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

Mohammad R. Rezaei[1,2]*, **Reza Saadati Fard**[3], **Jayson L. Parker**[1], **Rahul G. Krishnan**[1,2], **Milad Lankarany**[1]
[1] University of Toronto
[2] Vector Institute
[3] Worcester Polytechnic Institute
*mr.rezaei@mail.utoronto.ca

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 Agentic Medical Graph-RAG (AMG-RAG), a comprehensive framework that automates the construction and continuous updating of Medical Knowledge Graphs (MKGs), integrates reasoning, and retrieves current external evidence from the MKGs for medical QA. 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. By dynamically linking new findings and complex medical concepts, AMG-RAG not only boosts accuracy but also enhances interpretability for medical queries, which has a critical impact on delivering up-to-date, trustworthy medical insights (GitHub: https://github.com/MrRezaeiUofT/AMG-RAG).

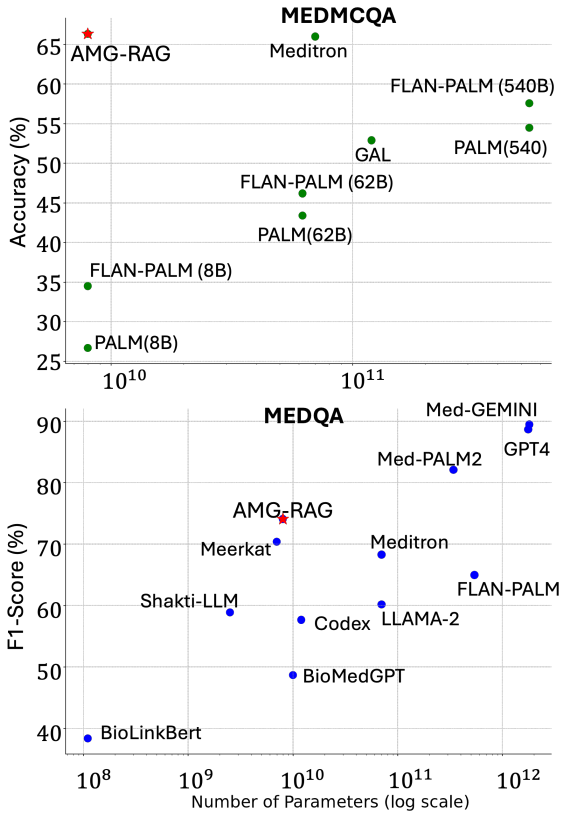

Figure 1: Performance versus parameter count on the MEDQA and MEDMCQA benchmarks. Our system, Agentic Medical Graph-RAG (AMG-RAG), attains an F1 of 74.1 % on MEDQA and an accuracy of 66.34 % on MEDMCQA, outperforming models that contain 10–100× more parameters. See Tables 1 and 2 for details.

1 Introduction

Medical knowledge is growing at an unprecedented rate: every day brings new research findings, revised clinical guidelines, and updated treatment protocols. Recent work shows that Large Language Models (LLMs) can already harness this ever-expanding corpus for medical Question Answering (QA) (Nazi and Peng, 2024; Liu et al., 2023).

Despite their promise, LLMs face two persistent challenges. First, they must remain *factually current* in a field where knowledge can become obsolete almost overnight (Rohanian et al., 2024; Yu et al., 2024). Second, they must correctly model the intricate relationships among information entities. Knowledge Graph (KG) provides a structured and interconnected view of information that supports nuanced reasoning (Huang et al., 2021), yet creating and maintaining them by hand is costly—especially in medicine, where new insights rapidly invalidate older facts (Yang et al., 2024).

We introduce an automated framework that constructs and continuously refines Medical Knowledge Graphs (MKGs) for medical QA. Our

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 12682–12701
November 4-9, 2025

LLM-driven agentic workflow, AMG-RAG, assisted by domain-specific search tools, generates graph entities enriched with metadata, confidence scores, and relevance indicators. This automation sharply reduces manual curation while keeping the graph aligned with the latest discoveries. In contrast to Retrieval Augmented Generation (RAG) systems that rely solely on vector similarity (Lewis et al., 2020), our graph-centric retrieval leverages explicit relationships to synthesise information across domains such as drug interactions, clinical trials, patient histories, and guidelines.

The AMG-RAG combines dynamically synthesised MKG with multi-step reasoning, guided by confidence scores and adaptive traversal strategies (Trivedi et al., 2022). This design yields more accurate and complete answers without incurring additional fine-tuning or inference costs. On the MEDQA and MEDMCQA benchmarks—both of which test evidence retrieval, complex reasoning, and multiple-choice comprehension—AMG-RAG achieves an F1 of 74.1% and an accuracy of 66.34%, respectively (Fig. 1). These results surpass those of similarly sized RAG approaches and even much larger state-of-the-art models, underscoring the benefits of a dynamically evolving MKG for medical QA. Our findings highlight the potential of automated, relationally enriched knowledge retrieval to enhance clinical decision-making by delivering timely and trustworthy insights (Zhou et al., 2023).

Contributions. Our contributions are threefold:

1. We developed an autonomous search and graph-building process powered by specialized LLM agents that continuously generate and refine MKGs through integrated workflows using search engines and medical textbooks.

2. Our system embeds confidence scoring mechanisms that explicitly model information uncertainty, providing transparent reliability assessments for medical information.

3. We created an adaptive graph traversal system that transcends traditional retrieval methods, enabling dynamic contextualization of medical knowledge.

2 Related Work

Medical QA has progressed mainly through three complementary lines of research: (i) domain-specific language models, (ii) retrieval-augmented generation, and (iii) knowledge-graph reasoning.

Domain-specific language models. BioBERT (Lee et al., 2020), PubMedBERT (Gu et al., 2021), and MedPaLM (Singhal et al., 2023) adapt transformer pre-training to biomedical corpora, delivering strong gains on entity recognition, relation extraction, and multiple-choice QA (Nazi and Peng, 2024; Liu et al., 2023). Yet, even these specialised models struggle to synthesise multi-hop relations (e.g., rare comorbidities or drug–gene interactions) and must be re-trained to absorb new discoveries (Rohanian et al., 2024; Yu et al., 2024).

Retrieval-Augmented Generation (RAG). RAG pipelines couple an LLM with an external evidence retriever, injecting fresh context at inference time (Lewis et al., 2020; Rezaei et al., 2024). Vendi-RAG (Rezaei and Dieng, 2025) and MMED-RAG (Xia et al., 2024) extend this paradigm to biomedical and multimodal sources, respectively. Chain-of-Thought (CoT) prompting further boosts reasoning: IRCoT (Trivedi et al., 2022) interleaves iterative retrieval with step-wise justification. Gemini's long-context model recently pushed MedQA scores beyond GPT-4 (Saab et al., 2024). Nevertheless, most RAG systems rely on static vector stores and cannot *explain* answers in terms of explicit biomedical relations.

Knowledge-graph reasoning. KG-Rank (Huang et al., 2021) and related work such as KG-RAG(Sanmartin, 2024) harness ontologies to re-rank evidence or enforce logical constraints, improving factual consistency in long-form QA (Yang et al., 2024) for RAG frameworks. Recent work by (Jiang et al., 2024) on knowledge-graph community retrieval for healthcare prediction shares a similar goal of graph-based reasoning in clinical settings. Their KARE framework demonstrates the value of structured knowledge representations for healthcare predictions. However, constructing and curating a high-coverage, up-to-date MKG remains labour-intensive, limiting scalability and freshness.

3 Method

We propose our framework, Agentic Medical Graph-RAG (AMG-RAG), bridges these threads by *dynamically* generating a confidence-scored Medical Knowledge Graph (MKG) that is *tightly coupled* to a Retrieval Augmented Generation

(RAG)+CoT pipeline. AMG-RAG features autonomous Knowledge Graph (KG) evolution through Large Language Model (LLM) agents extracting entities and relations from live sources with provenance tracking; graph-conditioned retrieval that maps queries onto the MKG to guide evidence selection; and reasoning over structured context where the answer generator utilizes both textual passages and traversed sub-graphs for transparent, multi-hop reasoning.

3.1 Retrieval Augmented Generation (RAG)

RAG is a framework designed to enhance Question Answering (QA) by integrating relevant external knowledge into the generation process. 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.

RAG with Chain-of-Thought (CoT). Enhancing RAG's performance can be achieved by integrating intermediate reasoning steps before 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.

RAG with Search. The RAGs's performance can be 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

Algorithm 1 AMG-RAG: KG-Based Medical QA Inference Pipeline

Require: Query q, Medical Knowledge Graph $\mathcal{G}$, Confidence Threshold τ, Document Limit M
Ensure: Final Answer $\hat{a}$ with Confidence $\hat{s}$

Medical Entity Recognition:
Extract medical terms: $\{n_1, n_2, \ldots, n_m\} \leftarrow \text{MER}(q)$, where $m \leq M$
Initialize reasoning traces: $\mathcal{C} \leftarrow \emptyset$
Initialize confidence: $s(n_i) \leftarrow 1.0$ for all terms n_i
for $i = 1$ to m **do** ▷ Process each medical term
  Graph Exploration:
  Retrieve node description: $d(n_i) \leftarrow \mathcal{G}.\text{getNodeData}(n_i)$
  Get connected nodes: $\mathcal{N}_i \leftarrow \{n_j : (n_i, n_j) \in \mathcal{G}\}$
  for each $n_j \in \mathcal{N}_i$ **do**
    Retrieve relationship: $(r_{ij}, s(r_{ij})) \leftarrow \mathcal{G}.\text{getEdge}(n_i, n_j)$
    Compute path confidence: $s_{\text{path}}(n_j) \leftarrow s(n_i) \cdot s(r_{ij})$
    if $s_{\text{path}}(n_j) \geq \tau$ **then**
      Add n_j to exploration queue with confidence $s_{\text{path}}(n_j)$
    end if
  end for
  Reasoning Trace Generation:
  $c_i \leftarrow \text{LLM}(n_i, \{d(n_j) : n_j \in \mathcal{N}_i, s_{\text{path}}(n_j) \geq \tau\})$
  $\mathcal{C} \leftarrow \mathcal{C} \cup \{c_i\}$
end for
Answer Synthesis:
$(\hat{a}, \hat{s}) \leftarrow G(\mathcal{C})$ where $\mathcal{C} = \{c_1, c_2, \ldots, c_m\}$
return $\hat{a}, \hat{s}$

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.

The suggested AMG-RAG framework dynamically creates a MKG and incorporates sophisticated reasoning abilities, overcoming the shortcomings of traditional methods. Our system utilizes structured medical knowledge and reasoning, ensuring flexibility to accommodate new data.

The AMG-RAG pipeline begins with question parsing, where an LLM agent extracts medical

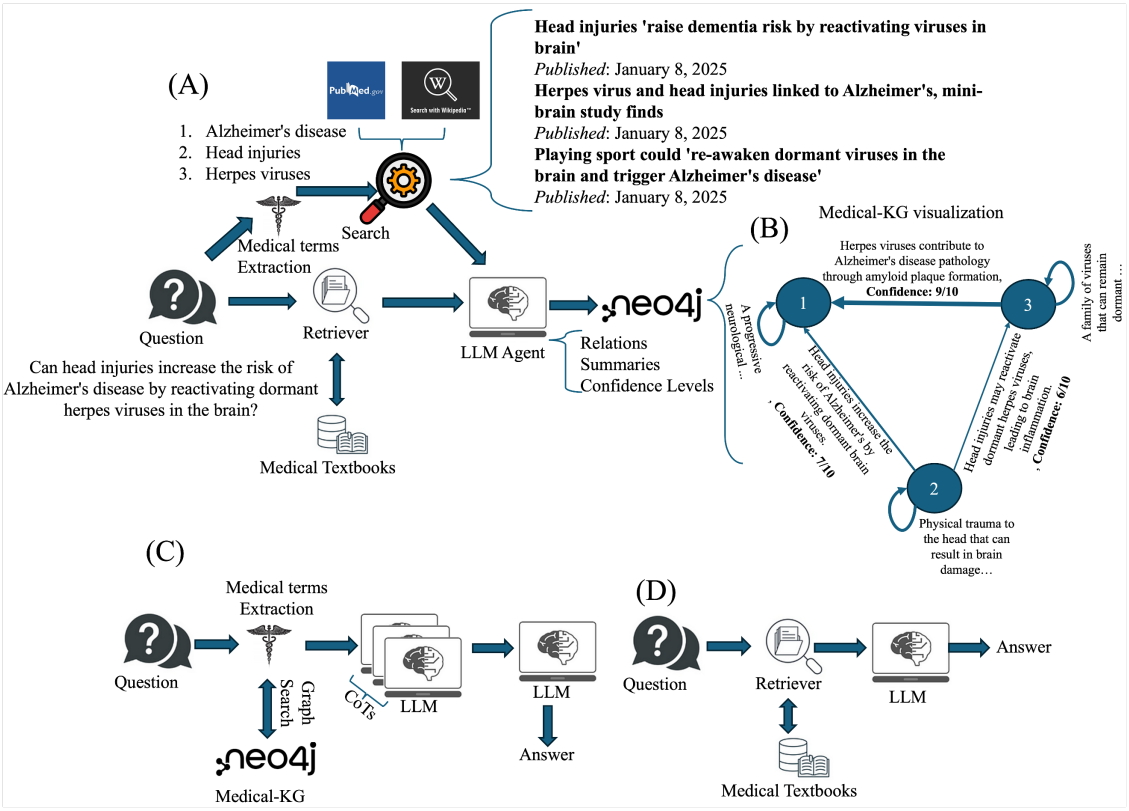

Figure 2: Model Schema. A) The pipeline for creating the MKG using search tools and an LLM agent. B) An example of the generated MKG in Neo4J, illustrating nodes and relationships derived from search results and contextual information. Our model successfully retrieved and utilized recent knowledge to accurately answer a medical question, highlighting the practical benefit of continuously updating the knowledge graph. Furthermore, we extended this evaluation by providing additional examples retrieved by our system using recent publications in Table 12. C) The AMG-RAG pipeline. D) A simplified RAG pipeline.

terms $\{\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_m\}$ from the user query $\mathbf{q}$:

$$\{\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_m\} = \text{LLM}(\mathbf{q}, M), \quad m \leq M. \tag{1}$$

Where M represents the upper limit of permissible medical terms that can be extracted. During node exploration, the system queries the KG for each term $\mathbf{n}_i$, applying a confidence threshold that filters relationships based on their reliability scores. The system propagates confidence through the KG by computing child confidence as:

$$s(\mathbf{n}_j) = s(\mathbf{n}_i) \cdot s(\mathbf{r}_{ij}), \quad \forall j \in \text{children of } i. \tag{2}$$

This design naturally models diminishing certainty over multi-hop reasoning chains and helps downrank low-reliability paths during inference.

Potential Failure Modes and Mitigation:

- **Over-pruning:** May occur if confidence thresholds are set too high, excluding rare but valid edges. We mitigate this through empirically-tuned thresholding that balances precision and recall.
- **Vanishing scores:** Propagation over long chains can lead to vanishingly small scores. We address this through hybrid traversal strategies—breadth-first to ensure local coverage, and depth-first for deeper exploration in sparse regions.

Our framework supports both breadth-first and depth-first exploration strategies, enabling flexible knowledge traversal based on query characteristics. The exploration continues until either cumulative confidence meets the threshold τ or the document limit M is reached, ensuring comprehensive yet focused information gathering.

The chain-of-thought generation phase synthesizes reasoning traces $\mathbf{c}_i$ for each entity by integrating information from connected nodes:

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

Finally, answer synthesis aggregates these reasoning traces to produce the final output $\hat{\mathbf{a}}$ with an associated confidence score:

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

This approach ensures that answers are comprehensive, interpretable, and anchored in reliable medical knowledge. The complete algorithm for the AMG-RAG inference procedure is presented in Algorithm 1.

3.3 Dynamic Generation of the Medical Knowledge Graph

The MKG represents a core innovation in our AMG-RAG framework, enabling structured reasoning through dynamically evolving knowledge representations. Unlike static knowledge bases, our approach constructs and continuously updates the graph based on incoming queries and newly discovered evidence, with confidence scores quantifying the reliability of each relationship.

Two-Phase Construction. Our MKG operates through complementary offline and online phases:

1. **Offline Foundation Building:** Prior to query processing, LLM agents construct a comprehensive base MKG from authoritative sources including PubMed articles and medical textbooks. This phase extracts medical entities with descriptions, identifies relationships with confidence scores, and stores the structured knowledge in Neo4j.

2. **Online Query-Driven Expansion:** During query processing, the system dynamically extends the MKG when gaps exist. It parses queries for medical terms, searches the existing graph, triggers real-time expansion for missing coverage, and integrates new knowledge while preserving graph structure.

An LLM agent identifies domain-specific terms from queries as defined in Equation 1. For each entity $\mathbf{n}_i$, specialized search tools retrieve contextual information:

$$\mathbf{d}(\mathbf{n}_i) = \text{MedicalSearch}(\mathbf{n}_i; \mathcal{S})$$

where $\mathcal{S}$ represents medical knowledge sources. These descriptions provide semantic definitions, clinical context, and recent research findings. Then, for entity pairs $(\mathbf{n}_i, \mathbf{n}_j)$, another LLM agent determines relationship type and assigns confidence scores:

$$(\mathbf{r}_{ij}, s_{ij}) = \text{LLM}_{\text{relation}}(\mathbf{d}(\mathbf{n}_i), \mathbf{d}(\mathbf{n}_j), \mathcal{E})$$

where $\mathcal{E}$ represents supporting evidence and $s_{ij} \in [0, 1]$ reflects evidence strength. Relationship types include short descriptions of the relation between entity $\mathbf{n}_j$ and $\mathbf{n}_i$. During graph traversal, confidence scores accumulate multiplicatively as defined in Equation 2.

This models uncertainty accumulation in multi-hop reasoning. Paths with $s_{\text{path}} < \tau$ are pruned to maintain reasoning quality.

Graph Maintenance. The MKG maintains consistency through incremental updates without full reconstruction. The final structure comprises nodes (medical entities with attributes), edges (typed relationships with confidence and provenance), and metadata (citations, timestamps, quality indicators).

This architecture enables the MKG to serve as both a persistent knowledge repository and a computational substrate for confidence-weighted reasoning, remaining current with evolving medical knowledge while maintaining interpretability and efficiency.

4 Experiments

The **MEDQA** dataset is a free-form, multiple-choice open-domain QA dataset 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

To address the challenges of inaccurate knowledge updating—such as those stemming from noisy retrieval results or LLM hallucinations—our AMG-RAG introduces a robust and dynamic approach to MKG construction. This is particularly critical in healthcare applications, where the absence of error detection and correction mechanisms in automated KG generation can compromise system reliability.

The dynamic update mechanism encompasses strategies resilient to errors, defined by the confidence level of the medical information retained within the MKG and among nodes i and j, denoted as $\mathbf{s}_{ij}$. This approach facilitates monitoring and reduces the spread of erroneous information during the refinement or reasoning stages. These protective measures allow the system to identify and rectify inconsistencies that may arise from external retrieved information.

The MKG is dynamically constructed for each question by integrating search items, contextual information, and relationships extracted from medical textbooks and search tools, including Wikipedia (Wiki-MKG) and PubMed (PubMed-MKG) queries. The ablation in Table 3 demonstrates that the created MKG based on PubMed (PubMed-MKG) is more effective in enhancing the performance of the AMG-RAG. This data is processed and structured within a Neo4j database. Key innovations in the knowledge graph include:

1. **Dynamic Node and Relationship Creation**: Nodes are instantiated based on retrieved entities and search terms, while relationships are constructed using predefined semantic templates aligned with medical ontologies.

2. **Bidirectional Relationships**: The graph includes both forward and reverse relationships between nodes to allow flexible traversal and comprehensive context understanding.

3. **Confidence-Based Relevance Scoring**: Each relationship is enriched with textual annotations and a quantitative confidence score that measures the reliability of the connection. This confidence score enables the system to down-rank or filter out uncertain associations, thereby mitigating the effects of noisy retrievals.

4. **Summarization with Reliability Indicators**: Each search item is paired with a concise summary derived from contextual sources. These summaries are accompanied by confidence scores that indicate their trustworthiness, allowing nuanced uncertainty modeling.

5. **Thresholding for Quality Control**: In our experiments, we applied a confidence threshold of 8 (on a 10-point scale) to retain only high-reliability nodes and edges. This value was empirically found to yield the best results in benchmark performance.

6. **Integration with Neo4j**: The complete graph is stored in a Neo4j database, leveraging its powerful graph query engine for efficient retrieval and analysis during inference.

A partial visualization of the MKG structure is shown in Figure 2.B. Additional complete examples with retrieved papers are provided in Table 12. This MKG forms the core knowledge source for the AMG-RAG inference pipeline.

As discussed in Appendix B, extensive validation was conducted through both human and machine evaluators. Clinical experts verified the correctness of the knowledge graph, and expert LLMs such as GPT-4 achieved high accuracy (e.g., 9/10) in validating the extracted knowledge. These results underscore the MKG's ability to support reliable and explainable medical reasoning within AMG-RAG.

The knowledge graph creation process in AMG-RAG operates independently from the QA process, allowing for continuous background updates of the MKG via search tools such as PubMedSearch or WikiSearch. This approach significantly reduces latency during question answering since the system frequently retrieves information from the pre-populated MKG rather than performing new

searches. By maintaining an updated MKG, AMG-RAG achieves a balanced minimum dependency on computational resources and search tools during the test phase.

Despite having only 8B parameters, it delivers competitive results compared to much larger models like Med-Gemini (1800B) and GPT-4 (1760B). Even in worst-case scenarios where relevant information is absent from the MKG, the additional search cost is still significantly lower than the resource requirements of much larger models.

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 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 8, 9, 10, and 11, 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 mechanisms.

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 reasoning, dynamically generated MKG, 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.

Impact of Search Tools on MKG creation and CoT Reasoning on AMG-RAG Performance. Figure 3 and Table 3 demonstrate the effect of integrating different search tools for creating the MKG on the performance of the AMG-RAG system applied to the MEDQA benchmark. Incorporating these external retrieval capabilities significantly enhances both accuracy and F1 scores, as they allow the model to access relevant and up-to-date evidence critical for answering complex medical questions. Among the two search tools for creating the MKG, PubMed-MKG consistently outperforms Wiki-MKG, likely due to its focused, domain-specific content that aligns closely with the specialized nature of medical QA tasks.

In addition to the integration of the dynamical MKG, the reasoning module plays a pivotal role in performance. As highlighted in Figure 3, ablating either CoT or MKG integration causes a considerable degradation in accuracy and F1 score. This demonstrates that structured multi-hop reasoning and medical knowledge grounding through the MKG are indispensable for the system's ability to deliver accurate and evidence-based answers."

Comparison Against Traditional RAG Models. Table 3 presents a comprehensive comparison of various RAG models evaluated on the MEDQA benchmark. This includes models with different retrieval mechanisms and model sizes, enabling a head-to-head evaluation of AMG-RAG with other state-of-the-art baselines such as Self-RAG (Asai et al., 2023), HyDE (Gao et al., 2022), GraphRAG (Edge et al., 2024), and MedRAG (Zhao et al., 2025). The results clearly show that AMG-RAG configured with the PubMed-MKG and an 8B LLM backbone achieves the highest accuracy of 73.92%, surpassing all competing models. Notably, ablation results indicate that removing search functionality or CoT reasoning significantly degrades accuracy (dropping to 67.16% and 66.69%, respectively), confirming the essential role of structured retrieval and reasoning components in complex question

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	✓	✗	✗

Table 1: Comparison of LLM models on the MEDQA Benchmark. Additional comparison with RAGs are provided in Table B

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

Table 2: Comparison of Models on the MedMCQA.

answering. Other baseline models such as Gemini-pro and PMC-LLaMA demonstrate weaker performance, further validating the efficacy of domain-aware retrieval and reasoning modules proposed in AMG-RAG. Importantly, the domain specificity and freshness of PubMedSearch provide a significant advantage in retrieving relevant knowledge that general-purpose search modules often fail to deliver.

Model	Size	Accuracy (%)
AMG-RAG	PubMed-MKG-8B	**73.92**
	Wiki-MKG-8B	70.62
	No-MK-8B	67.16
	No-MKG & CoT-8B	66.69
Self-RAG	8B	67.32
	HyDE-8B	68.32
RAG	Gemini-pro	64.5
	70B	56.2
	8B	64.3
GraphRAG	Gemini-pro	65.1
	70B	55.1
	8B	64.8
MedRag	70B	49.57
	13B	42.58
PMC-LLaMA	13B	44.38

Table 3: Comparison of MEDQA accuracy across various RAG models and retrieval strategies.

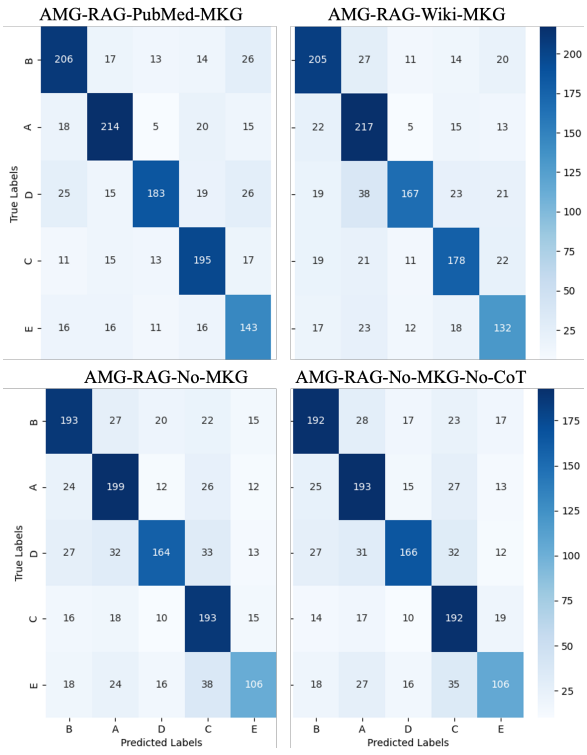

Figure 3: Confusion matrix for AMG-RAG with and without CoT and Knowledge Graph integration on the MEDQA dataset.

Comparison Against LLM Backbones. In addition to evaluating different retrieval strategies, we assess how the choice of LLM backbone influences performance in Table 4. This comparison highlights that AMG-RAG built on GPT4o-mini with PubMed-MKG achieves the best performance (73.92%). In contrast, performance declines when switching to LLaMA 3.1 or Mixtral, even when using the same retrieval pipeline. These results reinforce the importance of synergy between the language model and the retrieval mechanism. Larger models do not necessarily guarantee higher

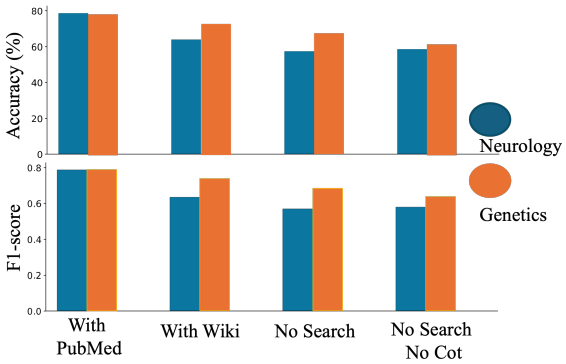

Figure 4: Performance comparison across different question domains in the Neurology and Genetics fields.

accuracy—domain alignment and reasoning ability, such as that of GPT4o-mini, are crucial for success on high-stakes tasks like medical QA.

Model	Config-Size	Accuracy (%)
GPT4o-mini	PubMed-MKG-8B	**73.92**
	No-MKG & CoT-8B	66.69
LLaMA 3.1	PubMed-MKG-8B	66.5
	No-MKG-8B	62.6
Mixtral	PubMed-MKG-8×7B	61.4
	No-MKG-8×7B	53.2
GPT 3.5	PubMed-MKG	65.2
	No-MKG	58.4

Table 4: AMG-RAG performance across different LLM backbones on the MEDQA benchmark.

Improving QA in Rapidly Changing Medical Domains with AMG-RAG. Figure 4 shows AMG-RAG's superior performance in rapidly evolving subfields like Neurology and Genetics. This advantage stems from real-time PubMed integration during inference, combined with structured reasoning and knowledge graph grounding, enabling precise answers to complex medical questions with enhanced interpretability and trustworthiness.

4.3 Latency Analysis and Deployment Considerations

To quantify the performance characteristics of AMG-RAG under realistic deployment conditions, we conducted comprehensive latency benchmarks using 100 randomly selected queries from the MEDQA dataset. Our analysis in Table 5 reveals that the offline mode is the optimal deployment scenario for clinical environments requiring predictable response times, while the 1.2s end-to-end performance falls well within acceptable bounds for interactive clinical decision support systems.

Dynamic search mode exhibits substantially higher latency due to real-time PubMed queries and on-the-fly entity extraction. However, this mode provides the most current medical knowledge by accessing recent publications and emerging research findings.

Scenario	MKG Lookup	End-to-End QA
Offline Mode	∼35 ms	∼1.2 s o
Dynamic Search	∼1.9 s	∼4.6–6.5 s
Hybrid	∼200–500 ms	∼3.1–4.8 s

Table 5: Latency Analysis Under Different Operating Modes

The hybrid mode offers a balanced compromise, with MKG lookup times ranging from 200–500ms depending on the extent of dynamic augmentation required. The variable performance range (3.1–4.8s end-to-end) reflects the adaptive nature of the system, where lookup complexity scales with the specificity and recency of medical terms.

4.4 Domain Generalizability and Future Applications

While our focus is medical QA, the AMG-RAG framework's principles are fundamentally domain-agnostic. The AMG-RAG framework can extend to legal research (Zhong et al., 2020), scientific literature analysis (Lo et al., 2019), and other domains requiring dynamic knowledge integration and multi-hop reasoning capabilities (Yang et al., 2018).

Financial Domain Example: Consider the query: "How do Federal Reserve interest rate hikes affect Apple's stock price?"

Knowledge Graph Structure: Fed Funds Rate → Consumer Borrowing Cost → Apple Product Demand → Apple Stock Price

Reasoning: Higher Fed rates increase consumer borrowing costs, reducing discretionary spending on Apple products, which pressures stock performance. Additional examples are provided in Appendix G.

5 Conclusion

We introduce AMG-RAG, an advanced QA system that dynamically constructs MKG while integrating sophisticated structured reasoning for medical QA. The system demonstrates 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.

6 Limitations

Despite AMG-RAG's advancements, our approach has certain limitations. Firstly, it relies on external search tools which introduce latency during the creation of the MKG. However, this occurs only once, when the MKG is built from scratch for the first time. Additionally, while the system 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. To some degree, our AMG-RAG framework tackles these concerns using a variety of mechanisms:

Reliability and Uncertainty. We implement confidence scoring for both entities and relationships in our MKG to validate information quality and quantify uncertainty in medical evidence.

Bias Mitigation Limitations. We acknowledge that bias mitigation remains challenging in medical QA systems. Current public benchmarks (MEDQA, MedMCQA) lack demographic annotations, limiting systematic bias evaluation. Our focus has been on demonstrating factual reliability and multi-hop reasoning capabilities. Future work will incorporate diversity-aware retrieval techniques using structured metadata (MeSH tags, demographic stratification) and graph-regularization mechanisms to ensure equitable knowledge representation.

Environmental Impact. Our 8B-parameter system offers significant environmental advantages over larger models (340B–1800B parameters), achieving $\sim$0.36 gCO_2e per query—a $10\breve{}20\times$ reduction compared to models like GPT-4 (4–7 gCO_2e per query). This efficiency stems from knowledge graph-guided retrieval that avoids extensive fine-tuning and full-document generation.

Clinical Deployment. While clinical deployment is beyond this paper's scope, AMG-RAG's confidence scoring and transparent reasoning provide foundations for clinical auditability. The system can be deployed on HIPAA-compliant infrastructure for future clinical applications. Our evaluation uses publicly available benchmarks to demonstrate research utility while avoiding clinical data experimentation complexities.

This work establishes AMG-RAG's technical validity through standardized evaluations, with clinical deployment and broader ethical considerations as important future directions.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

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*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Precise zero-shot dense retrieval without relevance labels, 2022. *URL https://arxiv.org/abs/2212.10496*.

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.

Pengcheng Jiang, Cao Xiao, Minhao Jiang, Parminder Bhatia, Taha Kass-Hout, Jimeng Sun, and Jiawei Han. 2024. Reasoning-enhanced healthcare predictions with knowledge graph community retrieval. *arXiv preprint arXiv:2410.04585*.

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.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S Weld. 2019. S2orc: The semantic scholar open research corpus. *arXiv preprint arXiv:1911.02782*.

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 and Adji Bousso Dieng. 2025. Vendi-rag: Adaptively trading-off diversity and quality significantly improves retrieval augmented generation with llms. *Preprint*, arXiv:2502.11228.

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*.

Diego Sanmartin. 2024. Kg-rag: Bridging the gap between knowledge and creativity. *arXiv preprint arXiv:2405.12035*.

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*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

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.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. 2025. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. *arXiv preprint arXiv:2502.04413*.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Jec-qa: a legal-domain question answering dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9701–9708.

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 specialists 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 6 and 7, 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 6, 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 6 and 7 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 8 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 9, 10, and 11

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:

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.

Table 6: Examples from the Medical Knowledge Graph (MKG) with Expert and Blind Analysis (Part 1)

- **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):

- **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

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.

Table 7: Examples from the Medical Knowledge Graph (MKG) with Expert and Blind Analysis (Part 2)

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).

Does Neo4j-Based Storage scale well? Neo4j's scalability for our medical knowledge graph storage is strategically robust, offering several key advantages for large-scale, relationship-intensive medical data. Its graph-based architecture is particularly well-suited for handling highly interconnected medical knowledge networks, supporting horizontal scaling that enables efficient performance even as the knowledge base grows. The cloud-based accessibility further enhances the framework's flexibility, allowing seamless knowledge sharing and distributed access without local storage constraints.

How large were the knowledge graphs? Our automatically constructed medical knowledge graphs demonstrate significant complexity and depth, comprising approximately 76,681 nodes and 354,299 edges. These nodes encompass a comprehensive range of medical entities including diseases, symptoms, treatments, drugs, anatomical structures, and clinical findings, all interconnected through semantically meaningful, typed relationships. This substantial scale not only reflects the intricate nature of medical knowledge but also enables more nuanced, multi-hop reasoning capabilities across diverse medical queries. The graph's architecture allows for dynamic expansion and refinement, ensuring that the knowledge representation remains both comprehensive and adaptable to emerging medical research and understanding.

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 8: 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

F Additional Results

G Across Domain Examples

Q1: How do Federal Reserve interest rate hikes affect Apple's stock price, and why?

Knowledge Graph: Fed Funds Rate $\xrightarrow{\text{raises}}$ Consumer Borrowing Cost $\xrightarrow{\text{reduces}}$ Apple Product Demand $\xrightarrow{\text{lowers}}$ Apple Stock Price

Reasoning: When the Fed raises interest rates, borrowing costs for consumers increase. This tends to reduce spending on discretionary items like Apple products. Lower sales then translate into weaker stock performance for Apple.

Answer: Yes—higher Fed rates increase consumer costs, reducing demand for Apple products and pressuring its stock price.

Q2: How does quantitative easing influence stock market returns, and through what mechanism?

Knowledge Graph: Quantitative Easing $\xrightarrow{\text{boosts}}$ Liquidity $\xrightarrow{\text{suppresses}}$ Interest Rates $\xrightarrow{\text{enhances}}$ Stock Returns

Reasoning: When central banks implement QE, they increase liquidity in the financial system. This drives down interest rates, making equities more attractive compared to fixed income. As a result,

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 9: 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

overall stock returns tend to rise while volatility decreases.

Answer: Yes—QE injects liquidity that lowers rates, boosting equity returns and reducing market volatility.

Q3: Will a Fed rate cut in Q3 2025 significantly boost the "Magnificent 7" tech stocks?

Knowledge Graph: Fed Rate Cut $\xrightarrow{\text{reduces}}$ Discount Rate $\xrightarrow{\text{increases}}$ PV of Future Earnings $\xrightarrow{\text{raises}}$ "Magnificent 7" Stock Prices

Reasoning: As of Q3 2025, a Federal Reserve rate cut would reduce the discount rate used in financial models, thereby increasing the present value of long-term earnings. The "Magnificent 7" tech companies—known for their strong growth trajectories—typically benefit the most from such revaluations. However, recent data indicates that this effect may be moderating due to already elevated valuations and shifting investor focus toward profitability and macro risks.

Answer: Yes—as of Q3 2025, a Fed rate cut is expected to support the 'Magnificent 7' tech stocks by boosting the present value of their future earnings, though the effect may be less pronounced compared to previous cycles.

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 10: 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 11: 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.

Example 1	**Question:** A 29-year-old man presents with infertility. He has been trying to conceive for over 2 years. His wife has no fertility issues. Exam shows tall stature, long limbs, sparse body hair, gynecomastia, and small testes. Labs reveal elevated FSH and a high estradiol:testosterone ratio. Cytogenetic analysis indicates a chromosomal abnormality. If inherited from the father, during which stage of spermatogenesis did this error most likely occur? **Choices:** A: Primary spermatocyte, B: Secondary spermatocyte, C: Spermatid, D: Spermatogonium, E: Spermatozoon **Answer:** A (Primary spermatocyte) **Reasoning:** This corresponds to an error in meiosis I during the father's spermatogenesis, consistent with Klinefelter syndrome due to paternal nondisjunction. **Retrieved Papers:** 1) Black et al., *The Genetic Landscape of Male Factor Infertility*, Uro, 2025. 2) Niyaz et al., *Chromosome Disorders in Sperm Anomalies*, 2025. 3) Leslie et al., *MNS1 variant and Male Infertility*, EJHG, 2020.
Example 2	**Question:** A 23-year-old woman is referred for genetic counseling after her brother is diagnosed with hereditary hemochromatosis. She is asymptomatic and her labs are normal. Which gene mutation is most consistent with hereditary hemochromatosis? **Choices:** A: BCR-ABL, B: BRCA, C: FA, D: HFE, E: WAS **Answer:** D (HFE gene) **Reasoning:** Most hereditary hemochromatosis cases in Northern European populations are caused by HFE mutations (C282Y, H63D). Even asymptomatic individuals with normal iron studies should be screened if they have an affected first-degree relative. **Retrieved Papers:** 1) Delatycki & Allen, *Population Screening for HH*, Genes, 2024. 2) Lou et al., *Utility of Iron Indices in HH Genotyping*, Clin. Biochem., 2025. 3) Lucas et al., *HFE Genotypes and Outcomes*, BMJ Open, 2024.

Table 12: Examples of AMG-RAG-generated answers with structured reasoning and citation-based grounding for clinical QA.