ChatASD: A Dialogue Framework for LLMs Enhanced by Autism Knowledge Graph Retrieval

Lei Chu[†]
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Shenzhen Guangdong China
l.chu@siat.ac.cn

Hongyan Wu[‡]
Shenzhen Institute of Advanced Technology
Chinese Academy of Sciences
Shenzhen Guangdong China
hy.wu@siat.ac.cn

Yi Pan[‡]
Shenzhen Institute of Advanced Technology
Shenzhen University of Advanced Technology
Shenzhen Guangdong China
panyi@suat-sz.edu.cn

ABSTRACT

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by developmental delays, communication difficulties, repetitive behaviors, and restricted interests. Large Language Models (LLMs) have demonstrated exceptional capabilities in various natural language tasks, particularly in providing personalized question-and-answer(Q&A) services, making them well-suited for constructing dialogue engines for autism Q&A systems. However, general LLMs often lack integrated autism knowledge during training, limiting their professional competency in autism consultation. Additionally, the automatic evaluation of scientific accuracy in autism medical knowledge Q&A remains underexplored. To address this gap, we propose ChatASD, an autism knowledge Q&A framework based on Graph Retrieval-Augmented Generation (GraphRAG) technology. This framework leverages LLMs and retrieves relevant information from medical literature to generate an autism knowledge graph, employing a combination of global and community queries to produce reliable responses. Compared to traditional methods, ChatASD effectively addresses the sparse distribution of autism knowledge, providing more accurate and comprehensive answersAutomatic efficacy evaluations and competitive experiments on system responses indicate our approach significantly improves reliability of autism-related professional knowledge queries.

CCS CONCEPTS

• **Information systems** → **Question answering;** • **Applied computing**→**Health informatics.**

KEYWORDS

Autism, Knowledge Graph, Retrieval-Augmented Generation, LLM, Question-and-Answer System

ACM Reference format:

Lei Chu, Hongyan Wu and Yi Pan. 2024. ChatASD: A Dialogue Framework for LLMs Enhanced by Autism Knowledge Graph Retrieval. In *Proceedings of The 15th ACM Conference on Bioinformatics, Computational Biology, and Health Informatic(ACM BCB). ACM, Shenzhen, Guangdong Province, P.R.China, 8 pages.* https://doi.org/10.1145/3698587.3701538.

[‡]Corresponding Authors

BCB '24, November 22–25, 2024, Shenzhen, China

ACM ISBN 979-8-4007-1302-6/24/11.
https://doi.org/10.1145/3698587.3701538

1 Introduction

Autism is a neurodevelopmental disorder that affects communication, learning, and behavior. It is characterized by difficulties in social interaction, restricted interests, and repetitive behaviors, impacting various aspects of life, including school and work. The United Nations continuously calls for greater understanding of autism and has designated April 2nd as World Autism Awareness Day to help autism patients obtain basic treatment and rehabilitation, ensuring their human rights and fundamental freedoms. However, the cutting-edge nature of autism medicine makes disseminating information challenging. Thus, a key research goal is to find new auxiliary treatment methods and enhance the dissemination of autism knowledge.

The emergence of LLMs such as ChatGPT and LLaMA has advanced medical AI research, providing advanced tools for researchers and acting as auxiliary tools and effective assistants throughout the disease treatment process. These advancements bring new perspectives to autism auxiliary treatment and knowledge dissemination. However, applying LLMs in the rigorous medical field has potential issues, such as hallucination, lack of domain-specific knowledge, and outdated information. Retrieval-Augmented Generation (RAG)[1] technology integrates and retrieves key information from knowledge bases, providing precise information support for LLMs to help generate more accurate and in-depth responses. Recent developments in RAG within the medical field focus on enhancing model output interpretability and reducing error rates in high-risk medical decisions. Although traditional RAG technology excels in explicit retrieval tasks and conceptual answers, its shortcomings in entity recognition and relationship extraction make it challenging to handle inductive, summarizing, or causal queries effectively. In the medical domain, there are often intricate connections between information, such as diseases and their symptoms, or medications and their side effects.

For autism patients and their families, professional medical institutions and special education facilities are costly and not

accessible to every family. Efficient, reliable, and interactive disease knowledge Q&A systems are an alternative, allowing patients and their families to obtain reliable, decision-supporting information at a lower cost. In autism information retrieval, the sparse distribution of knowledge information makes it difficult for traditional RAG methods to provide LLMs with prompts for summarizing or causal questions from a global perspective. For example, parents of affected children might be concerned about specific abnormalities in their child's brain structure and function. Traditional RAG systems might erroneously assert abnormalities occur in "altered brain connectivity patterns, abnormal brain development, or differences in brain volume," providing a retrieval basis. Traditional RAG-generated results are often partial and inaccurate, contradicting the expected medical facts: "Key structural abnormalities may occur in the thalamus, cerebellum, temporal lobe, and many other structures." This can lead to incorrect medical judgments due to incomplete information. Such issues often arise from the sparse distribution of related information, causing traditional RAG systems to only retrieve partial information for judgment, leading to new drawbacks—erroneous judgments based on incomplete, partial information.

This paper proposes a knowledge graph-based LLM retrieval-enhanced framework—ChatASD. The framework comprehensively collects autism-related medical literature, clinical guidelines, rehabilitation training guidance, and more, performs summary analysis, optimizes data distribution, and uses LLM to extract entities and relationships from text corpora to construct an autism knowledge graph. It pre-generates community summaries for all closely related entity groups. After user queries, the knowledge graph information enhances the LLM through an enhanced generation algorithm. Compared to traditional RAG methods that struggle to provide precise medical advice, our GraphRAG-based solution excels in targeted global searches within large, complex, and sparsely distributed autism corpora, accurately locating one or more related information sources. Specifically, the ChatASD system first performs a global query to check if community summaries (clusters of similar entity groups) already have specific descriptions of the question. If not, it calculates the correlation degree between different community key summaries and the query, locates the communities where related entities may exist, and sequentially performs community summary queries and specific entity queries. This approach significantly improves the comprehensiveness, accuracy, and professionalism of generated answers.

To further validate the effectiveness of the GraphRAG-based ChatASD framework, we invited a team of professional autism-related clinical doctors for quality assessment. Under their supervision, LLM learned relevant literature and treatment cases, designed multidimensional autism-related questions, and generated an initial set of questions. These were filtered by the team, and questions were classified and evaluated based on difficulty and relevance to autism, resulting in the autism-related knowledge Q&A datasets AQA-R (Autism Question and Answer - Regular) and AQA-P (Autism Question and Answer - Professional). Internal and external evaluation results consistently verified the superiority of the proposed ChatASD framework. Figure 1 shows the overall architecture of ChatASD.

Our contributions are as follows:

- To the best of our knowledge, our work is the first attempt to generate autism-related consultation Q&A dialogues based on LLMs. It is also the first practice of generating an autism knowledge graph using LLMs. This effectively expands the avenues for medical consultation for autism patients and reduces the economic burden on affected families.
- By using GraphRAG technology, we supplemented LLM with specific knowledge in the autism field. Benchmark tests prove that this approach's accuracy in answering autism-related questions surpasses the best existing general models, significantly reducing LLM hallucination risks and greatly enhancing performance.
- We have released the first public automatic evaluation benchmark for autism knowledge Q&A dialogues, including comprehensive evaluation metrics, datasets, and methods, whose professionalism has been evaluated by clinical doctors in the relevant field.

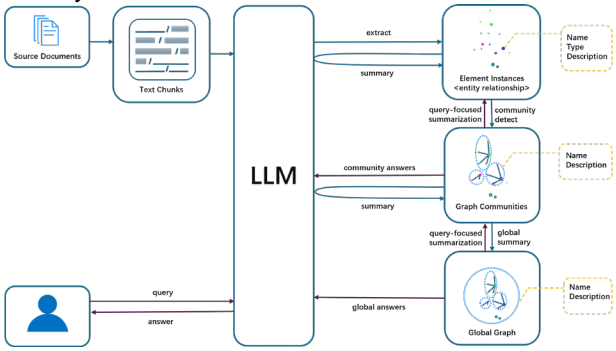

Figure 1: Architecture of proposed ChatASD framework

2 Related Work

2.1 Medical LLMs

The application of LLMs in dialogue generation and reconstruction has significantly improved medical consultations and patient interactions. For example, BioGPT leverages LLMs to enhance medical information retrieval, assist in clinical decision-making, and summarize medical literature. Similarly, Health-LLM incorporates domain-specific knowledge to boost the conversational capabilities of LLMs, offering personalized disease predictions and health advice based on patient interactions. These advancements highlight the critical role of LLMs in revolutionizing medical dialogue systems, where precision and reliability are essential.

In this work, we introduce LLMs into the autism knowledge Q&A domain, aiming to enhance the understanding of complex medical information and personalized queries, thus providing a more seamless and natural dialogue experience.

2.2 Application of RAG Technology in Medicine

In the medical field, RAG technology is crucial for enhancing the knowledge base of LLMs and minimizing error risks. RAG leverages verified medical databases and literature to ensure outputs are scientifically accurate. For instance, DISC-MedLLM employs RAG to align medical advice with the latest clinical guidelines and research findings. Similarly, MedRAG incorporates medical Q&A datasets and evaluates error risks through the MIRAGE assessment. This underscores the importance of augmenting LLMs with reliable external sources, particularly in high-stakes environments such as healthcare. The integration of these technologies advances AI applications in healthcare, improving safety and reliability in clinical and diagnostic settings.

In this study, we curated high-quality autism-related medical literature and guidelines to support an autism LLM dialogue system. This approach ensures the system delivers accurate and reliable information to researchers, healthcare professionals, patients, and their families.

2.3 Utilizing Knowledge Graphs to Assist LLMs

The integration of knowledge graphs (KGs) into LLMs structurally organizes factual knowledge, providing attributes that enhance LLMs' capabilities. There are two primary approaches in research combining KGs and LLMs: enhancing LLMs with KGs and employing LLMs to manage KG tasks.

The first approach enhances LLMs with KGs by integrating them into pre-training and inference phases. Examples include incorporating KGs into the objective function, LLM input, or additional fusion modules to optimize pre-training (e.g., GLM); using KGs during inference to update LLMs without retraining (e.g., RAG); and leveraging KGs to explain LLM learning and reasoning processes (e.g., MedLAMA).

The second approach utilizes LLMs for various KG tasks, such as embedding, completion, construction, graph-to-text generation, and answering. This involves encoding text descriptions of entities and relationships to enrich KG representations (e.g., KEPLER); improving KG completion by generating facts from text (e.g., KG-BERT); addressing entity discovery, coreference resolution, and relationship extraction in KG construction (e.g., JointGT); generating natural language descriptions of KG facts; and bridging natural language questions with KG-based answers using LLMs (e.g., QA-GNN).

In this study, we reviewed and summarized Graph-based RAG work[2] and implemented bidirectional reasoning collaboration between LLMs and KGs. At the data level, we utilized LLMs' text deconstruction capabilities to assist KGs in extracting entities and relationships from text corpora. At the inference level, we employed the constructed KG to provide LLMs with necessary related information through GraphRAG technology.

3 Methodology

In this paper, to equip LLMs with accurate professional domain knowledge related to ASD, we collected and processed publicly available or licensed autism medical literature, clinical guidelines, and rehabilitation training guidance from multiple credible data sources. Our analysis revealed that, compared to other well-established common diseases, autism research is more cutting-edge and involves more complex relationships between symptoms and causes. The research directions are also more diversified, covering fields such as behavioral science, psychology, genomics, and imaging, leading to a sparser distribution of professional terms in the semantic space. A single autism characteristic might result from several or even dozens of possible causes, necessitating a high-level understanding of the overall scientific knowledge related to autism for accurate answers. This poses a significant challenge to non-expert human responses and the answering engine of a Q&A system. The theory behind the combined reasoning of knowledge graphs and LLMs suggests that LLMs can assist in constructing knowledge graphs using domain-specific textual corpora in a higher-dimensional semantic space, thus achieving a comprehensive understanding of the target field.

Based on this foundation, we proposed a system framework named ChatASD, as shown in the Figure 1, consisting of two main components: the construction of an autism knowledge graph and the enhancement of LLM responses using GraphRAG. During the construction phase of the autism knowledge graph, we used GPT-4o as the reasoning model to deeply explore the semantic information behind the text, identify entities and extract relationships in the autism field, and build a layered autism knowledge graph. The GraphRAG component leverages the autism knowledge graph as the basis for associative retrieval, employing community search and localized search strategies to pinpoint and refine relevant information. This refined information is used as part of the prompts for LLM to generate professional, accurate, and comprehensive autism knowledge responses.

Additionally, we utilized graphical tools to achieve visual analysis of the autism knowledge graph, examining data sources from a graph structure perspective to further reveal the complex relationships and clustering characteristics between entities in the autism context, thereby increasing the reliability of analysis and evaluation results.

3.1 Data Collection

Internationally, the diagnostic criteria for autism are predominantly based on the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5-TR) [3] issued by the American Psychiatric Association (APA). Using this standard, we collected 241 autism-related articles from PubMed spanning from 1994 to 2024. Additionally, we included clinical guidelines, rehabilitation treatment guidance from professional institutions, and content from several highly credible websites in our factual knowledge dataset. The accuracy of this information was cross-verified through multiple sources of evidence and confirmed by clinical doctors.

3.2 Data Cleaning

To build a high-quality dataset, we enlisted clinical doctors to assist in filtering and selecting 186 autism-related medical articles from the initial text corpus. The filtering criteria excluded papers whose theories had been disproven, were severely outdated, or no longer met current standards. These curated documents covered various aspects of autism, including advancements in imaging, genomics, psychology, and behavioral science. Based on the primary focus of the literature, we categorized the documents into nine topics: overview, symptoms, diagnosis, causes, treatment, rehabilitation, prevention, prognosis, and other related issues. Some documents were assigned to multiple categories if they covered multiple topics. The classification results are detailed in the Appendix II.

3.3 Knowledge Graph Construction

After thoroughly collecting and filtering topic-related factual knowledge data on autism, we detail the core workflow of constructing the autism knowledge graph, outlining the high-level data flow and processes, and explaining the key design parameters, techniques, and implementation details at each step.

The specific steps are as follows. The example cases can be found in the Appendix II:

1. **Document Extraction**.

Given the significant differences in document formats and paragraph structures from different (and even the same) sources, we chose to use a lightweight tool called Marker to extract text content from PDF documents to minimize noise interference during information retrieval.

2. **Text Chunking**.

For the extracted text, we employed block segmentation techniques to decompose it into manageable text segments. According to HotPotQA research, selecting 600 tokens as the size of each text block effectively balances information concentration and processing speed. This block size not only covers a broader context to support more accurate understanding and answer generation but also avoids information loss or context overload issues due to overly large text blocks. Therefore, a block size of 600 tokens is considered an ideal compromise, ensuring sufficient contextual information without sacrificing processing efficiency.

3. **Element Instance Extraction**.

In this step, we used GPT-4o to identify and extract entities and their relationships from the source text blocks, including <name, type, description> tuples, ensuring high coverage and accuracy through multiple extraction rounds. To balance efficiency and quality, we used multiple "harvest rounds," up to a specified maximum, encouraging GPT-4o to detect any previously missed entities.

The extracted entities and relationships form the elements (nodes and edges) of the initial knowledge graph. For example, from the document "Early Intervention Strategies for Children with Autism," we might extract entities such as "autism," "early intervention," "behavioral therapy," and "speech therapy," and relationships like "early intervention includes behavioral therapy" and "speech therapy used for autism intervention."

4. **Graph Indexing and Community Detection.**

After extracting instances and relationships, we generated descriptive summaries for each entity and relationship using GPT-4 and converted them into descriptive texts for individual blocks. To eliminate duplicate entity elements and nodes, we further summarized matching groups using a language model. These summaries are semantically independent. For instance, for the "autism" entity, a descriptive summary might be "Autism is a neurodevelopmental disorder affecting social interaction, communication, and behavior patterns.".

Finally, using the Leiden algorithm[4], we divided the graph into several communities, each containing closely related nodes, to facilitate parallel processing and summary in subsequent steps. This allows for fine-grained analysis and evaluation.

5. **Generating Community Summaries.**

For the divided communities, we will instruct GPT-4o to generate detailed summaries in the form of reports. These summaries include descriptions of all entities and relationships within the community, providing a comprehensive understanding of the entire community. Community summaries can be used to answer global queries or can be browsed as individual topics. For each community, the summaries can be adapted to the context window size by initially adding element summaries and gradually replacing them with sub-community summaries.

6. **Knowledge Graph Generation.**

After consuming approximately 294,450 GPT-4o tokens, the construction of the autism knowledge graph was completed, resulting in a graph with 62,930 nodes and 103,593 edges. Filtering out nodes with a degree (sum of in-degree and out-degree) less than 10 (and connected to at least five other nodes), we retained a graph with 3,011 nodes (4.78%) and 12,946 edges (12.5%). The visualization results are shown below, with most nodes clustered around centers like "treatment/rehabilitation" (0.98%), "symptoms" (1.29%), and "causes" (1.87%). The largest community, "causes," includes sub-communities such as "behavioral science," "psychology," "genomics," and "imaging."

This distribution reveals that most entities are sparsely distributed across different locations and dimensions in the semantic space of autism. Such sparse distribution challenges traditional RAG systems in accurately locating highly relevant information for summarization, induction, or causal queries, preliminarily demonstrating the advantages of our GraphRAG work in the context of autism.

3.4 Retrieval-Augmented Generation

Combining the generated knowledge graph, the community summaries created in the previous step can be used in a multi-stage process to generate final answers when given a user query. The hierarchical structure of the community also allows users to select different levels of communities to answer their questions based on their needs. The system can decompose user inputs related to autism into multiple sub-questions, each targeting different sub-communities. For example, for the question "Are there specific abnormalities in brain structure and function in children with autism, and how are these abnormalities related to the symptoms of autism?" the system first performs a global query, retrieving symptom information already extracted from different

communities. If not found, it sequentially generates an initial answer list from summaries of three different related communities A: "Brain Structure Abnormalities," B: "Brain Function Abnormalities," and C: "Autism Symptoms." The system further integrates and summarizes these answers using LLM, generating a comprehensive, coherent global answer.

As shown in Table 1 in the Appendix I, we present examples of the case of user query and answer:

4 Evaluation

4.1 Dataset

There is no unified evaluation standard for question-and-answer dialogues using LLMs on the topic of autism. Existing evaluation frameworks for general RAG systems often focus on examining structured, independent indicators like contextual relevance achieved by retrievers, making it challenging to compare performance on specific topics with general LLMs. Given these limitations and to analyze ChatASD's performance in real autism consultation cases, we invited a team of professional autism-related clinical doctors and special education institutions for quality assessment. Under their supervision, we used GPT-4-ALL to learn relevant literature and treatment cases, coherently organizing and summarizing 687 autism-related questions. Based on the difficulty and relevance of the questions to autism, we developed evaluation standards and categorized the questions into two datasets: AQA-R (Autism Question and Answer - Regular) with 370 questions and AQA-P (Autism Question and Answer - Professional) with 317 questions. Compared to AQA-R, AQA-P contains more specialized questions, requiring a higher level of understanding of autism, making it suitable as the evaluation test set.

As shown in Table 3 in the Appendix I, examples of from the AQA-P question set are provided.

Under the supervision of these teams, GPT-4-ALL was used to study related literature and treatment cases, and the effective Q&A learned from these materials were coherently organized. The answers provided by professional doctors were used as standard answers. Subsequently, based on the difficulty and relevance to autism, questions were graded, and evaluation standards were established. This resulted in the creation of two autism-related knowledge Q&A datasets: AQA-R (Autism Question and Answer - Regular) and AQA-P (Autism Question and Answer - Professional). Compared to AQA-R, the questions in AQA-P are more specialized, consisting of questions that convey a higher level of understanding of autism topics, making it suitable as the evaluation test set for this study.

As shown in Table 2, there are examples of the AQA-P problems:

4.2 Evaluation Methodology

LLMs have proven to be excellent in natural language generation evaluations, often matching or surpassing human judgment[5, 6]. This method can generate reference-based metrics when known standard answers are available and can also measure the quality of generated text without references (e.g., fluency).

The Elo rating system is a method for calculating the relative skill levels of players, widely used in competitive games and sports events. The rating difference between two players can serve as a predictor of the match outcome. This system is highly suitable for our situation as we have multiple models that compete pairwise against each other. If the rating of model A is R_a and the rating of model B is R_b, then the exact probability formula for calculating the probability of model A winning using the base-10 logarithmic curve is:

$$E_A = \frac{1}{1+10^{(R_B-R_A)/400}}.$$

The model's match ratings can be linearly updated after each comparison. Suppose model A (with rating R_a is expected to score E_a points but scores S_a, the formula to update the rating of the model is:

$$R'_A = R_A + K \cdot (S_A - E_A).$$

For competitive output comparisons, using LLMs as the judge for Elo scores can closely approximate human expert judgment. Compared to traditional scoring, this approach significantly reduces the subjective randomness of LLM scoring, providing a more objective measure of different models' win rates in the context of autism.

In our evaluation, a neutral third-party model receives the questions, target metrics, and a pair of answers from different models. It is required to judge the quality of the answers based on the specified metrics and provide reasons for its judgment. If there is a clear difference in quality, the model returns the winner; otherwise, if the answers are essentially similar with negligible differences, it declares a tie. To minimize the randomness of LLMs, we repeat each comparison five times and use the average scores.

Considering the characteristics of autism-related questions, we evaluate the following models compatible with the GraphRAG API request format in four dimensions—accuracy, relevance, fidelity, and interactivity—using Elo scores to verify the differences in model capabilities. To facilitate the observation of these scores under a unified standard, we calculate the Z-Score to eliminate potential differences in distribution and range of Elo scores across dimensions. This standardization makes it easier to compare and understand each model's relative performance in different categories.

4.3 Evaluation Process

We automated the process of filling the AQA-P question set into prompts and used a recognized high-quality third-party neutral model as the judge model.

As shown in Table 3 in the Appendix I, we present examples of the evaluation of several LLMs using the AQA-P question set.

4.4 Experiment

Using the standardized platform of LMSYS Org (Large Model Systems Organization) Chatbot Arena, we randomly selected 120 questions from the AQA-P question set as a test set to conduct the above Elo tests on the top-ranked models as of July 2024: ChatGPT, Yi series models, including a total of five different-sized models.

To test the actual effect of the ChatASD framework, we pre-selected the medium-sized Yi model (Yi-medium) for ASD Knowledge Graph-based RAG and used another neutral open-source model, DeepSeek, as the LLM-as-a-judge model. This setup ensures that the base model used has reliable text understanding and answer generation capabilities for the generation task and provides a verifiable foundation to test the RAG capability based on the autism knowledge graph.

4.5 Case Study

Table 4 in the Appendix I is an example of a generated result for an autism-related question using the ChatASD framework with RAG based on the autism knowledge graph. Specifically, the question asks, "Are there specific abnormalities in brain structure and function in children with autism, and how are these abnormalities related to the symptoms of autism?" This high-level understanding question requires a comprehensive, summarizing, or causal answer, which cannot be directly obtained from a single specific document. The general nature of the question also implies that it cannot be answered by a simple description from a specific literature source. From a human perspective, this question is a trap due to the cutting-edge nature of autism research, requiring scientific judgment based on the latest findings and extensive reasoning to provide a comprehensive response. Our ChatASD framework queries text blocks related to specific brain structure abnormalities in autism from the knowledge graph (listing the locations of the nodes within the knowledge graph), and with the summarizing ability of the LLM, provides a reliable and specific description.

4.6 Result

The scoring and ranking showed a basic performance trend: GPT-4o, Yi-large, Yi-medium, Yi-spark, and GPT-3.5-turbo in descending order, aligning with the Chatbot Arena ranking, confirming the validity of this evaluation.

Experimental results indicate that before using our ChatASD framework enhancement, as shown in the Figure 2 and Table 5 in the Appendix I, the Yi-medium model's performance on autism-related questions across the four dimensions was not outstanding. Analysis of Yi-medium's answers revealed a tendency to provide direct results but lacked specific examples or detailed explanations. When asked to summarize and further reason, it tended to give general mechanism answers without in-depth explanations. This is likely due to insufficient pre-training knowledge on autism in the Yi series, leading to inadequate information for deeper reasoning. After enhancement with the ChatASD framework, the Yi-medium model showed significant improvements in accuracy, relevance, and fidelity, demonstrating the framework's reliable and stable performance in answering autism-related questions and effectively enhancing various capabilities of the model. This framework is plug-and-play and easy to use.

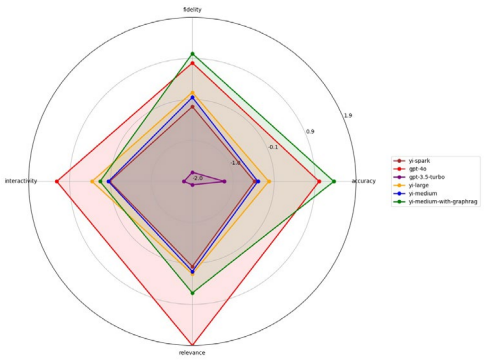

Figure 2: Evaluation Chart of Each Model's Performance

5 Conclusion

This study presents ChatASD, a novel framework utilizing knowledge graph-enhanced LLMs for autism Q&A systems. It involves data collection, knowledge graph construction, retrieval strategy, prompt engineering, and evaluation benchmarks. Experimental results show significant performance improvements in autism-related information retrieval. Practical tests confirm the system's accuracy and impartiality in providing disease information and treatment suggestions. Our findings offer crucial knowledge for autism patients and families, reducing medical expenses and conserving clinical resources. This innovation addresses issues like inaccessibility, inaccuracy, and lack of comprehensive information in autism care. We aim to inspire new trends in autism diagnosis and treatment through this work.

ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China (62273322 and U22A2041), Shenzhen Key Laboratory of Intelligent Bioinformatics (ZDSYS20220422103800001) and Shenzhen Science and Technology Program (CJGJZD20220517142000002).

REFERENCES

[1] Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., Yang, L., Zhang, W. and Cui, B. Retrieval-augmented generation for ai-generated content: A survey. arXiv preprint arXiv:2402.19473 (2024).

[2] Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S. and Larson, J. From local to global: A graph rag approach to query-focused summarization. arXiv preprint arXiv:2404.16130 (2024).

[3] APA APA Releases Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR). City, 2022.

[4] Traag, V. A., Waltman, L. and Van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. Scientific reports, 9, 1 (2019), 1-12.

[5] Wang, J., Liang, Y., Meng, F., Sun, Z., Shi, H., Li, Z., Xu, J., Qu, J. and Zhou, J. Is chatgpt a good nlg evaluator? a preliminary study. arXiv preprint arXiv:2303.04048 (2023).

[6] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D. and Xing, E. Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in Neural Information Processing Systems, 36 (2024).

Appendix I

Table 1: User Query and Answer

User Query	Community Answer	Global Answer
Are there specific brain abnormalities in children with autism, and how do these relate to autism symptoms?	1. Community A (Brain structure abnormalities): "Children with autism often have abnormal amygdala and hippocampus regions, affecting emotion and memory processing." 2. Community B (Brain function abnormalities): "Functional MRI shows atypical neural connectivity in children with autism, especially in social communication and sensory areas." 3. Community C (Autism symptoms): "Brain structure and function abnormalities are linked to autism's core symptoms: social difficulties, communication challenges, and repetitive behaviors."	Research shows children with autism have specific abnormalities in brain structure and function. Structural abnormalities occur in regions like the amygdala and hippocampus, which are vital for emotion and memory processing. Functional MRI studies reveal atypical neural connectivity patterns in autistic children, particularly in areas responsible for social communication and sensory processing. These brain abnormalities are closely related to autism's core symptoms: social difficulties, communication challenges, and repetitive behaviors. Understanding these neural differences provides insight into autism's biological underpinnings and highlights the importance of early diagnosis and intervention.

Table 2: Examples of AQA-P Problems

Category	Question
Causes	What are the pathogenic mechanisms of autism? What are the known genetic and environmental factors involved?
Diagnosis	What are the limitations in the accuracy and reliability of existing diagnostic tools and methods for autism? How can they be improved to enhance diagnostic efficiency?
Symptoms	What are the differences in symptom manifestation among individuals with autism at different ages? How do these symptoms develop over time?
Treatment	What ongoing clinical trials are evaluating the effectiveness and safety of new autism treatments?
Rehabilitation	What are the long-term impacts of rehabilitation interventions on individuals with autism? What factors influence the effectiveness of rehabilitation?
Prognosis	What is the prognosis for individuals with autism? What factors influence their long-term prognosis?
Prevention	Can the incidence of autism be reduced through early interventions or preventive measures? What are the potential preventive strategies?

Table 3: Examples of LLM Evaluation

Category	Content
Example Question	How do transcranial magnetic stimulation (TMS) and transcranial direct current stimulation (tDCS) work in treating autism? What mechanisms might they involve?
Generated Answer (yi-spark)	Transcranial magnetic stimulation (TMS) and transcranial direct current stimulation (tDCS) are non-invasive neuromodulation techniques that can affect brain activity through the skull. These techniques show potential in treating various neurological and psychiatric disorders, including ASD. TMS and tDCS might work by modulating neurotransmission, enhancing neuroplasticity, improving cognitive function, and alleviating symptoms. However, the effectiveness of these techniques in treating autism still requires more clinical trials for validation.
Generated Answer (gpt-4o)	Transcranial magnetic stimulation (TMS) and transcranial direct current stimulation (tDCS) as neuromodulation techniques are increasingly considered in the treatment of ASD. These techniques non-invasively modulate brain activity and might potentially improve certain symptoms in ASD patients. TMS and tDCS might function by regulating neuronal activity in specific cortical areas, enhancing neuroplasticity, improving cognitive functions, and alleviating symptoms.
LLM Decision	
Comprehensiveness	Winner: gpt-4o
Details for Comprehensiveness	yi-spark: Provides a broad overview of TMS and tDCS, mentioning their potential mechanisms and the need for further clinical trials. gpt-4o: Offers a detailed account of how TMS and tDCS might function, focusing on specific mechanisms like neuron activity regulation and neural plasticity enhancement.
Diversity	Winner: gpt-4o
Details for Diversity	yi-spark: Lists general mechanisms but lacks depth in explaining each mechanism. gpt-4o: Provides a more detailed and varied explanation of potential benefits and mechanisms.
Empowerment	Winner: gpt-4o
Details for Empowerment	yi-spark: Provides a general understanding of TMS and tDCS applications and emphasizes the need for further research. gpt-4o: Gives a structured and detailed explanation, empowering the reader with in-depth knowledge.
Directness	Winner: yi-spark

Details for	yi-spark: Directly lists mechanisms but lacks specific examples or detailed explanations.
Directness	gpt-4o: More detailed, which may reduce directness but increases informativeness.
Overall Verdict	gpt-4o is the preferred model for its detailed and informative answer.

Table 4: Examples of GraphRAG Response

Category	Content
Example Question	Are there specific brain abnormalities in children with autism, and how do these relate to autism symptoms?
Generated Answer (GraphRAG)	Children with autism show significant brain structure and function abnormalities. These are linked to symptoms like impaired social communication, repetitive behaviors, and cognitive dysfunction.
Structural Abnormalities	**1**. Brain Overgrowth and Cortical Development: Children with autism often exhibit brain overgrowth in the first year and structural differences in cortical development through early childhood [Data: Reports (3465, 3464, 3535, 4631, 2030, 369)]. **2**. Cerebellum: Shows decreased mitochondrial ETC complex gene expression, significant methylation differences, increased early growth, larger volumes in high-risk infants, and decreased gray matter volume (GMV) in some subregions. These link to motor control, coordination, and cognitive functions [Data: Reports (2123, 6302, 4517)]. **3**. Cortical Surface Area: Increased growth rate in the first year and significant expansion in regions like the left and right middle occipital gyrus, right cuneus, and right lingual gyrus [Data: Reports (3702)]. **4**. Thalamus: Reduced mitochondrial ETC gene expression, decreased choline-containing compounds, and significant cardiovascular alterations are noted in ASD patients [Data: Reports (2338)]. **5**. Temporal Lobe: Abnormalities include changes in proteins regulating mitochondrial dynamics, increased DNA oxidation, oxidative damage to lipids and proteins, and reduced enzyme activity controlling reactive oxygen species. Neuropathologic studies show cortical alterations and white matter microstructural disorganization [Data: Reports (2262, 2131)]. **6**. Amygdala: Often enlarged with unique growth trajectories and atypical dendritic spine density changes, linked to greater sensory sensitivity and socioemotional processing issues [Data: Reports (2122, 3492)]. **7**. Corpus Callosum: Abnormalities include lower white matter integrity and higher diffusion coefficients, correlated with cognitive and emotional performance [Data: Reports (2341, 2542, 782)]. **8**. Basal Ganglia: Observations pending completion.
Functional Abnormalities	**1**. Neurotransmission: GABA and glutamate abnormalities are linked to ASD. GABA-related genes are often under-expressed, leading to decreased inhibitory neurotransmitter levels, increased neuronal excitability, and seizures. Excess glutamate disrupts excitatory-to-inhibitory balance, causing seizures and cognitive dysfunction [Data: Reports (2121)]. **2**. Functional Connectivity: Children with ASD show different brain activity patterns during visual emotional cognition tasks, highlighting impairments in social, language, and cognitive functioning [Data: Reports (2184, 533, 3731)]. **3**. Resting-State Functional Connectivity: Abnormalities in resting-state connectivity are observed, crucial for understanding autism's neural mechanisms [Data: Reports (533, 3731)]. **4**. Default Mode Network (DMN): Differences in DMN connectivity between ASD and neurotypical individuals provide insights into the neural mechanisms underlying ASD [Data: Reports (3726)].
Implications for Symptoms	These structural and functional abnormalities are closely linked to autism symptoms. For instance, early brain overgrowth and cortical surface area differences are tied to early diagnosis and intervention, potentially improving outcomes [Data: Reports (3465, 3702)]. Cerebellum and basal ganglia abnormalities relate to motor control, coordination, and repetitive behaviors common in ASD [Data: Reports (2123, 2125)]. Functional connectivity issues, especially in the DMN and during visual emotional cognition tasks, are connected to social, language, and cognitive impairments [Data: Reports (2184, 3726)].
Summary	Specific abnormalities in brain structure and function in children with autism are extensive and varied, affecting multiple brain regions and systems. These abnormalities are closely linked to core autism symptoms, offering valuable insights into the disorder's neurobiological underpinnings and guiding future diagnostic and therapeutic strategies.

Table 5: Elo Scores and Z-Score

Model	Accuracy Elo Score	Accuracy Z-Score	Fidelity Elo Score	Fidelity Z-Score	Interactivity Elo Score	Interactivity Z-Score	Relevance Elo Score	Relevance Z-Score
gpt-3.5-turbo	1266.38	-1.24	1267.76	-1.78	1068.79	-1.8	1063.53	-1.92
gpt-4o	1639.48	1.02	1589.1	0.82	**1755.76**	**1.23**	**1679.12**	**1.9**
yi-spark	1384.26	-0.52	1459.81	-0.22	1469.69	-0.03	1487.28	0.02
yi-medium	1397.43	-0.44	1487.64	0	1477.88	0	1513.26	0.14
yi-large	1439.32	-0.18	1503.31	0.12	1564.98	0.39	1526.85	0.2
chatasd	**1696.14**	**1.37**	**1616.41**	**1.04**	1522.86	0.2	1624.39	0.65