Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Decoding Rarity: Large Language Models in the Diagnosis of Rare Diseases

Valentina Carbonari
Magna Graecia University
Pierangelo Veltri
University of Calabria
Pietro Hiram Guzzi
hguzzi@unicz.it

Magna Graecia University

Research Article

Keywords: rare diseases, large language models, diagnosis

Posted Date: April 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9356553/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Decoding Rarity: Large Language Models in the Diagnosis of Rare Diseases

Valentina Carbonari[1], Pierangelo Veltri[2], and Pietro Hiram Guzzi*[1]

[1]University of Catanzaro, valentina.carbonari@unicz.it, hguzzi@unicz.it
[2]University of Calabria, pierangelo.veltri@dimes.unical.it

April 8, 2026

Abstract

Large language models (LLMs) have shown promising capabilities in transforming rare disease research. This survey paper explores the integration of LLMs in analyzing rare diseases, highlighting significant strides and pivotal studies that leverage textual data to uncover insights and patterns critical for diagnosis, treatment, and patient care. While current research predominantly employs textual data, the potential for multimodal data integration combining genetic, imaging, and electronic health records is a promising frontier. We review foundational papers demonstrating the application of LLMs in identifying and extracting relevant medical information, simulating intelligent conversational agents for patient interaction, and formulating accurate and timely diagnoses. Furthermore, this paper discusses the challenges and ethical considerations inherent in deploying LLMs, including data privacy, model transparency, and the need for robust, inclusive data sets. As part of this exploration, we present a section on experimentation that utilizes multiple LLMs alongside structured questionnaires specifically designed for diagnostic purposes in the context of different diseases. We conclude with future perspectives on the evolution of LLMs towards truly multimodal platforms, which would integrate diverse data types to provide a more comprehensive understanding of rare diseases, ultimately fostering better outcomes in clinical settings.

Keywords: rare diseases; large language models; diagnosis.

Key Points

- Rare diseases affect a significant global population (3.5–5.9%), despite their low individual prevalence.

*Corresponding author: Pietro Hiram Guzzihguzzi@unicz.it

- Large Language Models (LLMs) show potential to accelerate diagnosis and knowledge extraction for rare diseases from unstructured biomedical data.
- Some of the reviewed studies apply LLMs to structured clinical questionnaires.
- Most existing applications remain unimodal, but integrating multimodality could provide higher accuracy and superior diagnostic results.
- Major challenges include hallucination risk, lack of domain-specific fine-tuning, data scarcity, and limited explainability.

1 Introduction

The European Union officially defines rare diseases as conditions affecting fewer than one in 2,000 individuals[1] and by the United States as those impacting fewer than 200,000 people nationwide at any given time[2]. Yet, rare diseases are collectively anything but rare: over 7,000 distinct disorders have been identified, affecting an estimated 3.5–5.9% of the global population, or more than 400 million people worldwide[3].

Despite this scale, patients often endure prolonged, complex diagnostic journeys. The low prevalence and clinical heterogeneity of these conditions frequently result in misdiagnoses or delayed recognition, as symptoms can mimic more common diseases or manifest atypically[4]. While diagnostic technologies and awareness have advanced, therapeutic development lags far behind. Today, only a small fraction of rare diseases have approved, effective treatments, leaving most patients without viable medical options[5].

This landscape presents profound challenges for clinicians, researchers, and healthcare systems. From a computational perspective, key hurdles include the inherent scarcity of patient data and the fragmentation of knowledge across diverse and siloed sources. Nevertheless, computer science has become a crucial ally. Bioinformatics and machine learning methods have facilitated the discovery of novel biomarkers, enhancing diagnostic precision and enabling better patient stratification[6, 7]. Meanwhile, in silico drug discovery and repurposing efforts have accelerated therapeutic identification, a critical advance given the limited financial incentives for orphan drug development[8, 9].

More recently, artificial intelligence, particularly the advent of large language models (LLMs), has opened new horizons for rare disease research[10]. These models, trained on extensive biomedical corpora, demonstrate a remarkable ability to understand, generate, and extract insights from unstructured clinical text, supporting decision-making and enabling richer cross-domain knowledge integration[11, 12]. Given the abundance of textual data—ranging from case reports to clinical notes—LLMs hold unique promise for rare diseases.

However, the application of LLMs in this domain remains fragmented and insufficiently characterized. A comprehensive evaluation is lacking, with open

questions surrounding their effective deployment, potential benefits, and methodological limitations in this sensitive, data-scarce context. A systematic survey is therefore critical to map current approaches, highlight successful applications, and identify challenges, thereby guiding clinicians, researchers, and practitioners toward informed and responsible use[13, 14, 15].

This review aims to provide that comprehensive synthesis, detailing how LLMs are leveraged to improve diagnosis, clarify disease mechanisms, and inform therapeutic strategies in rare diseases. We also explore emerging efforts to integrate LLMs with multimodal datasets—including genomic sequences, medical imaging, and electronic health records—underscoring the potential to advance personalized medicine and improve patient outcomes. By synthesizing recent advances and outlining key future directions, we chart a path forward for harnessing LLMs to meet the unique challenges of rare disease research and care.

2 Methodology Selection of Papers

The following section describes the different methodologies adopted to identify and analyze relevant scientific studies on applying traditional Large Language Models (LLM) in the context of rare diseases. The authors conducted a systematic review of the literature following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, which offer a structured approach for identifying, selecting, and including scientific literature. As shown in Figure 1, the PRISMA process was articulated in four phases.

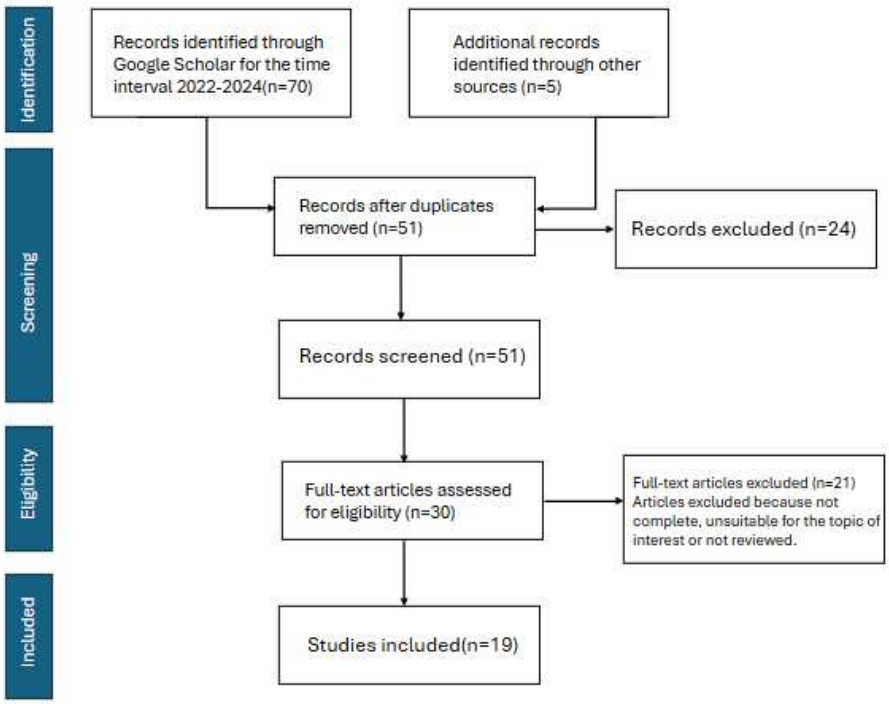

Figure 1: PRISMA flow diagram illustrating the systematic review process. Searches included combinations of "LLM", "Rare Diseases", "Data Analysis", and "Questionnaires" - (2022-2024).

Table 1 presents a summary of the main articles analyzed that focus on specific rare diseases. Each column of the table has been designed to highlight methodological elements relevant for the analysis. The "**Article**" column provides the full title of each included study, while "**Studied Disease**" indicates the rare disease being investigated, such as glaucoma, sarcoidosis, Kinsbourne disease, or amyloidosis. The "**Objective**" column follows, summarizing the main purpose of the study, which can range from evaluating diagnostic accuracy to the ability to extract information from textual data. The "**Type of Study**" field distinguishes between monomodal and multimodal studies; in this selection all studies are monomodal, that is, based on a single type of input. The "**Input Data**" column then specifies the nature of the data used, which can be structured questionnaires or unstructured texts (such as posts from social media or clinical forums).

Then there is "**Access Type**", which indicates whether the model used in the study is closed-source or not, and finally "**Pipeline**", which indicates whether the model was used in stand-alone mode or integrated into a larger system. In all the articles analyzed, the models were tested in closed environments and independently, without clinical pipelines or external tools to support them.

Article	Studied Disease	Objective	Type of Study	Input Data	Access Type	Pipeline
Assessment of a Large Language Model's Responses to Questions and Cases About Glaucoma and Retina Management [16]	Glaucoma	Evaluation of diagnostic accuracy and completeness	Monomodal	Questionnaire	Closed	Standalone
Understanding Sarcoidosis Using Large Language Models and Social Media Data [17]	Sarcoidosis	Extracting insight from patient Reddit data	Monomodal	Unstructured textual data	Closed	Standalone
High accuracy but limited readability of large language model-generated responses to frequently asked questions about Kienböck's disease [18]	Kienböck's disease	General knowledge, diagnosis and treatment	Monomodal	Questionnaire	Closed	Standalone
A multidisciplinary assessment of ChatGPT's knowledge of amyloidosis [19]	Amyloidosis	Clinical question-answering performance evaluation	Monomodal	Questionnaire	Closed	Standalone

Table 1: Summary of articles on specific rare disease-related LLM evaluations

3 Research in LLM and Rare Diseases

Research into the use of Large Language Models (LLMs) in the context of rare diseases is gaining momentum, but the field remains highly diverse. The variety in research goals, data sources, technical implementations, and clinical applications makes it difficult to directly compare studies. To address this issue, we propose a structured classification framework that can systematically analyze and interpret the literature.

This framework describes several key dimensions for categorizing LLM-based studies on rare diseases. Our aim is to provide a clearer understanding of the research landscape, highlight emerging trends, and identify areas that require further exploration or standardization.

Standardization encompasses several fundamental aspects. First, we consider the type of disease focus: whether the study targets a single rare disease, a class of disease (e.g. genetic disorders, ophthalmologic diseases), or adopts a general-purpose approach applicable to multiple rare conditions. Andrew et al., [20], present the evaluation of the performances of the LLMs in the diagnosis of pediatric rare diseases. Asfuroglu et al. [18] apply LLM to diagnose a rare muscoloskeletal disease; Huang et al. [16] apply LLM to the study of glaucoma. Finally, most approaches focus on the whole spectrum of rare diseases, and limitations are often related to the availability of specific datasets [21, 22, 23, 24, 25, 26, 27, 28, 29, 17, 30, 31, 32].

Second, we categorize the study's objectives, which may include diagnosis, prognosis, or research-oriented tasks such as gene prioritization and data extraction.

Large Language Models (LLMs) can be effectively adapted for the diagnosis of rare diseases by using a combination of techniques: prompting, retrieval-augmented generation (RAG), fine-tuning, and domain-specific pre-training [33].

Prompting techniques, such as role-based prompts and chain-of-thought instructions, help guide LLMs to systematically analyze patient phenotypes and suggest potential differential diagnoses. Dynamic few-shot prompting enhances

model performance by providing carefully selected clinical examples within the input, which improves the model's ability to generalize to rare and complex cases.

Retrieval-augmented generation further supports prompting by supplying LLMs with curated external medical knowledge at query time. This enriches the model's contextual understanding and enhances the accuracy of diagnostic suggestions.

In addition to input-level adaptations, fine-tuning and pre-training approaches contribute to deeper model specialization. Fine-tuning involves updating model parameters using rare disease-specific corpora, allowing LLMs to internalize patterns of symptom-disease associations and clinical reasoning processes. Domain-specific pre-training, which exposes models to extensive biomedical literature, structured phenotype ontologies, and curated case reports from the early stages of training, establishes a foundational understanding of rare disease contexts.

Large language models (LLMs) are being used to assist healthcare professionals by improving the precision and timeliness of diagnostic processes; they are capable of synthesizing vast amounts of clinical data—including patient history, imaging reports, and laboratory findings—to recognize disease patterns, reduce diagnostic errors, and even flag conditions at an early stage, which is particularly valuable for diseases with subtle or non-specific symptoms.

When applied to prognosis, LLMs support the ongoing evaluation of a patient's condition over time by analyzing longitudinal health records, identifying trends or risk factors indicative of disease progression, and generating predictive models that inform clinicians about likely future developments; this enables more proactive management strategies, timely interventions, and the personalization of follow-up plans based on individual risk profiles [34].

In research-focused applications, LLMs serve as powerful tools for uncovering complex genotype-phenotype associations by parsing large-scale genomic and clinical datasets, identifying meaningful correlations, and suggesting candidate genes for further study; additionally, these models can be employed to construct interactive systems that comprehend and respond to specialized biomedical questionnaires, facilitating automated data collection, hypothesis generation, and support for literature-based discovery in biomedical sciences [35].

In addition, we classify the input data modalities used, ranging from free-text clinical reports to structured questionnaires and laboratory results. The framework also captures the type of LLM employed, including general-purpose models such as GPT-4 and domain-specific fine-tuned models. Importantly, we note the availability and accessibility of these models, distinguishing between open-source, commercial, closed-source, and freely accessible non-open platforms.

Finally, we assess the role of the LLM within the broader computational pipeline, determining whether it acts as a standalone diagnostic tool or is integrated within a multicomponent system involving other data processing or inference modules. This comprehensive classification allows for a richer understanding of how LLMs are being adapted to the rare disease domain and provides a foundation for more consistent benchmarking and evaluation in the

future. Research into the use of Large Language Models (LLMs) in the context of rare diseases is gaining momentum, but the field remains highly diverse. The variety in research goals, data sources, technical implementations, and clinical applications makes it difficult to make direct comparisons between studies. To address this issue, we propose a structured classification framework that can systematically analyze and interpret the literature.

Standardization encompasses several fundamental aspects. First, we consider the type of disease focus: whether the study targets a single rare disease, a class of disease (e.g. genetic disorders, ophthalmologic diseases), or adopts a general-purpose approach applicable to multiple rare conditions. Second, we categorize the study's objectives, which may include diagnosis, prognosis, or research-oriented tasks such as gene prioritization and data extraction.

In addition, we classify the input data modalities used, ranging from free-text clinical reports to structured questionnaires and laboratory results. The framework also captures the type of LLM employed, including general-purpose models such as GPT-4 and domain-specific fine-tuned models. Importantly, we note the availability and accessibility of these models, distinguishing between open-source, commercial, closed-source, and freely accessible non-open platforms.

Finally, we assess the role of the LLM within the broader computational pipeline, determining whether it acts as a standalone diagnostic tool or is integrated within a multi-component system involving other data processing or inference modules. This comprehensive classification allows for a richer understanding of how LLMs are being adapted to the rare disease domain and provides a foundation for more consistent benchmarking and evaluation in the future.

3.0.1 Types of Input Data

The selected studies involving specific rare diseases use different types of input data, ranging from structured clinical questionnaires to unstructured online content. For example, Huang et al. [16] used 20 expert-generated clinical questions—10 each on glaucoma and 10 on retinal diseases—randomly selected from the "Ask an Eye Doctor" section of the American Academy of Ophthalmology website, along with 20 anonymized patient cases from ophthalmology clinics. These were used to assess the accuracy of GPT-4 responses for glaucoma and retinal diseases.

In contrast, Xi et al. [17] analyzed user-generated content on Reddit, mining sarcoidosis-related posts to explore patients' self-reported symptoms and their assessments of medications, diagnoses, and experiences. Asfuroğlu et al. [18] developed a frequently asked questions (FAQ) questionnaire on Kienböck's disease. The terms "Kienböck's disease" and "lunar avascular necrosis" were used to collect the most commonly searched terms from the three major search engines: Google, Bing, and Yahoo; then, relevant FAQs were identified from reliable medical websites. Finally, King et al. [19] evaluated ChatGPT's ability to answer a total of 98 amyloidosis-related questions from medical societies

and institutions. To provide a more comprehensive perspective for patients, questions from amyloidosis Facebook support groups were also included. Of the total questions, 56 were general amyloidosis topics, while 42 focused on specific subfields, including cardiology and neurology.

3.0.2 Prompting Strategies

Regarding prompting strategies, most of the studies reviewed employed simple, zero-shot or instruction-based prompts designed to generate medical responses from LLMs without the use of contextual examples. These prompts often mirrored typical clinical queries or patient questions. For example, in one study [16], the authors instructed GPT-4 to respond as if it were writing a clinical note, including the use of domain-specific abbreviations commonly used in ophthalmology, thus simulating the tone and format of real medical documentation.

In other cases, [18], the prompt structure consisted of a straightforward question-answer format, in which the model was presented with frequently asked questions about Kienböck's disease, drawn from patient-oriented web sources. Responses were generated without prior examples or guided context, relying on the model's in-built medical knowledge.

In another study [19], questions about amyloidosis were collected from both institutional resources and patient support communities. These were presented to the model in standard question form, without explicit role instructions.

3.0.3 Accuracy Measures

In the selected studies, different strategies were adopted to evaluate the accuracy and quality of the results generated by the LLM. In most cases, the evaluations were conducted by expert ratings using Likert scales, allowing clinicians to evaluate the relevance, correctness and completeness of the model responses [16, 19]. Some studies also included additional readability metrics, such as the Flesch-Kincaid Readability Index, which was used to quantify the comprehensibility of the responses provided by the model [18].

Other works have integrated classic natural language processing performance metrics, such as accuracy and F1 score, to quantify the level of alignment of the generated responses with the reference responses or with the expected results [19]. In contrast, qualitative approaches have also been employed, such as thematic analysis, used to interpret LLM-generated content derived from social media posts, with the aim of identifying patterns in patients' concerns and needs [17].

Although methodological diversity strongly depends on the research objectives, the common thread of all studies is the use of human experience-based assessments, both quantitative and qualitative.

4 Dataset

This section outlines key datasets referenced in the reviewed studies, emphasizing their core characteristics and accessibility.

4.1 Clinical Datasets

Clinical datasets are central to biomedical research, providing the foundation for developing and validating NLP systems tailored to medical applications. Typically derived from electronic health records (EHRs), clinical notes, laboratory findings, and demographic data, these resources capture the complexity of real-world healthcare settings. They support predictive modelling, enable automated phenotyping, and power clinical decision support tools. Among the datasets most frequently cited in the reviewed studies are MIMIC-III, MIMIC-IV, RareDis, RAMEDIS, and PUMCH (Table 2).

Source	**Typology**	**Material**
Ramedis	Open	Rare disease data & clinical cases
MIMIC-III	Open (with registration)	ICU patient data (clinical, demographics)
MIMIC-IV	Open (with registration)	ICU patient data (clinical, time-series)
RareDis	Open	Rare disease phenotypic data
PUMCH	Restricted (institutional access)	Patient records (clinical & genomic)

Table 2: Clinical Dataset

MIMIC-III (Medical Information Mart for Intensive Care III) [36] is one of the most widely used clinical datasets in medical NLP. Developed by the MIT Lab for Computational Physiology, it contains de-identified health records from over 40,000 patients admitted to Beth Israel Deaconess Medical Center between 2001 and 2012. The dataset includes both structured data—such as laboratory test results, medications, demographics, and vital signs—and unstructured clinical notes. MIMIC-III has been extensively adopted for patient phenotyping, disease classification, and predictive modelling [37].

MIMIC-IV [38] is the successor to MIMIC-III and extends its coverage to the years 2008–2019. It introduces refined data structuring, improved linkage between hospital departments, and enhanced consistency across patient records. Like its predecessor, it includes both structured and unstructured components and continues to serve as a cornerstone for reproducible research in medical artificial intelligence and clinical NLP [39].

Another frequently cited dataset is **RareDis** [40] id designed to support rare disease research. It consists of annotated clinical case descriptions sourced from the scientific literature and expert-curated repositories. Each case typically includes a narrative of symptoms and diagnostic conclusions.

RAMEDIS (Rare Metabolic Diseases Information System) [41] is a specialized database designed to support diagnosis and research in the field of rare metabolic disorders. It integrates clinical, biochemical, and genetic data and

offers multilingual access. Developed in the early 2000s, RAMEDIS is aligned with standardized disease classifications and has proven useful for both clinical decision support and interoperability in biomedical NLP applications.

Several studies also reference the so-called **PUMCH dataset**, derived from clinical records at the Peking Union Medical College Hospital. While not a standardized or publicly available dataset, it comprises institutional EHRs, diagnostic notes, and genomic data used internally for developing and validating AI models. Despite its restricted access, PUMCH has contributed significantly to rare disease informatics, particularly within studies focused on Chinese populations.

4.2 Biomedical Ontologies and Knowledge Bases

Biomedical ontologies and knowledge bases offer structured representations of complex medical and biological entities, enabling semantic interoperability, data harmonization, and automated reasoning. The main ontologies and knowledge bases identified across studies are summarized in Table 3.

Ontology	**Typology**	**Material**
ORDO	Open	Rare diseases classification
ORPHANET	Open	Rare diseases database
OMIM	Open (partial access)	Genetic disorders & phenotypes

Table 3: Ontology Table

ORDO (Orphanet Rare Disease Ontology) [42] provides a structured representation of the rare disease domain, capturing relationships between disorders, phenotypes, inheritance patterns, and associated genes. Developed in OWL (Web Ontology Language), ORDO is designed for semantic reasoning and classification and is frequently used to annotate patient records in rare disease registries, linking phenotypic and molecular data. Its machine-readable structure supports both clinical interpretation and automated disease matching in NLP pipelines.

Orphanet [43] is a comprehensive, curated knowledge base maintained by a multinational consortium coordinated by INSERM. It encompasses more than 6,000 rare diseases, offering detailed metadata on prevalence, diagnostic criteria, associated genes, clinical features, and therapeutic options. In biomedical NLP, Orphanet is widely used for lexicon development, annotation tasks, and extraction of structured knowledge from unstructured medical narratives. Its standardized nomenclature, including ORPHA codes, enhances data interoperability across studies.

OMIM (Online Mendelian Inheritance in Man) [44] is a foundational reference catalog of human genes and genetic disorders, currently comprising over 15,000 entries. It includes curated descriptions of gene-disease associations, inheritance patterns, mutation profiles, and phenotypic variability. In computational research, OMIM is frequently used for gene–disease mapping and serves

as a benchmark for genetic entity recognition and linkage tasks in biomedical NLP.

4.3 Textual Corpora and Bibliographic Sources

Text corpora and bibliographic resources play a crucial role in the development and evaluation of NLP systems aimed at understanding medical language, generating clinical summaries, and responding to domain-specific queries.

PubMedQA [45] is a curated dataset designed for biomedical question answering. It consists of clinical yes/no/maybe questions paired with corresponding PubMed abstracts. This resource is commonly used to assess the reasoning capabilities and factual consistency of NLP models in processing medical literature.

MedQA [46] comprises multiple-choice questions derived from official medical licensing examinations. The dataset presents complex, clinically relevant scenarios and has become a benchmark for evaluating the diagnostic reasoning performance of large language models, including GPT-based systems, in tasks related to medical education and clinical decision-making.

5 Challenges

Table 4 outlines the main limitations associated with using large language models (LLMs) for supporting rare disease diagnosis. These limitations can be broadly categorized into technical, clinical, and governance-related challenges.

First, LLMs are inherently prone to hallucinations, which means they can generate outputs that are linguistically coherent but factually incorrect or fabricated. This issue is particularly concerning in the context of rare diseases, where the scarcity of reliable training examples increases the risk of model errors. Due to the underrepresentation of rare disease data in general corpora, LLMs may produce inaccurate associations between clinical features and specific conditions.

Zero-shot prompting, where models are queried without domain-specific fine-tuning, faces unique challenges due to the complexity and specificity required for rare disease diagnosis. Generalization failures are common, as LLMs have limited capacity to extrapolate reliably to medical domains for which they have not been explicitly trained. Additionally, the lack of curated, high-quality datasets for rare diseases restricts both the models' initial learning and opportunities for external validation.

General-purpose LLMs are mainly trained on non-peer-reviewed, heterogeneous internet sources, which diminishes their reliability for clinical application. Consequently, the outputs often deviate from established evidence-based diagnostic standards, posing potential patient safety risks. While specialized medical LLMs show improved performance, they still suffer from insufficient coverage of rare diseases due to the limited availability of appropriate training materials.

Fine-tuning on small rare disease datasets carries the risk of overfitting, where the model may capture dataset-specific noise rather than generalizable patterns. This overfitting can amplify biases present in the training data, such as demographic underrepresentation, leading to inequities in diagnostic outcomes.

LLMs also struggle to recognize subtle clinical nuances that are often critical for identifying rare diseases. Specialized models require ongoing updating and curation to remain relevant as medical knowledge evolves. However, few studies have systematically validated LLM performance specifically for rare disease diagnostic applications, which limits confidence in their deployment.

Another significant barrier is the opacity of LLMs, as their internal decision-making processes are largely uninterpretable. This lack of explainability complicates their integration into clinical workflows and undermines clinician trust. Moreover, errors in model outputs may propagate through diagnostic chains, compounding clinical risks.

Current LLMs often fail to adequately capture demographic variability in the presentation of rare diseases, further exacerbating the risks of diagnostic disparities. Additionally, regulatory and ethical challenges persist, particularly concerning the deployment of unvalidated diagnostic tools in clinical settings.

Finally, while specialized medical LLMs demonstrate superior performance compared to general-purpose models, both types require rigorous validation, governance, and continuous oversight. Without these safeguards, the deployment of LLMs could perpetuate and potentially worsen diagnostic inaccuracies and disparities, especially among vulnerable populations affected by rare diseases.

In conclusion, careful design, domain-specific adaptation, external validation, and human-centered oversight are essential for the safe and effective integration of LLMs into rare disease diagnostic workflows.

First Challenge: Medical LLMs vs General LLMs. General purpose LLMs can often be inaccurate in diagnosing rare diseases due to their limited domain-specific data knowledge and lack of integration with structured clinical knowledge. In contrast, medical LLMs, especially those designed with access to curated ontologies and biomedical databases, demonstrate greater potential in terms of diagnostic accuracy. In support of this, recent studies [21] have proposed hybrid approaches that combine the language understanding capabilities of LLMs with structured phenotypic or clinical knowledge sources, integrating LLMs with phenotype-disease relationships and biomedical ontologies to improve diagnostic accuracy and reduce hallucinations.

Second challenge: Improving Results One of the main challenges in applying LLMs to clinical diagnosis, especially in the field of rare diseases, is to ensure reliable performance in real-world use cases. Recent studies have shown that diagnostic accuracy is closely related to the format and completeness of the input; specifically, narrative-style case descriptions perform better than prompts derived from structured clinical data. However, detailed narratives may often be unrealistic in clinical practice due to time constraints and privacy issues; this highlights the need to develop realistic prompting techniques that are consistent with typical clinical documentation practices.

6 Future Perspectives

Rare disease research faces a significant data scarcity challenge that limits the development of diagnostic tools, therapeutic strategies, and essential biological insights. In this context, Large Language Models (LLMs) offer a transformative opportunity for data augmentation. Future research should focus on systematically utilizing LLMs to simulate, expand, and enrich rare disease datasets with synthetic yet clinically plausible information [47].

Firstly, LLMs could be fine-tuned to generate synthetic patient cases that include structured phenotypic profiles, detailed clinical narratives, and plausible disease progression trajectories. Such augmented datasets could enhance the training of downstream machine learning models, particularly in scenarios where real-world case numbers are prohibitively small [28]. The RareBench initiative [28] already highlights the difficulties LLMs encounter when faced with limited, heterogeneous clinical examples and points toward dynamic few-shot prompting as a promising strategy for improving data diversity.

Secondly, retrieval-augmented generation (RAG) frameworks can be integrated to ensure that LLM-generated content is grounded in verified medical knowledge, thereby minimizing the risk of hallucinations. Future LLM-based augmentation systems must closely couple generation with retrieval from curated rare disease knowledge bases, dynamically constraining outputs to maintain clinical accuracy.

Third, LLMs have significant potential to expand phenotype ontologies and fill gaps in rare disease knowledge graphs. LLMs could propose novel associations of symptoms, patterns of comorbidity, or disease pathways that could otherwise remain underreported in the clinical literature. These hypotheses could then be validated experimentally, accelerating the pace of knowledge discovery.

Fourth, adaptive domain-specific pre-training strategies are essential. Rather than relying solely on generalist models, future research should explore lightweight, continual pre-training of LLMs on rare disease–focused corpora, including clinical reports, case studies, and genetic databases. This would enhance the biological and clinical relevance of synthetic outputs without necessitating a complete model retraining.

However, realizing this vision requires careful attention to validation and ethical considerations. Synthetic datasets must be rigorously evaluated for realism, diversity, and potential bias propagation. Furthermore, governance frameworks should be developed to delineate appropriate use cases for synthetic data, ensuring that augmented datasets do not inadvertently reinforce diagnostic inaccuracies or demographic disparities.

Finally, collaborations between clinical experts, data scientists, and AI ethicists will be crucial in co-developing standards for the responsible use of LLMs in rare disease research. Establishing transparent benchmarks for evaluating synthetic data quality, clinical plausibility, and downstream impact will be a priority.

With careful design, grounding, and oversight, LLM-driven data augmentation could substantially mitigate current data limitations, accelerate model

development, and ultimately improve outcomes for patients affected by rare diseases.

7 Conclusion

Even while big language models have advanced a lot, there is still much work to be done in terms of integrating multimodal data for rare diseases. Although rare diseases necessitate a more comprehensive approach, the majority of current research on LLM applications is unimodal, such as in natural language processing or medical imaging . These diseases require a technology that can combine information from genomic data, imaging investigations, laboratory results and textual descriptions. Such integration now represents an unmet need for LLMs' full potential to offer thorough insights for diagnosis and therapy. Our experimentation with different LLMs, however, showed promising results regarding their potential to assist in diagnosis. The application of LLMs in simulating interactions with patients and analysing medical texts has shown the ability to identify patterns and correlations that can help healthcare professionals.

Therefore, further development in this area is necessary to improve LLM's capacity to handle multimodal data. The diagnosis and treatment of rare diseases necessitate the integration of text, photos, structured data and even longitudinal patient histories, even if the existing models are excellent at text-based information processing. Multimodal AI frameworks, according to Ou et al. [48], have the potential to completely transform diagnostics by revealing intricate connections between various data sources that could otherwise go overlooked.

For instance, the early and more precise identification of rare illnesses may be made possible by combining phenotypic descriptions with genetic variants and imaging biomarkers. The lack of sizable, harmonized multimodal datasets and the processing power needed to train such models are two of the many challenges that impede this integration. As a result, creating LLMs in a multimodal setting presents both a technological challenge and an innovative chance to rethink the way that rare illness research and treatment are perceived.

Cross-disciplinary cooperation between AI researchers, physicians, geneticists, and bioinformaticians will be necessary to overcome these obstacles. According to Hasani et al. , this kind of cooperation is essential to guaranteeing that developments in AI technology are clinically significant and satisfy patient requirements.

Critical issues in data standardization, ethical issues, and converting AI-driven solutions into practical applications can be resolved by active cross-disciplinary collaboration. In order to emphasize therapeutic outcomes, Denecke et al. [49] advocate collaborative approaches, which would guarantee innovation but, more crucially, the usefulness and accessibility of AI technologies. Ou et al. [48] assert that interdisciplinary approaches have the best chance of breaking through bottlenecks including, among other things, clinical interpretation of outcomes and standardized dataset production in numerous modalities.

By encouraging cooperation and combining a variety of skills, the research

community may expand the realm of what is feasible using LLMs in rare disease scenarios.

Table 4: Limitations of Large Language Models (LLMs) in Supporting Rare Disease Diagnosis

Limitation	**Description**
Hallucination Risk	LLMs may produce plausible but factually inaccurate or fabricated outputs, increasing diagnostic risk.
Data Scarcity	Rare diseases are underrepresented in training data, limiting model exposure and learning.
Zero-Shot Prompting Challenges	Lack of domain-specific adaptation hampers model ability to handle rare disease cases effectively.
Limited Generalization	LLMs generalize poorly to highly specialized or unfamiliar medical scenarios.
Training Data Quality	Scarcity of curated datasets restricts model performance and validation opportunities.
Source Quality Limitations	General LLMs are trained on non-peer-reviewed sources, reducing clinical reliability.
Deviation from Medical Standards	Model outputs may not align with evidence-based diagnostic protocols.
Constraints of Specialized LLMs	Even fine-tuned models show limited rare disease representation.
Risk of Overfitting	Fine-tuning on small datasets can reduce model generalizability.
Amplification of Biases	Existing biases in clinical data may be reinforced in model outputs.
Misinterpretation of Clinical Nuances	Subtle symptomatology critical for rare diseases may be overlooked.
Need for Continuous Updating	Specialized models require ongoing curation to remain clinically relevant.
Lack of Systematic Validation	Few studies assess LLM performance specifically for rare disease diagnosis.
Opacity and Limited Explainability	LLMs' black-box nature complicates clinical decision-making and trust.
Error Propagation Risk	Errors in LLM outputs can cascade through diagnostic workflows.
Demographic Variability	LLMs insufficiently capture diverse disease presentations across populations.
Regulatory and Ethical Challenges	Unvalidated outputs pose risks for ethical clinical deployment and approval.
Performance Gap Between General and Specialized LLMs	General models perform markedly worse than domain-specific models, but both require scrutiny.
Risk of Diagnostic Disparities	Improper deployment could exacerbate inequities in rare disease diagnosis.
Need for Governance and Oversight	Human supervision and rigorous frameworks are essential for safe use.

Figure Legends

Figure 1. Flowchart of the diagnostic journey for rare disease patients. The path often includes multiple incorrect diagnoses, specialist referrals, and redundant testing before the correct diagnosis is reached, often several years after symptom onset.

Alt Text: A flowchart showing the typical diagnostic process for patients with rare diseases, including repeated misdiagnoses, referrals to various specialists, and unnecessary tests, often resulting in the correct diagnosis only years after symptoms first appear.

Figure 2. PRISMA flow diagram illustrating the systematic review process. Searches included combinations of "LLM", "Rare Diseases", "Data Analysis", and "Questionnaires" - (2022-2024).

Alt Text: A PRISMA flow diagram detailing the systematic review process, illustrating the identification, screening, eligibility, and inclusion of studies. The search strategy combined terms such as "LLM," "Rare Diseases," "Data Analysis," and "Questionnaires" covering literature from 2022 to 2024

8 Ethics approval and consent to participate

Not Applicable

9 Consent for publication

Not Applicable

10 Availability of data and materials

Not Applicable.

11 Competing interests

Authors declare that they do not ave competing interest.

12 Funding

Not Applicable.

13 Authors' contributions

V.C. is the main author. P.V and PHG supervised the research, wrote the manuscript. All the authors read and approved the manuscript.

14 Acknowledgements

Not Applicable.

Table Legends

Table 1. Summary of articles on specific rare disease-related LLM evaluations.

Alt Text: Description of 4 scientific articles analyzing 4 rare diseases and LLMs: amyloidosis (Input data: questionnaires), sarcoidosis (Input data: unstructured textual data), glaucoma (Input data: questionnaires) and Kienbock's disease (Input data: questionnaires). Each study has a closed access type, uses a standalone pipeline and employs a unimodal approach.

Table 2. Clinical Dataset.

Alt Text: Table listing datasets, typology and material. The datasets analyzed are: Ramedis (Rare disease data clinical cases) , MIMIC-III (ICU patient data), MIMIC-IV (ICU patient data), RareDis (Rare disease phenotypic data) and PUMCH (Clinical and genomic patient records). They are all open sources except PUMCH (institutional access).

Table 3. Ontology Table.

Alt Text: Table listing ontologies, typology and material. The ontologies analyzed are: ORDO (Rare diseases classification), ORPHANET (Rare disease database), OMIM (Genetic disorders and phenotype), MONDO (Disease ontology), HPO (Phenotype descriptions) and UMLS (Medical vocabularies and ontologies). They are all open sources except OMIM (partial access) and UMLS (free access with registration).

Table 4. PubMed, PubMedQA, MedQA Table.

Alt Text: Table listing Bibliographic Sources such as PubMed, PubMedQA and MedQA. PubMed provides open-access abstracts, PubMedQA offers open question-answer pairs based on abstracts, MedQA includes exam questions with both restricted and open access.

Table 5. Limitations of Large Language Models (LLMs) in Supporting Rare Disease Diagnosis.

Alt Text: Table outlining limitations of LLM in supporting rare disease diagnosis; including risks of inaccurate outputs, limited training data, poor generalization to rare cases, data and source quality issues, lack of domain adaptation, potential biases, difficulty interpreting clinical details, need for continuous updates, limited validation, and ethical and regulatory concerns.

References

[1] European Commission. *Rare diseases*. https://health.ec.europa.eu/system/files/2023-01/rare_diseases_en_0.pdf. Accessed: 2025-05-01. 2023.

[2] U.S. Food and Drug Administration. *Rare Diseases: About Rare Diseases*. https://www.fda.gov/patients/rare-diseases-fda. Accessed: 2025-05-01. 2023.

[3] A. Fehr and F. Prütz. "Rare diseases: a challenge for medicine and public health". In: *Journal of Health Monitoring* 8.4 (2023). Published 2023 Dec 13, pp. 3–6. DOI: 10.25646/11826.

[4] Arrigo Schieppati et al. "Why rare diseases are an important medical and social issue". In: *The Lancet* 371.9629 (2008), pp. 2039–2041.

[5] Fatoumata Faye et al. "Time to diagnosis and determinants of diagnostic delays of people living with a rare disease: results of a Rare Barometer retrospective patient survey". In: *European Journal of Human Genetics* 32.9 (2024), pp. 1116–1126.

[6] Srivatsan Rao, Gino Fung, and Sangeeta Rao. "Rare disease diagnosis: a review of web search, social media and large-scale data-mining approaches". In: *Rare Diseases* 5.1 (2017), e1355661.

[7] Tuan Nguyen, Thanh Tran, and Thao Nguyen. "Deep learning for genomics: A concise overview". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12.2 (2022), e1449.

[8] Samantha M Corsello et al. "The Drug Repurposing Hub: a next-generation drug library and information resource". In: *Nature medicine* 23.4 (2017), pp. 405–408.

[9] Kevin Yang et al. "Machine learning for drug discovery". In: *Nature Reviews Drug Discovery* 18.6 (2019), pp. 463–477.

[10] Sina Shool et al. "A systematic review of large language model (LLM) evaluations in clinical medicine". In: *BMC Medical Informatics and Decision Making* 25.1 (2025), p. 117.

[11] Karan Singhal et al. "Large language models encode clinical knowledge". In: *Nature* 620.7972 (2023), pp. 172–180.

[12] Wonjin Lee et al. "BioGPT: generative pre-trained transformer for biomedical text generation and mining". In: *Briefings in Bioinformatics* 24.1 (2023), bbac409.

[13] Tianran He et al. "A survey of large language models in medicine and healthcare". In: *arXiv preprint arXiv:2306.13542* (2023).

[14] Jie Zhao et al. "A survey of large language models for healthcare". In: *Journal of Biomedical Informatics* 144 (2023), p. 104408.

[15] Daniele Mercatelli et al. “Exploiting the molecular basis of age and gender differences in outcomes of SARS-CoV-2 infections”. In: *Computational and Structural Biotechnology Journal* 19 (2021), pp. 4092–4100.

[16] Andy S Huang et al. “Assessment of a large language model’s responses to questions and cases about glaucoma and retina management”. In: *JAMA ophthalmology* 142.4 (2024), pp. 371–375.

[17] Nan Miles Xi, Hong-Long Ji, and Lin Wang. “Understanding Sarcoidosis Using Large Language Models and Social Media Data”. In: *Journal of Healthcare Informatics Research* (2024), pp. 1–26.

[18] Zeynel Mert Asfuroğlu, Hilal Yağar, and Ender Gümüşoğlu. “High accuracy but limited readability of large language model-generated responses to frequently asked questions about Kienböck’s disease”. In: *BMC Musculoskeletal Disorders* 25.1 (2024), p. 879.

[19] Ryan C King et al. “A multidisciplinary assessment of ChatGPT’s knowledge of amyloidosis”. In: *medRxiv* (2023), pp. 2023–07.

[20] Judith Jeyafreeda Andrew et al. “Evaluating llms for temporal entity extraction from pediatric clinical text in rare diseases context”. In: *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health)@ LREC-COLING 2024*. 2024, pp. 145–152.

[21] Jinge Wu et al. “A hybrid framework with large language models for rare disease phenotyping”. In: *BMC Medical Informatics and Decision Making* 24.1 (2024), p. 289.

[22] Cathy Shyr et al. “Identifying and extracting rare diseases and their phenotypes with large language models”. In: *Journal of Healthcare Informatics Research* 8.2 (2024), pp. 438–461.

[23] Jian Yang et al. “RDguru: a conversational intelligent agent for rare diseases”. In: *IEEE Journal of Biomedical and Health Informatics* (2024).

[24] Junyoung Kim et al. “Assessing the utility of large language models for phenotype-driven gene prioritization in the diagnosis of rare genetic disease”. In: *The American Journal of Human Genetics* 111.10 (2024), pp. 2190–2202.

[25] Tassallah Abdullahi, Ritambhara Singh, Carsten Eickhoff, et al. “Learning to make rare and complex diagnoses with generative AI assistance: qualitative study of popular large language models”. In: *JMIR Medical Education* 10.1 (2024), e51391.

[26] Stefania Zampatti et al. “Innovations in Medicine: Exploring ChatGPT’s Impact on Rare Disorder Management”. In: *Genes* 15.4 (2024), p. 421.

[27] Cameron C Young et al. “Diagnostic accuracy of a custom large language model on rare pediatric disease case reports”. In: *American Journal of Medical Genetics Part A* 197.2 (2025), e63878.

[28] Xuanzhong Chen et al. “RareBench: Can LLMs Serve as Rare Diseases Specialists?” In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2024, pp. 4850–4861.

[29] Karthik Soman et al. “Zebra-Llama: A Context-Aware Large Language Model for Democratizing Rare Disease Knowledge”. In: *arXiv preprint arXiv:2411.02657* (2024).

[30] Lang Cao, Jimeng Sun, Adam Cross, et al. “An Automatic and End-to-End System for Rare Disease Knowledge Graph Construction Based on Ontology-Enhanced Large Language Models: Development Study”. In: *JMIR Medical Informatics* 12.1 (2024), e60665.

[31] David Oniani et al. “Large language models vote: Prompting for rare disease identification”. In: *arXiv preprint arXiv:2308.12890* (2023).

[32] Juanjo do Olmo et al. “Assessing dxgpt: Diagnosing rare diseases with various large language models”. In: *MedRxiv* (2024), pp. 2024–05.

[33] Jasmine Chiat Ling Ong et al. “A scoping review on generative AI and large language models in mitigating medication related harm”. In: *npj Digital Medicine* 8.1 (2025), p. 182.

[34] Hannah E Bergman et al. “Development of a comprehensive heart disease knowledge questionnaire”. In: *American journal of health education* 42.2 (2011), pp. 74–87.

[35] Philip Chung et al. “Large language model capabilities in perioperative risk prediction and prognostication”. In: *JAMA surgery* 159.8 (2024), pp. 928–937.

[36] Alistair EW Johnson et al. “MIMIC-III, a freely accessible critical care database”. In: *Scientific data* 3.1 (2016), pp. 1–9.

[37] Pietro Hiram Guzzi and Swarup Roy. *Biological network analysis: Trends, approaches, graph theory, and algorithms*. 2020.

[38] Alistair EW Johnson et al. “MIMIC-IV, a freely accessible electronic health record dataset”. In: *Scientific data* 10.1 (2023), p. 1.

[39] Pietro Hiram Guzzi, Francesco Petrizzelli, and Tommaso Mazza. “Disease spreading modeling and analysis: A survey”. In: *Briefings in Bioinformatics* 23.4 (2022), bbac230.

[40] Claudia Martínez-deMiguel et al. “The RareDis corpus: a corpus annotated with rare diseases, their signs and symptoms”. In: *Journal of biomedical informatics* 125 (2022), p. 103961.

[41] Thoralf Töpel et al. “RAMEDIS: the rare metabolic diseases database”. In: *Applied bioinformatics* 5 (2006), pp. 115–118.

[42] Drashtti Vasant et al. “Ordo: an ontology connecting rare disease, epidemiology and genetic data”. In: *Proceedings of ISMB*. Vol. 30. researchgate. net. 2014, pp. 1–4.

[43] Stéphanie Nguengang Wakap et al. "Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database". In: *European journal of human genetics* 28.2 (2020), pp. 165–173.

[44] Victor A McKusick. "Mendelian Inheritance in Man and its online version, OMIM". In: *The American Journal of Human Genetics* 80.4 (2007), pp. 588–604.

[45] Qiao Jin et al. "Pubmedqa: A dataset for biomedical research question answering". In: *arXiv preprint arXiv:1909.06146* (2019).

[46] Di Jin et al. "What disease does this patient have? a large-scale open domain question answering dataset from medical exams". In: *Applied Sciences* 11.14 (2021), p. 6421.

[47] Aya A Mitani and Sebastien Haneuse. "Small data challenges of studying rare diseases". In: *JAMA network open* 3.3 (2020), e201965–e201965.

[48] Peiling Ou et al. "Artificial intelligence empowering rare diseases: a bibliometric perspective over the last two decades". In: *Orphanet Journal of Rare Diseases* 19.1 (2024), p. 345.

[49] Kerstin Denecke et al. "Potential of large language models in health care: Delphi study". In: *Journal of Medical Internet Research* 26 (2024), e52399.