Adaptive Knowledge Graphs Enhance Medical Question Answering: Bridging the Gap Between LLMs and Evolving Medical Knowledge

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have greatly advanced medical Question Answering (QA) by leveraging vast clinical data and medical literature. However, the rapid evolution of medical knowledge and the labor-intensive process of manually updating domain-specific resources can undermine the reliability of these systems. We address this challenge with Adaptive Medical Graph-RAG (AMG-RAG), a comprehensive framework that automates the construction and continuous updating of Medical Knowledge Graphs (MKGs), integrates Chain-of-Thought (CoT) reasoning, and retrieves current external evidence (e.g., PubMed, WikiSearch). By dynamically linking new findings and complex medical concepts, AMG-RAG not only boosts accuracy but also enhances interpretability for medical queries.

Evaluations on the MEDQA and MEDMCQA benchmarks demonstrate the effectiveness of AMG-RAG, achieving an F1 score of 74.1% on MEDQA and an accuracy of 66.34% on MEDMCQA—surpassing both comparable models and those 10 to 100 times larger. Importantly, these improvements are achieved without increasing computational overhead, underscoring the critical impact of automated knowledge graph generation and external evidence retrieval in delivering up-to-date, trustworthy medical insights.

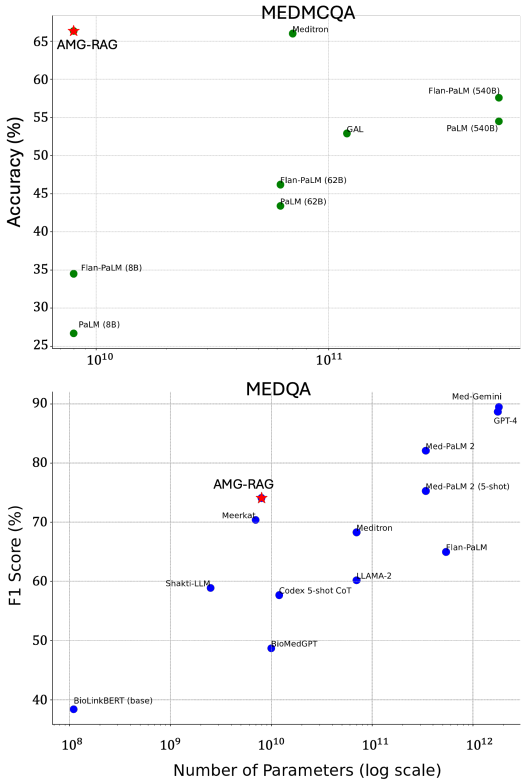

Figure 1: Performance vs parameter numbers for medical Question Answering (QA) on the MEDQA and MEDMCQA datasets. Adaptive Medical Graph-RAG (AMG-RAG) achieves an F1 score of 74.1% on MEDQA and an accuracy of 66.34% on MEDMCQA, surpassing both comparable models and those that are 10 to 100 times larger in size. See Tables 1 and 2 for more details.

1 Introduction

Medical knowledge expands at a tremendous pace, with new research findings, clinical guidelines, and treatment protocols emerging constantly. Large Language Models (LLMs) have already demonstrated their value in harnessing this vast and evolving information for medical question answering by processing large corpora of domain-specific literature and data (Nazi and Peng, 2024; Liu et al., 2023). However, a major challenge lies in ensuring that these models remain factually current and can accurately represent complex relationships among medical concepts (Rohanian et al., 2024; Yu et al., 2024). Traditional approaches to mitigating these issues involve the use of knowledge graphs, which offer structured, interconnected representations of medical information and can support more nuanced reasoning (Huang et al., 2021). Yet, constructing and maintaining such graphs is labor-intensive, time-consuming, and expensive. These burdens are particularly acute in a domain as dynamic as medicine, where new insights quickly render old

information out-of-date (Yang et al., 2024). To address this pressing issue, we propose an automated framework for constructing and continuously evolving Knowledge Graphs (KGs) specifically tailored to medical question answering. By leveraging LLMs agents and domain-specific search tools, our method autonomously generates graph Medical Knowledge Graphs (MKGs), enriched with descriptive metadata, confidence scores, and relevance indicators. In doing so, it drastically reduces the manual effort traditionally required to build and update knowledge graphs, while ensuring alignment with the latest medical advances. Unlike traditional Retrieval Augmented Generation (RAG) solutions that rely heavily on vector similarity for retrieval (Lewis et al., 2020), our knowledge-graph-based approach provides more sophisticated reasoning capabilities through shared attributes and explicit relationships. This facilitates accurate synthesis of information across diverse medical domains, ranging from drug interactions and clinical trial data to patient histories and treatment guidelines.

A central component of our solution is the integration of these evolving knowledge graphs into a RAG-based pipeline. New and updated graph entities are continuously fed into an LLM question answering module, ensuring that responses draw upon the most up-to-date and contextually relevant medical information (Singhal et al., 2022). Building on this dynamic architecture, we introduce an iterative pipeline, AMG-RAG, which combines insights from these automatically maintained graphs with traditional textual retrieval and multi-step chain-of-thought reasoning. By optimizing retrieval through confidence scoring and adaptive graph traversal, AMG-RAG demonstrates significantly improved accuracy and completeness in medical QA (Trivedi et al., 2022).

We assess AMG-RAG on the MEDQA and MEDMCQA benchmarks, which are designed to test evidence retrieval, complex reasoning, and multi-choice comprehension in the medical domain. Our model achieves an F1 score of 74.1% on MEDQA and an accuracy of 66.34% on MEDMCQA, outperforming both similarly sized RAG approaches and much larger state-of-the-art models (Fig. 1). Crucially, these gains do not necessitate additional fine-tuning or higher inference costs; instead, they result from seamlessly integrating knowledge graphs and domain-specific search tools. This efficient and scalable approach underscores the value of dynamically evolving knowledge retrieval in medical QA, offering an avenue for enhancing clinical decision-making by delivering reliable, relationally enriched insights (Zhou et al., 2023).

2 Related Work

Medical QA systems are essential for enhancing clinical decision-making, research, and patient care (Nazi and Peng, 2024; Liu et al., 2023; Rezaei et al., 2024). Over time, the field has evolved with various technological advancements addressing key challenges in processing medical information (Singhal et al., 2022). Domain-specific language models such as BioBERT (Lee et al., 2020), PubMedBERT (Gu et al., 2021), and MedPaLM (Singhal et al., 2023) have achieved significant success in biomedical tasks (Rohanian et al., 2024). However, these models often struggle with synthesizing complex relationships between medical entities and integrating data from diverse sources, particularly for rare conditions, drug interactions, and comorbidities (Zhou et al., 2023; Yu et al., 2024).

To overcome these challenges, RAG frameworks (Lewis et al., 2020) have enhanced LLMs by integrating external knowledge sources. Systems like MMED-RAG (Xia et al., 2024) have extended this paradigm to include multimodal data. The introduction of Chain-of-Thought (CoT) reasoning has further improved QA performance, with IRCoT (Trivedi et al., 2022) combining CoT reasoning with RAG for more sophisticated inference. Recent advancements, such as Gemini's multimodal and long-context reasoning capabilities, have set new benchmarks in MedQA, surpassing GPT-4 in performance (Saab et al., 2024). However, these systems often struggle to adapt to novel queries and dynamic data due to their rigid architectures.

KG-based approaches provide another avenue for advancing medical information processing. Systems like KG-Rank (Huang et al., 2021) utilize structured knowledge representations and ontologies to enable hierarchical reasoning and inference. By combining knowledge graphs with ranking and re-ranking techniques, these systems enhance the factual accuracy of long-form QA (Yang et al., 2024). However, KG-based systems face significant challenges in maintaining scalability and staying current with rapidly evolving biomedical discoveries.

Difference and Importance of AMG-RAG

Our AMG-RAG dynamically constructs relational medical KGs integrated with advanced search capabilities. Unlike traditional static systems, our approach extracts medical terms from queries, enriches them with real-time data, and utilizes LLMs to infer relationships. This dynamic mechanism ensures continuous alignment with emerging medical knowledge, addressing the limitations of static knowledge bases and pre-trained models. By combining dynamic KGs with CoT reasoning and RAG, our framework improves the adaptability and reliability of medical QA systems.

3 Method

Knowledge Graphs (KGs) offer structured frameworks for evidence-based reasoning. However, traditional KGs struggle to adapt to dynamic and evolving queries, as well as the continuous influx of new research and evidence in the medical domain. To overcome this limitation in the medical Question Answering (QA) pipeline, we propose a novel approach that automatically constructs medical-KGs using a combination of a Large Language Model (LLM) agent and a specialized medical search tool. This Medical Knowledge Graph (MKG) enhances medical QA by tailoring the graph's construction and querying processes to each specific query. This section outlines the methodologies used to develop the MKG and the Adaptive Medical Graph-RAG (AMG-RAG) pipeline, aimed at enhancing QA systems with advanced capabilities for retrieval, reasoning, and generation.

3.1 Retrieval Augmented Generation (RAG) for QA

Retrieval Augmented Generation (RAG) is a framework designed to enhance QA by integrating relevant external knowledge into the generation process. The framework combines retriever and generator components to ensure responses are grounded in evidence. Below, we outline various approaches within the RAG paradigm:

3.1.1 RAG

In the RAG approach, the retriever fetches a fixed number of relevant documents, $\{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n\} \in \mathbf{D}$, based on the query $\mathbf{q}$. Here, $\mathbf{D}$ represents the set of all domain-specific documents utilized. These documents are concatenated and passed directly to a LLM-based text generator, G, which produces the answer $\hat{\mathbf{a}}$:

$$\hat{\mathbf{a}} = G(\mathbf{q}, \{\mathbf{d}_1, \ldots, \mathbf{d}_n\}).$$

This approach is simple and computationally efficient but may struggle with domain-specific or complex queries that require additional supporting evidence.

3.1.2 RAG with Chain-of-Thought (CoT)

Enhancing RAG's performance can be achieved by integrating intermediate reasoning steps prior to producing the final response. The generator produces a chain of thought, $\mathbf{c}$, which serves as an explicit reasoning trace:

$$\{\mathbf{d}_1, \ldots, \mathbf{d}_k\} = \text{Retriever}(\mathbf{q}; \mathbf{D}),$$

$$\mathbf{c} = G(\mathbf{q}, \{\mathbf{d}_1, \ldots, \mathbf{d}_k\}), \quad \hat{\mathbf{a}} = G(\mathbf{c}).$$

This step-by-step approach enhances reasoning and interpretability, leading to improved accuracy in multi-hop reasoning tasks.

3.1.3 RAG with Search

The RAGs's performance can improved further by incorporating additional related documents retrieved from external sources, such as the internet, through a search tool. This variant integrates external search capabilities into the retrieval process. For a query $\mathbf{q}$, the retriever's results are combined with those from external search engines, providing more comprehensive evidence for the LLM to generate a response:

$$\{\mathbf{d}'_1, \ldots, \mathbf{d}'_m\} = \text{Search}(\mathbf{q}; \mathbf{D}'),$$

$$\hat{\mathbf{a}} = G(\mathbf{q}, \{\mathbf{d}_1, \ldots, \mathbf{d}_n, \mathbf{d}'_1, \ldots, \mathbf{d}'_m\}).$$

This additional search step significantly enhances performance, particularly in scenarios that require access to extensive and diverse knowledge.

3.2 Medical QA with AMG-RAG

In scenarios requiring domain expertise, such as medical or scientific QA, traditional methods often fail due to their inability to capture intricate domain-specific relationships or handle ambiguous queries. KG-driven approaches overcome these challenges by integrating explicit relationships and structured knowledge representations. This marks a significant advancement in intelligent QA systems, ensuring robustness and scalability across various applications.

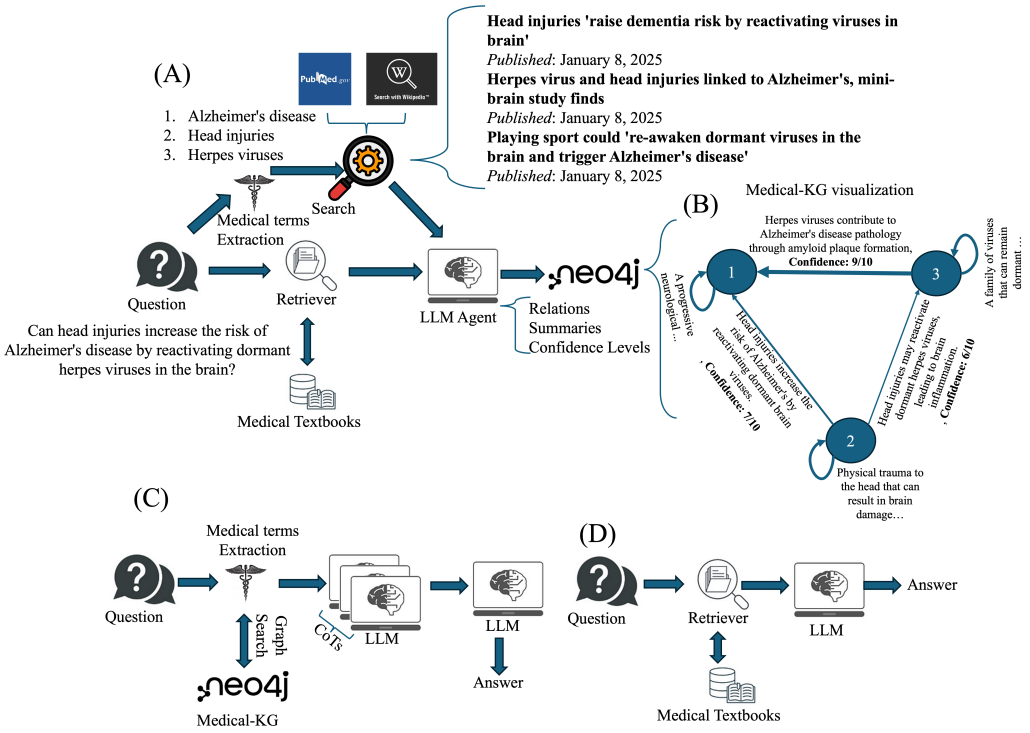

Figure 2: Model Schema. A) The pipeline for creating the MKG using search tools and an LLM agent. B) An example of the generated MKG in Neo4J, illustrating nodes and relationships derived from search results and contextual information. C) The AMG-RAG pipeline. D) A simplified RAG pipeline.

Below, we introduce the proposed AMG-RAG pipeline and outline the process for constructing the MKG, which is detailed in the following section.

The AMG-RAG pipeline consists of the following steps:

1. **Question Parsing:** Extract medical terms $\{\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_m\}$ from the user query $\mathbf{q}$ using an LLM agent:

   $$\{\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_m\} = \text{LLM}(\mathbf{q}, M), \quad m \leq M.$$

   The LLM agent is instructed to extract the medical terms $\{\mathbf{n}_i\}_{i=1}^{m}$ associated with the query $\mathbf{q}$.

2. **Node Exploration:** For each term $\mathbf{n}_i$, query the KG to retrieve relevant information while applying a confidence threshold. This threshold determines the minimum level of relationship confidence required to include information from the KG, filtering out relationships with confidence scores below the specified threshold. Details on the calculation of confidence scores can be found in Appendix A. Nodes and their children are examined iteratively, with parent confidence scores, $s(\mathbf{n}_i)$, multiplied by their relationship scores, $s(\mathbf{r}_{ij})$, to compute child confidence:

   $$s(\mathbf{n}_j) = s(\mathbf{n}_i) \cdot s(\mathbf{r}_{ij}), \quad \forall j \in \text{children of } i.$$

   Both Breadth-First Search (BFS) and Depth-First Search (DFS) strategies can be employed to explore child nodes. BFS prioritizes covering all immediate neighbors at the current depth before moving deeper, ensuring comprehensive breadth-wise exploration. In contrast, DFS delves deeply along one branch before backtracking, enabling a more targeted depth-first traversal. The exploration process continues until either the cumulative confidence meets or exceeds a threshold τ, or the maximum document limit M is reached.

3. **Chain of Thought Generation:** Generate a reasoning trace $\mathbf{c}_i$ for each entity $\mathbf{n}_i$ using an LLM, integrating information from nodes and their relationships:

   $$\mathbf{c}_i = \text{LLM}(\mathbf{n}_i, \{\mathbf{d}(\mathbf{n}_j) \mid j \in \text{connected nodes}\}).$$

4. **Answer Synthesis:** Aggregate reasoning traces $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_m\}$ and pass them to a fi-

nal answer generator, which produces the output $\hat{\mathbf{a}}$ along with an overall confidence score:

$$\hat{\mathbf{a}}, \hat{s} = G(\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_m\}).$$

This pipeline ensures that answers are comprehensive and grounded in the KG, with confidence scores providing interpretability and reliability.

Algorithm 1 KG-Based QA Inference Pipeline

Require: Query q, Knowledge Graph KG, Confidence Threshold τ, Max Iterations N
Ensure: Final Answer $\hat{\mathbf{a}}$ with Confidence s
Extract medical terms: $\{\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_m\} \leftarrow \text{ExtractTerms}(\mathbf{q})$
Initialize reasoning traces: $C \leftarrow \emptyset$
Initialize confidence: $\mathbf{s}_i \leftarrow 1.0$ for all terms $\mathbf{n}_i$
for $i = 1$ to m **do** ▷ Iterate over extracted terms
  Explore KG: Retrieve relevant nodes $\{\mathbf{d}_j\}$ and relationships $\mathbf{r}_{ij}$ for $\mathbf{n}_i$
  for each child node $\mathbf{n}_j$ of $\mathbf{n}_i$ in KG **do**
    Compute child confidence: $s_j \leftarrow s_i \cdot \mathbf{r}_{ij}$
    if $s_j \geq \tau$ **then**
      Include $\mathbf{n}_j$ in exploration set
    end if
  end for
  Generate Reasoning Trace: $\mathbf{c}_i \leftarrow \text{LLM}(\mathbf{n}_i, \{\mathbf{d}_j\})$
  Add $\mathbf{c}_i$ to reasoning traces: $C \leftarrow C \cup \{\mathbf{c}_i\}$
end for
Synthesize Answer: $\hat{a}, \hat{s} \leftarrow G(C)$
return $\hat{\mathbf{a}}, \hat{s}$ ▷ Return final answer with confidence

3.3 Dynamic Generation of the Medical Knowledge Graph

The construction of the MKG for QA represents a critical step toward enabling structured reasoning in our AMG-RAG. This approach extracts key entities and their interconnections from user queries, enriching them with information retrieved through external search tools. By organizing information into the MKG, we enable efficient, interpretable, and evidence-based QA. The methodology is as follows:

3.3.1 Node Extraction

Medical terms are identified within the user query $\mathbf{q}$ using an LLM agent named Medical Entity Recognizer (MER). These terms are treated as nodes, $\{\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_m\}$, in the KG. For each extracted term, a search tool (e.g., PubMed, or a specialized medical search engine) retrieves detailed descriptions $\mathbf{d}(\mathbf{n}_i)$, providing context for each node:

$$\mathbf{d}(\mathbf{n}_i) = \text{Search}(\mathbf{n}_i; \text{knowledge source}).$$

The retrieved descriptions form the foundational data for each node in the MKG, ensuring an accurate representation of medical terms and their attributes.

3.3.2 Relationship Inference

An LLM agent extracts relationships between nodes based on their descriptions and retrieved documents. The LLM analyzes pairs of nodes $(\mathbf{n}_i, \mathbf{n}_j)$ to determine potential relationships $\mathbf{r}_{ij}$ and their nature:

$$\mathbf{r}_{ij}, \mathbf{s}_{ij} = \text{LLM}(\mathbf{d}(\mathbf{n}_i), \mathbf{d}(\mathbf{n}_j)).$$

The agent generates a summary, infers the relationship type (e.g., causation, association), and assigns a confidence score. This process results in a KG rich in structure and semantics.

3.3.3 Knowledge Graph Construction

The nodes, descriptions, relationships, and confidence scores are integrated into the KG structure. The resulting graph supports medical QA by:

- Highlighting key medical concepts and their interrelations.
- Enabling efficient retrieval and reasoning over medical knowledge.
- Providing confidence metrics for each established relationship in the graph, serving as a source of reliability.

4 Experiments

The **MEDQA** dataset is a free-form, multiple-choice open-domain QA data set specifically designed for medical QA. Derived from professional medical board exams, this dataset presents a significant challenge as it requires both the retrieval of relevant evidence and sophisticated reasoning to answer questions accurately. Each question is accompanied by multiple-choice answers that demand a deep understanding of medical concepts and logical inference, often relying on evidence found in medical textbooks. For this study, the test partition of the MEDQA dataset, comprising approximately 1,200 samples, was used (Jin et al., 2021).

The **MedMCQA** dataset is another multiple-choice question-answering dataset tailored for medical QA. Unlike MEDQA, which is derived from board exam questions, MedMCQA offers a broader variety of question types, encompassing both foundational and clinical knowledge across diverse medical specialties. In this study, the MedMCQA development set, containing approximately 4,000 questions, was used to benchmark against other models (Pal et al., 2022a).

This study employed the MEDQA and MedMCQA datasets to benchmark and evaluate medical QA systems. These datasets serve as challenging testbeds for open-domain QA tasks due to their demands for multi-hop reasoning and the integration of domain-specific knowledge. The relevance of MEDQA in the real world, together with the diverse question styles and extensive development set of MedMCQA make them ideal for advancing the development of robust QA models capable of addressing medical inquiries. We utilize *GPT-4o-mini* as the backbone of the implementation for both MKG and AMG-RAG, leveraging its capabilities with approximately $\sim 8B$ parameters. This model serves as the core component, enabling advanced reasoning, RAG, and structured knowledge integration.

4.1 Medical Knowledge Graph

The KG was dynamically constructed by integrating search items, contextual information, and relationships derived from textbooks and search queries from the PubMed engine for each question in the dataset. This data was processed and stored in a Neo4j database. The key features of the knowledge graph include:

1. **Dynamic Node and Relationship Creation**: Nodes are dynamically generated for search items, and relationships between these nodes are established based on their relevance and predefined relationship types.

2. **Bidirectional Relationships**: To ensure a comprehensive representation, the graph includes both forward and reverse relationships between nodes, enhancing its utility for diverse queries.

3. **Relevance Scoring**: Each relationship is enriched with descriptive annotations and a confidence score, quantifying the strength of the association and aiding in prioritizing relevant connections.

4. **Summarization**: Concise summaries for each search item are included, derived from contextual data. A confidence score accompanies each summary to indicate its reliability.

5. **Integration with Neo4j**: The entire graph is stored in a Neo4j database, leveraging its graph-based query capabilities for efficient analysis and retrieval.

A snapshot of a portion of the knowledge graph is shown in Figure 2.B, illustrating its structure and relationships.

This MKG serves as the foundational information source for the AMG-RAG framework during the inference phase. The evaluation of MKG confirmed its robustness and reliability, with experts LLMs such as GPT-4 achieving high precision (e.g. 9/10). These results underscore the effectiveness of MKG in supporting medical reasoning and decision-making, as detailed in Appendix B.

4.2 Performance Comparison

Table 1 presents a comprehensive comparison of state-of-the-art language models on the MEDQA benchmark. The results highlight the critical role of advanced reasoning strategies in achieving higher performance, such as CoT reasoning, fine-tuning, and the integration of search tools. While larger models like Med-Gemini and GPT-4 achieve the highest accuracy and F1 scores, their performance comes at the cost of significantly larger parameter sizes. These models exemplify the power of scaling combined with sophisticated reasoning and retrieval techniques.

Significantly, AMG-RAG, despite having just 8 billion parameters, attains an F1 score of 74.1% on the MEDQA benchmark, surpassing models like Meditron, which possess 70 billion parameters without needing any fine tuning. This highlights AMG-RAG's exceptional efficiency and proficiency in utilizing CoT reasoning and external evidence retrieval. The model leverages tools such as PubMedSearch and WikiSearch to dynamically integrate domain-specific knowledge dynamically, thereby improving its ability to address medical questions. Examples of QA interactions, including detailed search items and reasoning for question samples, are provided in Appendix C. These examples are organized in Tables 6, 7, 8, and 9, drawn from the MEDQA benchmark.

On the MedMCQA benchmark, as shown in Table 2, AMG-RAG achieves an accuracy of 66.34%, even outperforming larger models like Meditron-70B and better than Codex 5-shot CoT. This result underscores AMG-RAG's adaptability and robustness, demonstrating that it can deliver competitive performance even against significantly larger models. Its ability to maintain high accuracy on diverse datasets further highlights the effectiveness of its design, which combines CoT reasoning with structured knowledge graph integration and retrieval

Table 1: Comparison of LLM models on the MEDQA Benchmark.

Model	Model Size	Acc. (%)	F1 (%)	Fine-Tuned	Uses CoT	Uses Search
Med-Gemini (Saab et al., 2024)	∼1800B	91.1	89.5	✓	✓	✓
GPT-4 (Nori et al., 2023)	∼1760B	90.2	88.7	✓	✓	✓
Med-PaLM 2 (Singhal et al., 2025)	∼340B	85.4	82.1	✓	✓	✗
Med-PaLM 2 (5-shot)	∼340B	79.7	75.3	✗	✓	✗
AMG-RAG	∼8B	73.9	74.1	✗	✓	✓
Meerkat(Kim et al., 2024)	7B	74.3	70.4	✓	✓	✗
Meditron (Chen et al., 2023)	70B	70.2	68.3	✓	✓	✓
Flan-PaLM (Singhal et al., 2023)	540B	67.6	65.0	✓	✓	✗
LLAMA-2 (Chen et al., 2023)	70B	61.5	60.2	✓	✓	✗
Shakti-LLM (Shakhadri et al., 2024)	2.5B	60.3	58.9	✓	✗	✗
Codex 5-shot CoT (Liévin et al., 2024)	–	60.2	57.7	✗	✓	✓
BioMedGPT (Luo et al., 2023)	10B	50.4	48.7	✓	✗	✗
BioLinkBERT (base) (Singhal et al., 2023)	–	40.0	38.4	✓	✗	✗

mechanisms.

Table 2: Comparison of Models on the MedMCQA.

Model	Model Size	Acc. (%)
AMG-RAG	∼8B	**66.34**
Meditron (Chen et al., 2023)	70B	66.0
Codex 5-shot (Liévin et al., 2024)	–	59.7
VOD (Liévin et al., 2023)	–	58.3
Flan-PaLM (Singhal et al., 2022)	540B	57.6
PaLM	540B	54.5
GAL	120B	52.9
PubmedBERT (Gu et al., 2021)	–	40.0
SciBERT (Pal et al., 2022b)	–	39.0
BioBERT (Lee et al., 2020)	–	38.0
BERT (Devlin, 2018)	–	35.0

Overall, AMG-RAG's results on MEDQA and MedMCQA benchmarks solidify its position as a highly efficient and effective model for medical QA. By leveraging CoT reasoning, search tools, and external knowledge sources, AMG-RAG not only closes the gap with much larger models but also sets a new standard for performance among smaller-sized models.

4.3 Impact of Search Tools and CoT Reasoning on AMG-RAG Performance

Figure 3 and Table 3 demonstrate the impact of integrating search tools such as PubMedSearch and WikiSearch on the performance of AMG-RAG when applied to the MEDQA dataset. The inclusion of these search capabilities significantly improves accuracy and F1 scores by providing access to relevant external evidence, which is critical for addressing medical questions. Among the search tools, PubMedSearch outperforms WikiSearch, likely due to its more focused and domain-specific content, which better aligns with the nature of medical QA tasks.

Additionally, the impact of CoT reasoning and MKG integration on AMG-RAG performance is highlighted in the same figure and table. The results reveal that the removal of either CoT reasoning or KG integration leads to a substantial drop in accuracy and F1 scores. This underscores the indispensable role of structured reasoning and domain-specific retrieval in enhancing the system's ability to generate accurate and evidence-backed answers.

Table 3: Performance metrics for AMG-RAG model with and without CoT and Knowledge Graph integration with different search tools for MEDQA dataset.

Model	Acc. (%)	F1-Score	Recall
PubMedSearch	73.92	0.7410	0.7392
WikiSearch	70.62	0.7067	0.7062
No Search	67.16	0.6696	0.6716
No Search & CoT	66.69	0.6655	0.6669

4.4 Improving QA in Rapidly Changing Medical Domains

Figure 4 illustrates the performance of various models across different question domains, including Neurology and Genetics. The AMG-RAG model consistently outperforms other approaches, showcasing its superior adaptability and robustness in these rapidly evolving and knowledge-intensive fields. This exceptional performance stems from its ability to seamlessly integrate external sources of information and evidence. By leveraging PubMed searches, the AMG-RAG model dynamically retrieves the latest medical research and continuously updates the MKG, ensuring that it remains relevant and up-to-date. This dynamic updating process not only enhances the model's ability to reason across multiple domains but also allows it to address complex, multi-hop questions with greater accuracy and depth.

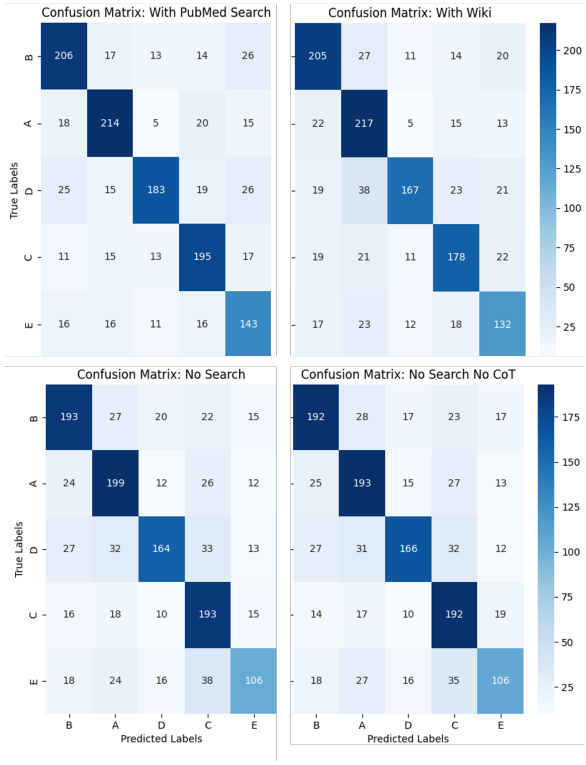

Figure 3: Confusion matrix for AMG-RAG with and without CoT and Knowledge Graph integration on MEDQA dataset.

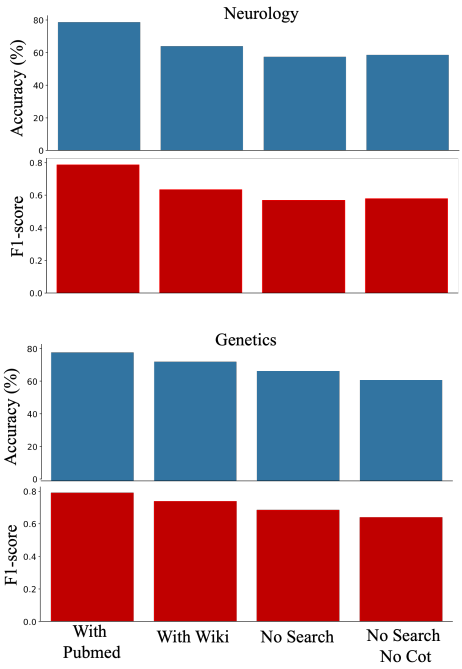

Figure 4: Performance comparison across different question domains in the Neurology and Genetics fields.

5 Conclusion

In this work, we introduce AMG-RAG, an advanced QA system that dynamically constructs MKG while integrating sophisticated reasoning and external domain-specific search tools. The model exhibits significant improvements in accuracy and reasoning capabilities, particularly for medical question-answering tasks, outperforming other approaches of similar model size or 10 to 100 times larger. Using structured knowledge representations and advanced reasoning frameworks, our approach establishes a new benchmark for QA in highly competitive and highly evolving domains such as medicine.

6 Limitations

Despite AMG-RAG advancements, our approach has certain limitations. Firstly it relies on external search tools to introduce latency during the creation of MKG. However, this occurs only once, when the MKG is built from scratch for the first time. Additionally, while the model performs exceptionally well in medical domains, its applicability to non-medical tasks remains unexplored.

Another limitation is the need for structured, authoritative sources of medical knowledge. Currently, AMG-RAG retrieves information from diverse sources, including research articles and medical textbooks. However, as emphasized in clinical decision-making, treatment guidelines serve as essential references for standardized diagnosis and treatment protocols (Hager et al., 2024). Future work on AMG-RAG should focus on integrating structured access to these sources to ensure compliance with evidence-based medicine.

7 Ethics Statement

The development of LLMs for medical QA requires careful ethical consideration due to risks of inaccuracy and bias. Ensuring the reliability of retrieved content is crucial, especially when integrating external knowledge sources. To mitigate these risks, we implement a confidence scoring mechanism into the MKG to validate the information. However, bias detection and mitigation remain active research areas.

Acknowledgements

References

Maciej Besta, Robert Gerstenberger, Emanuel Peter, Marc Fischer, Michał Podstawski, Claude Barthels, Gustavo Alonso, and Torsten Hoefler. 2023. Demystifying graph databases: Analysis and taxonomy of data organization, system designs, and graph queries. *ACM Computing Surveys*, 56(2):1–40.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622.

Hongcheng Huang and Ziyu Dong. 2013. Research on architecture and query performance based on distributed graph database neo4j. In *2013 3rd International Conference on Consumer Electronics, Communications and Networks*, pages 533–536. IEEE.

Xiaofeng Huang, Jixin Zhang, Zisang Xu, Lu Ou, and Jianbin Tong. 2021. A knowledge graph based question answering method for medical domain. *PeerJ Computer Science*, 7:e667.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Hyunjae Kim, Hyeon Hwang, Jiwoo Lee, Sihyeon Park, Dain Kim, Taewhoo Lee, Chanwoong Yoon, Jiwoong Sohn, Donghee Choi, and Jaewoo Kang. 2024. Small language models learn enhanced reasoning skills from medical textbooks. *arXiv preprint arXiv:2404.00376*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).

Valentin Liévin, Andreas Geert Motzfeldt, Ida Riis Jensen, and Ole Winther. 2023. Variational open-domain question answering. In *International Conference on Machine Learning*, pages 20950–20977. PMLR.

Jialin Liu, Changyu Wang, and Siru Liu. 2023. Utility of chatgpt in clinical practice. *Journal of Medical Internet Research*, 25:e48568.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2023. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442*.

Zabir Al Nazi and Wei Peng. 2024. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, page 57. MDPI.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022a. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022b. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Mohammad Reza Rezaei, Maziar Hafezi, Amit Satpathy, Lovell Hodge, and Ebrahim Pourjafari. 2024. At-rag: An adaptive rag model enhancing query efficiency with topic filtering and iterative reasoning. *arXiv preprint arXiv:2410.12886*.

Omid Rohanian, Mohammadmahdi Nouriborji, Samaneh Kouchaki, Farhad Nooralahzadeh, Lei Clifton, and David A Clifton. 2024. Exploring the effectiveness of instruction tuning in biomedical language processing. *Artificial intelligence in medicine*, 158:103007.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Syed Abdul Gaffar Shakhadri, Kruthika KR, and Rakshit Aralimatti. 2024. Shakti: A 2.5 billion parameter small language model optimized for edge ai and low-resource environments. *arXiv preprint arXiv:2410.11331*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509*.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. 2024. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*.

Rui Yang, Haoran Liu, Edison Marrese-Taylor, Qingcheng Zeng, Yu He Ke, Wanxin Li, Lechao Cheng, Qingyu Chen, James Caverlee, Yutaka Matsuo, et al. 2024. Kg-rank: Enhancing large language models for medical qa with knowledge graphs and ranking techniques. *arXiv preprint arXiv:2403.05881*.

Huizi Yu, Lizhou Fan, Lingyao Li, Jiayan Zhou, Zihui Ma, Lu Xian, Wenyue Hua, Sijia He, Mingyu Jin, Yongfeng Zhang, et al. 2024. Large language models in biomedical and health informatics: A review with bibliometric analysis. *Journal of Healthcare Informatics Research*, pages 1–54.

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, et al. 2023. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*.

A Confidence Scoring for the Relationships in the MKG

A confidence score, s_{ij}, is assigned to each inferred relationship, reflecting its strength and relevance. The scoring criteria are as follows:

- **10**: The target is directly and strongly related to the item, with clear, unambiguous relevance.
- **7-9**: The target is moderately to highly relevant to the item but may have some ambiguity or indirect association.
- **4-6**: The target has some relevance to the item but is weak or only tangentially related.
- **1-3**: The target has minimal or no meaningful connection to the item.

B Evaluating the Accuracy and Robustness of the Medical Knowledge Graph

The quality and reliability of the dynamically generated MKG are critical for its effectiveness in enhancing medical QA systems. To validate the accuracy, robustness, and usability of the MKG, a structured evaluation involving expert LLMs in the medical domain, such as GPT medical model, was conducted. This section outlines the methodology used to evaluate the MKG, emphasizing interpretability, clinical relevance, and robustness in real-world applications. Additionally, the role of medical experts in verifying the accuracy and applicability of the MKG is discussed, underscoring the necessity of human expertise in validating AI-driven medical knowledge representations.

To assess accuracy and robustness, a two-phase evaluation process was employed. In the first phase, a group of expert LLMs specializing in medical domains reviewed a subset of the MKG, including dynamically generated nodes, relationships, confidence scores, and summaries for various medical queries. They evaluated the accuracy of medical terms and concepts, the relevance of relationships between nodes, the reliability of node summaries, and the alignment of confidence scores with the perceived strength and reliability of the connections. Each LLM independently rated the graph components on a scale of 1 to 10. The results showed an average accuracy score of 8.9/10 for node identification, 8.8/10 for relationship relevance, and

8.5/10 for the clarity and precision of node summaries. Confidence scores generally aligned well with the LLMs's assessments, as illustrated in Tables 4 and 5, which highlight strong relationships across domains such as ophthalmology, cardiovascular treatments, and dermatology.

In the second phase, blind testing was conducted to evaluate usability and human-readability. Expert LLMs were tasked with answering complex medical queries requiring multi-hop reasoning, such as managing comorbidities or determining multi-drug treatment protocols. As shown in Table 4, relationships such as the co-usage of Ketotifen and Fluorometholone for allergic conjunctivitis or Labetalol and Nitroglycerin for acute hypertension demonstrated the MKG's ability to model clinically relevant associations effectively. The LLMs achieved a 89% accuracy rate in these test scenarios. Additionally, the LLMs rated the MKG 9.4/10 for interpretability and usability, underscoring its strength in visually and contextually representing complex medical relationships.

To further ensure the clinical relevance and practical applicability of the MKG, medical experts, including practicing physicians and clinical researchers, were involved in evaluating the generated relationships and summaries. Unlike LLMs, medical experts provided qualitative assessments, identifying potential discrepancies, overlooked nuances, and contextual dependencies that automated models might miss. The medical experts particularly assessed:

1. The correctness and completeness of medical relationships, ensuring they align with established clinical knowledge and best practices.

2. The validity of multi-hop reasoning paths, verifying whether inferred relationships reflected logical clinical decision-making processes.

3. The utility of the MKG in real-world medical applications, particularly in aiding diagnostic and treatment decision-making.

The feedback from medical experts was instrumental in refining the graph, addressing inconsistencies, and enhancing the confidence scores to better reflect real-world medical reliability. Notably, medical expert ratings aligned well with LLM evaluations but provided deeper insights into the contextual limitations of the graph. For example, while LLMs accurately linked Diltiazem and Nitroglycerin in cardiovascular treatment, medical experts highlighted additional considerations such as contraindications in specific patient populations, which were subsequently incorporated into the MKG.

The detailed evaluations in Tables 4 and 5 provide further insights into the graph's performance across diverse medical domains. For instance, the accurate representation of relationships between beta-blockers like Labetalol and Propranolol or the integration of treatments such as Diltiazem and Nitroglycerin for cardiovascular care highlight the MKG's capacity to support intricate clinical decision-making.

These results confirm that the MKG is both human-readable and usable by advanced LLMs, making it an invaluable tool for medical QA and decision-making. The graph's structured format, enriched with confidence scores and summaries, ensures a clear and interpretable representation of medical knowledge while enhancing the efficiency and accuracy of QA systems in addressing real-world medical scenarios. Moreover, the involvement of medical experts in the evaluation process enhances the credibility of the MKG, ensuring that AI-driven insights align with clinical expertise and practical healthcare applications.

C QA Samples with reasoning from MEDQA benchmark

This section presents a set of QA samples demonstrating the reasoning paths generated by our proposed AMG-RAG model when applied to the MEDQA dataset. These examples highlight how the model retrieves relevant content, structures key information, and formulates reasoning to guide answer selection.

Table 6 provides an example of how the model processes a clinical case question related to the management of acute coronary syndrome (ACS). The search items retrieved for possible answer choices (e.g., Nifedipine, Enoxaparin, Clopidogrel, Spironolactone, Propranolol) are accompanied by key content excerpts relevant to their roles in ACS treatment. Additionally, the reasoning pathways illustrate how the model synthesizes evidence-based knowledge to justify the selection of the correct answer (Clopidogrel), while also explaining why the alternative options are not suitable. Additional examples are also provided in Tables 7, 8, and 9

Table 4: Examples from the Medical Knowledge Graph (MKG) with Expert and Blind Analysis (Part 1)

Source Node	Relationship Type	Target Node	LLM Expert Analysis	Blind Analysis	Medical Expert Analysis
Botulism	Directly related as it is the target concept.	Myasthenia gravis	Rated 9.2/10 for relevance and clinical importance, considered highly accurate.	Demonstrated effective multi-hop reasoning with a 92% accuracy in identifying related conditions.	Rated 9.5/10 for relevance and accuracy, considered highly accurate.
Levodopa	Levodopa is a primary treatment for Parkinson's disease.	Parkinson's disease	Evaluated as highly reliable (9.6/10) for summarizing medical treatments and relationships.	Increases accuracy by 24% in answering queries about Parkinson's treatments and comorbidities.	Rated 10/10 for relevance and accuracy, considered highly accurate.
Zidovudine	Zidovudine is an antiviral drug used for HIV treatment.	HIV/AIDS	Experts rated it 9.4/10 for interpretability, highlighting the clear representation of the relationship.	Provided contextually accurate responses regarding drug interactions and side effects in queries.	Rated 10/10 for relevance and accuracy, considered highly accurate.
Inhibition of thymidine synthesis	Cross-linking of DNA is directly related to thymidine synthesis as both involve nucleic acid metabolism.	Cross-linking of DNA	Rated 9.2/10 for relevance to nucleic acid metabolism and DNA replication.	Demonstrated high accuracy in answering multi-hop queries related to DNA synthesis pathways.	Rated 9/10 for relevance and accuracy, considered accurate.
Hyperstabilization of microtubules	Cross-linking of DNA can be related to the stabilization of microtubules.	Cross-linking of DNA	Rated 9.0/10 for highlighting structural modifications affecting cellular functions.	Increases the accuracy by 20% in scenarios involving cellular structure interactions.	Rated 8/10 for moderated relevance.
Generation of free radicals	Free radicals can lead to oxidative damage, affecting DNA integrity and function.	Cross-linking of DNA	Rated 8.5/10 for its relevance to oxidative stress and DNA damage mechanisms.	Accurate in providing causal explanations for oxidative stress and DNA cross-linking.	Rated 7.5/10 for relevance.
Renal papillary necrosis	Allergic interstitial nephritis can lead to renal damage.	Allergic interstitial nephritis	Rated 9.0/10 for explaining the clinical progression of renal complications.	Effective in multi-hop reasoning for renal damage-related queries, achieving 91% accuracy.	Rated 10/10 for relevance and accuracy, considered highly accurate.

D Implementation Details for Dataset Ingestion and Vector Database

This section outlines the pipeline for dataset ingestion and vector database creation for efficient medical question-answering. The process involves document chunking, embedding generation, and storage in a vector database to facilitate semantic retrieval.

D.1 Dataset Processing and Chunking

The dataset, sourced from medical textbooks in the MEDQA benchmark, is provided in plain text format. Each document is segmented into smaller chunks with a maximum size of 512 tokens and a 100-token overlap. This overlap ensures context preservation across chunk boundaries, supporting multi-hop reasoning for long documents.

D.2 Embedding Model and Vector Storage

The system utilizes the **SentenceTransformer** model, specifically `all-mpnet-base-v2`, for generating dense vector representations of text chunks and queries. To optimize storage and retrieval, the embeddings are indexed in the **Chroma** vector database. Metadata, such as document filenames and chunk IDs, is also stored to maintain document traceability.

D.3 Batch Processing and Vector Database Population

To manage memory efficiently during ingestion, document chunks are processed in batches of up to 10,000. This ensures a smooth ingestion pipeline while preventing memory overflow. Each processed file is logged to avoid redundant computations, and error handling mechanisms are in place to manage failed processing attempts.

D.4 Query Answering Workflow

For retrieval, user queries (e.g., *"What are the symptoms of drug-induced diabetes?"*) are embedded using the `all-mpnet-base-v2` model. The top-ranked relevant chunks are retrieved based on their semantic similarity to the query using Chroma's similarity search mechanism. The system retrieves the top k relevant passages, which can be further processed in downstream QA models.

D.5 Key Configuration Details

The system is configured with the following parameters:

Table 5: Examples from the Medical Knowledge Graph (MKG) with Expert and Blind Analysis (Part 2)

Source Node	Relationship Type	Target Node	LLM Expert Analysis	Blind Analysis	Medical Expert Analysis
Ketotifen eye drops	Ketotifen eye drops are antihistamines used for allergic conjunctivitis, which may be used alongside Fluorometholone for managing eye allergies.	Fluorometholone eye drops	Rated 9.2/10 for relevance in managing allergic conjunctivitis.	Demonstrated 93% accuracy in multi-hop reasoning for ophthalmological conditions.	Rated 10/10 for relevance and accuracy, considered highly accurate.
Ketotifen eye drops	Latanoprost eye drops are used to lower intraocular pressure in glaucoma, while Ketotifen treats allergic conjunctivitis.	Latanoprost eye drops	Rated 9.0/10 for distinct yet complementary roles in ophthalmology.	Effective in identifying separate ophthalmic applications with 92% accuracy.	Rated 10/10 for relevance and accuracy, considered highly accurate.
Diltiazem	Nitroglycerin is relevant in discussions of cardiovascular treatments alongside Diltiazem.	Nitroglycerin	Rated 8.8/10 for contextual relevance to cardiovascular management.	Increases the accuracy for treatment-based queries by 20%.	Rated 9.5/10 for relevance and accuracy, considered highly accurate.
Labetalol	Labetalol is closely related to Propranolol, both managing hypertension.	Propranolol	Rated 9.5/10 for direct relevance in cardiovascular treatment protocols.	Highly interpretable responses for hypertension management, with 95% accuracy.	Rated 10/10 for relevance and accuracy, considered highly accurate.
Nitroglycerin	Nitroglycerin and Labetalol are often used in conjunction for managing hypertension and heart conditions.	Labetalol	Rated 8.7/10 for strong relevance in acute hypertension protocols.	Supported effective multi-drug therapy reasoning with 90% accuracy.	Rated 9/10 for relevance and accuracy, considered highly accurate.
Nitroglycerin	Nitroglycerin is often used with Propranolol in managing cardiovascular conditions like hypertension and angina.	Propranolol	Rated 9.0/10 for its importance in cardiovascular multi-drug therapy.	Demonstrated robust performance in connecting treatment protocols, with 93% query accuracy.	Rated 10/10 for relevance and accuracy, considered highly accurate.
Fluorometholone eye drops	Fluorometholone eye drops are corticosteroids that treat inflammation, complementing Ketotifen for allergic conjunctivitis.	Ketotifen eye drops	Rated 8.8/10 for their combined application in managing inflammation and allergies.	Improved query relevance for multi-drug therapy in eye care by 19%.	Rated 9.5/10 for relevance and accuracy, considered highly accurate.
Lanolin	Lanolin is used for skin care, particularly for sore nipples during breastfeeding.	Fluorometholone eye drops	Rated 8.5/10 for highlighting non-overlapping yet clinically useful contexts.	Demonstrated effective differentiation of clinical uses with high interpretability.	Rated 9/10 for relevance and accuracy, considered highly accurate.

- **Embedding Model:** `all-mpnet-base-v2` from SentenceTransformer.
- **Vector Database:** Chroma, stored persistently on disk for reusability.
- **Chunk Size:** 512 tokens per chunk, with a 100-token overlap for contextual consistency.
- **Batch Size:** Up to 10,000 chunks per batch to optimize ingestion efficiency.

D.6 Implementation and System Execution

The ingestion and query process is implemented using Python, leveraging `sentence-transformers` for embeddings and `Chroma` for vector storage. The ingestion pipeline reads and processes text files, splits them into chunks, generates embeddings, and stores them efficiently in the vector database. The querying process retrieves the top k most relevant text chunks to respond to user queries.

E Components Definition

E.1 Neo4j

As data complexity increases, traditional relational databases struggle with highly interconnected datasets where relationships are crucial. Graph databases, like Neo4j, address this challenge by efficiently modeling and processing complex, evolving data structures using nodes, relationships, and properties (Besta et al., 2023).

Neo4j, an open-source NoSQL graph database, enables constant-time traversals by explicitly storing relationships, making it ideal for large-scale applications such as social networks, recommendation systems, and biomedical research. Unlike relational models, Neo4j avoids costly table joins and optimizes deep relationship queries, enhancing scalability and performance (Besta et al., 2023).

Neo4j's architecture is centered around the property graph model, which includes(Huang and Dong, 2013):

Search Item/ Question Options	Key Content Highlighted	Reasoning Guiding the Answer
Nifedipine	Not typically used for acute coronary syndrome (ACS). Associated with reflex tachycardia.	Nifedipine is a calcium channel blocker effective for hypertension but does not address the antiplatelet needs of ACS patients.
Enoxaparin	Used for anticoagulation in ACS but mainly during hospitalization.	Enoxaparin is not continued after discharge when aspirin and another antiplatelet drug are prescribed.
Clopidogrel	Standard for dual antiplatelet therapy (DAPT) in ACS, especially post-percutaneous coronary intervention (PCI).	Clopidogrel complements aspirin in preventing thrombotic events post-angioplasty. Its use is supported by evidence-based guidelines.
Spironolactone	Useful in heart failure or reduced ejection fraction but not indicated for ACS management when EF is normal.	This patient's EF is 58%, so spironolactone is not necessary. Focus should be on antiplatelet therapy.
Propranolol	Effective for reducing myocardial oxygen demand but not part of standard DAPT.	While beneficial for stress-related heart issues, it does not address thrombotic risks in ACS management.

Table 6: Examples of Summary of search items for the question "A 65-year-old man is brought to the emergency department 30 minutes after the onset of acute chest pain. He has hypertension and asthma. Current medications include atorvastatin, lisinopril, and an albuterol inhaler. He appears pale and diaphoretic. His pulse is 114/min, and blood pressure is 130/88 mm Hg. An ECG shows ST-segment depressions in leads II, III, and aVF. Laboratory studies show an increased serum troponin T concentration. The patient is treated for acute coronary syndrome and undergoes percutaneous transluminal coronary angioplasty. At the time of discharge, echocardiography shows a left ventricular ejection fraction of 58%. In addition to aspirin, which of the following drugs should be added to this patient's medication regimen?" and Their Influence on the Correct Answer (Clopidogrel) and the reasoning paths

- **Nodes**: Entities representing data points.
- **Relationships**: Directed, named connections between nodes that define how entities are related.
- **Properties**: Key-value pairs associated with both nodes and relationships, providing additional metadata.

This model allows for intuitive representation of complex data structures and supports efficient querying and analysis. The system's internal mechanisms facilitate rapid traversal of relationships, enabling swift query responses even in large datasets (Huang and Dong, 2013).

Search Item/ Question Options	Key Content Highlighted	Reasoning Guiding the Answer
A history of stroke or venous thromboembolism	Contraindicated for hormonal contraceptives due to increased risk of thrombosis.	Copper IUDs do not carry the same thrombotic risk, making this option irrelevant for contraindication in IUD placement.
Current tobacco use	Increases cardiovascular risk with hormonal contraceptives but not with copper IUDs.	Tobacco use does not contraindicate IUD placement, though it may influence other contraceptive choices.
Active or recurrent pelvic inflammatory disease (PID)	Direct contraindication for IUD placement due to the risk of exacerbating infection and complications.	Insertion of an IUD can worsen active PID, leading to infertility or other severe complications.
Past medical history of breast cancer	Contraindicates hormonal contraceptives, but copper IUDs are considered safe.	This option does not contraindicate copper IUD placement, as it is non-hormonal and unrelated to breast cancer.
Known liver neoplasm	Contraindicates hormonal contraceptives but not copper IUDs.	Copper IUDs are safe for patients with liver neoplasms as they are free of systemic hormones.

Table 7: Examples of Summary of Search Items for the Question "A 37-year-old-woman presents to her primary care physician requesting a new form of birth control. She has been utilizing oral contraceptive pills (OCPs) for the past 8 years, but asks to switch to an intrauterine device (IUD). Her vital signs are: blood pressure 118/78 mm Hg, pulse 73/min and respiratory rate 16/min. She is afebrile. Physical examination is within normal limits. Which of the following past medical history statements would make copper IUD placement contraindicated in this patient?" and Their Influence on the Correct Answer (Active or recurrent pelvic inflammatory disease (PID)) and the Reasoning Paths

Search Item/ Question Options	Key Content Highlighted	Reasoning Guiding the Answer
Dementia	Typically presents as a gradual decline in cognitive function.	The sudden onset of symptoms after surgery and acute confusion makes dementia less likely.
Alcohol withdrawal	Requires significant and sustained alcohol use to cause withdrawal symptoms.	The patient's weekly consumption of one to two glasses of wine is insufficient to support this diagnosis.
Opioid intoxication	Oxycodone can cause sedation and confusion, but stable vital signs and lack of severe respiratory depression are inconsistent.	While oxycodone use is relevant, the observed fluctuating agitation and impulsivity are more consistent with delirium.
Delirium	Characterized by acute changes in attention and cognition with fluctuating levels of consciousness.	The patient's recent surgery, medication use, and fluctuating symptoms align strongly with a diagnosis of delirium.
Urinary tract infection (UTI)	Confusion in elderly patients can result from UTIs, but a normal urine dipstick test does not support this.	The absence of urinary findings on examination makes UTI less likely as the cause of symptoms.

Table 8: Examples of Search Items for the Question: "Six days after undergoing surgical repair of a hip fracture, a 79-year-old woman presents with agitation and confusion. Which of the following is the most likely cause of her current condition?" and Their Influence on the Correct Answer (Delirium) and the Reasoning Paths.

Search Item/ Question Options	Key Content Highlighted	Reasoning Guiding the Answer
Primary spermatocyte	Nondisjunction events during meiosis I often occur at this stage, leading to chromosomal abnormalities.	Klinefelter syndrome (47,XXY) is typically due to nondisjunction during meiosis, specifically at this stage.
Secondary spermatocyte	Meiosis II occurs here, dividing chromosomes into haploid cells, but errors at this stage are less likely to lead to 47,XXY.	The chromosomal abnormality associated with Klinefelter syndrome usually arises before this stage.
Spermatid	Spermatids are post-meiotic cells where genetic material is already finalized.	Errors at this stage would not result in a cytogenetic abnormality like 47,XXY.
Spermatogonium	Errors here affect the germline but are less likely to cause specific meiotic nondisjunction errors.	While germline mutations can occur, meiotic nondisjunction leading to Klinefelter syndrome occurs later.
Spermatozoon	These are fully mature sperm cells that inherit abnormalities from earlier stages.	By this stage, chromosomal errors have already been established.

Table 9: Examples of Search Items for the Question: "A 29-year-old man with infertility, tall stature, gynecomastia, small testes, and an elevated estradiol:testosterone ratio is evaluated. Genetic studies reveal a cytogenetic abnormality inherited from the father. At which stage of spermatogenesis did this error most likely occur?" and Their Influence on the Correct Answer (Primary spermatocyte) and the Reasoning Paths.