MIMIC-RD: Can LLMs differentially diagnose rare diseases in real-world clinical settings?

Zilal Eiz AlDin zelalae2@illinois.edu
Department of Computer Science, University of Illinois Urbana-Champaign, Champaign, IL, USA

John Wu
Department of Computer Science, University of Illinois Urbana-Champaign, Champaign, IL, USA

Jeffrey Paul Fung
University of Illinois College of Medicine, Peoria, IL, USA

Jennifer King
University of Illinois College of Medicine, Peoria, IL, USA

Mya Watts
University of Illinois College of Medicine, Peoria, IL, USA

Lauren O'Neill
University of Illinois College of Medicine, Peoria, IL, USA

Adam Richard Cross
University of Illinois College of Medicine, Peoria, IL, USA

Jimeng Sun jimeng@illinois.edu
Department of Computer Science, University of Illinois Urbana-Champaign, Champaign, IL, USA

Abstract

Despite rare diseases affecting 1 in 10 Americans, their differential diagnosis remains challenging. Due to their impressive recall abilities, large language models (LLMs) have been recently explored for differential diagnosis. Existing approaches to evaluating LLM-based rare disease diagnosis suffer from two critical limitations: they rely on idealized clinical case studies that fail to capture real-world clinical complexity, or they use ICD codes as disease labels, which significantly undercounts rare diseases since many lack direct mappings to comprehensive rare disease databases like Orphanet. To address these limitations, we explore MIMIC-RD, a rare disease differential diagnosis benchmark constructed by directly mapping clinical text entities to Orphanet. Our methodology involved an initial LLM-based mining process followed by validation from four medical annotators to confirm identified entities were genuine rare diseases. We evaluated various models on our dataset of 145 patients and found that current state-of-the-art LLMs perform poorly on rare disease differential diagnosis, highlighting the substantial gap between existing capabilities and clinical needs. From our findings, we outline several future steps towards improving differential diagnosis of rare diseases.

Keywords: Rare Disease, Agents, Data Mining

Data and Code Availability We use MIMIC-IV (Johnson et al., 2023b,a) for our rare disease benchmark construction. We offer our initial mining scripts through `https://github.com/zelal-Eizaldeen/rare_disease_pyHealth/tree/main` and benchmark data `https://github.com/jhnwu3/RDMA/blob/main/public_data/initial_diff_diagnosis_benchmark.json`.

Institutional Review Board (IRB) We received IRB approval from the relevant institutions to allow medical students to annotate whether a text entity was a rare disease or not.

1. Introduction

Rare diseases affect approximately 1 in 10 Americans, constituting a significant healthcare challenge despite their individual rarity (Virginia Tech, 2025). Accu-

© 2025 Z.E. AlDin, J. Wu, J.P. Fung, J. King, M. Watts, L. O'Neill, A.R. Cross & J. Sun.

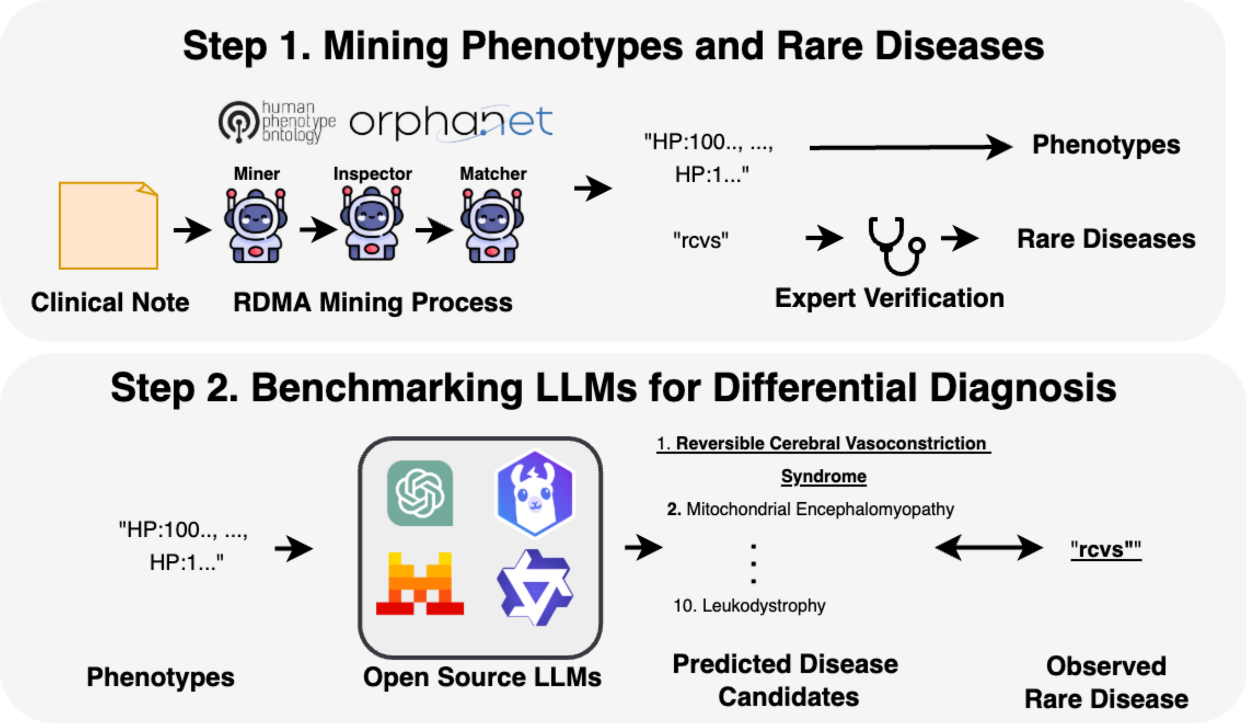

Figure 1: We directly mine rare disease mentions from clinical notes along with any phenotypes using RDMA (Wu et al., 2025) and we verify its mined outputs directly across 4 medical students, subsampling rare disease mentions with more than 3 annotators in agreement. We use this mining to benchmark a variety of LLM models for differential diagnosis. of rare diseases

rate diagnosis remains difficult due to the vast diversity and sparsity of these conditions (Auvin et al., 2018). Large language models (LLMs) are being actively explored for differential diagnosis due to their wide-recall abilities (McDuff et al., 2025a), leading to a plethora of work on LLM-based rare disease differential diagnosis. These approaches range from exploring large frontier models (Ao et al., 2025; Chen et al., 2024) to smaller agentic systems (Chen et al., 2025).

However, existing approaches suffer from several drawbacks. Studies by Ao et al. (2025) and Chen et al. (2024) focus purely on cleanly-curated clinical case studies with patient profiles that can differ vastly from typical clinical settings, such as those found in clinical notes from MIMIC-IV (Johnson et al., 2023b,a). Meanwhile, Chen et al. (2025) use ICD codes as a proxy for rare disease classification. While efforts have been made to map ICD codes to more granular rare disease Orphanet codes (Cavero-Carbonell et al., 2020), over 50% of Orphanet codes lack direct mappings, resulting in under-reporting of rare diseases within ICD-annotated systems.

To ensure better coverage, it is crucial to directly mine rare diseases from clinical notes and map them to specialized rare disease ontologies like Orphanet (Weinreich et al., 2008; Mazzucato et al., 2023). Fortunately, recent advancements in phenotype extraction with LLMs (Garcia et al., 2024) and rare disease mention extraction, such as RDMA (Wu et al., 2025), have achieved high precision results. Leveraging RDMA, we mine phenotypes mapped to the Human Phenotype Ontology (HPO) and rare diseases mapped to the Orphanet ontology. We validate the rare disease mentions (explicitly defined by Orphanet) with four medical students to ensure only true positive cases exist in our benchmark.

Our findings demonstrate that:

- LLMs severely underperform in rare disease differential diagnosis despite higher reported performance in prior work (Chen et al., 2025)
- Key challenges exist in the relationship between patient phenotype presentation and the ability to correctly predict rare diseases

We hope this work serves as an entry point for better discussions surrounding the use of LLMs for differential diagnosis as a key component of future healthcare research.

2. Methodology

Mining rare diseases and phenotypes. We use RDMA (Wu et al., 2025) to extract rare disease mentions from clinical notes. The approach follows a three-step process: LLMs first extract potential rare disease entities, then verify whether they represent actual rare diseases, and finally map them to their respective ontologies. To balance reliability and diversity in our entity capture, we run this pipeline twice—once at low temperature ($T = 0.01$) for consistent results and once at high temperature ($T = 0.7$) for broader coverage.

Next, four medical students manually verify each extracted rare disease mention within its clinical context, determining whether it truly represents a rare disease according to Orphanet criteria (Weinreich et al., 2008). Annotators make binary ”yes or no” decisions on whether each mined entity constitutes a genuine rare disease observation for the patient. We then create a high-agreement subset for final evaluation, requiring consensus from at least three annotators.

For patients with verified rare disease mentions, we extract phenotypes using the same three-step RDMA pipeline, mapping them directly to the Human Phenotype Ontology (Gargano et al., 2024). As a final preprocessing step, we address cases where entities map to both Orphanet and the Human Phenotype Ontology, making them both rare diseases and phenotypes. We remove these overlapping phenotypes (less than 2.5% of cases) from each patient's phenotype set to avoid redundancy.

Differential diagnosis setup. Effectively, given a patient's list of observed phenotypes, LLMs are asked to predict a list of the 10 most likely rare diseases in order of likelihood. To benchmark their performance, we check if an observed rare disease falls within their top k predictions, hence the Hit@k metric ($k = 1, 5, 10$).

3. Results

Baselines. We explore 5 readily available open source models. OpenBio-LLM 70B (Ankit Pal, 2024), Qwen 3 32B (Yang et al., 2025), Llama 3.3 70B (Grattafiori et al., 2024), Mistral 3.1 24B (AI, 2025), and Mixtral 70B (Jiang et al., 2024).

Real-world clinical setups are substantially more complex. We sampled 1,000 patients from MIMIC-IV (Johnson et al., 2023b) and mined their clinical notes for rare disease mentions. Using RDMA, we identified 223 patients who contained rare disease references. Our annotators then reviewed these cases and confirmed rare diseases in 145 patients with high inter-rater agreement. From the discharge summaries of these 145 confirmed cases, we extracted associated phenotypes. Our benchmark contains an average of approximately 128 phenotypes per patient—substantially more than existing public benchmarks including RAMEDIS (Töpel et al., 2010), MME (Philippakis et al., 2015), HMS (Ronicke et al., 2019), and Lirical (Robinson et al., 2020), as shown in Figure 2. In contrast with our mined phenotypes, our dataset contains far fewer observed rare diseases, averaging approximately 1 per patient.

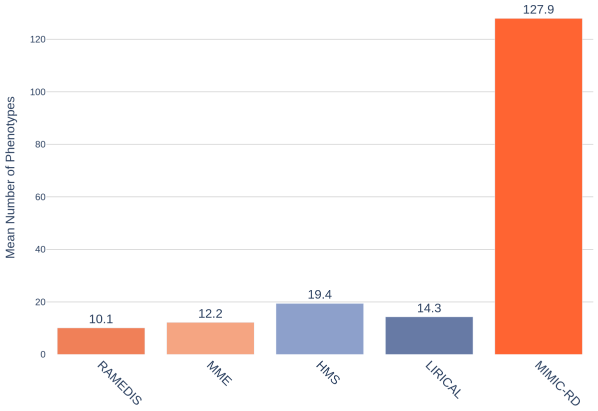

Figure 2: Mined from clinical notes with thousands of words, MIMIC-RD offers substantially greater numbers of phenotypes per patient for each rare disease. As a key implication, this dramatic increase in observations makes differential diagnosis a more complicated task as LLMs have to consider upwards of 10x more phenotypes in its differential diagnosis.

Open-source LLMs still highly underperform in differential diagnosis rare disease tasks. Even when phenotypes and rare diseases are extracted from the same clinical note, state-of-the-art open-source LLMs fail to predict approximately 60% of observed rare diseases given their relevant phenotypes. Our rare disease hit-rates in Tables 2 and 3 are substantially lower than those reported in

(Chen et al., 2025), showing approximately 35% lower Hit@10 scores.

Table 1: MIMIC-RD Statistics: We observe high inter-annotator agreement, a massive phenotype to disease ratio in the MIMIC hospital setting.

Statistic	**Value**
# of Patients	145
# of Unique Diseases	120
# of Diseases	192
Avg. # of Phenotypes Per Patient	127.93
Avg. # of Rare Diseases Per Patient	1.32
Mean Pairwise Cohen Kappa	**0.71**
Fleiss' Kappa	**0.71**

Model	Hit@1	Hit@5	Hit@10
Llama 3.3 70B	**20.31%**	**35.94%**	**40.10%**
Mistral 24B	13.02%	27.60%	33.33%
OpenBioLLM 70B	9.47%	17.89%	26.32%
Mixtral 70B	6.84%	15.79%	22.63%
Qwen 32B	18.23 %	30.20 %	37.50 %

Table 2: Differential-diagnosis recall performance (Hit@k) across all 192 observed diseases. Llama 3.3 (Grattafiori et al., 2024) emerges as a remarkably strong performer in rare disease differential diagnosis, substantially outperforming its biomedically fine-tuned Llama 3 counterpart (Ankit Pal, 2024). This suggests that biomedical fine-tuning does not necessarily improve performance on related tasks such as rare disease differential diagnosis.

Patient profiles typically have at least one overlapping phenotype with a rare disease profile. Regardless of whether an LLM correctly proposes a rare disease within its top 10 candidates, patient phenotype profiles typically contain at least one phenotype that appears in the corresponding rare disease's profile on Orphanet (Weinreich et al., 2008). This suggests that the majority of failed rare disease predictions are not due to missing relevant phenotype information, but rather the models' inability to appropriately rank the correct rare disease above other conditions in their top 10 predictions.

LLM diagnosis performance appears to be aligned with documented phenotype presentations. When comparing patient phenotype profiles to documented clinical presentations, we observe distinct patterns in Figure 3: patients with correctly predicted rare diseases show substantial phenotype overlap with their disease's documented profile. In contrast, LLMs typically fail to identify rare diseases whose phenotype profiles are more ambiguous or contain phenotypes that co-occur less frequently with the observed conditions.

Model	Hit@1	Hit@5	Hit@10
Llama 3.3 70B	**19.31%**	**30.34%**	**35.17%**
Mistral 24B	13.79%	24.83%	31.03%
OpenBioLLM 70B	11.03%	17.93%	25.52%
Mixtral 70B	7.59%	15.17%	20.69%
Qwen 32B	18.62%	27.59%	34.48%

Table 3: Differential-diagnosis recall performance (Hit@k) across 145 patients. Llama 3.3 (Grattafiori et al., 2024) again emerges as a remarkably strong performer in rare disease differential diagnosis, substantially outperforming its biomedically fine-tuned Llama 3 counterpart (Ankit Pal, 2024). At the patient level, the relative performance ranking across models remains consistent.

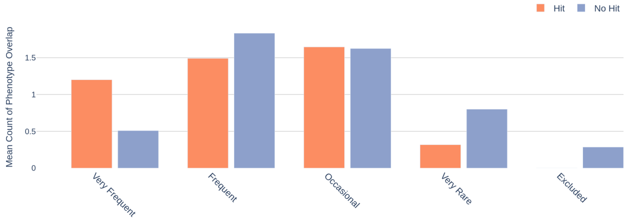

Figure 3: Comparison of phenotype overlap between "Hit@10" and "No Hit" patients with documented clinical phenotype presentations of rare diseases. Here, we plot the average number of phenotype overlaps that a patient's phenotype profile has with well-documented phenotypes for each rare disease. Such documented phenotypes have classified incidence rates to indicate their frequency of co-occurrence with a rare disease as defined by Orphanet (Weinreich et al., 2008) Specifically, these frequencies are rated as "very frequent" with 99-80% co-occurrence, frequent with "79-30%", "occasional" with "29-5%", very rare.

4. Discussion

Mining a substantially larger phenotype-rare disease patient dataset. LLMs enable the extraction of substantially more phenotype and rare disease information from clinical text across electronic healthcare systems. This approach can create more comprehensive benchmarks and, more importantly, improve our understanding of co-occurrence relationships between observed rare diseases and phenotypes. Additionally, building deep predictive models for differential diagnosis could help uncover nonlinear relationships between specific phenotype types and rare diseases.

Narrowing the number of disease candidates through multimodality. Our rare disease phenotype characterization mapping from Orphanet (Weinreich et al., 2008) reveals a wide range of rare diseases associated with each phenotype, from 1 to over 1,026 diseases. Given this variability, incorporating additional modalities is crucial for effective differential diagnosis to narrow down potential rare diseases. Lab tests and imaging such as X-rays can significantly reduce the number of candidate diseases. Fortunately, MIMIC-IV (Johnson et al., 2023b) contains a wealth of different modalities within its datasets. Exploring these modalities could substantially improve diagnostic performance and assist clinicians who may lack the time to review complete patient histories.

Differential diagnosis agents. While initial attempts have been made to construct rare disease diagnostic agents (Chen et al., 2025), significant opportunities remain for evaluating automated differential diagnosis frameworks that can support physicians in providing more comprehensive patient care (McDuff et al., 2025b).

References

Mistral AI. Mistral small 3.1, 2025. URL https://mistral.ai/news/mistral-small-3-1.

Malaikannan Sankarasubbu Ankit Pal. Openbiollms: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B, 2024.

Guangyu Ao, Min Chen, Jing Li, Huibing Nie, Lei Zhang, and Zejun Chen. Comparative analysis of large language models on rare disease identification. *Orphanet Journal of Rare Diseases*, 20(1): 150, 2025.

Stéphane Auvin, John Irwin, Paul Abi-Aad, and Alysia Battersby. The problem of rarity: estimation of prevalence in rare disease. *Value in Health*, 21(5):501–507, 2018.

C Cavero-Carbonell, J Rico, LJ Echevarría-González de Garibay, M García-López, S Guardiola-Vilarroig, LA Maceda-Roldán, and O Zurriaga. From icd10 to orphacodes: paving the way towards improved identification systems for rare diseases. *European Journal of Public Health*, 30(Supplement_5):ckaa166–494, 2020.

Xuanzhong Chen, Xiaohao Mao, Qihan Guo, Lun Wang, Shuyang Zhang, and Ting Chen. Rarebench: Can llms serve as rare diseases specialists?, 2024. URL https://arxiv.org/abs/2402.06341.

Xuanzhong Chen, Ye Jin, Xiaohao Mao, Lun Wang, Shuyang Zhang, and Ting Chen. Rareagents: Advancing rare disease care through llm-empowered multi-disciplinary team, 2025. URL https://arxiv.org/abs/2412.12475.

Brandon T Garcia, Lauren Westerfield, Priya Yelemali, Nikhita Gogate, E Andres Rivera-Munoz, Haowei Du, Moez Dawood, Angad Jolly, James R Lupski, and Jennifer E Posey. Improving automated deep phenotyping through large language models using retrieval augmented generation. *medRxiv*, pages 2024–12, 2024.

Michael A Gargano, Nicolas Matentzoglu, Ben Coleman, Eunice B Addo-Lartey, Anna V Anagnostopoulos, Joel Anderton, Paul Avillach, Anita M Bagley, Eduard Bakštein, James P Balhoff, et al. ¡? mode longauthoraffil?¿ the human phenotype ontology in 2024: phenotypes around the world. *Nucleic acids research*, 52(D1):D1333–D1346, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Alistair Johnson et al. MIMIC-IV-Note: Deidentified free-text clinical notes, 2023a. URL https://doi.org/10.13026/1n74-ne17.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1): 1, 2023b.

Monica Mazzucato, Laura Visonà Dalla Pozza, Paola Facchin, Cèline Angin, Francis Agius, Clara Cavero-Carbonell, Virginia Corrochano, Katerina Hanusova, Kurt Kirch, Deborah Lambert, et al. Orphacodes use for the coding of rare diseases: comparison of the accuracy and cross country comparability. *Orphanet Journal of Rare Diseases*, 18 (1):267, 2023.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, et al. Towards accurate differential diagnosis with large language models. *Nature*, pages 1–7, 2025a.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, et al. Towards accurate differential diagnosis with large language models. *Nature*, pages 1–7, 2025b.

Anthony A Philippakis, Danielle R Azzariti, Sergi Beltran, Anthony J Brookes, Catherine A Brownstein, Michael Brudno, Han G Brunner, Orion J

Buske, Knox Carey, Cassie Doll, et al. The matchmaker exchange: a platform for rare disease gene discovery. *Human mutation*, 36(10):915–921, 2015.

Peter N Robinson, Vida Ravanmehr, Julius OB Jacobsen, Daniel Danis, Xingmin Aaron Zhang, Leigh C Carmody, Michael A Gargano, Courtney L Thaxton, Guy Karlebach, Justin Reese, et al. Interpretable clinical genomics with a likelihood ratio paradigm. *The American Journal of Human Genetics*, 107(3):403–417, 2020.

Simon Ronicke, Martin C Hirsch, Ewelina Türk, Katharina Larionov, Daphne Tientcheu, and Annette D Wagner. Can a decision support system accelerate rare disease diagnosis? evaluating the potential impact of ada dx in a retrospective study. *Orphanet journal of rare diseases*, 14(1):69, 2019.

Thoralf Töpel, Dagmar Scheible, Friedrich Trefz, and Ralf Hofestädt. Ramedis: a comprehensive information system for variations and corresponding phenotypes of rare metabolic diseases. *Human mutation*, 31(1):E1081–E1088, 2010.

Virginia Tech. One in 10 Americans is living with a rare disease. Virginia Tech News, 02 2025. URL news.vt.edu/articles/2025/02/research_fralinbiomed_rarediseaseday2025_0228.html. Accessed: 2025-04-02.

Steffanie S Weinreich, R Mangon, JJ Sikkens, ME En Teeuw, and MC Cornel. Orphanet: a european database for rare diseases. *Nederlands tijdschrift voor geneeskunde*, 152(9):518–519, 2008.

John Wu, Adam Cross, and Jimeng Sun. Rdma: Cost effective agent-driven rare disease discovery within electronic health record systems, 2025. URL https://arxiv.org/abs/2507.15867.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.