Rare Disease Differential Diagnosis with Large Language Models at Scale: From Abdominal Actinomycosis to Wilson's Disease

Anonymous ACL submission

Abstract

Large language models (LLMs) have demonstrated impressive capabilities in disease diagnosis. However, their effectiveness in identifying rarer diseases, which are inherently more challenging to diagnose, remains an open question. Rare disease performance is critical with the increasing use of LLMs in healthcare settings. This is especially true if a primary care physician needs to make a rarer prognosis from only a patient conversation so that they can take the appropriate next step. To that end, several clinical decision support systems are designed to support providers in rare disease identification. Yet their utility is limited due to their lack of knowledge of common disorders and difficulty of use.

In this paper, we propose RareScale to combine the knowledge LLMs with expert systems. We use jointly use an expert system and LLM to simulate rare disease chats. This data is used to train a rare disease candidate predictor model. Candidates from this smaller model are then used as additional inputs to black-box LLM to make the final differential diagnosis. Thus, RareScale allows for a balance between rare and common diagnoses. We present results on over 575 rare diseases, beginning with Abdominal Actinomycosis and ending with Wilson's Disease. Our approach significantly improves the baseline performance of black-box LLM by over 17% in Top-5 accuracy. We also find that our candidate generation performance is high (*e.g.* 88.8% on gpt-4o generated chats).

1 Introduction

Rare diseases, often overlooked in the medical landscape, impact approximately 5.9% of the global population. For those affected, the journey to diagnosis is fraught with challenges, with an average wait of four to five years. Patients typically endure three misdiagnoses and consultations with at least five doctors during this process. This prolonged diagnostic odyssey is often due to limited provider knowledge, implicit biases, or symptom overlap with common conditions (Kliegman and Brett J Bordini, 2017; Office, 2021). Only in a few instances does the absence of specific tests or the extreme rarity of the disease present the challenge.

Clinical decision support systems focusing on rare disease diagnoses emerged in the 1980s as a tool to address this challenge Miller et al. (1982); Buchanan and Shortliffe (1985); Barnett et al. (1987); Lauritzen and Spiegelhalter (1988); Jackson (1998). Practical use of these systems, however, has been constrained by several factors (Miller, 1994), including lack of integration with physician workflows. Additionally, they are often custom-built explicitly for rare diseases and exclude common conditions.

Recent advances in large language models (LLMs; OpenAI (2024); Anthropic (2024); Meta (2024); Jiang et al. (2023)) achieve state-of-the-art performance on a variety of tasks (Zheng et al., 2023; Wang et al., 2019; Hendrycks et al., 2020), including in high-stakes healthcare applications (Chen et al., 2023; Thawakar et al., 2024; Nair et al., 2024). Studies on rare disease diagnosis using LLMs show similar promise but have been on smaller, curated datasets of clinical vignettes that often include laboratory and imaging information to diagnose (Hu et al., 2023; Shyr et al., 2024; Sandmann et al., 2024; Chen et al., 2024; Yang et al., 2024b,a; Mehnen et al., 2023; Shyr et al., 2024; do Olmo et al., 2024).

Furthermore, the challenge with these studies is that the first point of contact for the patient is the primary care physician. For an untrained physician in rare diseases, the overlap of symptoms (*e.g.*, *pulmonary legionellosis* usually manifests with symptoms similar to flu or pneumonia) makes it easy to overlook without specialized knowledge. While rare diagnoses often require additional information beyond the initial symptom presentation, they cannot be diagnosed if not considered in the first

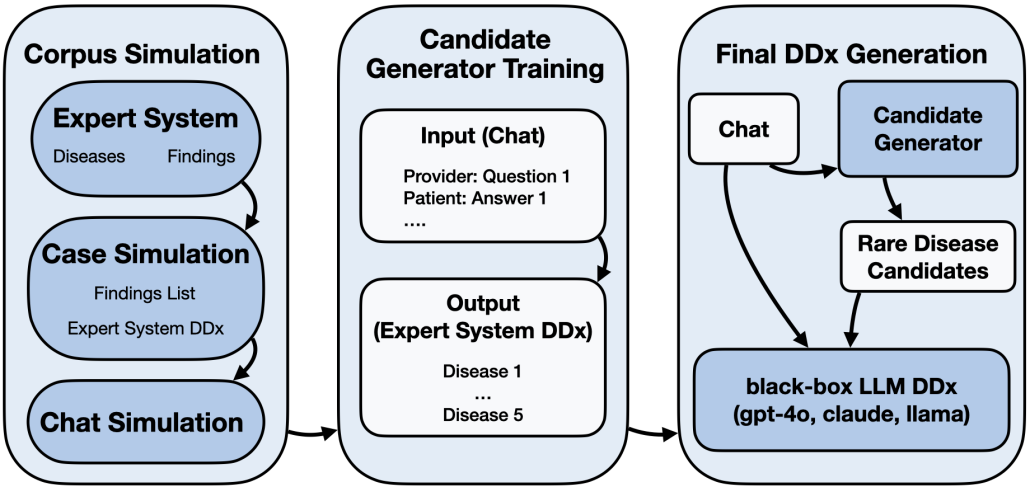

Figure 1: An overview of RareScale. Our approach consists of three stages. First, we simulate a corpus of rare disease chats (§3 and Figure 2). Then, we train a candidate generation LLM (§4.1). Finally, we perform inference to generate a final DDx (§4.2).

place. This illustrates the criticality of including relevant rare diseases in the differential diagnosis, even before any lab work, directly based on patient-provider history-taking conversations. We found that applying black-box LLMs to the task of rare disease diagnosis shows only moderate performance (*e.g.* 56.8% Top-5 accuracy using gpt-4o). This indicates that while LLMs have some knowledge of rare diseases, they also may struggle in certain cases.

In this paper, we improve the diagnostic capability of LLMs in identifying rare diseases based on two key insights. First, **given LLMs have some knowledge of rare diseases, can we more explicitly target this knowledge?** Second, **can knowledge from rare expert systems be used to better evaluate and inform LLMs?**

We propose RareScale, which is designed to improve differential diagnoses of rare diseases directly from history-taking dialogues. Figure 1 provides an overview of our approach.

We utilize a rare disease expert system and a medical case simulator to generate a broad set of structured clinical cases with closed-world assumptions within the expert system. These cases consist of structured findings, that a patient with the given rare disease may exhibit. These cases are then used as inputs to the history-taking conversation simulator that generates free-form provider questions and patient answers that conform to those cases. This approach covers 575 rare diseases, beginning with Abdominal Actinomycosis and ending with Wilson's Disease.

Given that the expert system excludes common diseases, we develop a rare disease candidate generation model instead of a full diagnosis system. For each case, we generate a list of candidate rare diseases to prompt a larger, black-box LLM, which combines its general diagnostic capabilities with the specialized knowledge of the smaller model.

Our results show statistically significant performance improvements compared to relying solely on LLMs like GPT-4o. For example, we find that the Top-5 accuracy improves from 56.7% to 74.1% on gpt-4o generated chats. We believe this illustrates the potential of integrating curated expert knowledge sources and the power of LLMs.

2 Problem Setup

When a patient visits a medical provider with a new, unknown health condition, the first step is understanding the issue deeply. This process, known as history-taking, consists of collecting all relevant information about the patient's current health issue, including positive and negative findings (or symptoms), via a series of questions. These details are used to make treatment decisions if a confident diagnosis can be made. Alternatively, they inform the selection of tests or imaging to help further diagnose.

What if a possible diagnosis is not even considered because it's so rare, and the medical provider is unaware of it? This paper aims at proposing a method RareScale as a step towards removing that educational barrier. Given the history-taking conversation between the provider and the patient,

the goal is to identify possible rare diseases that need to be considered. When doing this, we must balance the fact that the disease is rare and that the patient is more likely to have a common disease.

Overview of the approach Figure 1 provides the overview of RareScale: It trains a smaller expert LLM § 4 to generate candidate rare diseases using a simulated labeled conversational dataset using a combination of expert systems and LLMs (§ 3). The candidates generated from this model are used as additional inputs to the black-box LLM to enable better weighing rare and common diseases. This leads to improved efficacy over using only black-box LLMs.

3 Corpus Simulation

The use of synthetic data from large language models has been widely studied (Li et al., 2023; Yu et al., 2023; Liu et al., 2024). However, our task requires generating history-taking conversational data with labeled differential diagnosis. Since identifying differential diagnosis is the task we are trying to solve with RareScale, we took a different approach to generate labeled history taking chats to avoid simply evaluating an LLM based on it's existing knowledge.

As shown in Figure, 2, we first simulate an example using expert system knowledge. We start with a seed disease and generate a structured patient case (represented as a list of findings). Such a generation can have ambiguity about the final diagnosis. Therefore, we use the same expert system to assign a full differential diagnosis. The structured case is then used to simulate a history-taking conversation as detailed in § 3.2 so that LLM needs to only generate a patient-facing question with the provided finding, while maintaining the conversational flow. We repeat this process independently for all diseases. The resulting dataset of history-taking conversation and differential diagnosis pairs is used to train rare diseases candidate generation model (§ 4).

3.1 Generation of structured cases

We use a rare disease expert system, evolved from Internist-1 (Miller et al., 1982) and QMR (Miller and Masarie, 1990), to generate structured cases using expert-curated diagnostic rules based on a knowledge base of diseases, findings, and their relationships. Findings include symptoms, signs, lab results, demographics, or medical history. In this study, we focus on patient-answerable findings and *only* focus on symptoms. Each finding-disease link is defined by *evoking strength* and *frequency*, scored from 1 to 5 by a team of medical experts based on the studies for that disease. *Evoking strength* measures the association between a finding and a disease, while *frequency* indicates how often a finding occurs in patients with the disease. Each finding also has a disease-independent *import* variable, indicating its global importance.

Our simulation algorithm, adapted from medical case simulator Parker and Miller (1989); Ravuri et al. (2018), starts by sampling a seed disease and sequentially constructing a set of findings. First, demographic variables are sampled, followed by predisposing factors, and then other findings are made to decrease frequency relative to the disease. Each finding is randomly determined to be present or absent, with impossible findings excluded and high co-occurrence findings prioritized. The simulation concludes once all findings in the knowledge base for that disease are considered, operating under a closed-world assumption that limits diseases and findings to those in the knowledge base. Unlike previous approaches to simulation, we also use the expert system to compute differential diagnosis after the first six findings are sampled. We use this to consider findings from other diagnoses in the differential diagnosis to prioritize sampling additional negative findings that overlap with the seed disease. This allows the sample to be slightly more targeted at the seed disease. In turn, we widen the gap between the seed disease and the remainder of the differential diagnosis.

Each simulation includes a differential diagnosis of up to 5 diseases as ranked by expert system scoring. A DDx may consist of multiple diseases in cases where a final diagnosis may not be ascertained without laboratory testing. We maintain this differential diagnosis list for training in §4. We iterate through 630 rare diseases in the expert system, using each as a seed disease. We keep the generated structured case if the seed disease is top-scoring in the differential to reduce the effect of the noise in the simulation process. We set a minimum of 50 valid simulations out of 200 attempts. Any diseases falling below that are excluded (*e.g.* Friedreich's ataxia), as this typically indicates that the disease is unidentifiable from symptoms alone. This results in 575 diseases in our dataset. We include an example in Appendix §3.

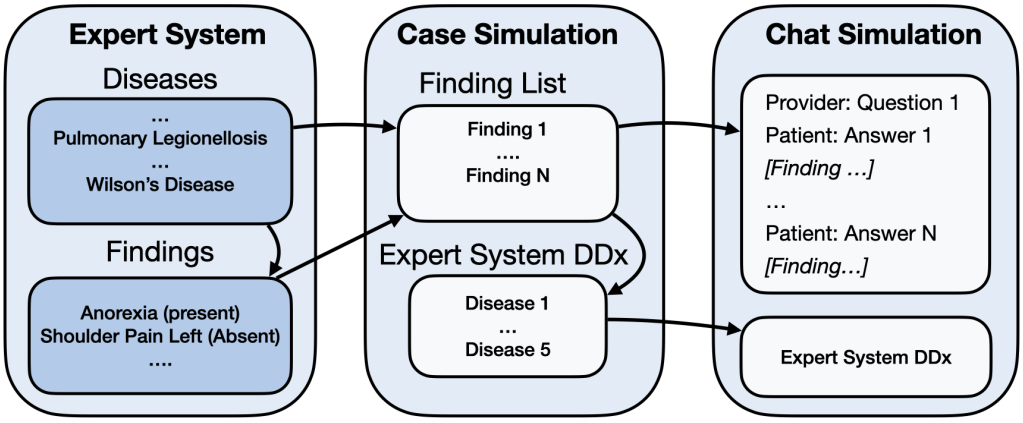

Figure 2: RareScale's Corpus Simulation Pipeline (§3). Using an expert system, we create a set of structured case simulations, which are then used to guide an LLM in history taking chat generation.

3.2 Generation of chats from structured cases

For each (structured case, DDx) pair, we start with creating a complete demographic profile. This includes name, gender, age, race, education, and location by first selecting from the structured case. If not present in the case, we randomly generate from the LLM. We incorporate location and education to introduce additional variability in patient language. This diversifies the profiles and reduces sensitivity to names and locations across different simulations, which could lead to misleading patterns.

We use the structured case along with the expanded demographic profile to anchor LLM-powered history-taking chat simulation. We prompt an LLM to create a conversation that closely adheres to that case. The LLM-based chat simulator has access to the disease name, finding set, and demographic profile. For each finding, we include a definition if the expert system provides one so the LLM's understanding of the term aligns with the expert system. For each disease, we persist a database of messages that correspond to each finding (*e.g.* "I feel hot" corresponds to *fever (present)*). We include the previous message in the prompt for each subsequent generation and ask the LLM to generate a different one so it does not repeat.

We enforce a format for the chat, which begins with a "system" message with the patient's demographic information. The provider simulator starts the conversation open-ended, and the patient simulator reports the symptoms drawn from the findings set. For each patient utterance, the LLM outputs both the message and which findings are included. This is to encourage the message to be consistent with the findings. We prompt LLM for the patient simulator to report only at most, three findings in a single message and also prompt the LLM for the provider simulator not to generate leading questions.

For the simulation, we use two different LLMs to generate our chats – OpenAI's Gpt-4o and Anthropic's Claude-3.5-sonnet – and use the same approach unless noted. All components of the generation pipeline use only one of the LLMs. For chats generated by GPT-4o, we found that doing this process in a single prompt is sufficient (see Appendix Prompt 1). For Claude, we found that this generated chats which were very terse. Therefore, we prompt Claude with the same information and ask it to generate one patient-provider turn at a time (see Appendix Prompt 2). The findings provided to Claude are separated into findings already included and those that need to be added.

For chats generated by both LLMs, we finally run a checker prompt that ensures that the resulting chats include all symptoms in the finding set and edits the chat if not all are included. If the first edit fails, we try two more times. If all other attempts fail, we exclude this chat from our dataset. This results in a corpus of rare disease chats – 28,589 using GPT-4o and 14,573 using Claude (further statistics are included in Appendix Table 5). We include a full example chat using GPT-4o in Appendix Figure 4 and using Claude in Figure 5.

4 Rare Disease Differential Diagnoses

The RareScale scalable synthetic generation pipeline provides us with a sizable dataset of 575 rare disease history-taking conversations. However, our expert system doesn't provide information on a variety of common diseases, such as the Common

Cold. A real-world diagnostic system will need to account for these diseases. Given the high diagnostic performance of many LLMs (Liu et al., 2025; Rutledge, 2024; Zhou et al., 2024; Tu et al., 2024), it would be challenging to meet their performance even with a corpus of similar non-rare chats.

We therefore combine existing LLM diagnostic performance with a specialized rare-disease model. To do this, we train a rare disease candidate generation system which proposes a set of at most 5 rare diseases. This list is recall-oriented, to allow several possibilities to be considered. These rare disease candidates are then passed through a prompt of a larger LLM to produce a final diagnostic list. The combination can leverage the niche expertise of an expert system while also leveraging the broader knowledge of general LLMs.

4.1 Candidate Generation

To generate rare disease candidates for a given history-taking conversation, we train a smaller LLM to generate a list of at most 5 rare diseases. We train using Llama 3.1 7b as our base model (Meta, 2024) using the chat text as input (training details in Appendix §A.1). We exclude the structured findings and all other information used in the previous section. Note that these history-taking chats do not include a final diagnosis from the provider, so the model must make inferences from the chat alone. As the target output, we use the differential diagnosis list produced by our expert systems which contains up to five diseases. We only include the disease names ordered by likelihood per the expert system score but do not include the scores themselves.

4.2 Diagnosis Generation

For the final step of generating the differential diagnosis for the conversation, we use a larger black-box general LLMs[1] (gpt-4o, claude, llama 3.3 70b) to fuse common and rare diseases seamlessly. We leave the role of identifying the relevant common diseases to the LLM, but infuse the candidate rare diseases through inferencing on the smaller model we trained according to § 4.1

The prompt (Appendix Prompt 3) uses multi-stage instructions, instructing the LLM to generate a list of 5 diagnoses without considering the rare list. Then, it is instructed to consider the rare disease candidate list. Finally, it selects whichever from the two sections is most appropriate, including discarding all rare candidates if best. The prompt is also flexible, so we can remove instructions for incorporating candidate rare diseases or include other means of candidate rare diseases when available. We study these variations in § 5.1.

[1]We explored using the publicly-available MEDITRON-70B(Chen et al., 2023), but it could not follow the DDx instructions in the prompt, instead outputting unrelated text.

5 Evaluation and Results

To understand the performance of RareScale, we frame our evaluation around three questions. First, does the RareScale end-to-end approach improve rare disease diagnoses over black-box LLMs (§5.1)? Second, how does our expert candidate generation model perform at producing an accurate list of rare diseases for each case (§5.2)? Finally, how accurate are the simulated chats as judged by experts (§5.3)? We include details on dataset creation in Appendix §A.2.

5.1 DDx Improvements with RareScale

We compare the end-to-end RareScale to two baselines. First, in **base gpt-4o**, we generate a DDx with no external candidate list. **gpt-4o rare candidates** uses the RareScale prompt described in § 4.2 without any candidate list. Similar to RareScale, **gpt-4o rare candidates** uses rare disease candidate list, but obtained using a separate LLM prompt. In contrast, RareScale uses a smaller trained model to generate the candidate list. For all three methods, we use the same prompt setup (§ 4.2) except for the additional candidate list and instructions.

To be able to compare the generated DDx to the ground seed disease, we use judge prompts adopted from Tu et al. (2024) (Appendix Prompts 4 and 5), with gpt-4o as judge model. Previously, these prompts were shown to be well-calibrated with the human expert as a judge. The first judge prompt returns a binary judgment, iterating through the DDx list and taking the first match present, if any. The second judge prompt compares the full DDx list to the seed disease and returns a degree of similarity (unrelated to exact match) We use Top-1 accuracy, Top-5 accuracy, and mean reciprocal rank (MRR) as the metrics.

Results Table 1 shows the performance of the combination of our best-performing candidate generation model (gpt-4o + claude, see §5.2) with gpt-4o serving as a DDx generator. We find that RareScale produces the best performance across

Differential Diagnosis	gpt-4o test set (n=3403)			claude test set (n=2868)		
	Top-5	Top -1	MRR	Top-5	Top -1	MRR
baseline	56.80%	28.65%	0.390	56.69%	30.65%	0.406
gpt-4o rare candidates	52.66%	25.95%	0.357	55.47%	29.04%	0.388
RareScale candidates	**74.38%**	**33.12%**	**0.471**	**71.41%**	**33.23%**	**0.461**

Table 1: Performance on generated gpt-4o ddx task. All metrics for RareScale on both datasets (see bolded) are significant using a two-sided Wilcoxon signed-rank test with $p < 0.01$ compared to the no candidates baseline.

DDx LLM	Exact	Extremely Rel.	Relevant	Somewhat Rel.	Unrelated
baseline gpt-4o	22.8%	19.9%	4.9%	21.0%	31.3%
RareScale gpt-4o	55.8%	8.8%	2.3%	12.8%	20.2%
baseline claude	19.2%	16.9%	3.9%	14.5%	45.6%
RareScale claude	56.8%	10.7%	1.6%	10.6%	20.4%
baseline Llama 3.3 70b	20.3%	19.3%	5.3%	21.7%	33.5%
RareScale Llama 3,3 70b	47.3%	12.2%	3.3%	15.4%	21.9%

Table 2: We compare LLM baseline DDx generation performance to LLMs with addition of RareScale candidates. We report the LLM as judge results across several categories of similarity, ranging from Exact Match to Unrelated. We combine gpt-4o and claude test sets for this analysis.

the board, including statistically significant gains over the no candidate baseline. Notably, we also find that performance is worse when using gpt-4o as the rare disease candidate generator, reinforcing the utility of a specialized model.

We believe the most crucial improvement is seen in the Top-5 performance, as adding the rare disease candidate in the DDx would allow a provider to consider that disease. The degree of improvement – 17.42% – illustrates this impact. While we also see improvements in MRR performance, we also observe that the rare disease candidate is not very likely to be in the first position.

Including the rare disease candidate but not over-relying on the candidate list is likely the preferred behavior (see § 5.2). In many cases, a common disease should be considered first, but the inclusion of a rare possibility can help a provider to best identify next steps. Using only our smaller rare disease candidate generation model would likely overproduce rare diagnoses, whereas providing a black box LLM with options allows it to weigh the entire picture. Alternatively, the LLM may not be able to properly diagnose a specific rare condition because it has no knowledge of it to begin with (see the analysis for further discussion).

In Table 2, we compare LLM performance on the differential diagnosis task with and without RareScale broken down by similarity level. We combine the gpt-4o and claude chats for this analysis. Since this prompt uses varying levels of granularity and compares the entire DDx, the distribution is different than with the binary prompt, which only compares diseases one-to-one.

In the case of gpt-4o, we see broad improvements when adding RareScale. This includes a 33% improvement in the number of Exact Matches. Yet we see even larger performance improvements when applying RareScale to claude-sonnet's DDx capabilities. We see a 25.2% reduction in unrelated diagnoses and a 37% increase in exact matches. Surprisingly, we also see similar performance gains on Llama 3.3 70b, including a 27% increase in exact matches and an 11.7% reduction in unrelated matches. While there is a gap between the closed- and open-weight models, the margin is small.

Analysis We break down the results of the RareScale gpt-4o dataset results in Table 1 into performance by disease hierarchical category as taken from our expert system. A disease can align with one or more categories. The full results are shown in Appendix Table 4. The group with the largest number of diseases – Infectious diseases – performs slightly less than the average. This could potentially be due to the overall broader number of diseases in this category, many of them closely

Training Dataset	Training Size	gpt-4o test set (n=3403)			claude test set (n=2868)		
		Top-5	Top-1	MRR	Top-5	Top-1	MRR
claude	8837	48.37%	34.12%	0.4007	64.92%	45.64%	0.5371
gpt-4o	21782	88.04%	63.88%	0.7410	44.18%	28.45%	0.3490
gpt-4o downsampled	8813	70.88%	47.90%	0.5742	37.20%	23.25%	0.2884
gpt-4o + claude	30619	88.80%	64.21%	0.7463	77.82%	56.35%	0.6526

Table 3: Evaluation on the candidate generation task, with MRR, Top-5 and Top-1 Accuracy. We evaluate on models only trained on claude data, gpt-4o data, and both, and evaluate separately on claude and gpt-4o test sets. We include a model trained on a downsampled set of gpt-4o data that approximates the size of the claude training set.

related. Other large groups, such as *Neoplastic disease*, *Impaired cardiovascular function*, and *non-infectious inflammatory diseases* have roughly average performance. On the lower end, as expected, *Congenital disorders due to abnormal fetal development* performs worst at 53.3% Top-5 accuracy, whereas *Disorders of smooth muscle contraction and/or relaxation* performs best at 94.4%, although both are small categories.

5.2 RareScale Candidate Generation

We measure the performance of RareScale candidate generation by comparing it to the seed disease. The candidate generator outputs a list of at most 5 candidates, and we compare to the seed disease using exact match. As before, we use MRR and Top-1 and Top-5 accuracy. Since GPT-4o and Claude required different approaches to generate the chat data, we considered different dataset variations to understand differences in the performance.

Results Table 3 reports the performance of RareScale on candidate generation. Using the combined training set during training leads to the highest performance – 88.8% Top-5 on gpt-4o, and 77.82% on claude on the test sets. When only trained on one of the data sources, performance is best when applied to data from the same source. In our initial experiments, we found that including roughly 40 to 50 training examples per disease would achieve the same performance as more samples (*e.g.* 100), but fewer would result in worse performance.

While the gpt-4o test set performs similarly on gpt-4o trained models and combined training data, the claude chat test set sees a sizable performance bump when using all training data. Since our claude dataset is smaller in size, we hypothesize that the additional signal from the gpt-4o training examples helps improve performance on the claude test set. To illustrate this, we train a model on only gpt-4o data but randomly (per-seed disease basis) downsampled to roughly the same amount of data in claude's training set. We find a similar pattern – the training data reduction (40.5% of the original size) results in a 17% drop in Top-5 accuracy, and similar reductions for other metrics. This suggests the lower claude performance is due to dataset size instead of worse corpus simulation performance.

While we only consider exact matches in Table 3, we also run the similarity judge prompt (Prompt 5) on the gpt-4o + claude trained model, gpt-4o test set results. In only 5.2% of cases, the model returns a DDx list unrelated to the seed diagnosis. In the remaining cases, non-exact matches have some degree of similarity. We also include validation results in Appendix Table 7

5.3 Expert Evaluation of Chats

We evaluate whether our synthetically generated chats are possible manifestations of the seed rare disease. For this, we use an external reviewing service that provides medical annotators and use third-year medical students for this task as they have been recently trained in similar tasks. They can also use additional resources (*e.g.* medical references) during annotation. We also included hard negative examples during annotations to serve as distractors. We include details in Appendix §A.2.1.

Results On our annotation set of 651 cases (with 73 annotation-specific negatives), we found the overall agreement rate to be 88.6% with a Cohen's kappa of 0.53. For the expected positive cases, *i.e.* the disease was indicated for that patient case during simulation, the agreement rate was 90.48%. For the negative cases, the agreement rate was 73.97%. The high agreement rate of the positive examples used for training illustrates the performance of our synthetic generation pipeline. We include addi-

tional discussion in Appendix Section A.3.

6 Deployment Considerations

We show that RareScale significantly improves the ability to identify potential rare diseases over baseline LLM performance. We also know that most patients at a primary care clinic are likely to have a common disease. RareScale aims at balancing common diseases with rare diseases so that the primary care physician does not oversubscribe to the possibility of a rare disease. Such rare diseases are hard to rule out without expensive testing and added mental toll.

When deploying in practice, there are several possible approaches. For instance, a system could only include rare diseases in cases where common diseases are ruled out first. For example, *pulmonary legionellosis* presents similarly to the flu or pneumonia, and a provider would likely treat one of the common conditions first. But when flu or pneumonia management fails, they can use RareScale to consider rarer conditions. Alternatively, RareScale could be used to gather more information from the patient and exclude rarer conditions more quickly. We believe that this is an open problem that may require physician training and clinical testing.

7 Related Work

Using LLMs for diagnosis is increasingly powerful (Liu et al., 2025; Rutledge, 2024; Zhou et al., 2024; Tu et al., 2024). Previous rare disease investigations (Hu et al., 2023; Shyr et al., 2024; Sandmann et al., 2024; Chen et al., 2024; Yang et al., 2024a; Mehnen et al., 2023; Shyr et al., 2024; do Olmo et al., 2024) focus on generating diagnoses from smaller sets of vignettes, which are concise summaries of a patient's health issue. Vignettes are commonly used in narrow settings. Comparatively, using the history-taking chat replicates a primary care setting where confirmatory testing has not been performed. Previous chat-related work focused on a much smaller set (Yang et al., 2024b).

Prompting methods such as chain of thought (Wei et al., 2022) may improve rare disease performance to a degree but cannot enable the integration of external knowledge. Other techniques such as retrieval augmented generation (Lewis et al., 2020) are unlikely to capture the complex nuances of symptoms as stated by patients. LLM performance on emerging data has also been studied (Mitchell et al., 2021; Gu et al., 2024). A related approach is Yao et al. (2024), which uses a smaller expert model, but their approach seeks to update the original model. (Gupta et al., 2024) illustrates that model editing can also lead to the model losing knowledge. Other work (Zhao et al., 2024) studies LLM performance in other rare data situations, and studies the related issue of hallucinations(Fabbri et al., 2022; Min et al., 2023).

8 Conclusion

We propose RareScale as a multi-component approach to improve rare disease diagnosis. Based on curated rare disease expert systems, we use a medical case simulator to generate structured cases and differential diagnoses for each rare disease. In turn, we prompt black-box LLMs to generate history-taking conversations of these cases. This allows us to benchmark the performance of LLMs on rare disease history-taking conversations. We show that explicitly training an LLM for disease candidate generation, which can then provide suggestions to a black-box LLM, achieves statistically significant performance improvements compared to baselines.

While this paper focused on the rare disease diagnosis, we hope RareScale can transfer to other settings. For example, one could leverage a location-specific or population-specific expert system instead of using a rare-specific expert system. This could especially be useful in settings where the real-world disease distribution differs significantly and at the tail of the data distribution, which black-box training data can not corroborate.

Although our approach scales the number of rare diseases to 575, many rare diseases (NORD, 2025) are left unaddressed. An inherent limitation of our approach is that expert systems rely on the labor of highly-trained medical experts who review medical literature and curate knowledge. This allows us to use the expert system as a surrogate source of knowledge instead of directly integrating with rare disease literature. Yet expanding an expert system to all rare diseases with human experts alone is likely impossible. Future work should focus on using medical literature with techniques beyond retrieval augmented generation, which likely struggles to capture the subtlety and relative importance of the various studies. Alternatively, explorations that rely on electronic health record systems could remove the need for the expert system while still using a smaller model tuned for the rare disease candidate generation to be used as in RareScale.

9 Limitations

One limitation of our work is that we are limited to the rare diseases and symptoms provided in our rare disease expert system. While our approach does cover a broader set of rare diseases than existing work, it does not cover the vast set of diseases present in the world. For example, (NORD, 2025) lists over 10,000. Yet rare diseases highly vary in prevalence, with only about 100 rare diseases accounting for 80% of diagnoses(Evans and Rafi, 2016). Others are so rare that only a few cases have ever been identified. Expanding to these extremely rare diseases is a remaining challenge.

We are also limited by the number of expert annotations that are feasible for this task. While the results were positive, we couldn't annotate a larger amount without spending excessively more time and money. Additionally, we are unable to release the datasets due to legal requirements. While this is a common issue in the medical domain, we hope that future work can provide open datasets for broader use.

References

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

G. O. Barnett, J. J. Cimino, J. A. Hupp, and E. P. Hoffer. 1987. DXplain: An Evolving Diagnostic Decision-Support System. *JAMA*, 258(1):67–74.

B. G. Buchanan and E. H. Shortliffe. 1985. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley.

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*, pages 4850–4861.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

Juanjo do Olmo, Javier Logro~no, Carlos Masc´ıas, Marcelo Mart´ınez, and Juli´an Isla. 2024. Assessing dxgpt: Diagnosing rare diseases with various large language models. In *medRxiv*.

William Evans and Imran Rafi. 2016. Rare diseases in general practice: recognising the zebras among the horses. *British Journal of General Practice*, 66:550–551.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. 2024. Model editing can hurt general abilities of large language models. *arXiv preprint arXiv:2401.04700*.

Akshat Gupta, Anurag Rao, and Gopala Anumanchipalli. 2024. Model editing at scale leads to gradual and catastrophic forgetting. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15202–15232, Bangkok, Thailand. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Xiaoyan Hu, An Ran Ran, Truong X Nguyen, Simon Szeto, Jason C Yam, Carmen KM Chan, and Carol Y Cheung. 2023. What can gpt-4 do for diagnosing rare eye diseases? a pilot study. *Ophthalmology and Therapy*, 12(6):3395–3402.

P. Jackson. 1998. *Introduction to Expert Systems*, 3rd edition. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Robert M Kliegman and James J Nocton Brett J Bordini, Donald Basel. 2017. How doctors think: Common diagnostic errors in clinical judgment-lessons from an undiagnosed and rare disease program. In *Pediatric clinics of North America*.

S. L. Lauritzen and D. J. Spiegelhalter. 1988. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(2):157–224.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020.

Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. 2024. Best practices and lessons learned on synthetic data for language models. *arXiv preprint arXiv:2404.07503*.

Xiaohong Liu, Hao Liu, Guoxing Yang, Zeyu Jiang, Shuguang Cui, Zhaoze Zhang, Huan Wang, Liyuan Tao, Yongchang Sun, Zhu Song, Tianpei Hong, Jin Yang, Tianrun Gao, Jiangjiang Zhang, Xiaohu Li, Jing Zhang, Ye Sang, Zhao Yang, Kanmin Xue, and Guangyu Wang. 2025. A generalist medical language model for disease diagnosis assistance. *Nature Medicine*, pages 1–11.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Lars Mehnen, Stefanie Gruarin, Mina Vasileva, and Bernhard Knapp. 2023. Chatgpt as a medical doctor? a diagnostic accuracy study on common and rare diseases. *MedRxiv*, pages 2023–04.

Meta. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

R. A. Miller. 1994. Medical diagnostic decision support systems–past, present, and future: a threaded bibliography and brief commentary. *Journal of American Medical Informatics Association*.

R. A. Miller and F. E. Masarie. 1990. Quick Medical Reference (QMR): A Microcomputer-based diagnostic decision-support system for General Internal Medicine. *Proc Annu Symp Comput Appl Med Care*, pages 986–988.

R. A. Miller, H. E. Pople, and J. D. Myers. 1982. Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. *N. Engl. J. Med.*, 307(8):468–476.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2021. Fast model editing at scale. *arXiv preprint arXiv:2110.11309*.

Varun Nair, Elliot Schumacher, Geoffrey Tso, and Anitha Kannan. 2024. DERA: Enhancing large language model completions with dialog-enabled resolving agents. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 122–161, Mexico City, Mexico. Association for Computational Linguistics.

NORD. 2025. Nord rare disease database. `https://rarediseases.org/rare-diseases/`. Accessed: 2025-01-30.

US Government Accountability Office. 2021. Rare diseases although limited, available evidence suggests medical and other costs can be substantial.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

R. C. Parker and R. A. Miller. 1989. Creation of realistic appearing simulated patient cases using the INTERNIST-1/QMR knowledge base and interrelationship properties of manifestations. 28:346–51.

Murali Ravuri, Anitha Kannan, Geoffrey J. Tso, and Xavier Amatriain. 2018. Learning from the experts: From expert systems to machine-learned diagnosis models. In *Proceedings of the 3rd Machine Learning for Healthcare Conference*, volume 85 of *Proceedings of Machine Learning Research*, pages 227–243. PMLR.

Geoffrey W. Rutledge. 2024. Diagnostic accuracy of gpt-4 on common clinical scenarios and challenging cases. *Learning Health Systems*, 8.

Sarah Sandmann, Sarah Riepenhausen, Lucas Plagwitz, and Julian Varghese. 2024. Systematic analysis of chatgpt, google search and llama 2 for clinical decision support tasks. *Nature Communications*, 15(1):2050.

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.

Omkar Chakradhar Thawakar, Abdelrahman M. Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Khan. 2024. XrayGPT: Chest radiographs summarization using large medical vision-language models. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 440–448, Bangkok, Thailand. Association for Computational Linguistics.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, Shekoofeh

Azizi, Karan Singhal, Yong Cheng, Le Hou, Albert Webson, Kavita Kulkarni, S Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. 2024. Towards conversational diagnostic ai. *Preprint*, arXiv:2401.05654.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. *SuperGLUE: a stickier benchmark for general-purpose language understanding systems*. Curran Associates Inc., Red Hook, NY, USA.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Jian Yang, Liqi Shu, Huilong Duan, and Haomin Li. 2024a. Rdguru: a conversational intelligent agent for rare diseases. *IEEE Journal of Biomedical and Health Informatics*.

Jian Yang, Liqi Shu, Mingyu Han, Jiarong Pan, Lihua Chen, Tianming Yuan, Linhua Tan, Qiang Shu, Huilong Duan, and Haomin Li. 2024b. Rdmaster: A novel phenotype-oriented dialogue system supporting differential diagnosis of rare disease. *Computers in Biology and Medicine*, 169:107924.

Zihan Yao, Yu He, Tianyu Qi, and Ming Li. 2024. Scalable model editing via customized expert networks. *arXiv preprint arXiv:2404.02699*.

Da Yu, Arturs Backurs, Sivakanth Gopi, Huseyin Inan, Janardhan Kulkarni, Zinan Lin, Chulin Xie, Huishuai Zhang, and Wanrong Zhang. 2023. Training private and efficient language models with synthetic data from llms. In *Socially Responsible Language Modelling Research*.

Wenting Zhao, Tanya Goyal, Yu Ying Chiu, Liwei Jiang, Benjamin Newman, Abhilasha Ravichander, Khyathi Chandu, Ronan Le Bras, Claire Cardie, Yuntian Deng, et al. 2024. Wildhallucinations: Evaluating long-form factuality in llms with real-world entity queries. *arXiv preprint arXiv:2407.17468*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Shuang Zhou, Zidu Xu, Mian Zhang, Chunpu Xu, Yawen Guo, Zaifu Zhan, Sirui Ding, Jiashuo Wang, Kaishuai Xu, Yi Fang, Liqiao Xia, Jeremy Yeung, Daochen Zha, Genevieve B. Melton, Mingquan Lin, and Rui Zhang. 2024. Large language models for disease diagnosis: A scoping review. *Preprint*, arXiv:2409.00097.

A Appendix

A.1 Training Details

We train for 10 epochs using supervised fine-tuning only on 8 Nvidia A100s. Each training run, using the huggingface transformers library (v4.43.3), took 2-3 hours depending on the setup. We used the AdamW 8-bit optimizer (Loshchilov and Hutter, 2017), a learning rate of $2e^{-05}$, and a batch size of 4. We selected these parameters based on early experiments on subsets of our larger disease set.

A.2 Evaluation Details

Following the approach in Section 3, we generate chats using gpt-4o and claude-sonnet 3.5. For each, we create train, validation, and test splits as detailed in Appendix Table 5. We also provide the average and standard deviation of the number of findings and the number of messages. To ensure that both the validation and test sets are distinct from the training set, we removed all examples from the training set that have the exact same structured case in the validation or test sets. We generated a smaller dataset for claude due to the additional cost and time of running multiple prompts per chat.

A.2.1 Expert Evaluation

We provide the annotators with the chat, list of findings, and the seed disease. They are asked *"Assume the patient has a rare disease and more common conditions have been ruled out. Is this specific rare disease a possible diagnosis for the patient? This rare disease doesn't need to be the most likely diagnosis, but just a possible diagnosis."*. They are prompted to respond yes or no, and provide a 1-2 sentence explanation. They are allowed to use any external reference material they choose.

One challenge in this annotation task is that our dataset does not include negative examples as they are not required for training. However, only annotating positive examples may lead to annotators blindly answering "yes". Therefore, we generate negative training examples solely for expert evaluation and ask annotators to review datasets that are 90% positives and 10% negatives.

Generating a chat where a rare disease is not possible is a challenging task. Providing annotators with cases that are obviously negative would not be informative. Generating cases where a disease is negative but close requires conclusively ruling out a disease, which is challenging in a history taking setting where lab work isn't performed. To achieve a close approximation, we regenerate examples from our expert system and take diagnoses that were present in earlier simulation phases but removed later when additional findings were added. These "discarded diagnoses" are likely to be negative diagnoses for the findings.

We separately generate chats for these findings and discarded diagnosis using the process described in Section 3. As a second filtering step, we provide the chat and discarded diagnosis to a gpt-4o prompt, and prompt it to provide the same binary judgment and explanation as we do with the annotators. Finally, we manually filter out chats where the explanation rests mostly on likelihood (*i.e.* disease A is unlikely because disease B is more likely), and retain cases where there is a clear separation. However, we want to emphasize that even in these cases, it is hard to completely rule out a rare disease. All cases were reviewed by at least one annotator – cases where the original annotator disagreed with the initial label was reviewed by a second labeler.

A.3 Expert Evaluation Results

While the agreement rate for negative cases is lower than the agreement for positive cases, it is inherently challenging to create believably negative cases. It also does not impact RareScale's performance given that we only add negatives to ensure that annotators do not blindly choose yes. We include a selection of annotator reasoning in Appendix Table 6. In many cases where the annotator disagrees with our baseline label, they note that the full symptom set isn't present for the disease. While that may make it unlikely, that does not mean it can be ruled out entirely because an atypical presentation could occur or further symptoms may arise. This tension is inherent when dealing with rare diseases.

Dx Category	Top K Accuracy	Top 1 Accuracy	MRR	N
Congenital disorders due to abnormal fetal development	0.533	0.433	0.458	30
Immune system disorders	0.548	0.190	0.295	42
End organ damage secondary to other disorders	0.583	0.083	0.222	36
Disorders with excess or abnormal fluid accumulation	0.603	0.256	0.370	78
Disorders of the hematopoietic and lymphatic systems	0.667	0.000	0.306	6
Drug induced injury alias adverse drug effects	0.699	0.192	0.361	73
Degenerative disorders	0.704	0.241	0.391	108
Infectious disease alias infections	0.711	0.278	0.424	862
Neoplastic disease	0.721	0.236	0.384	330
Disorders involving cysts stones or calculi	0.722	0.611	0.650	18
Impaired cardiovascular function	0.738	0.359	0.486	390
Disorders due to toxic or chemical or radiation injury	0.745	0.398	0.523	98
Musculoskeletal disorders	0.750	0.292	0.467	24
Non-infectious inflammatory disease	0.753	0.326	0.477	384
Metabolic disorders	0.778	0.333	0.482	144
Bleeding disorders and coagulopathies	0.783	0.267	0.449	60
Disorders of thorax cardiovascular system and lymphatic ducts	0.783	0.522	0.590	23
Endocrine disease	0.788	0.417	0.547	132
Fibrosis or scarring of visceral organ	0.796	0.245	0.436	49
Neuropsychiatric disorders	0.803	0.439	0.573	66
Disorders secondary to trauma	0.806	0.361	0.519	36
Impaired fluid flow within hollow viscus or viscera non-vascular	0.833	0.583	0.649	24
Multisystem disorders	0.833	0.667	0.708	6
Disorders associated with pregnancy	0.833	0.542	0.663	24
Inherited congenital or degenerative disorders	0.847	0.458	0.588	144
Miscellaneous mechanical disorders	0.861	0.542	0.655	72
Kidney and urinary tract disorders	0.875	0.458	0.654	24
Disorders due to mechanical tear or trauma or visceral erosion	0.875	0.667	0.764	24
Disorders of abdomen digestive system and/or nutrition	0.889	0.389	0.573	18
Disorders due to nutritional and/or vitamin deficiency	0.917	0.611	0.738	36
Electrophysiological neurological disorders	0.917	0.542	0.687	24
Disorders of smooth muscle contraction and/or relaxation	0.944	0.722	0.801	18

Table 4: Performance from RareScale, gpt-4o in Table 1 broken down by disease category. Note that a disease may fall into multiple categories.

Findings List:
Abdomen pain left upper quadrant (absent)
Abdomen pain non-colicky (absent)
Anorexia (present)
Chest pain lateral sharp (present)
Chest pain substernal exacerbation with cough (present)
Cigarette smoking history (present)
Dyspnea abrupt onset (present)
Headache severe (present)
Myalgia (present)
Onset abrupt (absent)
Rigors (present)
Shoulder pain left (absent)
larger
DDx List:
pulmonary legionellosis **(seed disease)**
pneumococcal pneumonia
mycoplasma pneumonia
brucellosis

Patient-Doctor Chat:
System: Patient is Tyler, a 24-year-old male.
Doctor: Hi Tyler, I'm here to help you today. Can you tell me what's been going on?
Patient: Hey doc, I've been feeling really off lately.
Doctor: Can you describe what you mean by off? Are you experiencing any pain or discomfort?
Patient: Yeah, there's this sharp pain on the side of my chest, and it gets worse when I breathe.
[chest pain lateral sharp-present]
Doctor: That sounds uncomfortable. Have you noticed any other types of chest pain, maybe when you cough?
Patient: Not really sharp pain, but it does hurt under my sternum when I cough.
[chest pain substernal exacerbation with cough-present]

Figure 3: Example expert system simulation to chat

generation model	split	size	findings	messages
gpt-4o	train	21782	11.88 ± 1.91	17.54 ± 3.58
	val	3404	11.69 ± 2.08	17.31 ± 3.78
	test	3403	11.66 ± 2.05	17.27 ± 3.69
claude	train	8837	11.88 ± 1.90	14.54 ± 3.47
	val	2868	11.69 ± 2.04	14.11 ± 3.38
	test	2868	11.70 ± 2.08	14.30 ± 3.48

Table 5: Data statistics for train, validation, and test sets. We include the number of chats, the average number of findings and standard deviation, and the average number of messages and standard deviation.

Synthetic	Annotator	Disease	Reasoning
yes	yes	histoplasma meningitis	This disease is possible because the patient's history of lymphoma and transplant makes them immunocompromised, which increases their risk of histoplasmosis meningitis. The histoplasmosis meningitis could cause this patient's symptoms of nausea/vomiting, fever/chills, and severe headache.
yes	yes	cerebral malaria	This disease is possible because urinary and bowel incontinence, intractable headache, visual loss/retinopathy, and fever point to an infectious cerebral process that could be caused by cerebral malaria.
yes	yes	neurogenic osteoarthropathy alias charcot_joint_disease	This could be possible charcot joint disease, which is set off by trauma to a neuropathic extremity and can cause joint pain. Although Charcot's is technically most common in the foot, it can also extend to any major joint like the knees, shoulder, hip, etc.
yes	yes	glaucoma acute angle closure	Blurriness with rainbow rings/halos, nausea/vomiting, headache, decreased vision, and sudden onset are all symptoms of acute angle-closure glaucoma.
yes	yes	cytomegalovirus infection disseminated	This disease is possible because the patient is immunocompromised by their organ transplant, and their symptoms of vision changes, diarrhea, fever, chills, vomiting, and myalgias are consistent with a disseminated cytomegalovirus infection.
yes	no	cutaneous anthrax	Lack of bump/ulcer/eschar. though presence of pruritis and exposure to possible animals infected as a vet, symptoms are nonspecific.
yes	no	herpes zoster	This disease is not possible because although herpes zoster can have peripheral symptoms including this patient's myalgia, headache, and abdominal pain, this patient does not have the characteristic dermatomal rash, skin changes, or skin-level pain of herpes zoster.
no	no	pulmonary aspergillosis invasive type	This disease is not possible because the patient is not immunocompromised, has not had recent surgeries or pneumonia, or chemotherapy, and does not have a cough, which I would expect from pulmonary aspergillosis since it is an opportunistic pulmonary infection.
no	no	henoch schonlein syndrome alias henoch-schonlein purpura	This disease is not possible because the patient does not report the purpura around the legs/gluteus that is characteristic for Henoch Schonlein purpura. The patient is also not in the typical age group for this disease, which primarily is in pediatric populations.
no	yes	bronchial asthma	This disease is possible because of the patient's history of dyspnea at rest and worse with activity. The urinary frequency could also be related because there are positive associations between bronchial asthma and increased urge to urinate.

Table 6: Sample explanations from the annotation task. Note that the disease column indicates which disease they were asked to annotate against – this is the expected disease for the positive synthetic labels, but is not expected for the negative ones.

	gpt-4o chats			**claude chats**		
	Top K	**Top-1**	**MRR**	**Top K**	**Top-1**	**MRR**
gpt-4o only	88.28%	63.13%	0.7384	47.25%	31.80%	0.3817
claude only	48.38%	34.43%	0.4025	65.93%	45.71%	0.5415
combined	88.57%	64.10%	0.7444	79.64%	57.46%	0.6675

Table 7: Candidate generation metrics for validation set.

gpt-4o Generated Chat for disseminated intravascular coagulation

System: Patient is George, a 67-year-old male.

[age greater than 55 (present) , sex male (present)]

Doctor: Hi George, I'm here to help you today. Can you tell me what's been going on?

Patient: Well, it all started pretty suddenly. I've been feeling really off.

[onset abrupt (present)]

Doctor: I see. Can you describe some of the symptoms you've been experiencing?

Patient: I've had a fever on and off. It's been bothering me.

[fever (present)]

Doctor: Alright, fever can be quite uncomfortable. Have you noticed any changes in your appetite?

Patient: I haven't really felt like eating much lately. Just no appetite.

[anorexia (present)]

Doctor: Got it. Have you experienced any issues with bleeding or bruising?

Patient: Yeah, even a small bump seems to make me bleed a lot. It's strange.

[bleeding excessive after minor trauma (present)]

Doctor: That does sound unusual. Have you had any nosebleeds recently?

Patient: No, I haven't had any nosebleeds.

[epistaxis or epistaxis severe history (absent)]

Doctor: Okay, and how about your breathing? Any issues there?

Patient: Yes, I've been feeling short of breath even when I'm just sitting around.

[dyspnea at rest (present)]

Doctor: That's important to note. Have you had any recent surgeries?

Patient: Yes, I had surgery a few weeks ago.

[surgery recent history (present)]

Doctor: Thanks for sharing that. Have you experienced any vomiting?

Patient: Yes, I've been throwing up, and sometimes it looks really dark, like coffee grounds.

[vomiting coffee ground (present) , vomiting recent (present)]

Doctor: I see. Have you ever had a reaction to a blood transfusion?

Patient: Yeah, I've had a reaction before.

[transfusion reaction history (present)]

Doctor: Alright. Have you experienced any seizures?

Patient: Yes, I've had some seizures.

[seizures grand mal single or intermittent (present)]

Doctor: And how about any back pain?

Patient: No, my back's been fine.

[back pain midline severe (absent)]

Figure 4: Example gpt-4o generated chat. The simulated findings are in italics. Note that we do not provide the findings as input to train candidate generation model

Claude Generated Chat for porphyria acute intermittent

System: Sarah is a 35-year-old female in Seattle, Washington.
[sex female (present) , age 26 to 55 (present)]

Doctor: Hello Sarah, I'm Dr. Smith. What brings you to see me today?

Patient: Hey doctor. I've been dealing with these terrible stomach pains lately. They come and go in waves, and they're really intense, especially in my upper right side. It started about two days ago.
[abdomen pain present (present) , abdomen pain acute (present) , abdomen pain colicky (present)]

Doctor: I'd like to know more about these pains. Have you had any pain in other parts of your stomach, like around your belly button or on the left side?

Patient: No, not really. The pain is mostly just in my upper right side. It doesn't really go to my belly button area or the left side at all.
[abdomen pain periumbilical (absent) , abdomen pain left upper quadrant (absent)]

Doctor: How long have you been experiencing these stomach pains? Have you had similar episodes in the past months?

Patient: No, this is really the first time it's happened. Like I said, it just started a couple days ago. I haven't had any ongoing stomach issues before this.
[abdomen pain chronic (absent)]

Doctor: Have you noticed any issues with eating? Any trouble swallowing certain foods?

Patient: Actually, yeah. I've been having a hard time swallowing, especially with meat and bread. They seem to get stuck, and I have to really work to get them down. It's making eating pretty difficult.
[dysphagia solids (present)]

Doctor: Have you noticed any vision changes recently? Anything unusual when you're looking at objects?

Patient: Yeah, actually it's weird - I keep seeing double. Like when I look at my computer screen at work, I see two screens. It's really making it hard to code.
[eyes diplopia (present)]

Doctor: Have you experienced any muscle cramps recently, like muscles tightening up and staying tight for a minute or longer?

Patient: No, I haven't had any muscle cramps like that. The stomach pain is bad enough, but my muscles feel normal.
[muscle cramps (absent)]

Doctor: Could you tell me about your family background? Where are your ancestors from?

Patient: Oh yeah, my family is from Sweden and Norway. Both my parents' families came from there a few generations back.
[ethnic background northern european (present)]

Doctor: Have you noticed any connection between your symptoms and your menstrual cycle?

Patient: Now that you mention it, yes. I started my period right when these symptoms began. It seems like the pain and other symptoms got worse when my period started.
[menses precipitation or exacerbation of disease history (present)]

Figure 5: Example claude generated chat. The simulated findings are in italics.

```
Create a history taking conversation where a patient has the following medical profile, and is chatting with a doctor.
Include only the history taking part of the chat, do not include any diagnosis or include the doctor confirming the findings.
Be sure to use language appropriate to the patient's profile, and do not make the patient text technical unless indicated.  Use informal terms when appropriate
Ensure all findings are explicitly discussed, both present and absent statuses.
At the end of the chat, all findings must be included in a previous message.

The first message should consist of a generic system message which  explicitly states the patient's name, age, gender (stated directly), race if present.  Do not include any findings beyond that list.
The second message should consist of a generic greeting from the doctor.  The doctor does not have any information at this point.
Do not include more than two findings in a question answer pair.
The patient is also unlikely to be able to explain their condition in a coherent manner.  Do not assume that they will clearly offer up the findings, and the doctor may have to ask multiple questions before they get to a finding.
The doctor should not ask leading questions.

Disease: {{disease}}

--Patient Profile--
{{demographics_str}}
--End Profile--

--Findings--
{{findings_str}}
--End Findings--

Output format in yaml
--Format--
chat:
 - role : "System"
   msg : [MSG ABOUT GENDER, AGE, RACE]
   findings :
    - [FINDING 0 , STATUS]
    - [FINDING 1 , STATUS]
    ...
 - role : "Doctor"
   msg : [MSG]
   findings : null
 - role : "Patient"
   msg : [MSG]
   findings :
    - [FINDING 0 , STATUS]
    - [FINDING 1 , STATUS]
    ...
...
--End Format--

Chat:
```

Prompt 1: Prompt for generating chats with gpt-4o.

```
Create a history taking conversation where a patient has the following medical profile, and is chatting with a doctor.
Include only the history taking part of the chat, do not include any diagnosis or include the doctor confirming the findings.
Be sure to use language appropriate to the patient's profile, and do not make the patient text technical unless indicated.  Use informal terms when appropriate.
Ensure all findings in the findings list are explicitly discussed, both present and absent statuses.

You will generate one turn of messages at a time.  A turn of messages is defined as a Doctor asking a question and a patient response.
In the first turn of messages, there should be the following;
The first message of the first turn should consist of a generic system message which  explicitly states the patient's name, age, gender (stated directly), race if present.  Do not include any findings beyond that list.
The second message of the first turn should consist of a generic greeting from the doctor.  The doctor does not have any information at this point.
The third message of the first turn should consist of the first findings they report.

All previous turns should consist of a provider question and a patient response.

At the end of the chat, all findings must be included in a previous message.

Do not include more than two findings in a question answer pair.
The patient is also unlikely to be able to explain their condition in a coherent manner.  Do not assume that they will clearly offer up the findings, and the doctor may have to ask multiple questions before they get to a finding.
The doctor should not ask leading questions.
Do not assume that the doctor can see the patient. Express everything only through text.

When available, a definition and a previous message from a different patient. for each finding is provided.  Follow the definition of the finding provided.
You must create a different message - including the content - than the previous message if possible given the definition.
---Example---
tongue protrusion with marked deviation:
  definition: upon protrusion of the tongue, there is significant digression away from the midline, toward either the left or right side
  previous message: Yes, actually. I'd say its towards the right side of my tongue.
  status: present
The next message should be discuss one of the other options in the definition (e.g., left side).  For example, "Hmm, maybe on the left side of my tongue."
---End Example---

Disease: {{disease}}

--Patient Profile--
{{demographics_str}}
--End Profile--

Output format in yaml
--Format--
msg_turn:
[CONDENSED]
--End Format--
```

Prompt 2: Prompt for generating chats with claude.

```
The following is a history taking conversation between a patient and a doctor.

Do the following steps;
First, generate 5 differential diagnoses that are likely given the history taking chat. For each, briefly discuss the reasoning. Ignore the candidate diagnoses for this step. Output the result in a section beginning with "--First pass--" and ending with "--/First pass--"

{% if candidate_diagnoses -%}

Second, consider the following candidate diagnoses. You are not required to use them, and should discard them if not appropriate. For each, briefly discuss the reasoning. Output the result in a section beginning with "--Candidate pass--" and ending with "--/Candidate pass--"

--Candidate Diagnoses--
{{candidate_diagnoses}}
--End Candidate Diagnoses--

Finally, consider the diagnoses from 1) and 2), and pick the 5 most appropriate ones. You must discard any that aren't appropriate, either from step 1 or 2.

{% else %}
Finally, consider the diagnoses from 1) and pick the 5 most appropriate ones. You must discard any that aren't appropriate.

{% endif -%}

Output in yaml described at the end, where [DDx #] is replaced with the disease name.
Include a likelihood rating of high, medium, or low in [LIKELIHOOD]. Include a short description of your reasoning in [REASON #].


--Chat--
{{chat_formatted}}
--End Chat--

--Begin Output--
differential_diagnosis:
  [DDx 0]:
      likelihood: [LIKELIHOOD]
      reasoning:
        - "[REASON 0]"
        - "[REASON 1]"
        ...
  [DDx 1]:
      likelihood: [LIKELIHOOD]
      reasoning:
        - "[REASON 0]"
        - "[REASON 1]"
        ...
--End--
```

Prompt 3: Prompt for generating differential diagnosis. Note that the text between the if statements is only included if candidates are provided.

```
Is our predicted diagnosis correct (Y/N)? It is okay if the predicted diagnosis is more specific/detailed.
Predicted diagnosis: {{prediction}}, True diagnosis: {{gold_label}}
Answer [Y/N]:
```

Prompt 4: Prompt for judging binary differences between differential diagnoses (Tu et al., 2024).

```
How close did the differential diagnosis (DDx) come to including the DIAGNOSIS
    from the answer key?


DDX: {{ddx}}
DIAGNOSIS: {{final_diagnosis}}

CHOICES:
-  (Unrelated): Nothing in the DDx is related to the diagnosis.
- (Somewhat Related): DDx contains something that is related, but unlikely to
    be helpful in determining the  diagnosis.
- (Relevant): DDx contains something that is closely related and might have
    been helpful in determining the  diagnosis.
- (Extremely Relevant): DDx contains something that is very close, but not an
    exact match to the  diagnosis.
- (Exact Match): DDx includes the  diagnosis.

Choice: Best from the CHOICES
Chosen Dx: diagnosis in DDX which was matched to DIAGNOSIS
location of Dx: location of the chosen Dx in DDX
Rationale: rationale for the choice
```

Prompt 5: Prompt for judging relative differences between differential diagnoses (Tu et al., 2024).