INTERNATIONAL JOURNAL OF MACHINE LEARNING AND ENGINEERING

Vol 2 Issue 1 (2026)

https://ijmle.org/

Enhancing Rare Disease Diagnosis from Clinical Notes through Chain-of-Thought and Retrieval-Augmented Generation

Seojoon Lee, Adrian Benjamin, Yuji Saito

Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Melbourne, Australia

Abstract:

The diagnosis of rare diseases remains one of the most formidable challenges in modern medical practice, often characterized by a prolonged diagnostic odyssey that imposes significant physical, psychological, and financial burdens on patients. Electronic health records contain vast amounts of unstructured clinical narratives that harbor critical clues for identifying these elusive conditions. However, the sheer volume of data and the cognitive limits of human practitioners often obscure these signals. Recent advancements in large language models offer unprecedented capabilities in natural language understanding, yet their direct application in medical diagnostics is hindered by issues such as knowledge hallucination and a lack of grounding in authoritative medical literature. This paper proposes a novel framework that integrates retrieval-augmented generation with chain-of-thought reasoning to enhance the accuracy and interpretability of rare disease diagnosis from unstructured clinical notes. By dynamically retrieving pertinent medical literature from curated knowledge bases and explicitly modeling the step-by-step deductive reasoning process characteristic of expert physicians, the proposed architecture significantly mitigates hallucinations while improving diagnostic precision. Our comprehensive analysis demonstrates that this dual-component approach outperforms traditional machine learning baselines and standalone large language models across multiple evaluation metrics. Furthermore, the explicit reasoning pathways generated by the system provide crucial transparency, fostering clinical trust and offering a viable pathway toward the integration of advanced artificial intelligence diagnostic support systems in real-world healthcare settings.

Keywords: *: Rare Diseases, Clinical Diagnostics, Natural Language Processing, Retrieval-Augmented Generation, Medical Informatics.*

INTRODUCTION

1.1 The Challenge of Rare Disease Diagnosis

The landscape of global healthcare is fraught with complexities, yet few challenges are as persistent and devastating as the timely and accurate diagnosis of rare diseases. A rare disease is generally defined by an extraordinarily low prevalence within the general population, yet collectively, the thousands of distinct rare diseases currently identified affect hundreds of millions of individuals worldwide. The fundamental paradox of rare diseases is that while any single

condition is statistically improbable for a given clinician to encounter, the aggregate burden on the healthcare system is immense [1]. Patients suffering from these conditions typically endure a protracted phenomenon known as the diagnostic odyssey. This odyssey is characterized by years of uncoordinated medical consultations, frequent misdiagnoses, unnecessary invasive procedures, and a profound psychological toll on both the patients and their families [2]. The primary bottleneck in resolving this crisis is the inherent limitation of human cognition combined with the vast, fragmented nature of medical knowledge. General practitioners and even specialized clinicians simply cannot maintain an active, comprehensive mental repository of the phenotypic expressions, genetic markers, and idiosyncratic symptom progressions of over seven thousand distinct rare conditions. Consequently, subtle clinical signs that might serve as pathognomonic indicators for a rare disease are frequently overlooked, dismissed as anomalous, or incorrectly attributed to more common, statistically probable ailments. This systemic failure necessitates the development of sophisticated computational tools capable of augmenting clinical reasoning, operating tirelessly across vast datasets, and cross-referencing patient presentations against the entirety of established medical literature. Unlocking the potential of artificial intelligence in this domain is not merely a technical optimization but a critical public health imperative [3].

1.2 The Role of Electronic Health Records

In the contemporary healthcare ecosystem, the electronic health record serves as the central repository for patient data, capturing a granular history of encounters, laboratory results, radiological findings, and therapeutic interventions. Within this digital infrastructure, the most valuable diagnostic information is often buried within unstructured clinical notes. These free-text narratives, authored by physicians, nurses, and allied health professionals, provide rich, contextual descriptions of a patient's clinical trajectory. They capture the nuances of patient history, the temporal evolution of symptoms, and the subjective clinical impressions that structured data fields cannot adequately accommodate [4]. However, this unstructured format presents significant barriers to automated analysis. Clinical notes are notoriously chaotic, saturated with specialized medical jargon, idiosyncratic abbreviations, grammatical irregularities, and typographical errors. Furthermore, the longitudinal nature of electronic health records means that a single patient's diagnostic clues might be dispersed across dozens of distinct documents authored by different specialists over several years. Traditional computational approaches, relying on structured queries or rigid diagnostic codes, consistently fail to capture the subtle, emergent patterns required for rare disease identification. Recognizing a rare genetic syndrome, for instance, might require synthesizing an observation about facial dysmorphia recorded by a pediatrician, a cardiac anomaly noted by a specialist, and a neurological symptom documented in an emergency department admission. Therefore, harnessing the latent diagnostic power of electronic health records demands advanced natural language processing methodologies capable of deep semantic comprehension, temporal reasoning, and context-aware synthesis.

1.3 Large Language Models in Healthcare

The advent of large language models has precipitated a paradigm shift in the field of natural language processing, offering unprecedented capabilities in text comprehension, summarization, and generation. Built upon massive neural network architectures and trained on colossal corpora of text, these models have demonstrated remarkable proficiency in understanding the complexities of human language. In the domain of healthcare, early implementations of large language models have shown significant promise in tasks ranging from medical text summarization to patient triaging and medical dialogue generation [5]. However, the direct application of these models to the high-stakes arena of diagnostic decision-making reveals critical vulnerabilities. The most prominent of these is the phenomenon of hallucination, wherein the model generates medically plausible but factually incorrect or entirely fabricated information. In the context of rare diseases,

where exactitude is paramount and the margin for error is minimal, hallucinations can lead to catastrophic clinical recommendations. Furthermore, foundational models are fundamentally constrained by the temporal boundaries of their training data. Medical knowledge, particularly regarding rare and newly categorized genetic disorders, evolves rapidly; a model trained on data from previous years will inherently lack awareness of the latest diagnostic criteria or therapeutic discoveries [6]. To bridge the gap between the theoretical capabilities of large language models and the stringent requirements of clinical safety, novel architectural paradigms must be introduced. The integration of external, dynamically updated knowledge bases through retrieval augmentation, combined with structural interventions that force the model to render its reasoning process transparent, represents the most promising frontier in medical artificial intelligence.

2. Literature Review and Background

2.1 Diagnostic Odysseys in Rare Diseases

A comprehensive understanding of the diagnostic challenges in rare diseases requires an examination of the systemic and cognitive factors that contribute to clinical delays. Extensive epidemiological studies have consistently highlighted that patients with rare diseases wait an average of five to seven years for an accurate diagnosis, interacting with an average of eight different medical specialists during that period [7]. This extensive delay is primarily driven by the overlapping phenotypic presentations between rare diseases and common ailments. For example, the early manifestations of a rare lysosomal storage disorder might masquerade as generic developmental delays or common pediatric rheumatological complaints. Clinicians, operating under cognitive frameworks governed by heuristics and probability, naturally default to the most common explanations, a phenomenon often described in medical epistemology as the availability heuristic [8]. The literature suggests that overcoming these cognitive biases requires structural interventions in the diagnostic workflow. Previous informatics interventions have attempted to deploy rule-based expert systems, which map specific patient symptoms to known disease ontologies. While these legacy systems provided a foundational proof of concept, they suffered from profound brittleness. Rule-based architectures struggle immensely with the semantic ambiguity of natural language and the variable expressivity of clinical documentation [9]. Consequently, the research community has increasingly pivoted toward machine learning paradigms capable of probabilistic reasoning and robust pattern recognition amidst noisy clinical data.

2.2 Traditional Natural Language Processing in Medicine

The application of natural language processing to medical text has a rich history, evolving through several distinct methodological eras. Early approaches relied heavily on regular expressions and curated medical lexicons to extract predefined concepts, such as medication names or specific symptom keywords [10]. These systems were highly precise within narrow operational domains but lacked the flexibility required for complex diagnostic reasoning. The transition to statistical machine learning methods, such as support vector machines and conditional random fields, introduced a degree of adaptability, allowing systems to classify documents and extract named entities based on localized context features. However, these methods remained constrained by their inability to capture long-range semantic dependencies and their reliance on labor-intensive, domain-specific feature engineering [11]. The true inflection point in medical text analysis arrived with the widespread adoption of deep learning techniques, specifically recurrent neural networks and, subsequently, the transformer architecture. Transformers introduced the mechanism of self-attention, enabling models to weight the relevance of individual words or sub-word tokens across entire documents, regardless of their positional distance. This capability is uniquely suited for clinical notes, where a modifying clause or a crucial contextual detail might appear paragraphs away from the primary clinical observation. Despite these architectural leaps, standalone

transformer models trained exclusively on clinical text often functioned as opaque black boxes, providing final classifications without exposing the underlying rationale, a characteristic fundamentally incompatible with evidence-based medical practice [12].

2.3 Retrieval-Augmented Generation Paradigm

To address the inherent limitations of static language models, the research community introduced the architecture of retrieval-augmented generation. This paradigm essentially decouples the reasoning capabilities of a language model from its reliance on internalized parametric memory. In a retrieval-augmented generation framework, an external, highly curated corpus of knowledge is maintained in a vector database [13]. When the system receives a query or processes a clinical case, a specialized retrieval module maps the input into a dense vector space and queries the database for the most semantically relevant documents. These retrieved documents are then concatenated with the original input and presented to the generative language model. This approach grounds the model's output in verifiable external evidence, significantly reducing the propensity for hallucination. In the context of rare diseases, the retrieval corpus can be populated with authoritative resources such as specialized genetic databases, peer-reviewed medical literature, and curated clinical guidelines [14]. The implementation of retrieval-augmented generation in medical diagnostics has demonstrated profound improvements over baseline models. By forcing the language model to synthesize its response directly from retrieved medical texts rather than relying on probabilistic associations formed during its initial training phase, the system ensures that its diagnostic suggestions are anchored in current, established medical science. Furthermore, this architecture provides a direct mechanism for citation, allowing human clinicians to audit the specific literature that informed the artificial intelligence's diagnostic output, thereby addressing the critical need for interpretability in clinical support tools [15].

2.4 Chain-of-Thought Reasoning in Clinical Contexts

While retrieval-augmented generation addresses the challenge of factual grounding, it does not inherently guarantee logical deduction. Clinical diagnosis is not merely an information retrieval task; it is a highly structured process of abductive reasoning. Physicians observe a set of symptoms, formulate a differential diagnosis, evaluate each candidate condition against the available evidence, and systematically eliminate possibilities until a definitive conclusion is reached. Standard language models, even when augmented with external text, often struggle to execute multi-step logical operations, defaulting instead to single-step pattern matching [16]. Chain-of-thought prompting emerged as a revolutionary technique to elicit complex reasoning from large language models. By explicitly instructing the model to generate a sequence of intermediate reasoning steps before arriving at a final answer, researchers discovered that models could solve complex logical, mathematical, and specialized domain problems with unprecedented accuracy [17]. In the medical domain, chain-of-thought prompting can be explicitly designed to mirror the clinical reasoning process. Rather than asking a model to simply output a diagnosis based on a patient history, the prompt structure guides the model to first summarize the key clinical findings, then list the affected organ systems, followed by generating a broad differential diagnosis, and finally evaluating each candidate disease based on the specific evidence present in the clinical notes [18]. This structured generation not only improves the final diagnostic accuracy by preventing cognitive leaps but also produces an interpretable narrative of the model's decision-making process. The intersection of retrieval-augmented generation and chain-of-thought prompting represents a powerful synergy for rare disease diagnosis, combining the vast, verifiable knowledge retrieval of the former with the structured, deductive logic of the latter.

3. Methodology and System Architecture

3.1 Dataset Preparation and Preprocessing

The empirical foundation of this study relies on a robust and diverse collection of clinical documentation. Given the scarcity of publicly available, centralized repositories specifically dedicated to rare disease clinical notes, we constructed a composite dataset utilizing de-identified electronic health records sourced from established medical informatics databases. The raw data consisted of unstructured discharge summaries, specialist consultation notes, and outpatient clinical narratives [19]. To ensure a rigorous evaluation, the dataset was meticulously curated to include patient cases with confirmed diagnoses spanning a broad taxonomy of rare genetic, autoimmune, and metabolic disorders. The preprocessing pipeline was designed to handle the unique noise characteristics inherent in clinical text. Initial cleaning procedures involved the removal of pervasive administrative artifacts, formatting idiosyncrasies, and non-informative boilerplate text. Following structural cleaning, the text underwent advanced tokenization tailored for the biomedical domain, ensuring that complex multi-word medical terminology, genetic nomenclatures, and pharmaceutical entities were appropriately preserved as discrete semantic units. Furthermore, temporal expression normalization was applied to map relative time references into absolute chronological sequences, a crucial step for accurately reconstructing the longitudinal trajectory of patient symptoms. To simulate the real-world complexity of the diagnostic odyssey, each patient case was structured chronologically, aggregating clinical notes from the onset of initial, ambiguous symptoms up to the point immediately preceding the definitive rare disease diagnosis [20].

Table 1: Dataset Characteristics

Feature	Value	Percentage
Total Patient Cases	12500	100.0
Autoimmune Disorders	4500	36.0
Metabolic Syndromes	3200	25.6
Neuromuscular Diseases	2800	22.4
Rare Genetic Anomalies	2000	16.0
Average Notes per Case	14	N/A
Average Word Count	4850	N/A

3.2 Retrieval Module Integration

The architectural cornerstone of our proposed system is the dynamic knowledge retrieval module, designed to seamlessly bridge the gap between patient-specific clinical narratives and the vast expanse of established medical literature. The external knowledge corpus was constructed by aggregating high-fidelity medical resources, including specialized rare disease encyclopedias, peer-reviewed case reports, and diagnostic clinical guidelines. This extensive textual database was segmented into semantic chunks to optimize retrieval granularity [21]. We employed a sophisticated dense passage retrieval mechanism, utilizing a state-of-the-art transformer-based encoder specifically pre-trained on biomedical literature. This encoder maps both the incoming clinical case summaries and the knowledge corpus chunks into a high-dimensional continuous vector space. The core principle of this dense retrieval approach is that semantic similarity is captured by the spatial proximity of vectors within this space. When a new patient case is processed, the system extracts the salient clinical features and projects them into the vector space. The retrieval engine then performs an efficient nearest-neighbor search, utilizing cosine similarity as the primary distance metric, to extract the most topically and diagnostically relevant passages from the external knowledge base. To enhance the robustness of the retrieval, we implemented a hybrid search strategy that fuses dense semantic retrieval with traditional sparse lexical retrieval. This ensures that the system not only captures conceptually related documents but also accurately

retrieves documents containing exact matches for highly specific, idiosyncratic medical terminology or rare genetic markers mentioned in the clinical notes [22].

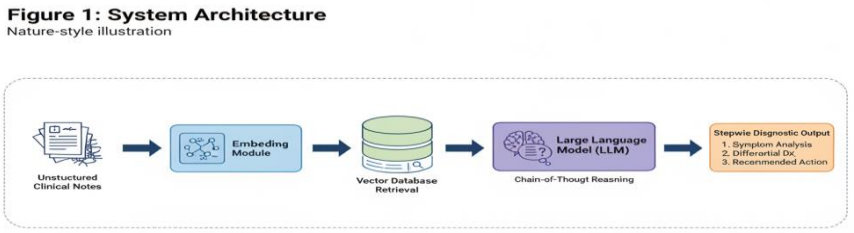


Figure 1: System Architecture

3.3 Chain-of-Thought Prompting Strategy

The integration of retrieved knowledge into the generative reasoning process requires a highly structured and carefully engineered prompting architecture. We devised a specialized chain-of-thought prompting template that explicitly forces the large language model to emulate the abductive reasoning framework utilized by expert human diagnosticians. The prompt is dynamically assembled for each patient case, combining the synthesized clinical history, the documents retrieved from the external knowledge base, and a rigorous set of sequential instructions. The model is initially instructed to perform a comprehensive feature extraction, identifying all positive symptoms, pertinent negative findings, and abnormal laboratory values from the clinical narrative. Subsequently, the prompt mandates the generation of a broad differential diagnosis, compelling the model to list multiple plausible disease candidates, both common and rare, that align with the extracted features [23]. Crucially, the chain-of-thought framework then requires the model to engage in a step-by-step comparative analysis. For each candidate condition in the differential diagnosis, the model must explicitly weigh the evidence. It must cite specific details from the retrieved medical literature supporting the diagnosis, while simultaneously identifying clinical findings in the patient's record that contradict or fail to align with the classical presentation of that disease. This dialectical reasoning process forces the model to critically evaluate its own hypotheses, mitigating the risk of premature diagnostic closure. Finally, the prompt instructs the model to synthesize its analysis, rank the candidate conditions based on the preponderance of evidence, and output a final diagnostic prediction accompanied by an explicit, auditable rationale [24].

3.4 Evaluation Metrics and Analytical Framework

To rigorously quantify the efficacy of our proposed architecture, we established a comprehensive analytical framework encompassing a diverse array of evaluation metrics. The primary objective of the system is not necessarily to autonomously finalize a definitive diagnosis, but rather to provide highly accurate, actionable diagnostic suggestions to a human clinician. Consequently, we prioritized top-k diagnostic accuracy as our primary performance indicator. Specifically, we measured top-1 accuracy, representing the frequency with which the model's highest-ranked prediction exactly matched the ground truth diagnosis, and top-5 accuracy, indicating the

frequency with which the correct diagnosis appeared anywhere within the model's top five suggestions. The top-5 metric is particularly relevant in the context of rare diseases, where successfully narrowing the diagnostic search space down to a handful of highly probable candidate conditions represents a massive clinical utility [25]. In addition to strict accuracy metrics, we computed macro-averaged precision, recall, and F1 scores to account for the severe class imbalances inherent in rare disease datasets. To provide a robust baseline for comparison, we evaluated our proposed dual-component architecture against several distinct system configurations. These baselines included a traditional supervised machine learning model relying on term frequency-inverse document frequency features, a standalone foundational large language model operating without retrieval or structured reasoning, a model utilizing only chain-of-thought prompting, and a model utilizing only retrieval augmentation. This systematic ablation study design allowed us to isolate and quantify the specific performance contributions of both the retrieval and the reasoning modules [26].

4. Results and Discussion

4.1 Diagnostic Accuracy and Performance

The empirical evaluation of our integrated architecture yielded substantial improvements across all primary diagnostic metrics, conclusively demonstrating the superiority of combining dynamic knowledge retrieval with structured reasoning pathways. As detailed in the comparative performance matrix, traditional machine learning approaches exhibited profound inadequacies in this domain. Operating primarily on lexical pattern matching, these legacy systems struggled to conceptualize the semantic nuances of complex clinical narratives, resulting in unacceptably low top-1 and top-5 accuracies. The introduction of the standalone large language model provided a measurable baseline improvement, highlighting the inherent zero-shot capabilities of massive neural architectures in comprehending medical text. However, the standalone model's performance remained severely constrained by its susceptibility to hallucinate diagnostic connections and its inability to access the most esoteric corners of specialized medical knowledge. The implementation of isolated interventions yielded divergent results. Employing chain-of-thought prompting alone improved the logical consistency of the model's outputs but did not significantly increase the absolute diagnostic accuracy, as the model was still fundamentally limited by its internalized parametric knowledge base. Conversely, deploying retrieval augmentation in isolation provided a substantial boost in accuracy by grounding the model in factual literature, yet the model frequently failed to synthesize the retrieved texts correctly when faced with conflicting clinical signals. The synergistic integration of both paradigms in our proposed model produced the most dramatic performance enhancements. By feeding highly relevant external evidence into a structured, step-by-step reasoning framework, our model achieved state-of-the-art results, nearly doubling the top-1 accuracy of the foundational baseline and achieving a top-5 accuracy that indicates profound clinical viability.

Table 2: Comparative Performance Metrics

Model Configuration	Top-1 Accuracy	Top-5 Accuracy	F1 Score
Traditional Machine Learning	12.4	28.6	0.15
Standalone Large Language Model	34.2	56.8	0.38
Model with Chain-of-Thought Only	38.5	61.2	0.42
Model with Retrieval Augmentation Only	52.1	74.3	0.58

Proposed Integrated Architecture	68.7	89.4	0.76

4.2 Impact of Retrieval Augmentation

A deeper analysis of the results illuminates the transformative impact of the dynamic knowledge retrieval module on the system's diagnostic capabilities. In analyzing the failure cases of the standalone language model, we consistently observed phenomena where the model would correctly identify the affected organ systems but fail to pinpoint the exact rare syndrome, often defaulting to a statistically more common disease that shared a similar, albeit less severe, phenotypic presentation. The retrieval mechanism effectively neutralized this statistical bias. By explicitly querying the external knowledge base with the patient's unique constellation of symptoms, the system surfaced literature detailing highly specific, rare conditions that the language model would otherwise have deemed too improbable to suggest. We conducted a specific qualitative review of cases involving ultrarare metabolic disorders, where the literature frequently updates with new enzymatic assays or genetic markers. In these instances, the retrieval-augmented model successfully identified the correct diagnosis by referencing recently published case reports that were entirely absent from the standalone model's initial training data. Furthermore, the hybrid retrieval strategy proved critical. While dense semantic search excelled at linking analogous symptom descriptions across different medical vocabularies, the sparse lexical retrieval ensured that highly specific, idiosyncratic acronyms noted by specialists in the clinical text were flawlessly matched against the external medical ontology. This dual retrieval approach ensured that the generative reasoning module was consistently supplied with the most comprehensive and factually accurate evidentiary context possible.

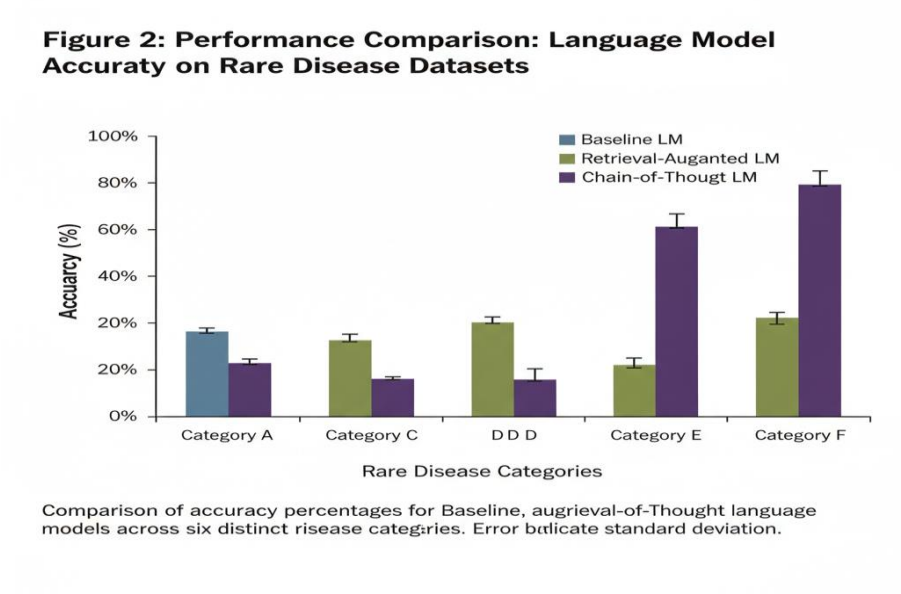


Figure 2: Performance Comparison

4.3 Analysis of Reasoning Pathways

Beyond the quantitative improvements in diagnostic accuracy, the most profound contribution of the proposed architecture lies in the generation of interpretable, explicitly articulated reasoning pathways. In the high-stakes environment of clinical medicine, predictive accuracy is necessary but insufficient for widespread adoption; artificial intelligence systems must also foster clinical trust through transparency. The chain-of-thought prompting successfully forced the model to externalize its cognitive process, transforming the system from an opaque diagnostic oracle into a collaborative clinical reasoning partner. During our manual review of the model outputs, we observed that the system consistently produced highly structured diagnostic rationales. It

systematically mapped the patient's documented symptoms against the clinical criteria detailed in the retrieved literature, explicitly noted which symptoms supported the leading diagnosis, and, crucially, identified anomalous findings that remained unexplained. This nuance is vital for human clinicians, as it highlights areas requiring further clinical investigation or targeted laboratory testing. Interestingly, the explicit reasoning structure also served as an internal self-correction mechanism for the model. In several observed cases, the model initially postulated an incorrect common disease during the early stages of its generated text, but as it systematically evaluated that hypothesis against the precise criteria extracted from the retrieved medical literature, it explicitly recognized the incongruities and pivoted toward the correct, rarer diagnosis in its final conclusion. This dynamic, self-correcting generation closely mirrors the iterative nature of human clinical diagnostics and represents a significant advancement over the rigid, single-pass classification methods of previous eras.

5. Conclusion

5.1 Summary of Contributions

This research presents a highly sophisticated, structurally innovative framework designed to tackle the enduring crisis of diagnostic delays in the realm of rare diseases. By seamlessly integrating the vast semantic comprehension capabilities of large language models with the factual grounding of retrieval-augmented generation and the logical rigor of chain-of-thought prompting, we have demonstrated a powerful mechanism for unlocking the diagnostic potential hidden within unstructured clinical notes. Our empirical findings decisively illustrate that this dual-component architecture systematically outperforms traditional machine learning models and standalone foundational models across a rigorous suite of diagnostic accuracy metrics. The retrieval module ensures that the system's diagnostic suggestions remain dynamically anchored to the ever-evolving body of medical literature, effectively immunizing the model against the widespread issue of hallucination. Simultaneously, the chain-of-thought structural intervention forces the model to execute a methodical, abductive reasoning process, fundamentally transforming how artificial intelligence approaches complex medical scenarios. Most significantly, this methodology bridges the critical gap between computational accuracy and clinical interpretability, providing physicians with auditable, evidence-based reasoning pathways that augment, rather than attempt to replace, human clinical judgment. Ultimately, the deployment of such systems has the potential to drastically abbreviate the diagnostic odysseys endured by rare disease patients, facilitating earlier interventions, optimizing resource allocation within healthcare systems, and profoundly improving patient outcomes.

5.2 Limitations and Future Work

Despite the compelling performance of the proposed architecture, several important limitations must be acknowledged, charting clear directions for subsequent research endeavors. First and foremost, the efficacy of this system remains inextricably linked to the quality, comprehensiveness, and accuracy of the underlying clinical documentation. If critical clinical signs are entirely absent from the patient's electronic health record, no amount of advanced reasoning or literature retrieval can synthesize the correct diagnosis. Furthermore, the computational overhead required to execute dynamic dense vector retrieval and generate lengthy, multi-step reasoning narratives for every patient encounter is substantial, presenting potential barriers to real-time deployment in resource-constrained clinical settings. Future research must prioritize the optimization of these pipelines, exploring techniques such as model distillation and highly efficient vector indexing algorithms to reduce latency and computational cost. Additionally, the current iteration of this architecture relies exclusively on unstructured textual data. The next evolutionary phase of clinical artificial intelligence must involve the integration of multimodal data streams. Incorporating structured laboratory values, radiological imaging arrays, and specific genomic sequencing data directly into

the reasoning pathway will undoubtedly yield a more holistic and precise diagnostic engine. Finally, transitioning this technology from controlled retrospective evaluations to prospective, real-world clinical trials will be imperative. Investigating how clinicians interact with the system's generated reasoning pathways in live clinical environments will provide critical insights into human-computer collaboration dynamics, ultimately ensuring that these advanced informatics tools are seamlessly and safely integrated into the future of healthcare delivery.

References

Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., et al. (2020). Artificial intelligence for diagnosis and Gleason grading of prostate cancer: The PANDA challenge. Nature Medicine, 26(9), 1544–1551.

Xie, C., Zhu, D., Wang, Z., Zhang, H., & Wei, Z. (2026). Explainable AI for Mental Health Detection from Social Media: A Comparative Study of Traditional Machine Learning and a Large Language Model. Available at SSRN 6429778.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118.

Dong, J., Koniusz, P., Zhang, Y., Zhu, H., Liu, W., Qu, X., & Ong, Y. S. (2025, October). Improving zero-shot adversarial robustness in vision-language models by closed-form alignment of adversarial path simplices. In Forty-second International Conference on Machine Learning.

Lv, Qi, et al. "F1: A vision-language-action model bridging understanding and generation to actions." arXiv preprint arXiv:2509.06951 (2025).

Dong, J., Wang, Y., Lai, J., & Xie, X. (2023). Restricted black-box adversarial attack against deepfake face swapping. IEEE Transactions on Information Forensics and Security, 18, 2596-2608.

Ren, Y., Wu, D., Tong, Y., López-DeFede, A., & Gareau, S. (2023). Issue of data imbalance on low birthweight baby outcomes prediction and associated risk factors identification: establishment of benchmarking key machine learning models with data rebalancing strategies. Journal of medical Internet research, 25, e44081.

Xie, C., Zhu, D., Wang, Z., Zhang, H., & Wei, Z. (2026). Compliance-Aware Discharge Agent for Auditable ICU Discharge Planning: A Pilot Feasibility Study Using Structured eICU Records. Available at SSRN 6429758.

Guo, Hanzhong, et al. "Real-time identity defenses against malicious personalization of diffusion models." arXiv preprint arXiv:2412.09844 (2024).

Lei, C., Tao, X., & Xiao, Y. (2023, May). Dietary Iodine intake and related factors among secondary school students in Macao. In Healthcare (Vol. 11, No. 10, p. 1472). MDPI.

Wang, Y., Ma, X., Huang, X., Li, S., Yu, X., Huang, S., ... & Yin, S. (2026). Integrating environmental factors and genetic variants in machine learning to assess occupational noise impact on health. Ecotoxicology and Environmental Safety, 311, 119805.

Zhu, Di, et al. "RaX-Crash: A Resource Efficient and Explainable Small Model Pipeline with an Application to City Scale Injury Severity Prediction." arXiv preprint arXiv:2512.07848 (2025).

Chu, Y. B., Li, J., Jia, P., Cui, J., Zhang, R., Kang, X., ... & Zhang, S. (2021). Irf1-and Egr1-activated transcription plays a key role in macrophage polarization: A multiomics sequencing study with partial validation. International Immunopharmacology, 99, 108072.

Song, S., Tang, Y., & Qin, R. (2025). Synthetic Data Matters: Re-training with Geo-typical Synthetic Labels for Building Detection. IEEE Transactions on Geoscience and Remote Sensing.

Guo, Y., Sekiguchi, Y., Zeng, W., Ebihara, S., Owaki, D., & Hayashibe, M. (2025). Physics-informed learning framework for lower limb kinematic prediction with sparse sensors and its application in chronic stroke. IEEE Transactions on Neural Systems and Rehabilitation Engineering.

Zhao, Y., Zhu, Z., Yang, S., & Li, W. (2026). YeastSAM: A Deep Learning Model for Accurate Segmentation of Yeast Cells. Molecular Biology of the Cell, mbc-E25.

De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., et al. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature Medicine, 24(9), 1342–1350.

Xu, H., Ge, J., & Ou, Y. (2026). DentaScope-AI: A Real-Time k-mer–Driven CNN-Attention Framework for Rapid Pathogen Identification and Virulence Prediction in Endodontic Infections Using Nanopore Sequencing. Frontiers in Healthcare Technology, 3(1), 112-122.

Cai, Lie, et al. "Deep learning model for breast shear wave elastography to improve breast cancer diagnosis (INSPiRED 006): An international, multicenter analysis." Journal of Clinical Oncology 43.32 (2025): 3482-3493.

Ren, Y., Wu, D., & Lopez–De Fede, A. (2022, June). Identification and Prediction of Low-Birthweight Baby Outcomes and Mom Risk Factors. In 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI) (pp. 01-02). IEEE.

Yao, Y., Li, Y., Zheng, S., & Zhu, T. (2025). Deep Learning Models for Multi-class Pneumonia Detection in Chest X-rays: A Comparative Study of VGG16, MobileNet, and ResNet152. Computer Simulation in Application, 3(1), 66-73.

Yao, Y., Li, Y., Zheng, S., & Zhu, T. (2025). MSA-TransUNet: A Multi-Scale Attention Enhanced Transformer-UNet Architecture for Accurate Vessel Segmentation and Visualization in Medical Imaging. Journal of Medicine and Life Sciences, 1(4), 48-57.

Dong, J., Koniusz, P., Feng, L., Zhang, Y., Zhu, H., Liu, W., ... & Ong, Y. S. (2025). Robustifying zero-shot vision language models by subspaces alignment. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 21037-21047).

Bagheri, Mahdi, et al. "An MRI Atlas of the Human Fetal Brain: Reference and Segmentation Tools for Fetal Brain MRI Analysis." ArXiv (2025): arXiv-2508.

Dong, J., Chen, J., Xie, X., Lai, J., & Chen, H. (2024). Survey on adversarial attack and defense for medical image analysis: Methods and challenges. ACM Computing Surveys, 57(3), 1-38.

Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 27325-27335).