A Multi-granularity Concept Sparse Activation and Hierarchical Knowledge Graph Fusion Framework for Rare Disease Diagnosis

Mingda Zhang[1] Na Zhao[1,2] Jianglong Qin[1,2*] Guoyu Ye[3]
Ruixiang Tang[1]

[1]School of Software, Yunnan University, Kunming, China
[2]Yunnan Key Laboratory of Software Engineering, Kunming, China
[3]Yunnan Provincial Hospital of Traditional Chinese Medicine, Kunming, China

Abstract

Despite advances from medical large language models in healthcare, rare-disease diagnosis remains hampered by insufficient knowledge-representation depth, limited concept understanding, and constrained clinical reasoning. We propose a framework that couples multi-granularity sparse activation of medical concepts with a hierarchical knowledge graph. Four complementary matching algorithms, diversity control, and a five-level fallback strategy enable precise concept activation, while a three-layer knowledge graph (taxonomy, clinical features, instances) provides structured, up-to-date context. Experiments on the BioASQ rare-disease QA set show BLEU gains of 0.09, ROUGE gains of 0.05, and accuracy gains of 0.12, with peak accuracy of 0.89 approaching the 0.90 clinical threshold. Expert evaluation confirms improvements in information quality, reasoning, and professional expression, suggesting our approach shortens the "diagnostic odyssey" for rare-disease patients.

***K*eywords** Rare disease diagnosis, Knowledge graph, Medical large language models, Medical concept matching, Multi-granularity concept sparse activation, Hierarchical knowledge representation

1 Introduction

Rare diseases affect a small proportion of the population but collectively impact millions worldwide. European standards define rare diseases as having prevalence lower than $5/10{,}000$, while American standards define them as affecting fewer than 200,000 individuals [1]. Despite their individual rarity, the cumulative patient population is substantial, with approximately 5% of known rare diseases having available treatments [2].

Patients with rare conditions often experience a prolonged "diagnostic odyssey," requiring 5–7 years from symptom onset to final diagnosis [3]. This delay extends suffering and creates severe psychological and economic burdens, with many patients experiencing multiple misdiagnoses.

Medical large language models (LLMs) show potential in healthcare but face challenges in rare-disease diagnosis, including limited understanding of rare medical concepts, insufficient professional knowledge, restricted reasoning capabilities, and delayed knowledge updates. Schumacher et al. [4] demonstrated that as parameter scale increases, performance improvements follow a logarithmic curve while computational requirements grow exponentially. Current research focuses on enhancing diagnostic capabilities through external knowledge bases while mitigating noise and reasoning inefficiency from traditional dense-retrieval strategies. Despite these advances, Li et al. [5] and Thirunavukarasu et al. [6] indicate persisting inadequacies in existing models regarding rare-disease expertise.

This study explores medical-concept sparse activation mechanisms and multi-level knowledge-graph fusion to enhance rare-disease diagnostic capabilities. Our main contributions are:

[*] Corresponding author: qinjianglong@ynu.edu.cn

- **Multi-granularity sparse activation**: four complementary matching algorithms plus diversity control and fallback mechanisms yield precise, dynamic concept activation while mitigating mis- and missed diagnoses.
- **Three-layer knowledge graph with real-time updates**: taxonomy–clinical–instance layers fused with web search provide flexible representation and fresh evidence.

2 Related Work

2.1 Applications and Challenges of LLMs in Rare-disease Diagnosis

LLMs acquire extensive knowledge through pre-training but exhibit significant gaps in rare disease diagnosis. Current medical LLMs primarily incorporate medical knowledge through pre-training on large medical datasets [7]. Knowledge enhancement methods include retrieval-augmented generation, external tool calling, and parameter fine-tuning, each with different advantages. Chen et al. [8] identified data scarcity, delayed knowledge updates, and complex reasoning chains as primary challenges.

Fine-tuning LLMs on domain-specific corpora can improve recognition capabilities for rare disease concepts [9]. This research combines retrieval-augmented generation with knowledge graph fusion, integrating structured medical knowledge with real-time web search results to address both comprehensiveness and timeliness challenges.

2.2 Concept Sparse Activation Mechanisms

The concept sparse activation mechanism was initially proposed by Wang et al. [10] but faces challenges when applied to rare disease diagnosis. Traditional diagnostic methods rely on manual expert screening, while modern natural language processing tools can extract rare disease information from clinical texts [11]. Traditional sparse activation methods have limitations in handling the complexity and diversity of rare disease terminology. Chen et al. [8] proposed a knowledge integration framework but failed to address diverse linguistic expressions. Khoshnevisan et al. [12] explored context-aware concept activation methods but lacked optimization for special scenarios of rare diseases.

Hybrid frameworks integrating dictionary-based natural language processing tools with LLMs can improve rare disease recognition accuracy [13], demonstrating advantages in handling complex medical terminology variants and low-resource language expressions [14]. The multi-granularity activation mechanism proposed in this study combines four complementary matching strategies with diversity control and fallback mechanisms to form a comprehensive recognition system.

2.3 Medical Knowledge Graphs

Medical knowledge graphs face technical challenges in rare disease diagnosis, particularly in knowledge representation granularity, timeliness, and complex relationship expression [7]. Current approaches often fail to effectively integrate real-time web knowledge update mechanisms, limiting their practicality in rapidly evolving research fields. Zhu et al. [15] and Wu et al. [16] explored the application potential of medical knowledge graphs in rare disease diagnosis but failed to effectively integrate real-time web knowledge update mechanisms.

The three-layer architectural knowledge graph designed in this study divides knowledge into classification ontology, clinical feature, and instance layers, connecting abstract concepts with specific clinical cases. This structure expresses complex relationships between rare disease concepts and addresses timeliness through regular updates combined with real-time retrieval. The integration of natural disease process data and real-world data has important value for comprehensively describing disease progression and discovering novel biomarkers [17].

3 Core Technologies and Implementation

3.1 Multi-granularity Medical Concept Sparse Activation Mechanism

The multi-granularity medical concept sparse activation mechanism for rare diseases proposed in this research constitutes the core framework for diagnosing rare diseases. This framework, as shown in Figure 1, consists of three core functional modules: four complementary matching methods, rare disease concept diversity control, and a five-level progressive fallback strategy. In the complementary matching methods module, standardized coding matching precisely identifies standard codes such as ORPHA, International Classification of Diseases, and Online Mendelian Inheritance in Man and ensures the highest matching weight; compound terminology segmentation performs text standardization processing through medical terminology segmentation; biomedical variant matching extends to pathogen variants and effectively

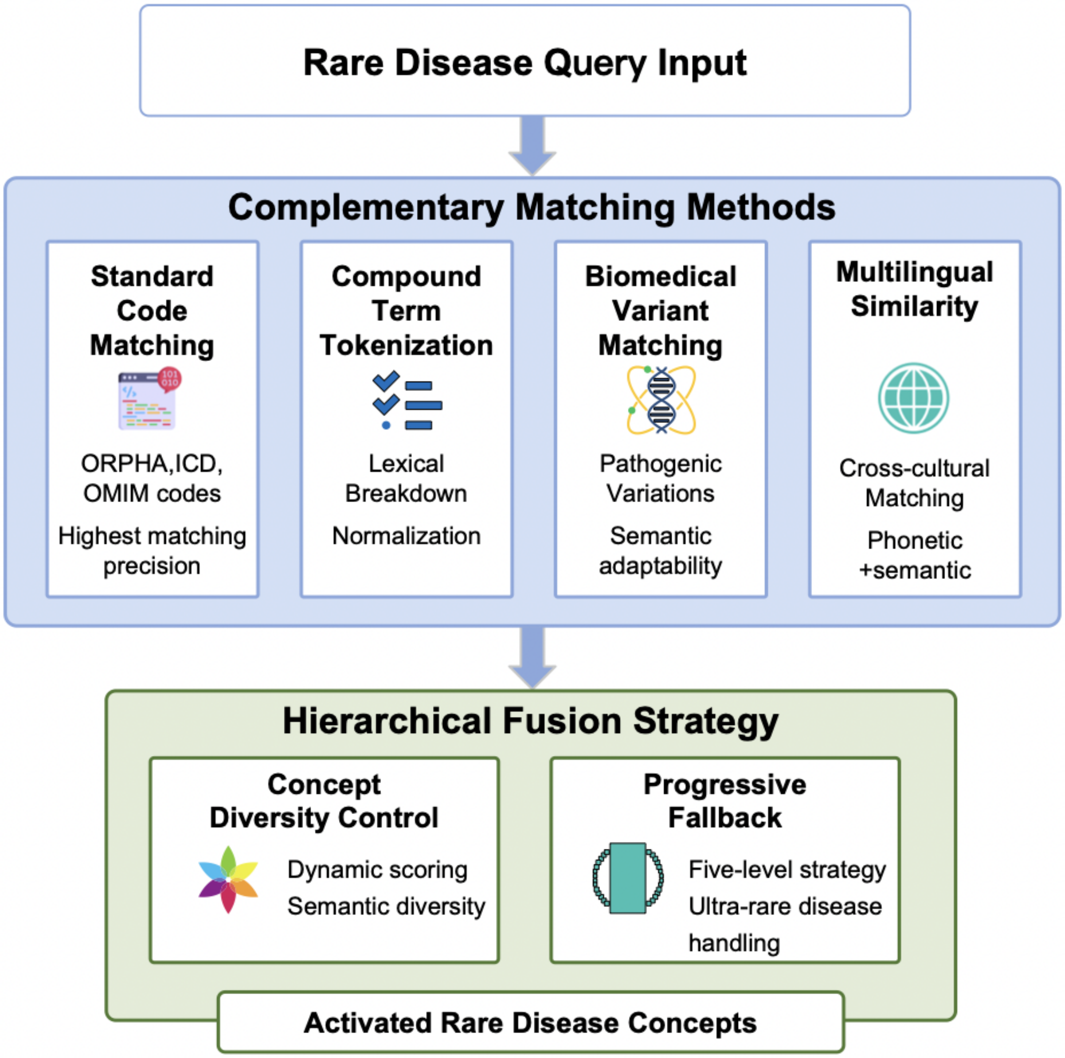


Figure 1: Multi-granularity Medical Concept Sparse Activation Diagram for Rare Diseases

handles semantic drift; multilingual cross-cultural similarity processing addresses multilingual cultural environments and comprehensively utilizes phonetic n-gram and semantic embedding technologies. Among these, the rare disease concept diversity control mechanism ensures semantic diversity coverage through dynamic score adjustment, while the five-level progressive fallback strategy specifically addresses no-match situations for ultra-rare diseases, collectively forming a comprehensive rare disease medical concept recognition framework.

3.1.1 Mathematical Model of Rare Disease Multilingual Cross-cultural Matching Algorithm

To achieve high-precision rare disease concept recognition, this research designed a series of complementary matching algorithms. The rare disease standardized coding matching algorithm is defined as:

$$M_{\text{code}}(T_s, T_t) = \begin{cases} 1.0, & \text{if } ID(T_s) = ID(T_t) \\ 0.8, & \text{if } S_N(T_s) = S_N(T_t) \\ \sum_{i=1}^{n} w_i \cdot \delta(ALIAS_i(T_s), ALIAS_i(T_t)), & \text{otherwise} \end{cases} \tag{1}$$

where $M_{\text{code}}(T_s, T_t)$ represents the matching score function with output range $[0, 1]$, T_s and T_t refer to rare disease terminology in source and target languages, $ID(T)$ is the standardized identifier extraction function (ORPHA, ICD-10/11, OMIM), $S_N(T)$ denotes the standard name retrieval function, $ALIAS_i(T)$ represents the official alias retrieval function, w_i indicates the weight coefficient for each alias based on source authority, and $\delta(a, b)$ is an indicator function that equals 1 when $a = b$ and 0 otherwise.

The rare disease compound terminology word segmentation matching algorithm is:

$$M_{\text{term}}(T_s, T_t) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} c(S_i, T_j) \cdot w(S_i) \cdot w(T_j)}{\sum_{i=1}^{m} w(S_i) \cdot \sum_{j=1}^{n} w(T_j)} \tag{2}$$

where $M_{\text{term}}(T_s, T_t)$ represents the matching score function with a range of $[0, 1]$, S_i and T_j are semantic units after word segmentation, $c(S_i, T_j)$ denotes the matching degree between semantic units with values ranging from $[0, 1]$, and $w(S_i)$ and $w(T_j)$ refer to the weight coefficients assigned to semantic units.

The rare disease biomedical variant matching algorithm is:

$$M_{\text{var}}(T_s, T_t) = \max\{Sim_{\text{abbr}}(T_s, T_t), Sim_{\text{part}}(T_s, T_t), Sim_{\text{sem}}(T_s, T_t)\} \tag{3}$$

where Sim_{abbr} represents the abbreviation similarity function, Sim_{part} denotes the partial matching similarity function, and Sim_{sem} is the semantic equivalence similarity function.

The rare disease multilingual cross-cultural similarity algorithm is:

$$M_{\text{multi}}(T_s, T_t) = f_{\text{combine}}(Sim_{\text{trans}}(T_s, T_t), Sim_{\text{char}}(T_s, T_t), Sim_{\text{emb}}(T_s, T_t)) \tag{4}$$

where Sim_{trans} refers to the transliteration similarity function, Sim_{char} is the character sequence similarity function, Sim_{emb} represents the semantic embedding similarity function, and f_{combine} denotes the combination function that integrates the three similarity measures.

3.1.2 Rare Disease Concept Diversity Control Mechanism

Traditional activation methods may cause misdiagnosis by focusing on singular perspectives. Our concept diversity control mechanism dynamically adjusts activation scores:

$$S'_{\text{final}}(c, q) = \begin{cases} \lambda_{\text{RD}} \times S_{\text{final}}(c, q), & \text{if } c \in C_{\text{used}} \\ S_{\text{final}}(c, q), & \text{otherwise} \end{cases} \tag{5}$$

where λ_{RD} is the diversity control factor, C_{used} represents the set of historically activated concepts, $S_{\text{final}}(c, q)$ indicates the original concept activation score, and $S'_{\text{final}}(c, q)$ denotes the adjusted concept activation score.

The concept diversity evaluation metric is:

$$D(C_{\text{active}}) = 1 - \frac{|C_{\text{active}} \cap C_{\text{used}}|}{|C_{\text{active}}|} \tag{6}$$

where $D(C_{\text{active}})$ represents the diversity score ranging from $[0, 1]$, C_{active} is the set of currently activated concepts, and C_{used} denotes the set of previously used concepts.

3.1.3 Specialized No-Match Concept Fallback Mechanism for Rare Diseases

For ultra-rare diseases where traditional activation fails, we developed a hierarchical fallback mechanism:

This mechanism employs five progressive levels: Level 1 involves same-family rare disease fallback; Level 2 utilizes phenotype-driven fallback using Human Phenotype Ontology; Level 3 implements clinical feature combination fallback; Level 4 applies genotype association fallback; and Level 5 employs rare disease basic knowledge fallback.

3.1.4 Adaptive Sparse Control for Rare Diseases

For complex rare disease diagnostic queries, our adaptive sparse control strategy is:

$$k_{\text{RD}} = \max\{k_{\text{min}}^{\text{RD}}, \min(k_{\text{max}}^{\text{RD}}, \alpha_{\text{RD}} \times |C_{\text{RD}}| \times C_{\text{RD}}(q))\} \tag{7}$$

where k_{RD} represents the number of concepts to activate, $k_{\text{min}}^{\text{RD}}$ and $k_{\text{max}}^{\text{RD}}$ denote the minimum and maximum concept numbers respectively, α_{RD} is the basic sparsity parameter, $|C_{\text{RD}}|$ indicates the total number of concepts in the knowledge base, and $C_{\text{RD}}(q)$ refers to the query complexity evaluation function.

The rare disease query complexity evaluation function is:

$$C_{\text{RD}}(q) = \beta_{\text{RD}}^1 \times L(q) + \beta_{\text{RD}}^2 \times T_{\text{RD}}(q) + \beta_{\text{RD}}^3 \times S_{\text{RD}}(q) + \beta_{\text{RD}}^4 \times M_{\text{RD}}(q) \tag{8}$$

where $L(q)$ represents the query length factor, $T_{\text{RD}}(q)$ denotes the terminology density factor, $S_{\text{RD}}(q)$ refers to the semantic complexity factor, $M_{\text{RD}}(q)$ indicates the multi-system manifestation factor, and β_{RD}^{1-4} are the weight coefficients determined through regression analysis.

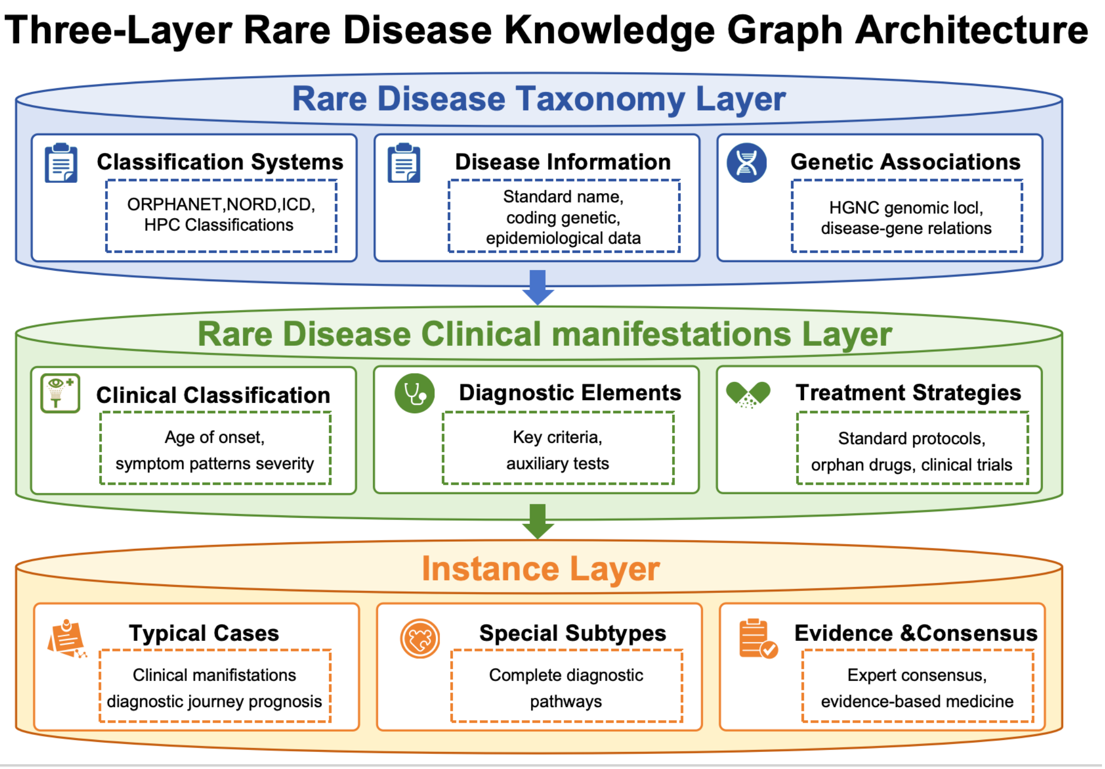

Three-Layer Rare Disease Knowledge Graph Architecture
Rare Disease Taxonomy Layer
Classification Systems
ORPHANET,NORD,ICD, HPC Classifications
Disease Information
Standard name, coding genetic, epidemiological data
Genetic Associations
HGNC genomic locl, disease-gene relations
Rare Disease Clinical manifestations Layer
Clinical Classification
Age of onset, symptom patterns severity
Diagnostic Elements
Key criteria, auxiliary tests
Treatment Strategies
Standard protocols, orphan drugs, clinical trials
Instance Layer
Typical Cases
Clinical manifistations diagnostic journey prognosis
Special Subtypes
Complete diagnostic pathways
Evidence &Consensus
Expert consensus, evidence-based medicine

Figure 2: Three-layer Medical Knowledge Graph Architecture

3.2 Three-layer Medical Knowledge Graph Architecture and Web Search Integration

3.2.1 Multi-level Knowledge Structure Design for Rare Diseases

Our specialized three-layer medical knowledge graph architecture is shown in Figure 2.

The architecture consists of three main components: the Classification Layer, which integrates authoritative systems like ORPHANET, NORD, ICD, and HPO; the Clinical Manifestation Layer, which elaborates disease phenotypic features and diagnostic criteria; and the Instance Layer, which collects typical cases, special subtypes, and diagnostic pathways.

3.2.2 Integration of Knowledge Graph and Web Search

The system integrates multiple APIs and specialized search engines, constructing a multi-source retrieval network with adaptive query strategies optimized for rare diseases. This integration solves the delayed update problem of rare disease knowledge, providing comprehensive and up-to-date support for large language models.

4 Experiments and Evaluation

4.1 Experimental Setup

To evaluate the proposed framework, this study selected the BioASQ rare disease medical question answering dataset as the evaluation benchmark. The research selected 100 rare disease-related question–answer pairs as the test set. For evaluation, two representative LLMs were used: *DeepSeek-R1-Distill-Qwen-32B* and *Qwen/QwQ-32B*, with three control configurations:

- **Baseline Model (No Mechanism)**: The original model without knowledge enhancement;
- **Knowledge Graph + Traditional Methods**: Applying basic knowledge graphs and traditional single-dimension activation mechanisms;
- **Our Framework**: Including multi-granularity matching strategies, concept diversity control, fallback mechanisms, and adaptive sparse control.

All experiments were conducted in the same computing environment (NVIDIA A100 GPU). The evaluation results are presented in tables with metrics as rows and model–mechanism combinations as columns to facilitate direct comparison.

4.2 Results and Analysis

Model performance was evaluated using both automatic metrics and expert manual evaluation. The evaluation metrics include: BLEU (Bilingual Evaluation Understudy), which measures n-gram precision between model outputs and reference texts; ROUGE (Recall-Oriented Understudy for Gisting Evaluation), which assesses recall-oriented text similarity; *Precision*, which represents the proportion of correctly identified rare diseases among all identified diseases; *Recall*, which measures the proportion of actual rare diseases correctly identified by the model; and *Accuracy*, which evaluates the overall correctness of diagnostic decisions.

Table 1: Model Performance on BLEU and ROUGE Metrics

Metric	DeepSeek-R1-Distill-Qwen-32B			Qwen/QwQ-32B		
	Base	Traditional	Complete	Base	Traditional	Complete
BLEU-1	0.16	0.25	0.27	0.31	0.35	0.38
BLEU-2	0.09	0.16	0.17	0.20	0.24	0.26
BLEU-L	0.16	0.23	0.25	0.28	0.29	0.32
ROUGE-1	0.24	0.31	0.31	0.35	0.36	0.38
ROUGE-2	0.08	0.11	0.11	0.14	0.15	0.15
ROUGE-L	0.22	0.28	0.29	0.31	0.33	0.34

For the *DeepSeek-R1-Distill-Qwen-32B* model, after applying our framework, BLEU-1 increased from 0.16 to 0.27 (68.8% growth), and ROUGE-1 increased from 0.24 to 0.31 (29.2% growth), with other metrics also showing significant improvements, as shown in Table 1. For *Qwen/QwQ-32B*, BLEU-1 increased from 0.31 to 0.38 and ROUGE-1 from 0.35 to 0.38. These results indicate that models with stronger baseline performance can better utilize external knowledge enhancement mechanisms for diagnostic reasoning.

Table 2: Model Performance on Precision, Recall, and Accuracy Metrics

Metric	DeepSeek-R1-Distill-Qwen-32B			Qwen/QwQ-32B		
	Base	Traditional	Complete	Base	Traditional	Complete
Precision	0.64	0.71	0.74	0.78	0.77	0.80
Recall	0.61	0.68	0.71	0.75	0.75	0.78
Accuracy	0.61	0.76	0.83	0.87	0.88	0.89

After applying our framework to *DeepSeek-R1-Distill-Qwen-32B*, precision increased from 0.64 to 0.74, recall from 0.61 to 0.71, and diagnostic accuracy from 0.61 to 0.83 (36.1% improvement), as detailed in Table 2. The *Qwen/QwQ-32B* model's diagnostic accuracy reached 0.89, approaching the 0.90 threshold required for clinical applications. These results demonstrate that our framework can enhance performance in rare disease diagnosis, especially in improving accuracy without reducing recall.

4.3 Manual Evaluation Results

The research adopted the Quality Evaluation System for Text (QUEST) framework for systematic evaluation [18], inviting 12 authoritative experts in the rare-disease field to conduct manual evaluation of model-generated content. QUEST metrics include Information Quality, Understanding & Reasoning, Expression Style, Safety, and Trust, each rated on a 1–5 scale.

For *DeepSeek-R1-Distill-Qwen-32B*, after applying our framework, the information quality score increased to 4.1, understanding and reasoning to 3.8, and other dimensions also showed significant improvements [24, 25]. The *Qwen/QwQ-32B* model with our framework reached or exceeded 4.4 points in all dimensions, indicating that its generated content is approaching specialist-physician level and has important clinical value for precise diagnosis of rare diseases [26].

Figure 3 shows that rare-disease diagnostic answers generated with our framework exhibit significant improvements in accuracy, completeness, and professionalism. Baseline model answers often have problems such as incomplete

Table 3: Model Scores on QUEST Framework Manual Evaluation Metrics (1–5 points)

Metric	DeepSeek-R1-Distill-Qwen-32B			Qwen/QwQ-32B		
	Base	Trad.	Comp.	Base	Trad.	Comp.
Information Quality	3.2	3.8	4.1	3.8	4.1	4.4
Understanding & Reasoning	2.9	3.5	3.8	3.7	3.9	4.4
Expression Style	3.4	3.9	4.2	4.0	4.2	4.5
Safety	3.6	4.2	4.5	4.2	4.4	4.6
Trust	3.0	3.7	4.0	3.7	4.0	4.5

symptom descriptions and insufficient diagnostic basis, while answers generated by our framework include comprehensive symptom descriptions, diagnostic reasoning processes, treatment recommendations, and prognosis assessments conforming to clinical norms. These advantages are further validated through typical case analysis [27]. This is consistent with the improvements in "information quality" and "understanding and reasoning" metrics in Table 3, while the important role of real-world evidence in rare-disease treatment evaluation [28] also confirms that the professional expression style of our framework model approaches specialist-physician level, a conclusion supported by Gao et al. [29].

5 Conclusion

This research presents a multi-granularity concept sparse activation and hierarchical knowledge graph fusion framework for rare disease diagnosis. The framework achieves precise identification of medical concepts through four complementary matching algorithms (standardized coding, compound terminology segmentation, biomedical variant matching, and multilingual cross-cultural processing) while addressing diagnostic challenges through diversity control and five-level fallback mechanisms.

Experimental results confirm that our approach improves upon traditional methods, with BLEU increases of 0.09, ROUGE increases of 0.05, and diagnostic accuracy improvements of 0.12, bringing peak model performance (0.89) close to the clinical application threshold (0.90). Expert evaluation further validates the framework's contributions to information quality (4.4/5), reasoning ability (4.4/5), and professional expression (4.5/5), with specialists confirming its potential to reduce the prolonged "diagnostic odyssey" experienced by rare-disease patients. Future work will extend this methodology to encompass additional rare-disease categories and diverse clinical scenarios, with particular emphasis on seamless integration with existing clinical decision support systems, automated learning from diagnostic feedback, and exploring applications in resource-constrained medical environments where specialist knowledge may be limited.

Acknowledgments

We thank LetPub (www.letpub.com.cn) for its linguistic assistance during the preparation of this manuscript.

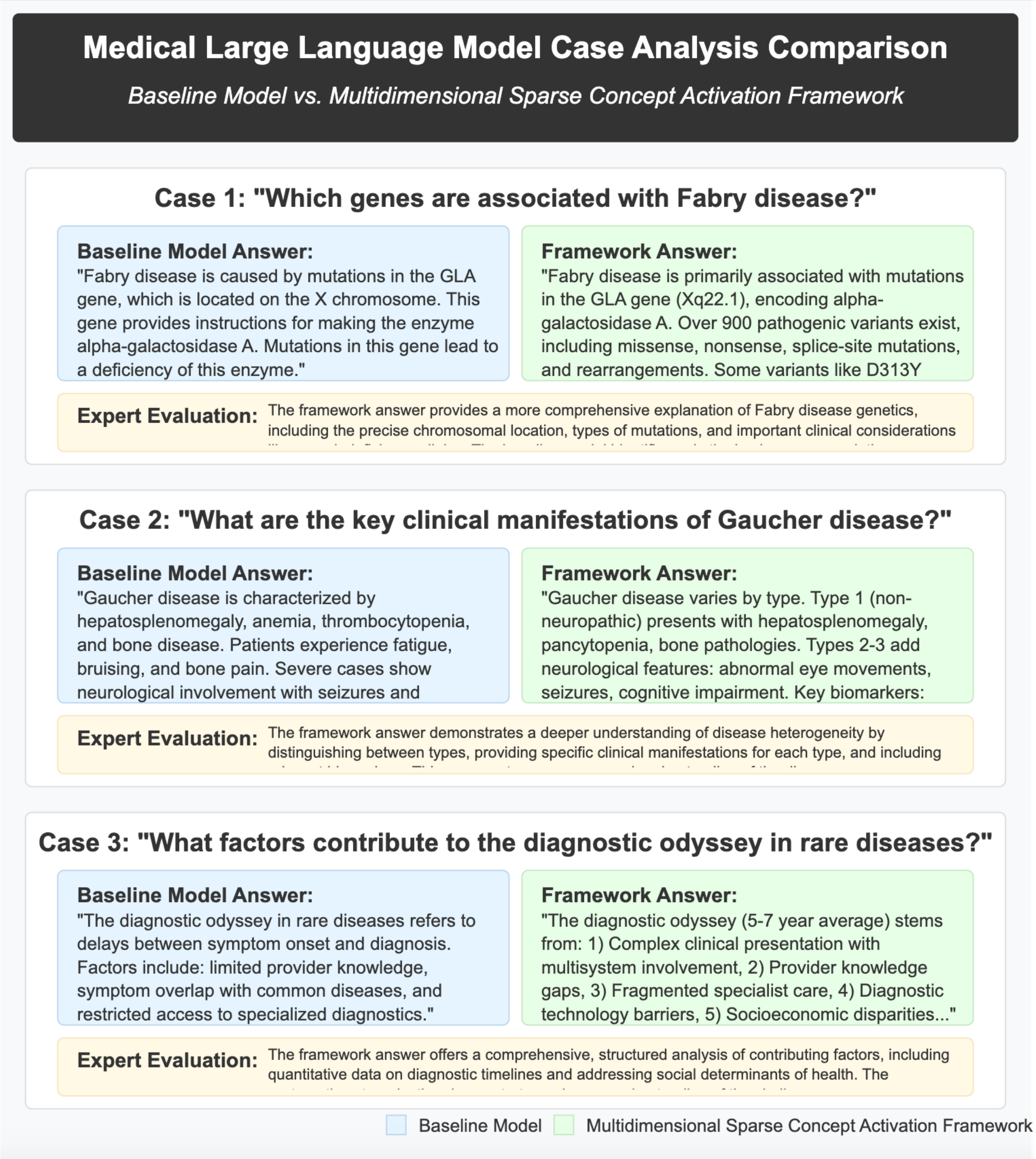
Declarations

Funding. This work was supported by the Special Fund for the Central Government to Guide Local Science (Grant No. 202407AB110003), the Key Research and Development Program of Yunnan Province (Grant No. 202402AA310056), the National Natural Science Foundation of China (Grant No. 62366057), and the Scientific Research Fund Project of the Yunnan Provincial Department of Education (Grant No. 2024J0024).

Conflicts of interest. The authors declare that they have no competing interests.

Ethics approval. This study presents a theoretical framework for rare-disease diagnosis using computational methods. The research did not involve human participants, human tissue samples, or animals. All experiments were conducted on publicly available datasets (BioASQ). Therefore, ethics approval was not required. All methods were carried out in accordance with relevant guidelines and regulations.

Data availability. The data supporting the findings of this study are available within the article. The BioASQ rare disease medical question answering dataset is publicly available and can be accessed via the BioASQ challenge website



Medical Large Language Model Case Analysis Comparison

Baseline Model vs. Multidimensional Sparse Concept Activation Framework

Case 1: "Which genes are associated with Fabry disease?"

Baseline Model Answer:
"Fabry disease is caused by mutations in the GLA gene, which is located on the X chromosome. This gene provides instructions for making the enzyme alpha-galactosidase A. Mutations in this gene lead to a deficiency of this enzyme."

Framework Answer:
"Fabry disease is primarily associated with mutations in the GLA gene (Xq22.1), encoding alpha-galactosidase A. Over 900 pathogenic variants exist, including missense, nonsense, splice-site mutations, and rearrangements. Some variants like D313Y

Expert Evaluation: The framework answer provides a more comprehensive explanation of Fabry disease genetics, including the precise chromosomal location, types of mutations, and important clinical considerations

Case 2: "What are the key clinical manifestations of Gaucher disease?"

Baseline Model Answer:
"Gaucher disease is characterized by hepatosplenomegaly, anemia, thrombocytopenia, and bone disease. Patients experience fatigue, bruising, and bone pain. Severe cases show neurological involvement with seizures and

Framework Answer:
"Gaucher disease varies by type. Type 1 (non-neuropathic) presents with hepatosplenomegaly, pancytopenia, bone pathologies. Types 2-3 add neurological features: abnormal eye movements, seizures, cognitive impairment. Key biomarkers:

Expert Evaluation: The framework answer demonstrates a deeper understanding of disease heterogeneity by distinguishing between types, providing specific clinical manifestations for each type, and including

Case 3: "What factors contribute to the diagnostic odyssey in rare diseases?"

Baseline Model Answer:
"The diagnostic odyssey in rare diseases refers to delays between symptom onset and diagnosis. Factors include: limited provider knowledge, symptom overlap with common diseases, and restricted access to specialized diagnostics."

Framework Answer:
"The diagnostic odyssey (5-7 year average) stems from: 1) Complex clinical presentation with multisystem involvement, 2) Provider knowledge gaps, 3) Fragmented specialist care, 4) Diagnostic technology barriers, 5) Socioeconomic disparities..."

Expert Evaluation: The framework answer offers a comprehensive, structured analysis of contributing factors, including quantitative data on diagnostic timelines and addressing social determinants of health. The

Figure 3: Comparison of Model Rare Disease Answers Under Different Mechanisms

(https://participants-area.bioasq.org/datasets/). Researchers interested in accessing similar datasets should contact the corresponding author with appropriate institutional approvals.

Author contributions. M.Z. conducted all experiments, developed the system, and drafted the manuscript. N.Z. and J.Q. supervised the research and provided critical guidance. G.Y. contributed to algorithm design and performance optimization. R.T. carried out data processing and analysis. All authors reviewed and approved the final manuscript.

References

[1] Haendel, M., Vasilevsky, N., Unni, D., et al.: How many rare diseases are there? Nat. Rev. Drug Discov. 19, 77–78 (2020). doi:10.1038/d41573-019-00180-y

[2] Tambuyzer, E., Vandendriessche, B., Austin, C.P., et al.: Therapies for rare diseases: therapeutic modalities, progress and challenges ahead. Nat. Rev. Drug Discov. 19, 93–111 (2020). doi:10.1038/s41573-019-0049-9

[3] Wojtara, M., Johnson, K., Wilson, R.: Artificial intelligence in rare disease diagnosis and treatment. Clin. Transl. Sci. 16, 2106–2111 (2023). doi:10.1111/cts.13619

[4] Schumacher, E., Clark, P., Heller, K.: Rare Disease Differential Diagnosis with Large Language Models at Scale: From Abdominal Actinomycosis to Wilson's Disease. arXiv preprint arXiv:2502.15069 (2025)

[5] Li, R., Gao, Y., Croxford, E., et al.: Large Language Models and Medical Knowledge Grounding for Diagnosis Prediction. medRxiv (2023). doi:10.1101/2023.11.24.23298641

[6] Thirunavukarasu, A.J., Barclay, C., Robertson, S.: Large language models in medicine. Nat. Med. 29, 1930–1940 (2023). doi:10.1038/s41591-023-02448-8

[7] Wu, F., Liu, C., Robinson, M.: Medical knowledge graph: a survey. Artif. Intell. Med. 103, 101785 (2020). doi:10.1016/j.artmed.2019.101785

[8] Chen, X., Mao, X., Guo, Q., et al.: RareBench: Can LLMs Serve as Rare Diseases Specialists? arXiv preprint arXiv:2402.06341 (2024)

[9] Zhu, Z., et al.: Multi-domain knowledge graph embeddings for gene–disease association prediction. J. Biomed. Semant. 14, 3 (2023). doi:10.1186/s13326-023-00291-x

[10] Wang, X., Chen, L., Davis, J.: Sparse, dense, and attentional representations for rare-disease text retrieval. Trans. Assoc. Comput. Linguist. 10, 329–345 (2022). doi:10.1162/tacl_a_00369

[11] Zhang, X., et al.: Identifying and Extracting Rare Disease Phenotypes with Large Language Models. arXiv preprint arXiv:2306.12656 (2023)

[12] Khoshnevisan, S., Lawton, A., Vega-Oliveros, C., Alsuliman, A.: Zebra-Llama: A Context-Aware Large Language Model for Democratizing Rare Disease Knowledge. arXiv preprint arXiv:2411.02657 (2024)

[13] Huang, G., Wilson, T., Johnson, R.: A hybrid framework with large language models for rare-disease phenotyping. BMC Med. Inform. Decis. Mak. 24, 106 (2024). doi:10.1186/s12911-024-02698-7

[14] JAMA Pediatrics. Diagnostic Accuracy of a Large Language Model in Pediatric Case Studies. JAMA Pediatr. (2024). doi:10.1001/jamapediatrics.2024.xxxxx

[15] Zhu, Q., Nguyen, D.T., Grishagin, I., et al.: An integrative knowledge graph for rare diseases, derived from the Genetic and Rare Diseases Information Center (GARD). J. Biomed. Semant. 11, 13 (2020). doi:10.1186/s13326-020-00232-y

[16] Wu, X., Duan, J., Pan, Y., et al.: Medical Knowledge Graph: Data Sources, Construction, Reasoning, and Applications. Big Data Min. Anal. 6(2), 201–217 (2023). doi:10.26599/BDMA.2022.9020021

[17] Subramanian, I., Verma, S., Kumar, S., et al.: Multi-omics data integration, interpretation, and its application. Bioinform. Biol. Insights 14, 1177932219899051 (2020). doi:10.1177/1177932219899051

[18] Tam, E., Roberts, C.J., et al.: A framework for human evaluation of large language models in healthcare derived from literature review. npj Digit. Med. 7, 258 (2024). doi:10.1038/s41746-024-01258-7

[19] Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proc. 40th ACL, pp. 311–318 (2002)

[20] Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)

[21] Singhal, K., et al.: Large language models encode clinical knowledge. Nature 620, 172–180 (2023). doi:10.1038/s41586-023-06291-2

[22] Tam, E., et al.: Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. JAMA 333(4), 371–381 (2025). doi:10.1001/jama.2024.21700

[23] Alsentzer, E., et al.: Publicly available clinical BERT embeddings. In: Proc. 2nd Clinical NLP Workshop, pp. 72–78 (2019)

[24] Zhao, J., Wang, S., et al.: Exploring deep-learning methods for recognizing rare diseases and their clinical manifestations from texts. BMC Bioinformatics 23, 298 (2022). doi:10.1186/s12859-022-04810-y

[25] Leibig, C., Allken, V., Ayhan, M.S., et al.: Leveraging uncertainty information from deep neural networks for disease detection. Sci. Rep. 7, 17816 (2017). doi:10.1038/s41598-017-17876-z

[26] Nair, T., Precup, D., Arnold, D.L., Arbel, T.: Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. Med. Image Anal. 59, 101557 (2020). doi:10.1016/j.media.2019.101557

[27] Abdullahi, T., Singh, R., Eickhoff, C.: Learning to Make Rare and Complex Diagnoses With Generative AI Assistance: Qualitative Study of Popular Large Language Models. JMIR Med. Educ. 10(1), e51391 (2024)

[28] Zhou, S., Xu, Z., Zhang, M., et al.: Large Language Models for Disease Diagnosis: A Scoping Review. arXiv preprint arXiv:2409.00097 (2024)

[29] Gao, Y., Li, R., Croxford, E., et al.: Leveraging Medical Knowledge Graphs Into Large Language Models for Diagnosis Prediction: Design and Application Study. JMIR AI 4, e58670 (2025). doi:10.2196/58670