tausky et al., 2025). However, most evidence on LLM performance comes from general medical knowledge tests or synthetic cases, rather than real-world clinical notes written before a diagnosis is known.

To fill this gap in the literature, we investigate a clinically important question: when presented with real-world pediatric pulmonology notes, do LLMs exhibit clinically appropriate diagnostic reasoning for PCD? What systematic failure modes emerge? We use real, de-identified pediatric clinical notes from patients who were evaluated for PCD, including both confirmed cases and controls in whom PCD was ruled out. Overall, this study provides an early assessment of how LLMs may behave when applied to real pediatric diagnostic reasoning for a

<sup>1</sup>[today.com/health/mom-chatgpt-diagnosis-pain-rcna101843](https://today.com/health/mom-chatgpt-diagnosis-pain-rcna101843). Accessed on 04/05/2026

rare respiratory disease<sup>2</sup>. Our key contributions:

- We evaluate how multiple open-source LLMs reason about suspected PCD using real first-visit pediatric pulmonology notes collected before the final diagnosis.
- We conduct quantitative evaluation against expert-validated gold-standard labels using sensitivity, specificity, and accuracy.
- We perform clinician-led qualitative review of models' explanations and PCD testing recommendations to assess clinical reasoning.

## 2. Related Work

Diagnosing primary ciliary dyskinesia (PCD) remains inherently complex, with studies showing that different guideline-based diagnostic algorithms can yield contradictory diagnoses in a significant subset of patients (Nussbaumer et al., 2021). This diagnostic uncertainty has motivated growing interest in computational approaches that support rare disease recognition. Recent studies have explored the use of large language models (LLMs) in the context of rare diseases (Hodges et al., 2025), motivated by the long diagnostic delays and limited specialist availability that many patients face (Chen et al., 2024; Shyr et al., 2024; Abdullahi et al., 2024; Wu et al., 2024; Hodges et al., 2025). Several works have shown that LLMs can support rare disease diagnosis, education (Niyonkuru et al., 2025), and clinical reasoning by synthesizing information across symptoms, histories, and medical literature. Benchmarks such as RareBench (Chen et al., 2024) and qualitative evaluations on complex diagnostic cases suggest that general-purpose LLMs are often able to include rare diseases in their diagnostic considerations, even when these conditions are infrequently encountered in routine clinical practice.

Several studies have further demonstrated that zero-shot LLMs, which do not require task-specific fine-tuning, can perform comparably to more complex pipelines for rare disease tasks such as generating plausible differential diagnoses or extracting clinically relevant information from unstructured clinical text (Wu et al., 2024; Shyr et al., 2024; Ríos-Hoyo et al., 2024). These findings motivate our use of a zero-shot setting, which better reflects real-world constraints and avoids reliance on curated training datasets that are difficult to obtain for rare conditions.

In contrast to the growing literature on rare diseases broadly, there has been relatively little work

focused specifically on PCD. Existing studies related to PCD have primarily examined how traditional machine learning or information retrieval approaches (Rother et al., 2015; Burns et al., 2025) can be used to identify patients who may have been missed during routine clinical care (Zhou et al., 2025). For example, prior work has shown that searching EMRs for characteristic symptom patterns can help uncover patients with undiagnosed PCD, highlighting the presence of early clinical signals that are often documented but not recognized at the time (Zhou et al., 2025). However, these approaches do not directly evaluate how language-based models reason over individual clinical notes in a diagnostic context. While these studies demonstrate the feasibility of identifying potential PCD cases using structured data or retrieval methods, our work shifts the focus to LLM-based reasoning using unstructured clinical notes, which more closely reflects real-world diagnostic settings.

## 3. Methods

Figure 1 shows the pipeline used in this study. Below, we explain the dataset, LLMs used, and their implementation details.

### 3.1. Data

This study uses de-identified clinical notes from pediatric patients seen at Children's Hospital of Atlanta (CHOA), collected under an approved Institutional Review Board (IRB) protocol (#STUDY00002645). The dataset includes two groups: children who were later confirmed to have PCD [ $n=28$ , mean age= 4.85 years ( $\pm 5.94$ ), sex= 12 M, 16 F] and children who were evaluated for PCD but ultimately did not have the disease [ $n=30$ , mean age= 7.58 years ( $\pm 1.87$ ), sex= 13 M, 17 F] to serve as controls. **Confirmed PCD cases** were identified via a pre-existing list maintained in the EMRs of children with PCD being cared for at CHOA, which included 44 patients. Manual chart review was performed for each patient to identify the first clinical encounter with a pediatric pulmonologist at the institution, to reflect the child's initial symptom onset and diagnostic evaluation before PCD was confirmed or ruled out. Patients were excluded if they were referred with a pre-existing diagnosis of PCD or if their first clinical encounter available in the EMRs followed a confirmatory diagnosis ( $n=11$ ), as this would remove diagnostic uncertainty.

**Control cases** were selected from a broader pre-existing list in the EMRs of patients referred to the PCD specialty clinic, which was cross-referenced with the confirmed PCD list to identify patients in whom PCD was suspected but not confirmed. Manual chart review was similarly performed. Patients

<sup>2</sup>Preliminary version of this work was accepted at the ATS 2026 International Conference (Rajwal et al., 2026).

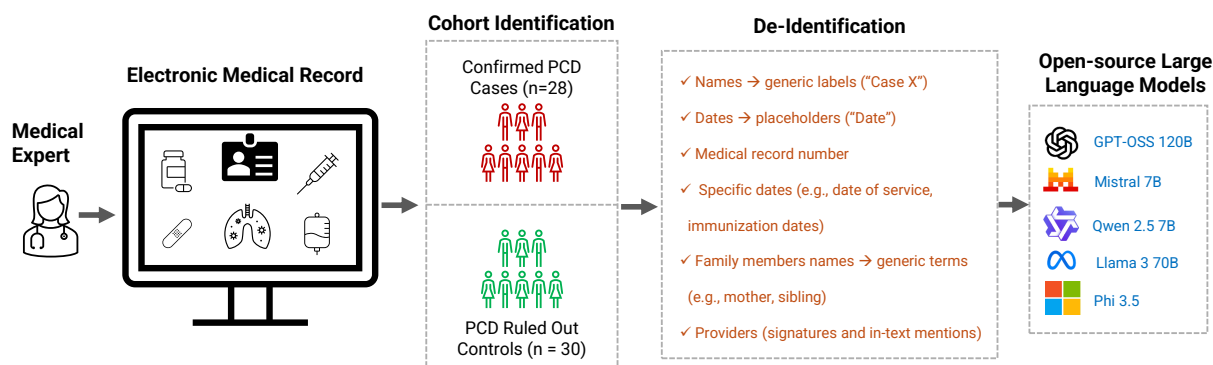


Figure 1: Study pipeline showing cohort curation and de-identification of first-visit clinical notes, zero-shot inference with five open-source LLMs. Subsequently, quantitative evaluation and clinician-led qualitative review of model outputs are reported in this study.

were included if their workup was ultimately negative for PCD, and excluded if their workup was still ongoing with uncertain diagnosis.

For the selected patients, the identifying information was removed from their encounter note such as medical record number, date of birth, names (of the patient, replaced by "Case" and a case number, as well as family member or provider names), and specific dates. Each selected note was used in its entirety as input to the LLMs and for expert annotation, without additional filtering of clinical content beyond standard de-identification.

### 3.2. Large Language Models Used

We evaluated five open-source LLMs: GPT-OSS-120B (Agarwal et al., 2025), Llama-3-70B (AI@Meta, 2024), Qwen-2.5-7B (Yang et al., 2024), Mistral-7B (Mistral AI, 2024), and Phi-3.5-mini (Abdin et al., 2024). We chose these models to reflect diversity in open-source LLM design and scale. Their differing behaviors are consistent with known trade-offs, such as higher sensitivity in smaller models and more conservative reasoning in larger ones. Each model was given the same zero-shot prompt instructions to assess a single clinical note per patient. Open-source models were used to enable local inference on secure, regulatory-compliant university servers without transmitting clinical data to external systems.

We utilized a zero-shot prompting approach to produce structured JSON output containing: (1) a recommendation for further PCD evaluation (Yes/No/Uncertain), (2) justification for that recommendation, (3) recommended diagnostic tests, and (4) justification for testing decisions. Below, we present the prompt structure:

#### Instruction Header

You are a clinician. Follow the schema exactly. Be concise and deterministic. Rules:  
 - Output JSON with exactly these keys:

```
"pcd_recommendation", "justification_pcd",
"testing_recommended",
"justification_tests".
- "pcd_recommendation" must be "Yes", "No", or
"Uncertain".
- Each justification in around 90 words.
- No extra text. After JSON, write END_OF_OUTPUT.
...
```

#### Key Meanings

```
...
- "pcd_recommendation": Clinical impression -
does this patient need further PCD evaluation?
(Yes/No/Uncertain)
- "justification_pcd": Brief reasoning citing find-
ings from the note.
- "testing_recommended": Diagnostic tests you
would order next. Use "None" if not needed.
- "justification_tests": Why those tests (or why
none).
...
```

#### Sharing Clinical Note

```
...
Below is a de-identified clinical note from a patient's
first pulmonology encounter. Read it and respond per
the schema.
[NOTE STARTS]
{text}
[NOTE ENDS]
...
```

#### Output Formatting Instructions

```
...
Reply ONLY in the following JSON structure:
{{
  "pcd_recommendation": " ",
  "justification_pcd": " ",
  "testing_recommended": " ",
  "justification_tests": " "
}}
END_OF_OUTPUT
```

In addition to individual model evaluations, we constructed an ensemble using a majority voting strategy across the five LLMs used in this study. For each clinical note, the final ensemble prediction was defined as the most frequently predicted label (Yes/No/Uncertain) across models. In cases of ties, the prediction was resolved as “Uncertain.”

### 3.3. Implementation Details

Open-source LLM weights were downloaded from the HuggingFace Model Hub and stored locally on University server. Entire code<sup>3</sup> was implemented in Python 3.10.15. Experiments were executed using SLURM with a consistent hardware configuration of one NVIDIA Ampere GPU, 3 CPU cores, and 80 GB of system memory. Across models, inference time ranged from 5 to 15 minutes per 30 cases, with an average runtime of about 10 minutes.

### 3.4. Evaluation & Manual Analysis

We evaluate models using clinically meaningful metrics, including sensitivity and specificity for recommending further PCD evaluation. We report 95% confidence intervals (CIs) for all metrics, computed using the Wilson score interval for binomial proportions (sensitivity, specificity, accuracy) and non-parametric bootstrapping with 10,000 resamples ( $F_1$ -score). We also perform manual annotation to review LLM outcomes. One PCD clinician based at CHOA manually reviewed the LLM output using an annotation guideline<sup>4</sup> (co-developed with a second PCD clinician) for the justification of the PCD recommendation, as well as any suggested diagnostic tests and their explanations. The annotator rated diagnostic explanations as clinically sound, partially sound, clinically weak, or clinically concerning, and noted any hallucinations. Suggested diagnostic tests were labeled as appropriate, incomplete, excessive, inappropriate, or potentially harmful. Explanations for test recommendations were separately rated based on whether they clearly justified the suggested tests. The annotator was blinded to model identity and based all judgments only on the provided note and model output.

## 4. Results

Figure 2 presents the distributions of model responses, illustrating clear differences in how models recommend further evaluation for PCD across confirmed and non-PCD cases. Models were more likely to recommend further evaluation for confirmed cases, though they varied in how strongly they favored a positive recommendation versus an

uncertain one. Mistral-7B and Phi-3.5-mini showed the strongest tendency toward recommending evaluation, identifying nearly all confirmed cases, while GPT-OSS-120B produced substantially fewer positive recommendations. The ensemble (majority voting) balanced these behaviors, recommending evaluation for most confirmed cases while reducing extreme responses seen in individual models. Figure 3 shows agreements between the models.

Table 1 summarizes performance excluding uncertain responses. The number of uncertain responses varied across models, ranging from 2/58 (3.4%) for GPT-OSS-120B to 17/58 (29.3%) for Llama-3-70B. Sensitivity ranged from 0.48 to 1.00, whereas specificity remained below 0.50 for all models. Phi-3.5-mini achieved the highest overall accuracy (0.75), followed closely by Qwen 2.5 7B (0.70). The ensemble (majority-vote) achieved perfect sensitivity (1.00) with an accuracy of 0.73, matching the strongest individual models in identifying confirmed cases but without improving specificity beyond that range. We additionally report a stricter setting where *Uncertain* predictions are treated as incorrect. Figure 4 shows examples of hallucinations. The expert annotator’s evaluation of LLM responses is shown in Figure 5.

## 5. Discussion

### 5.1. Diagnostic Performance

Across models, the dominant pattern is high sensitivity paired with weak specificity, consistent with a screening-oriented behavior rather than definitive decision-making. Although this is expected, given that all patients were referred for PCD evaluation, it reflects a consistent tendency of the models to favor sensitivity over specificity in clinically ambiguous settings. When uncertain predictions are excluded, sensitivity ranges from 0.48 to 1.00, while specificity remains low for all models (0.095 to 0.483). The highest accuracy observed is 0.750 (Phi-3.5-mini), and the ensemble achieves 0.727 with sensitivity of 1.000, but specificity remains limited at 0.429. These results indicate that while models rarely miss confirmed cases under this evaluation, they frequently flag non-PCD cases for further evaluation, which could increase unnecessary testing or referrals if outputs are used without clinical oversight.

Figure 2 and Table 1 together suggest that the “Uncertain” option plays a meaningful role in how models manage risk. Models vary in how often they select this option, and these differences influence the apparent trade-off between missed cases and false positives. Performance metrics computed after excluding “Uncertain” therefore reflect behavior conditional on a model choosing to make a decision,

<sup>3</sup>GitHub Code: [pcd-diagnosis-llm-study](#)

<sup>4</sup>Annotation Guideline: [Annotation-Guideline.pdf](#)

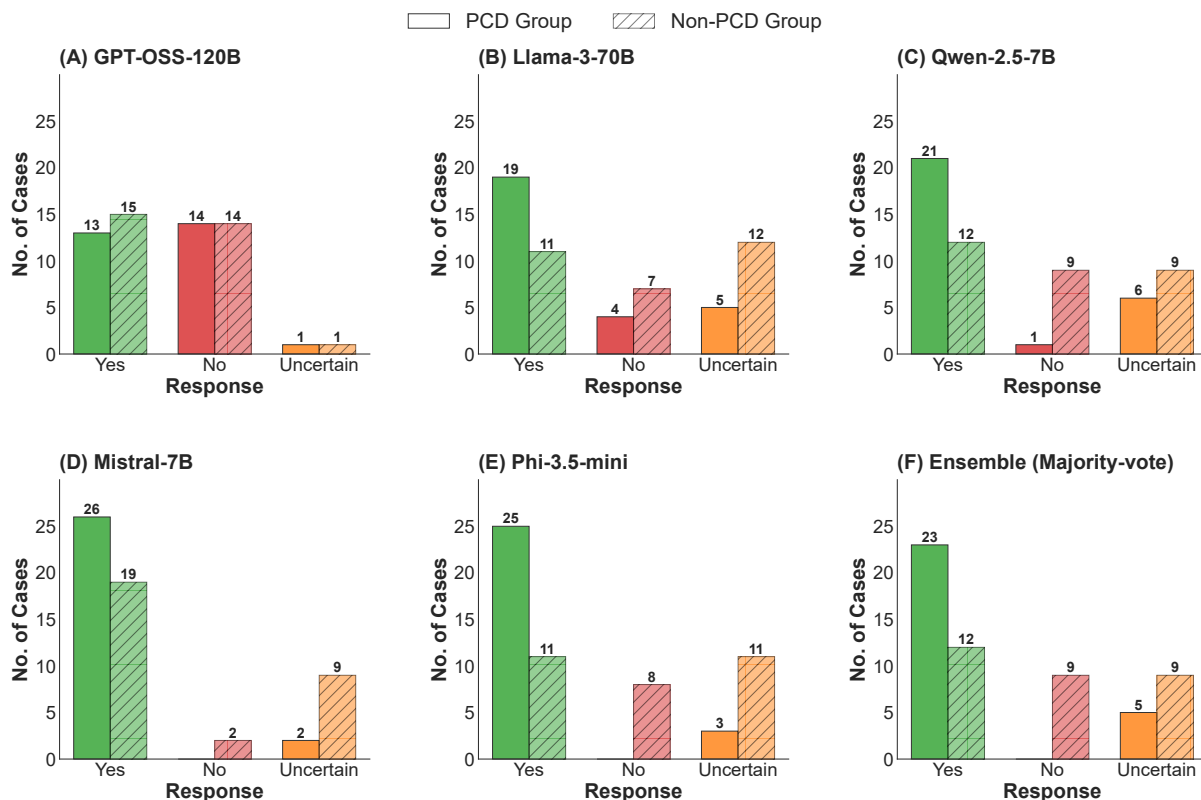


Figure 2: Distribution of model recommendations (“Yes”, “No”, “Uncertain”) across confirmed PCD cases (n=28) and non-PCD controls (n=30) for each LLM and a majority-vote ensemble of five models. Counts are shown separately for each group.

Model	Evaluated Cases (Abstention Rate)	Sensitivity	Specificity	Accuracy	F <sub>1</sub> Score
Approach 1: Committed cases only (uncertain excluded)					
GPT-OSS-120B	56 (3.4%)	0.481 [0.307, 0.660]	0.483 [0.314, 0.656]	0.482 [0.357, 0.610]	0.473 [0.314, 0.613]
Llama-3-70B	41 (29.3%)	0.826 [0.629, 0.930]	0.389 [0.203, 0.614]	0.634 [0.481, 0.764]	0.717 [0.600, 0.824]
Qwen-2.5-7B	43 (25.9%)	0.955 [0.782, 0.992]	0.429 [0.245, 0.635]	0.698 [0.549, 0.814]	0.764 [0.678, 0.851]
Mistral-7B	47 (19.0%)	1.000 [0.871, 1.000]	0.095 [0.027, 0.289]	0.596 [0.453, 0.724]	0.732 [0.667, 0.800]
Phi-3.5-mini	44 (24.1%)	1.000 [0.867, 1.000]	0.421 [0.231, 0.637]	0.750 [0.606, 0.854]	0.820 [0.746, 0.893]
Ensemble (Majority-vote)	44 (24.1%)	1.000 [0.857, 1.000]	0.429 [0.245, 0.635]	0.727 [0.582, 0.837]	0.793 [0.719, 0.868]
Approach 2: Pessimistic bound (uncertain = wrong prediction)					
GPT-OSS-120B	58	0.464 [0.295, 0.642]	0.467 [0.302, 0.639]	0.466 [0.343, 0.592]	0.456 [0.298, 0.596]
Llama-3-70B	58	0.679 [0.493, 0.821]	0.233 [0.118, 0.409]	0.448 [0.327, 0.575]	0.543 [0.424, 0.656]
Qwen-2.5-7B	58	0.750 [0.566, 0.873]	0.300 [0.167, 0.479]	0.517 [0.392, 0.641]	0.600 [0.486, 0.703]
Mistral-7B	58	0.929 [0.774, 0.980]	0.067 [0.018, 0.213]	0.483 [0.359, 0.608]	0.634 [0.564, 0.709]
Phi-3.5-mini	58	0.893 [0.728, 0.963]	0.267 [0.142, 0.444]	0.569 [0.441, 0.688]	0.667 [0.582, 0.750]
Ensemble (Majority-vote)	58	0.821 [0.644, 0.921]	0.300 [0.167, 0.479]	0.552 [0.425, 0.673]	0.639 [0.541, 0.732]

Table 1: Performance with 95% CI. Approach 1 excludes *Uncertain* cases (committed predictions only, i.e., yes/no). Abstention rate = % of *Uncertain* predictions; Approach 2 treats *Uncertain* as wrong prediction.

rather than end-to-end system behavior. In an unstructured clinician review of cases labeled as “uncertain,” the underlying source of uncertainty was noted to be variable, with some labels seeming to reflect an incomplete discussion of PCD features in the source note, some reflecting the source note’s suspicion for an alternate diagnosis, and still others directly referring to inconclusive PCD diagnostic

testing (for example, genetic variants of uncertain significance). In practice, any clinical workflows need explicit guidance on how to handle uncertain outputs, as they represent a substantial fraction of model responses. Additional structured review of these cases in the future may help identify patterns that could be leveraged to minimize uncertain output, or at minimum help interpret this response.

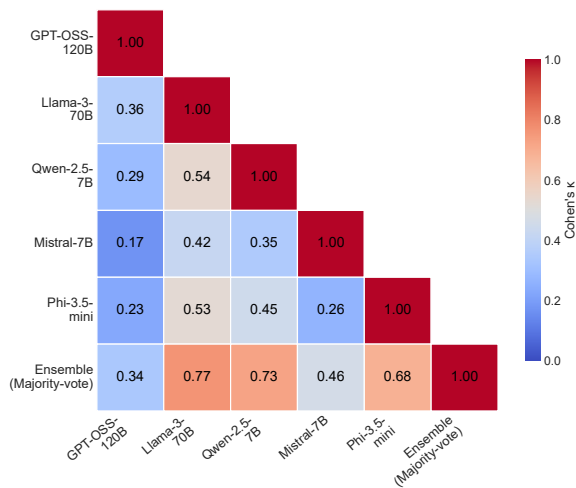


Figure 3: Inter-model agreement for PCD recommendation heatmap across all cases (N=58).

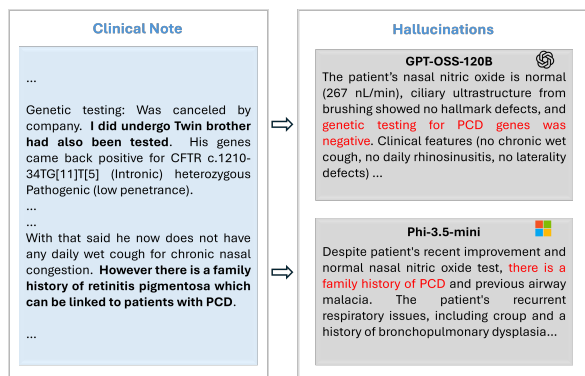


Figure 4: Hallucinations in PCD justification by two LLMs. Note that the part of clinician's note shown is for reference. The text marked in red shows misinterpretation by model.

## 5.2. Qualitative Performance

Manual annotations by domain experts and further analyses highlight that factual grounding does not imply strong clinical reasoning (Fig. 5). This pattern is consistent with prior work showing that LLMs can produce correct answers while generating inconsistent reasoning (Rajwal et al., 2025). Hallucination rates were low across models (0.0–8.6%), yet "clinically sound" diagnostic explanations accounted for only 20.7–48.3% of outputs, while "partially sound" explanations consistently represented 34.5–44.8%. "Clinically weak" explanations reached as high as 37.9% for some models. Notably, models with zero hallucinations, such as Llama-3-70B and Mistral-7B, still produced "clinically weak" or "concerning" explanations in a substantial fraction of cases, showing that avoiding unsupported statements is not sufficient for reliable clinical reasoning.

A large portion of clinically weak labels were applied due to incongruous outputs for further PCD

evaluation recommendation (Yes/No/Uncertain) and the justification of that answer. Specifically, models frequently answer "Yes" to recommend PCD evaluation, but their justifications may not include discussion of PCD. Indeed, some justifications of a "Yes" PCD evaluation recommendation assert a wholly different diagnosis as the likely underlying etiology, particularly if PCD was not discussed in the encounter note (Appendix A.1). Furthermore, in less severe instances, justifications would highlight symptoms not specific or "classic" for PCD as supportive of the diagnosis, which was a common reason for the "partially sound" label (Table 2). In more severe instances, models cited symptoms altogether unrelated to PCD – such as neuromuscular weakness – as supportive of PCD diagnosis, which was a common reason for the "clinically concerning" label. Though LLMs have previously been shown to respond well to clinical vignettes and have shown initial promise in extracting clinically relevant information from unstructured clinical text (Wu et al., 2024; Shyr et al., 2024; Ríos-Hoyo et al., 2024), our findings suggest that when presented with real-world clinical information in the form of an early encounter note—which may include information both contributory and non-contributory to the ultimate diagnosis, or discuss a patient who is struggling with multiple diagnoses—limitations remain in picking out clinically relevant information pertaining to a specific diagnosis, in this case PCD.

GPT-OSS-120B was a notable exception to this theme, as reflected in its high percentage of clinically sound justifications as compared to other models. It not only had more dependable logical alignment of "Yes" or "No" PCD evaluation recommendation and correlating justification, but also more consistently provided PCD-focused justifications and testing, even if PCD was not discussed in the source note (Table A.1). This finding is particularly impactful in the context of our diagnostic performance evaluation, in which GPT-OSS-120B had the lowest scores across performance metrics. We suspect that, were those PCD evaluation recommendations of "Yes" converted to "No" in cases where the model's justification did not actually imply a suspicion for PCD, performance would have been more similar across models. This underscores the importance of continued deeper manual analysis when assessing LLMs' clinical reasoning abilities, as it may reveal fundamentally flawed reasoning underlying seemingly appropriate recommendations.

**Performance on confirmed PCD cases** was particularly revealing. In this subset, clinically sound explanations dropped to as low as 10.7%, while clinically weak explanations rose to 42.9%. At the same time, incomplete testing recommendations were common, ranging from 39.3% to 60.7%

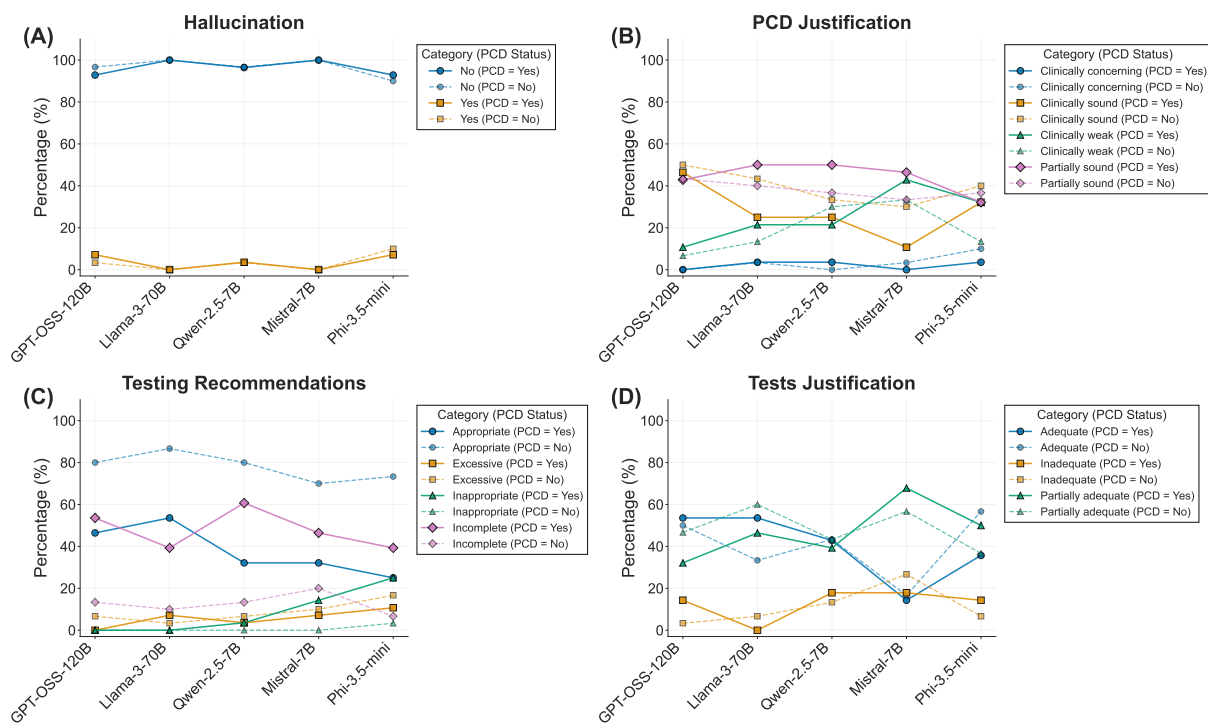


Figure 5: Expert evaluation of five LLMs across four metrics: (A) hallucinations in PCD justifications, (B) strength of clinical reasoning, (C) completeness of testing recommendations, and (D) adequacy of testing justifications. Solid lines indicate PCD-positive cases; dashed lines indicate PCD-negative controls. Colors denote annotation categories. A medical expert systematically evaluated LLM-generated clinical impressions, justifications, and testing recommendations. Overall, models rarely hallucinate (A), but their reasoning and test recommendations are often only partly correct or incomplete (B–D).

across models. This pattern suggests that models struggle most when prioritization and selection of high-yield next steps are required, rather than general symptom summarization. Similar to the observed incompatibility of some PCD evaluation recommendation responses and their justifications, the "incomplete" testing label was often applied to cases where a "Yes" PCD evaluation response was given, but no PCD-specific testing was recommended. In these cases, models frequently recommended workup that was reasonable for ruling out other pulmonary diseases, but uninclusive of important PCD diagnostic tests.

In regard to this discordance, both for PCD evaluation justifications and testing recommendations, one important underlying factor was the source note, as justifications and testing were observed to commonly reflect the note's assessment and plan. Therefore, if PCD was not considered or discussed by the provider in the source note, the model would produce explanations and testing recommendations that aligned more with those etiologies discussed in the note – even when the initial answer for PCD evaluation recommendation was "Yes" (A.1). Even GPT-OSS-120B was impacted, despite its higher percentage of clinically sound, PCD-focused justifications, and was susceptible to

answering "No" to further PCD evaluation in PCD cases where this disease was not discussed in the initial note. This may in part account for its lower sensitivity. Conversely, if the source note contained adequate discussion of PCD, the models were better able to produce clinically sound output. Overall, these findings suggest that the models – particularly when biased by information that already includes proposed diagnoses and testing – are less suited to critically evaluate this input and provide independent conclusions. In the future, prompt adjustment to encourage disagreement with the note, if warranted, may help mitigate some of this issue.

Although inappropriate testing was uncommon overall, it reached 25.0% for Phi-3.5-mini and 14.3% for Mistral-7B in confirmed cases, which could increase clinical burden if outputs are followed without review. Aside from unneeded testing, inappropriate testing labels were applied in instances where the model's recommendation included orders copied from the source note that were not testing, for example, medications or clinic nursing protocols. They were also applied when recommendations failed to consider the patient's age. For example, nasal nitric oxide measurement is a common screening test for PCD, however it cannot be performed in children under 4 years.

LLM	LLM Response	Expert's Evaluation
	PCD: No	×
GPT-OSS-120B	Justification: The infant's chronic cough and wheeze are explained by mild persistent asthma and upper airway congestion, with documented response to inhaled corticosteroids and albuterol. There is no history of situs inversus, persistent neonatal respiratory failure, chronic sinusitis, or otitis media that would raise suspicion for primary ciliary dyskinesia. The clinical picture aligns with asthma rather than a motile cilia disorder.	Clinically weak
	Testing: None	Incomplete
	Justification: Given the clear asthma diagnosis and lack of PCD red flags, additional PCD-specific testing (e.g., nasal nitric oxide, ciliary ultrastructure) is not indicated at this time.	Inadequate
Llama-3-70B	PCD: No	×
	Justification: The patient's symptoms and medical history do not suggest primary ciliary dyskinesia (PCD). The patient's cough and wheezing are likely related to mild asthma and upper airway congestion, which are being managed with Pulmicort and albuterol. There is no mention of situs inversus, nasal polyposis, or other characteristic features of PCD.	Partially sound
	Testing: None	Incomplete
Qwen-2.5-7B	Justification: Given the patient's symptoms and diagnosis of mild asthma, no further testing is recommended at this time. The patient's symptoms are being managed with current medications and therapies, and there is no indication for additional testing.	Partially adequate
	PCD: Yes	✓
	Justification: The patient has a history of wheezing, cough, and upper airway symptoms, which are suggestive of asthma. However, the patient also has persistent upper airway noise and mucus production, which may indicate underlying airway inflammation or other conditions. A detailed PCD evaluation is recommended to rule out other causes and to optimize her asthma management.	Clinically sound
Mistral-7B	Testing: PFTs, chest CT, and possibly allergy testing	Incomplete
	Justification: PFTs will help assess the degree of airway obstruction and responsiveness to bronchodilators. Chest CT can provide more detailed imaging to evaluate the extent of airway inflammation and other structural abnormalities. Allergy testing can help identify potential triggers for her symptoms.	Adequate
	PCD: Yes	✓
Phi-3.5-mini	Justification: The patient has a history of prematurity, prolonged NICU stay, RSV infection, and recurrent cough with mucus production. She has required CPAP and has had a chest CT scan to look for CPAM and TEF. She has also had a history of wheezing and improved with albuterol. The patient's cough persists despite Pulmicort and still sounds junky and rattly.	Partially sound
	Testing: Pulmonary function tests (PFTs) and echocardiogram	Inappropriate
	Justification: To further evaluate the patient's lung function and potential cardiac involvement.	Inadequate
Phi-3.5-mini	PCD: Yes	✓
	Justification: The patient has a history of respiratory symptoms including cough, wheezing, and upper airway congestion. Despite being on Pulmicort, symptoms persist, suggesting suboptimal control. The chest CT scan indicates consolidations and ground-glass opacities, which could be indicative of underlying asthma or bronchitis. The patient's growth and dietary changes may also impact reflux, which can exacerbate respiratory symptoms. Given the chronic nature of the symptoms and the findings on imaging, further evaluation is warranted.	Partially sound
	Testing: Spirometry, allergy testing, and possibly a trial of inhaled corticosteroids (ICS) via MDI for better control.	Inappropriate
	Justification: Spirometry will assess the degree of airflow obstruction and help confirm the diagnosis of asthma. Allergy testing may identify specific triggers contributing to the patient's symptoms. A trial of ICS via MDI could provide better symptom control and adherence compared to nebulizer use, which is currently causing resistance due to the patient's age and difficulty with nebulizer administration.	Adequate

Table 2: LLM responses and annotator evaluation for a sample case. The top row shows a partial excerpt of the input clinical note. Appendix A.1 shows another case with *clinically weak* justifications.

**Review of test justifications** further clarify limitations. Adequate justification of recommended tests ranged from 15.5% to 51.7% overall, with partially adequate explanations dominating for several models. In confirmed PCD cases, adequate justification was only 14.3% for Mistral-7B and 35.7% for Phi-3.5-mini, compared with 53.6% for GPT-OSS-120B and Llama-3-70B. This gap matters clinically because decision support tools must explain not only what to do, but why a given action is appropriate for a specific patient. The partially adequate label was commonly applied due to limited rationale given for a specific test. For outputs that did expound more on testing justification, this label was often given due to recurring, subtle inaccuracies observed. For example, models did not have a solid

grasp on which PCD diagnostic tests measure ciliary structure versus ciliary function, and there was a tendency to label a certain test as "necessary" for diagnosis of PCD, when in reality no one test is considered a gold standard test for its diagnosis (Knowles et al., 2016; Nussbaumer et al., 2021).

**Inadequate test justifications**, meanwhile, were overall rare, but did reach as high as 26.7% for Mistral-7B. Inadequate labels were given for vague outputs failing to justify testing in any meaningful way, as well as for misleading reasoning. For the latter, one inconsistent but notable mistake was asserting that tests used to diagnose other diseases are diagnostic of PCD. In particular, *CFTR* genetic testing and/or sweat chloride testing for cystic fibrosis (CF) — a disease with features similar to PCD

— are often pursued in PCD workup to rule out CF. However, these tests were at times either implied or definitively proclaimed to diagnose PCD. This issue was more prevalent with Qwen 2.5 7B and Phi-3.5-mini, which both made the clinically concerning mistake of calling sweat testing the "gold standard for PCD diagnosis." For example, in one confirmed PCD case, GPT-OSS-120B provided the following justification: "*Nasal NO is a sensitive screening tool; genetic testing can identify pathogenic mutations; HSVM and TEM provide functional and structural confirmation of ciliary dysfunction*". This was annotated as 'adequate'. In contrast, for the same case, Phi-3.5-mini stated: "*The sweat chloride test measures the amount of salt in sweat, which is typically elevated in PCD*", which was annotated as inadequate due to inaccurately asserting sweat testing as a test for PCD. Together with the more subtle inaccuracies above, these findings emphasize the need for oversight from an experienced clinician in testing prioritization and interpretation, both to grasp nuances and to avoid fundamental misunderstanding of tests recommended by LLMs.

### 5.3. Clinical Implications

Taken together, these findings suggest that no single model is uniformly best across all dimensions, even across models spanning approximately 7B to 120B parameters. Phi-3.5-mini and the ensemble achieve the highest sensitivity, but Phi-3.5-mini also shows the highest hallucination rate (8.6%) and notable inappropriate testing in confirmed cases. GPT-OSS-120B produces the highest proportion of clinically sound explanations (48.3%) and adequate test justifications (51.7%), with no inappropriate testing, but lower sensitivity. Llama-3-70B shows zero hallucinations and the highest overall rate of appropriate testing (70.7%), but weaker diagnostic explanation quality. Overall, current models appear best suited for early signal amplification and case review rather than independent clinical decision-making. Improving prioritization, justification quality, and consistency across models remains a key challenge for future work.

## 6. Conclusion & Future Work

Using real first-visit pediatric pulmonology notes, this study shows that several openly available language models can spot early signs that should trigger a closer look for primary ciliary dyskinesia, even before the diagnosis is known. Across models, the main strength was catching most true cases, but they also flagged many children who did not have the disease, which could lead to extra referrals or testing if used without clinician judgment. A clinician review showed that even though model outputs

stayed close to the note, their explanations and test choices were often only partly convincing, especially when choosing the most useful next steps that mattered most. In particular, despite high sensitivity when affirmatively recommending additional PCD evaluation, explanations of that recommendation – as well as recommended testing and justification of that testing – often did not appropriately reflect PCD as the disease process in question, which is a concerning weakness if the goal is to use these models as clinical decision making aids. Overall, these systems look more suitable as a cautious "early warning" aid to prompt review, not as a standalone decision maker. Our results justify larger studies across more hospitals and longer patient records to judge real-world usefulness.

## 7. Limitations

While this study demonstrates that LLMs can identify diagnostic signals from routine clinical notes, the findings should be interpreted as an initial step. The sample size was limited to 58 patients from a single pediatric hospital, which may affect generalizability to other settings and populations. We analyzed only the first pulmonology note per patient to reflect early referral, but later clinical information was not included. The models were provided with only one provider note. A task more reflective of clinical practice would be to review multiple provider notes, or even the lifetime accumulated notes and results within an EMR, to guide PCD diagnostic decision making. However, the volume of chart review, input limitations, as well as previous studies suggesting that larger volumes of information may actually impede LLM performance (Tang et al., 2023), made this approach impractical. Given this, we opted to utilize the first encounter note with a pulmonology specialist, as it tends to be a critical juncture where the possibility of disease exists, but diagnostic uncertainty remains – therefore an optimal time where providers or patients might turn to LLMs for guidance on evaluation. The evaluation focused on open source models with a single instruction setup, and model explanations were reviewed by one clinician. Future work should include larger and more diverse datasets, multiple institutions, longitudinal notes, additional models and prompting strategies, and review by multiple clinicians to better assess robustness and clinical usefulness.

## 8. Ethical Considerations

Models were evaluated retrospectively and not used for real clinical decisions. We emphasize that such systems are intended for research on clinical text understanding and should not be used

as standalone diagnostic tools without appropriate clinical oversight. The authors declare no competing interest.

## 9. Bibliographical References

- Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. [Phi-3 technical report: A highly capable language model locally on your phone.](#)
- Tassallah Abdullahi, Ritambhara Singh, and Carsten Eickhoff. 2024. [Learning to Make Rare and Complex Diagnoses With Generative AI Assistance: Qualitative Study of Popular Large Language Models.](#) *JMIR Medical Education*, 10(1):e51391. Company: JMIR Medical Education Distributor: JMIR Medical Education Institution: JMIR Medical Education Label: JMIR Medical Education Publisher: JMIR Publications Inc., Toronto, Canada.
- Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. 2025. [gpt-oss-120b & gpt-oss-20b model card.](#) *arXiv preprint arXiv:2508.10925*.
- AI@Meta. 2024. [Llama 3 model card.](#)
- Laura Behan, Audrey Dunn Galvin, Bruna Rubbo, Sarah Masefield, Fiona Copeland, Michele Manion, Bernhard Rindlisbacher, Beatrice Redfern, and Jane S Lucas. 2016. [Diagnosing primary ciliary dyskinesia: an international patient perspective.](#) *European Respiratory Journal*, 48(4):1096–1107.
- Gully Burns, Carey Kauffman, Michele Manion, Ruth-Anne Pai, Carlos Milla, Michael G. O'Connor, Adam J. Shapiro, and Heidi Bjornson-Pennell. 2025. [Feasibility of machine learning analysis for the identification of patients with possible primary ciliary dyskinesia.](#) *Orphanet Journal of Rare Diseases*, 20(1):516.
- Xuanzhong Chen, Xiaohao Mao, Qihan Guo, Lun Wang, Shuyang Zhang, and Ting Chen. 2024. [RareBench: Can LLMs Serve as Rare Diseases Specialists?](#) In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '24*, pages 4850–4861, New York, NY, USA. Association for Computing Machinery.
- Sudeshna Das, Yao Ge, Yuting Guo, Swati Rajwal, JaMor Hairston, Jeanne Powell, Drew Walker, Snigdha Peddireddy, Sahithi Lakamana, Selen Bozkurt, et al. 2025. [Two-layer retrieval-augmented generation framework for low-resource medical question answering using reddit data: proof-of-concept study.](#) *Journal of Medical Internet Research*, 27:e66220.
- Karamarie Fecho, Stanley C Ahalt, Michael Knowles, Ashok Krishnamurthy, Margaret Leigh, Kenneth Morton, Emily Pfaff, Max Wang, and Hong Yi. 2022. [Leveraging open electronic health record data and environmental exposures data to derive insights into rare pulmonary disease.](#) *Frontiers in Artificial Intelligence*, 5:918888.
- Frank M. Hodges, Sundareswar Pullela, Guy Cohen, Akshay Mulgund, Jared C. Roach, and Stephen A. Ramsey. 2025. [Using AI to Improve Diagnosis and Treatment of Rare Diseases: A Chat Agent for Equitable and Accessible Healthcare.](#) In *Artificial Intelligence in*

- Medicine*, pages 186–191, Cham. Springer Nature Switzerland.
- Michael R. Knowles, Maimoona Zariwala, and Margaret Leigh. 2016. [Primary ciliary dyskinesia](#). *Clinics in Chest Medicine*, 37(3):449–461. Rare and Orphan Lung Diseases.
- Mistral AI. 2024. [Mistral7b](#). Hugging Face Model. Apache-2.0 License.
- Enock Niyonkuru, J Harry Caufield, Leigh C Carmody, Michael A Gargano, Sabrina Toro, Patricia L Whetzel, Hannah Blau, Mauricio Soto Gomez, Elena Casiraghi, Leonardo Chimirri, Justin T Reese, Giorgio Valentini, Melissa A Haendel, Christopher J Mungall, and Peter N Robinson. 2025. [Leveraging generative AI to assist biocuration of medical actions for rare disease](#). *Bioinformatics Advances*, 5(1):vbaf141.
- Mirjam Nussbaumer, Elisabeth Kieninger, Stefan A Tschanz, Sibel T Savas, Carmen Casaulta, Myrofora Goutaki, Sylvain Blanchon, Andreas Jung, Nicolas Regamey, Claudia E Kuehni, et al. 2021. [Diagnosis of primary ciliary dyskinesia: discrepancy according to different algorithms](#). *ERJ open research*, 7(4).
- PCD Foundation. 2026. Pcd foundation registry. <https://www.pcdfoundation.org/registry/>. Accessed: 2026-03-29.
- PCD Support UK. 2026. National pcd registry. <https://pcdsupport.org.uk/pcd-registry/>. Accessed: 2026-03-29.
- Swati Rajwal, Eugene Agichtein, and Abeed Sarker. 2024. [EM\\_Mixers at MEDIQA-CORR 2024: Knowledge-enhanced few-shot in-context learning for medical error detection and correction](#). In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 590–595, Mexico City, Mexico. Association for Computational Linguistics.
- Swati Rajwal, Mary Ellen Fain, Kiri Diaz-Asper, George Lucas Silva, Abeed Sarker, and Lokesh Guglani. 2026. Assessing the performance of large language models for diagnostic reasoning for primary ciliary dyskinesia in pediatric patients. *American Journal of Respiratory and Critical Care Medicine*, 212(Abstracts).
- Swati Rajwal, Shivank Garg, Reem Abdel-Salam, and Abdelrahman Zayed. 2025. [Do biased models have biased thoughts?](#) In *Second Conference on Language Modeling*.
- Ann-Katrin Rother, Nicolaus Schwerk, Folke Brinkmann, Frank Klawonn, Werner Lechner, and Lorenz Grigull. 2015. [Diagnostic support for selected paediatric pulmonary diseases using answer-pattern recognition in questionnaires based on combined data mining applications—a monocentric observational pilot study](#). *PLoS One*, 10(8):e0135180.
- Alejandro Ríos-Hoyo, Naing Lin Shan, Anran Li, Alexander T. Pearson, Lajos Pusztai, and Frederick M. Howard. 2024. [Evaluation of large language models as a diagnostic aid for complex medical cases](#). *Frontiers in Medicine*, 11. Publisher: Frontiers.
- Alyssa K Salinas, Gerard Smits, Stephen B Spurgin, Tanya M Martínez-Fernandez, and Yadira M Rivera-Sánchez. 2025. [Referral rates and diagnostic evaluation for primary ciliary dyskinesia in patients with laterality defects who meet primary ciliary dyskinesia evaluation criteria](#). *Pediatric pulmonology*, 60(6):e71129.
- Cathy Shyr, Yan Hu, Lisa Bastarache, Alex Cheng, Rizwan Hamid, Paul Harris, and Hua Xu. 2024. [Identifying and Extracting Rare Diseases and Their Phenotypes with Large Language Models](#). *Journal of Healthcare Informatics Research*, 8(2):438–461.
- Liyan Tang, Zhaoyi Sun, Betina Ilday, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. 2023. [Evaluating large language models on medical evidence summarization](#). *NPJ digital medicine*, 6(1):158.
- Osnat Tausky, Eytan Kaplan, Gili Kadmon, Yulia Gendler, Elhanan Nahum, Shai Yitzhaki, and Avichai Weissbach. 2025. [Large language model as a clinical decision support tool in the initial management of critically ill children: a pilot evaluation](#). *European Journal of Pediatrics*, 184(12):757.
- Qihong Wei, Zhengxiong Yao, Ying Cui, Bo Wei, Zhezhen Jin, and Ximing Xu. 2024. [Evaluation of chatgpt-generated medical responses: A systematic review and meta-analysis](#). *Journal of Biomedical Informatics*, 151:104620.
- Jinge Wu, Hang Dong, Zexi Li, Haowei Wang, Runci Li, Arijit Patra, Chengliang Dai, Waqar Ali, Phil Scordis, and Honghan Wu. 2024. [A hybrid framework with large language models for rare disease phenotyping](#). *BMC Medical Informatics and Decision Making*, 24(1):289.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei

Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Maimoona A. Zariwala, Michael R. Knowles, and Margaret W. Leigh. 1993–2025. [Primary ciliary dyskinesia](#). In Margaret P. Adam, Ghayda M. Mirzaa, Roberta A. Pagon, Stephanie E. Wallace, Laura J. H. Bean, Kevin W. Gripp, and Anne Amemiya, editors, *GeneReviews®*. University of Washington, Seattle. Updated 2025.

Wangji Zhou, Qiaoling Chen, Yaqi Wang, Anhui Guo, Aohua Wu, Xueqi Liu, Jinrong Dai, Shuzhen Meng, Christopher Situ, Yaping Liu, et al. 2025. [An electronic medical record retrieval system can be used to identify missed diagnosis in patients with primary ciliary dyskinesia](#). *Journal of internal medicine*, 297(1):93–100.

## Appendix

### A. Sample case with LLM output

Table A.1 shows another sample case with predictions and justifications by each LLM. Another case is shown in Table 2.

Case X is a XX-month-old female who presents with transfer from XX facility due to concern for possible tracheal deviation, persistent oxygen requirement, and poor oral feeding. Consult requested for home oxygen and outpatient pulmonary follow-up. She is on XX LPM oxygen and being treated for BPD with inhaled corticosteroid (XX dose, XX puffs BID) and albuterol every XX hours. She is also on diuretics (XX). Mild reflux diagnosed on pH probe. Currently on reflux medications (XX) and receiving NG feeds. No desaturation episodes reported. Per review of outside records, she was extubated to CPAP at XX days of life and remained on CPAP for approximately XX duration before transition to high-flow nasal cannula, now on low-flow nasal cannula. She previously received diuretic therapy for approximately XX duration, then transitioned to current regimen prior to transfer. History of pulmonary hypertension treated with inhaled nitric oxide for XX days, now resolved. Continues on low-flow oxygen. Transferred for evaluation of suspected tracheal deviation to rule out airway abnormality.

LLM	LLM Response	Expert's Evaluation
<b>GPT-OSS-120B</b>	PCD: No	×
	Justification: The infant has chronic lung disease secondary to prematurity (BPD) with a mild right tracheal deviation likely due to a prominent thymus. There is no history of recurrent sinusitis, otitis media, or bronchiectasis, and no laterality defects or unexplained persistent cough that would suggest primary ciliary dyskinesia. The clinical picture is fully explained by prematurity, BPD, and gastroesophageal reflux, making PCD evaluation unnecessary at this time.	Clinically sound
	Testing: None	Incomplete
	Justification: No additional diagnostic work-up for PCD is indicated because the patient's symptoms are attributable to BPD and prematurity. Routine follow-up imaging (repeat echocardiogram) and clinical monitoring are already planned, and there are no findings that would warrant nasal nitric oxide measurement, ciliary biopsy, or genetic testing for PCD.	Adequate
<b>Llama-3-70B</b>	PCD: Yes	✓
	Justification: The patient has a history of bronchopulmonary dysplasia (BPD) and is on oxygen therapy, which increases the risk of primary ciliary dyskinesia (PCD). Additionally, the patient has a mild tracheal deviation, which could be related to PCD. Further evaluation is necessary to rule out PCD.	Clinically concerning
	Testing: Nasal nitric oxide measurement, high-speed video microscopy analysis of cilia, and genetic testing	Appropriate
	Justification: These tests are recommended to evaluate the patient's ciliary function and structure, which can help diagnose PCD. Nasal nitric oxide measurement is a non-invasive test that can indicate PCD, while high-speed video microscopy analysis of cilia can assess ciliary beat frequency and pattern. Genetic testing can identify mutations associated with PCD.	Adequate
<b>Qwen-2.5-7B</b>	PCD: Uncertain	×
	Justification: The patient has a history of BPD, mild tracheal deviation, and potential pulmonary hypertension. While the tracheal deviation could be secondary to a prominent thymus, further evaluation is warranted to rule out other causes and to monitor the patient's condition.	Clinically weak
	Testing: CT of the chest with contrast	Incomplete
	Justification: To better evaluate the tracheal deviation and rule out other causes, a CT of the chest with contrast is recommended. This will provide more detailed imaging to assess the trachea and other structures in the chest.	Adequate
<b>Mistral-7B</b>	PCD: Yes	✓
	Justification: The patient has a history of BPD, is on oxygen, and has a mildly dilated RV, thickened RV wall, and a probable left coronary cameral fistula to the RV on echocardiogram, suggesting possible pulmonary hypertension.	Clinically weak
	Testing: Right Heart Catheterization	Excessive
	Justification: To confirm the diagnosis of pulmonary hypertension and guide treatment	Adequate
<b>Phi-3.5-mini</b>	PCD: Yes	✓
	Justification: The patient has a history of BPD, mild tracheal deviation, and a probable coronary fistula, which could contribute to respiratory symptoms. The presence of chronic lung disease and the need to rule out pulmonary hypertension further support the need for PCD evaluation.	Clinically weak
	Testing: Echocardiogram with contrast, repeat chest CT, and possibly a cardiac MRI.	Excessive
	Justification: The echocardiogram with contrast is recommended to further evaluate the suspected coronary fistula and to rule out pulmonary hypertension. A repeat chest CT is suggested to monitor the chronic lung disease and tracheal deviation. A cardiac MRI may provide additional information on the coronary fistula and cardiac structure.	Adequate

Table A.1: LLM responses and annotator evaluation for a sample case. The top row shows a partial excerpt of the input clinical note.