Black Sea Journal of Engineering and Science
doi: 10.34248/bsengineering.1849342

Open Access Journal
e-ISSN: 2619 – 8991

Research Article
Volume 9 - Issue 2: 549-561 / March 2026

DISSECTING MEDICAL RAG: WHY RERANKING MATTERS MORE THAN COMPLEXITY IN QUESTION ANSWERING

Hakan EMEKCİ[1]*, Daniel Quillan ROXAS[1]

[1]*TED University, Applied Data Science Department, 06420, Ankara, Türkiye*

Abstract: Retrieval-Augmented Generation (RAG) systems integrate large language models with information retrieval to ground responses in factual data. This study systematically evaluates the contribution of each RAG component in a medical question answering system through comprehensive ablation analysis. We designed a hierarchical RAG architecture with six key components: hierarchical intent classification, query rewriting, two-stage retrieval (dense retrieval with FAISS + cross-encoder reranking using Clinical-Longformer), and specialist routing. We conducted systematic ablation studies across seven configurations on 476 medical questions from MedQA benchmarks. Each configuration was evaluated independently using GPT-4o mini as an LLM judge across four metrics: context relevance, completeness, faithfulness, and correctness (1-5 Likert scale), with each metric assessed through separate evaluation calls to minimize inter-metric bias. Statistical significance was validated through paired t-tests with effect size calculations (Cohen's d). The full system achieved an overall score of 3.64/5.0. Systematic ablation revealed two critical components: reranking (removal: -0.24 overall, $P<0.001$, $d = -0.44$) and specialists (removal: -0.17 overall, $P<0.001$, $d = -0.29$), both showing small but statistically significant effect sizes. Surprisingly, hierarchical intent classification degraded performance when included (+0.09 when removed, $p = 0.010$ for completeness), suggesting simpler query processing may be preferable. Query rewriting showed minimal impact (-0.04 overall), while raw query inclusion significantly affected completeness (-0.15, $P<0.001$). Reranking and specialist components are essential for medical RAG systems, with statistical significance confirmed across 476 queries. The counterintuitive finding that hierarchical intent classification degrades performance ($P<0.05$) suggests that architectural complexity does not always improve RAG system quality. These results provide evidence-based guidance for designing medical question answering systems, showing that reranking infrastructure and domain expertise are more critical than sophisticated query understanding techniques.

Keywords: Artificial intelligence, Information retrieval, Reranking, Retrieval-augmented generation

***Corresponding author:** TED University, Applied Data Science Department, 06420, Ankara, Türkiye
E mail: hakan.emekci@tedu.edu.tr (H.EMEKCİ)
Hakan EMEKCİ https://orcid.org/0000-0002-4074-5600
Daniel Q. ROXAS https://orcid.org/0009-0000-4484-6751

Received: December 28, 2025
Accepted: January 28, 2026
Published: March 15, 2026

Cite as: Emekci, H., & Roxas, Q. R. (2026). Dissecting medical RAG: Why reranking matters more than complexity in question answering. *Black Sea Journal of Engineering and Science*, *9*(2), 549–561.

1. Introduction

Recent advances in large language models (LLMs) show a strong potential for applications in the medical field (Thirunavukarasu et al., 2023). However, challenges such as hallucinations, bias, and potential misinformation are major concerns. To counteract this, more effective methods, such as retrieval augmentation, have been developed (Lewis et al., 2020)

The medical domain presents unique challenges for information retrieval and question answering systems. Medical terminology is highly specialized, queries can belong to multiple subdomains, and the consequences of providing inaccurate information are more potentially harmful than in general knowledge domains. Jin et al. (2021) conducted a detailed survey of biomedical question answering approaches, highlighting the complexity of this task. Question answering systems in medical domains struggle with limited annotated data and lexical gaps that slow the development of patient-focused systems (Jin et al., 2022). Several approaches have been proposed to address these challenges. Knowledge-infused models incorporate relevant medical entities and use synthetic data generation to improve answer accuracy (Manas et al., 2021). Semantic grounding using knowledge graphs can support query refinement in medical document retrieval (Selmi et al., 2022), while heterogeneous knowledge retrieval enhances medical question-answering by using diverse information sources (Zhao et al., 2024).

Hierarchical approaches have proven effective in complex information classification tasks. Zhang et al. (2025) demonstrated that LLMs-embedded taxonomy frameworks significantly outperform flat classification methods by maintaining consistency between hierarchical levels. Their work shows that the use of taxonomic structures leads to more accurate information organization and retrieval, with improvements of up to 16% for items incorrectly classified at one level but correctly identified at other levels. This suggests that hierarchical LLM-based approaches can effectively capture the relationships between classes in complex

domains.

Although existing approaches show promise, important challenges remain in developing systems that reflect real-world healthcare information needs and ensure the reliability of the response. Kim et al. (2025) note that many medical QA systems do not communicate confidence levels or consider the reliability of answer sources, which are critical considerations in clinical settings where information accuracy directly impacts patient care.

To address these limitations, we present a medical retrieval-augmented generation system designed with a hierarchical architecture and domain-specific components, specifically focused on answering the questions of the clinical guidelines. Unlike general-purpose question answering systems that may lack domain expertise, our system integrates:

- A centralized coordinator that manages the query processing pipeline
- A two-stage hierarchical classifier based on the CMID taxonomy (Chen et al., 2020)
- An intent-aware query rewriter that transforms natural language into structured search statements
- A two-stage retrieval architecture combining efficient dense retrieval (FAISS [Johnson et al., 2021] with S-PubMedBert-MS-MARCO [Deka et al., 2022] embeddings) with precise cross-encoder reranking (Clinical-Longformer [Beltagy et al., 2020; Li et al., 2023]), implementing cascade ranking principles that balance computational efficiency with retrieval accuracy
- Domain-specific specialist agents for different medical areas

This architecture processes complex medical queries across multiple domains while maintaining factual accuracy and relevance. The two-stage retrieval design is particularly important: while bi-encoder models enable efficient first-pass retrieval across large document collections, cross-encoder reranking provides better relevance estimation through direct query-document interaction modeling, a critical requirement for medical applications where precision is paramount. By incorporating specialist knowledge into the response generation process, we address the domain expertise gap identified in previous research.

Through systematic ablation studies on 476 medical questions from the MedQA benchmark (Jin et al., 2022), we quantify the impact of each RAG component on response quality using rigorous statistical validation (paired t-tests, effect sizes). Our experiments reveal two critical findings: (1) reranking emerges as the most important component (-0.24 overall impact, $P<0.001$, Cohen's $d=-0.44$), primarily affecting relevance and faithfulness, and (2) specialist agents provide substantial benefits (-0.17 overall, $P<0.001$, $d=-0.29$). Counterintuitively, hierarchical intent classification degrades performance when included (+0.09 when removed for completeness, $P=0.010$), challenging the assumption that sophisticated query understanding mechanisms necessarily improve RAG systems. These findings align with Yang et al.'s (2025) work on MedAide, which demonstrates that specialized multi-agent collaboration enhances medical proficiency.

This research contributes to ongoing efforts to develop effective methodologies that make use of the potential of LLMs while addressing unique challenges in the medical domain. By elaborating situations and improving context understanding (Sun et al., 2019) and reducing hallucination through retrieval-augmented generation (Chu et al., 2025), our work advances toward more reliable and clinically useful medical question answering systems for clinical guideline queries.

1.2. Literature Review

Medical question answering systems have evolved greatly over the past decade. The BioASQ challenge introduced by Tsatsaronis et al. (2015) established a framework for evaluating biomedical semantic indexing and question answering capabilities, providing benchmarks that have guided subsequent research. This challenge highlighted the complexity of medical questions and the need for specialized approaches to handle domain-specific terminology and knowledge.

Ben Abacha and Demner-Fushman (2019) studied question understanding for consumer health inquiries, noting that medical questions often contain peripheral information that complicates accurate response generation. Their work on question summarization demonstrated how extracting the core information needs from verbose queries can improve retrieval performance. This finding influenced our approach to query rewriting based on intent classification.

Recent work by Maharjan et al. (2024) has shown that well-designed prompting strategies can be as effective as fine-tuning for medical question answering with open-source LLMs. Their OpenMedLM platform achieved strong performance on medical benchmarks through prompt engineering alone. This evidence supports our use of intent-specific prompting templates within our specialist components.

Traditional IR systems often treat all queries uniformly, applying the same retrieval and ranking functions regardless of user intent. However, medical information needs vary across different query types. Our hierarchical intent classification implements the concept of query understanding, which has deep roots in IR theory. Belkin's Anomalous State of Knowledge (ASK) model suggests that users cannot always articulate their information need precisely (Belkin et al., 1982). Our two-stage classification identifies the type of medical knowledge gap (disease, medicine, treatment, or other) for more targeted retrieval and response generation.

Rather than using generic term co-occurrence or relevance feedback, our intent-aware query rewriting adapts reformulation based on detected intent. For example, medicine queries emphasizing side effects receive different expansion patterns than disease queries

focusing on symptoms. This aligns with research showing that context-aware query reformulation outperforms intent-agnostic methods.

The efficiency-effectiveness trade-off in information retrieval has led to cascade ranking architectures, where computationally inexpensive methods narrow the candidate pool before applying more sophisticated but costly ranking functions. Our two-stage retrieval implements this principle: FAISS (Johnson et al., 2021) with dense retrieval provides efficient first-pass ranking across 20,833 documents, while Clinical-Longformer (Beltagy et al., 2020; Li et al., 2023) cross-encoder reranking applies expensive query-document interaction modeling only to the top-20 candidates.

This design follows Robertson's Probability Ranking Principle (Robertson, 1997), which states that documents should be ranked by their probability of relevance. While embedding similarity provides an efficient approximation of relevance, cross-encoders model the interaction between query and document text directly, providing superior relevance estimation at the cost of computational complexity. Medical queries particularly benefit from this precision-focused second stage, as the high-stakes nature of medical information demands accuracy over recall in the result set.

Recommender systems and personalized search adapt results to user context or preferences. Our specialist framework applies analogous principles through intent-based personalization. Rather than personalizing based on user history or demographics, we personalize based on the detected information need type. This approach recognizes that a medicine query requires fundamentally different information organization and emphasis than a treatment query, even when addressing the same underlying condition.

Some medical queries exhibit polyrepresentation, which means they have multiple valid interpretations or information needs. When our classifier detects distributed probability across multiple intent types (entropy above threshold), the system consults multiple specialists in parallel. This approach draws from diversity-in-ranking research, particularly Maximal Marginal Relevance (MMR) (Carbonell & Goldstein, 1998), which balances relevance with diversity to ensure complete coverage. In the medical domain, this hedging strategy addresses uncertainty while maintaining factual grounding through retrieval.

Intent classification forms a core component of our system. Chen et al. (2020) developed the CMID dataset, which categorizes Chinese medical questions into 4 main types and 36 subtypes. This hierarchical taxonomy provides a comprehensive framework for classifying medical intents, which we adapted for our work. While Chen's research focused primarily on classification accuracy, our work extends this by integrating intent classification into the retrieval and response generation pipeline. In the broader context of intent recognition, Casanueva et al. (2020) demonstrated that dual sentence encoders can achieve strong intent classification performance with limited training data. Their work on efficient intent detection influenced our two-stage classification approach, which uses different models for main intent and subtype identification.

Fu et al. (2022) applied hierarchical networks to clinical-trial-outcome predictions, showing that multi-level classification can capture the relationships between different aspects of medical information. This provides precedent for our hierarchical approach to medical intent classification, though in a different application domain.

A distinctive feature of our system is its specialist-based architecture. Most directly relevant to our approach, Yang et al. (2025) proposed MEDAide, an LLM-based omni medical multi-agent collaboration framework for specialized healthcare services. MEDAide performs query rewriting through retrieval-augmented generation to accomplish accurate medical intent understanding, uses a contextual encoder to recognize fine-grained intents, and activates specialized paramedical agents based on intent relevance. This approach shares important similarities with our specialist framework, particularly in the use of intent-based routing to specialized components. However, while MEDAide focuses on agent collaboration for complex intent understanding, our work emphasizes the integration of hierarchical intent classification with retrieval-augmented generation and provides ablation studies that quantify the contribution of specialist components.

Simonds et al. (2024) similarly demonstrated that routing prompts to domain-specialized LLMs can improve performance over general-purpose models. While their MoDEM system shares our goal of using specialization through routing, it uses a flat classification approach rather than the hierarchical intent taxonomy we implement.

The question of whether specialized or general models perform better in medical contexts was examined by Dorfner et al. (2025), who found that biomedically fine-tuned LLMs do not consistently outperform general-purpose models on unseen medical tasks. This suggests that architectural specialization, rather than domain-specific pretraining alone, may offer advantages for medical applications, a hypothesis our ablation studies help to evaluate.

Our work contributes to the growing field of retrieval-augmented generation (RAG) for healthcare applications. Jeong et al. (2024) proposed Self-BioRAG, a framework for generating explanations, retrieving domain-specific documents, and self-reflecting on generated responses in the biomedical domain. Their work emphasizes the importance of domain-specific components in medical RAG systems, aligning with our specialist-based approach.

Gu et al. (2021) conducted a thorough evaluation of domain-specific language model pretraining for biomedical NLP, showing that domain-specific pretraining from scratch can outperform continual

pretraining of general-domain models. While their work focused on model pretraining rather than architectural design, it supports the value of domain specialization in medical NLP tasks.

In the context of medical information retrieval, Lu et al. (2023) developed Entity-BERT for entity recognition in electronic medical records, demonstrating improvements in retrieval performance through better entity identification. This research informed our approach to query rewriting, which similarly aims to identify key elements in user queries to improve retrieval accuracy.

Our evaluation framework builds on established approaches in medical QA assessment. Ben Abacha et al. (2023) studied evaluation metrics for medical text generation, comparing automatic metrics to human judgments. Their finding that factual correctness is particularly important in medical contexts influenced our decision to include it as a primary evaluation dimension.

Mishra et al. (2014) provided a systematic review of text summarization in the biomedical domain, identifying key dimensions for evaluation. Their framework, which considers input characteristics, purpose, output format, method, and evaluation approach, informed our multi-faceted assessment methodology.

Wei et al. (2024) addressed the challenge of hallucination detection in medical AI systems, proposing methods for evaluating the factuality of generated content. This research underscores the importance of factual correctness in medical applications and supports our inclusion of it as a primary evaluation criterion.

Building on this prior work, our research makes several contributions:

• We present a hierarchical intent classification system for medical questions that integrates with both retrieval and response generation, grounded in IR principles of query understanding and personalization

- We implement a two-stage retrieval architecture following cascade ranking principles, combining efficient dense retrieval with precise cross-encoder reranking
- We develop a specialist framework that processes different types of medical queries through domain-specific components
- We conduct systematic ablation studies on 476 medical questions from MedQA benchmarks that quantify the impact of different RAG components on response quality, with statistical validation through paired t-tests and effect size calculations
- We provide evidence that reranking is the most critical component (-0.24 overall, P<0.001), followed by specialists (-0.17, P<0.001), while revealing the counterintuitive finding that hierarchical intent classification degrades performance (+0.09 when removed, P=0.010)

Our work provides actionable, evidence-based guidance for medical RAG system design by quantifying component contributions with statistical rigor, demonstrating that reranking infrastructure and domain-specific specialists should be prioritized over sophisticated query understanding mechanisms.

2. Materials and Methods

2.1. System Architecture

Our medical RAG system is built on a hierarchical classification and specialist-based architecture consisting of five core components. The central coordinator handles the query processing pipeline and routes queries to appropriate specialists based on intent classification. A two-stage hierarchical classifier identifies main types and subtypes using the CMID (Chinese Medical Intent Dataset) taxonomy, incorporating 4-class main intent and 36-class subintent classification. The intent-aware rewriter transforms raw natural language queries into structured search statements optimized for retrieval. A two-stage retrieval system implements cascade ranking with dense retrieval and cross-encoder reranking. Finally, domain-specific specialist agents generate responses for different medical areas.

2.2. Processing Flow

The system processes queries through a sequential pipeline that ensures accurate intent classification, relevant context retrieval, and factual response generation. The process begins when a user submits a natural language medical query to the system. This raw query is preserved throughout the pipeline to ensure that responses remain faithful to the original information need.

The hierarchical classifier processes the query using a two-stage approach. First-stage softmax identifies the main intent type across 4 classes: disease, medicine, treatment, and other. Second-stage softmax identifies the specific subtype within the main type, selecting from a total of 36 subtypes distributed across the main classes. Conversation intent is classified separately to distinguish informational queries from conversational exchanges. The classification process for main intent types is formalized in equation 1, and the classification process for subtypes is formalized in equation 2:

$$P(t_i|q) = \frac{exp(f_\theta(q,t_i))}{\Sigma_{j=1}^{4} exp(f_\theta(q,t_i))} \quad (1)$$

$$P(s_k|q,t_i) = \frac{exp\left(g_\phi(q,t_i,s_k)\right)}{\Sigma_{l=1}^{n_i} exp\left(g_\phi(q,t_i,s_l)\right)} \quad (2)$$

where $P(t_i|q)$is the probability of main intent type t_i given query q, $P(s_k|q,t_i)$ is the probability of subtype s_k given the main type t_i, n_i is the number of subtypes for main type t_i, and f_θ and g_ϕ are the classification models are the classification models where θ and ϕ represent the learnable weight parameters of the neural network models for main-type and sub-type classification, respectively.

Once the intent is classified, the query undergoes intent-aware rewriting. Intent information is passed to the rewriter, which applies domain-specific transformations to improve retrieval. The output consists of structured

search statements optimized for the vector store. The rewriting process can be represented as $q' = R(q, t, s)$ where q' is the rewritten query, q is the original query, t is the main intent type, and s is the subtype.

Based on the classification results, the coordinator determines whether to use a single specialist or multiple specialists for multi-intent queries, whether to use the RAG pipeline or provide a direct response, and which specific specialist(s) to invoke. For queries requiring additional information, the rewritten query is used for initial retrieval from the vector store, with reranking applied to improve document selection. The raw query is retained as a fallback if the rewritten query fails to retrieve relevant documents.

2.3. Step-by-Step Classification Example

To illustrate how the hierarchical classification system operates in practice, we present an example using a query from our evaluation dataset. Figure 1 shows the general processing pipeline of our medical question answering system.

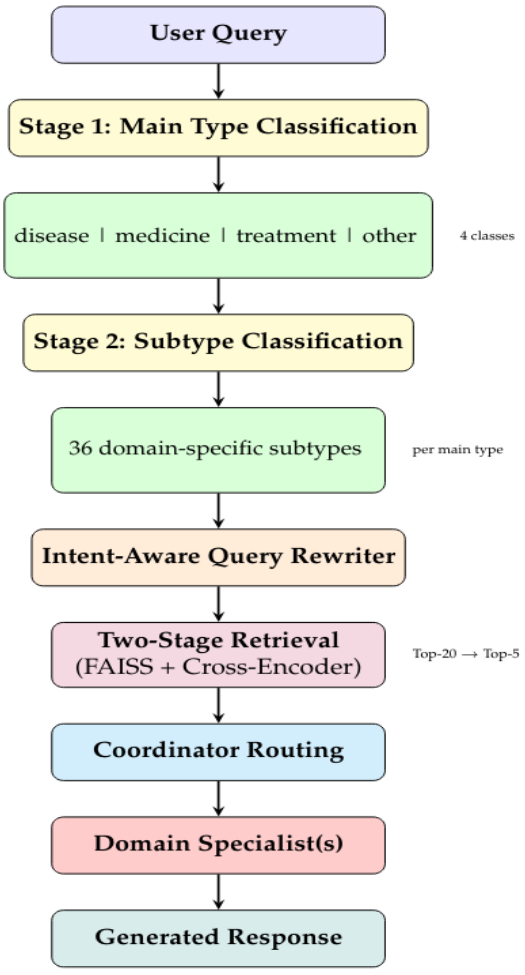

Figure 1. General system processing pipeline. User queries undergo hierarchical intent classification (4 main types, 36 subtypes), intent-aware rewriting, two-stage retrieval, and routing to appropriate domain-specific specialists based on detected intent.

Consider the query from our evaluation dataset: "What are the treatments for Back Pain?"

2.3.1. Stage 1: main type classification

The system evaluates this query against all 5 main intent categories using LLM-based scoring with softmax normalization. The query explicitly asks about treatment options, indicating clear treatment intent. Table 1 shows the actual probability distribution from our evaluation.

Table 1. Probability distribution for back pain treatment

Main Type	Probability	Interpretation
treatment	0.9801	Explicitly asks about treatments
disease	0.0066	Minor disease context
medicine	0.0066	Not medication-specific
other	0.0066	Not a general question
non_medical	0.0001	Clearly medical inquiry

With very high confidence (0.9801) in the treatment category and low entropy, the system classifies this as TREATMENT intent.

2.3.2. Stage 2: subtype classification

Given main_type = "treatment", the system evaluates against 10 treatment-specific subtypes. The query asks about treatment approaches for back pain, focusing on methodologies. Table 2 shows the actual conditional probability distribution.

Table 2. Treatment subtype probability distribution

Subtype	Prob.	Reasoning
method	0.4917	Focuses on treatment approaches
curative_effect	0.2459	Outcomes partially relevant
indication	0.1229	Context of when to treat
others	0.1395	Other treatment aspects

Based on this classification, the coordinator routes the query to the Treatment Specialist with method context. The specialist:

- Applies domain-specific prompting templates optimized for treatment methodology queries
- Uses intent-aware rewriting to reformulate the query, emphasizing treatment approaches and methodologies
- Retrieves relevant documents about back pain management from clinical guidelines through the two-stage retrieval pipeline (FAISS dense retrieval followed by Clinical-Longformer reranking)
- Generates a complete response covering conservative treatments (physical therapy, medications), interventional procedures, and surgical options with appropriate clinical evidence

This example demonstrates how probability

distributions guide routing decisions. High-confidence classifications (low entropy) trigger single-specialist processing with focused prompting, while distributed probabilities (high entropy) trigger multi-specialist consultation for complete coverage.

2.4. Document Retrieval and Two-Stage Ranking

The document retrieval process implements a cascade ranking architecture, balancing efficiency and effectiveness following established IR principles. Our two-stage approach uses the complementary strengths of bi-encoder and cross-encoder architectures.

2.4.1. Stage 1: dense retrieval with FAISS

The first stage uses dense retrieval to efficiently narrow the search space. We employ the S-PubMedBert-MS-MARCO model (pritamdeka/S-PubMedBert-MS-MARCO) (Deka et al., 2022), a sentence-transformer variant of PubMedBERT fine-tuned on the MS-MARCO passage ranking dataset for improved retrieval performance. This domain-specific embedding model, pre-trained on biomedical literature, ensures accurate representation of medical terminology and concepts in the vector space.

The vector store is implemented using Facebook AI Similarity Search (FAISS) (Johnson et al., 2021), which provides efficient similarity search for dense vectors, enabling fast retrieval from our collection of 18 English medical textbooks from MedQA. The retrieval process uses cosine similarity scoring (equation 3):

$$\text{score}(q', d_i) = \frac{E(q') \cdot E(d_i)}{||E(q')|| \cdot ||E(d_i)||} \quad (3)$$

Where $E(\text{q}')$ is the PubMedBERT embedding function. This stage retrieves the top-20 documents by similarity score, balancing recall (capturing relevant documents) with computational efficiency for the subsequent reranking stage.

2.4.1. Stage 2: cross-encoder reranking

The second stage applies a more sophisticated but computationally expensive cross-encoder model to rerank the top-20 candidates. We use Clinical-Longformer (yikuan8/Clinical-Longformer) (Beltagy et al., 2020; Li et al., 2023), which was specifically trained for clinical text understanding with a 4096-token context window.

Unlike the bi-encoder in Stage 1, which encodes query and documents independently, the cross-encoder processes [query, document] pairs jointly (equation 4):

$$\text{score}_{rerank}(q', d_i) = h_\psi([q'; d_i]) \quad (4)$$

where h_ψ is the cross-encoder model with parameters ψ, and [q'; d_i] represents the concatenation of query and document. This joint encoding enables the model to capture fine-grained interactions between query and document text, providing superior relevance estimation compared to independent embeddings.

The reranking stage selects the top-5 documents by reranker score for context provision to specialist components. This cascade design follows the efficiency-effectiveness principle: the bi-encoder efficiently reduces 20,833 documents to 20 candidates (99.9% reduction), while the cross-encoder applies expensive interaction modeling only to this small candidate set.

2.4.3. Stage 3: fallback mechanism

The system maintains robustness through a fallback mechanism. If the rewritten query fails to retrieve relevant documents (zero results from Stage 1), the system automatically falls back to the original raw query. This ensures that even with unusual phrasing or highly specialized terminology that challenges the query rewriter, the system can still provide relevant context rather than failing completely.

2.5. Specialist Framework

The specialist framework is a key innovation in our system that enables the generation of domain-specific responses. The system includes four primary specialist types aligned with our 4-class main intent taxonomy. The Disease Specialist handles disease information, symptoms, causes, and other aspects across 12 subtypes including definition, pathogeny, clinical manifestation, and prevention. The Medicine Specialist manages medication queries about usage, effects, and related topics across 7 subtypes including effect, applicable disease, drug contraindication, usage, and side effects. The Treatment Specialist addresses procedure and intervention information across 10 subtypes including method, cost, effective time, and recovery. The Other Medical Specialist handles miscellaneous medical topics including device usage and multi-question queries.

Each specialist is implemented as a class extending a base specialist interface. They contain domain-specific knowledge and prompt templates, are dynamically loaded based on the identified intent, and maintain a consistent interface while implementing domain-specific logic.

Our implementation supports weighted multi-subtype processing, allowing specialists to generate complete responses that address multiple aspects of a query based on the subtype probability distribution. This ensures that responses cover all relevant aspects proportional to their likelihood of relevance.

2.6. Hierarchical Intent Classification

The hierarchical classification approach offers several advantages. The two-stage softmax approach aligns with the CMID taxonomy structure. Main type classification (4-class) narrows the decision space for subtype identification (36-class), making the classification more accurate and efficient. Full probability distributions are maintained for uncertainty analysis, which helps in determining when to consult multiple specialists or when to request clarification from the user.

The system handles intents at multiple levels. At the cross-domain level, the coordinator integrates responses from different specialists. At the within-domain level, each specialist handles multiple subtypes. Weighted response generation ensures that response detail is proportional to intent probabilities, allowing the system to focus on the most relevant aspects of the query.

2.7. Knowledge Base and Datasets

Our system is specifically designed to answer questions using clinical guidelines from authoritative health organizations. The knowledge base consists of Medical textbooks covering 18 subjects used for USMLE preparation, including anatomy, biochemistry, pathology, pharmacology, physiology, and other core medical disciplines (Jin et al., 2021)

Our taxonomy design was inspired by the structure of the CMID (Chinese Medical Intent Dataset), which includes four main intent classes and 36 subtypes. For our evaluation, we curated a dataset specifically focused on questions where our clinical guideline knowledge base provides relevant context.

We used 476 medical questions from the MedQA benchmark (Jin et al., 2021), which provides expert-reviewed reference answers for general medical information queries. The original MedQA dataset contained 500 questions, from which we removed 24 redundant duplicates to create our evaluation set of 476 unique questions. This dataset provides substantial statistical power for detecting component effects through ablation analysis.

Each question includes expert-annotated reference answers with "must-have" information (critical medical facts that should be included) and "nice-to-have" information (additional helpful context). This annotation structure enables precise assessment of response completeness and accuracy. The dataset spans diverse medical topics including diseases, treatments, procedures, and diagnostic information, providing broad coverage for evaluating medical RAG system performance.

2.9. Evaluation Framework

To assess the impact of different system components, we implemented a comprehensive evaluation framework with seven ablation configurations: (1) full system with all components enabled, (2) no hierarchical intent classification, (3) no query rewriting, (4) no specialists, (5) no reranking, (6) no raw query inclusion, and (7) minimal system with all features disabled.

Each response was evaluated using four independent metrics on a 1-5 Likert scale: Relevance assessed how well retrieved context matched the query; Completeness measured coverage of reference answer key points and must-have information; Faithfulness evaluated groundedness in retrieved context (avoiding hallucinations); and Correctness assessed factual accuracy compared to reference answers. The overall score was calculated as the average of these four metrics.

We employed GPT-4o mini (OpenAI, 2024) as an LLM judge with independent metric evaluation: each metric was assessed through a separate API call with metric-specific prompting. This approach minimizes inter-metric bias that can occur when evaluating multiple aspects simultaneously. The total evaluation comprised 476 queries × 7 configurations × 4 metrics = 13,328 independent LLM judge calls.

Statistical validation was performed using paired t-tests comparing each ablation configuration against the full system, with effect sizes calculated using Cohen's d. Statistical significance was assessed at $P<0.05$, with Bonferroni correction applied for multiple comparisons. This rigorous statistical approach ensures that reported component effects are reliable and reproducible.

3. Results

We conducted systematic ablation studies across seven configurations to quantify the contribution of each RAG component in our medical question answering system. The results reveal that reranking and specialist components are statistically essential ($P<0.001$), while hierarchical intent classification unexpectedly degrades performance.

3.1. Overall Performance Across Configurations

Our ablation study on 476 medical questions from MedQA benchmarks revealed significant performance variations across system configurations. The full system achieved an overall score of 3.64/5.0. Removing reranking resulted in the largest performance drop to 3.40/5.0 (-6.6%, $P<0.001$, Cohen's $d = -0.44$), followed by specialist removal at 3.46/5.0 (-4.9%, $P<0.001$, $d = -0.29$). Surprisingly, removing hierarchical intent classification improved performance to 3.67/5.0 (+0.8%, $p = 0.010$ for completeness), the best performing configuration. Query rewriting showed minimal impact at 3.60/5.0 (-1.1%), while removing raw query inclusion scored 3.61/5.0 (-0.8%). The minimal system with all features disabled achieved 3.46/5.0 (-4.9%), matching the no_specialists configuration.

3.2. Detailed Metric Analysis

Table 3 presents detailed scores for relevance, completeness, faithfulness, and correctness across all system configurations. Reranking showed the strongest impact on relevance (relevance drops by 0.47 when removed, $P<0.001$) and faithfulness (drops by 0.54, $P<0.001$). Specialists primarily affected relevance (-0.45) and faithfulness (-0.43). The counterintuitive finding that removing hierarchical intent improved completeness (+0.09, $p = 0.010$) suggests simpler query processing may be preferable for this task.

Table 4 presents statistical significance tests comparing each ablation configuration against the full system baseline through paired t-tests. Reranking and specialists show highly significant impacts ($P<0.001$), confirming their critical role. The hierarchical intent finding is counterintuitive but statistically significant ($p = 0.010$ for completeness improvement).

Table 3. Performance metrics across seven system configurations on 476 medical questions. Lower relevance scores across all modes highlight the persistent challenge of medical retrieval, while consistently high correctness suggests robust answer generation once context is successfully retrieved*

Configuration	Relevance	Completeness	Faithfulness	Correctness	Overall
no_hierarchical_intent	2.32	4.68	2.81	4.86	3.67
full_system	2.36	4.58	2.83	4.77	3.64
no_raw_query	2.37	4.44	2.93	4.71	3.61
no_query_rewriting	2.21	4.64	2.75	4.79	3.6
no_specialists	1.91	4.72	2.39	4.84	3.46
minimal_system	1.90	4.71	2.4	4.84	3.46
no_reranking	1.88	4.60	2.29	4.80	3.40

*Each metric evaluated on 1-5 Likert scale through independent LLM judge calls.

Table 4. Statistical significance tests (paired t-tests) comparing ablation configurations vs full system on 476 queries

Configuration	Overall	P-value	Cohen's d	Effect	Sig.
No Reranking	-0.239	<0.001	-0.442	small	***
No Specialists	-0.173	<0.001	-0.287	small	***
Minimal System	-0.173	<0.001	-0.287	small	***
No-Query Rewriting	-0.038	0.036	-0.132	negligible	*
No Raw Query	-0.022	0.575	-0.035	negligible	ns
No-Hierarchical Intent	0.025	0.149	0.090	negligible	ns

Effect sizes calculated using Cohen's d. *** P<0.001, ** P<0.01, * P<0.05, ns = not significant.

Figure 2 presents a visual comparison of overall scores across all seven system configurations. Removing hierarchical intent classification yields the best performance (3.67), while removing reranking yields the worst (3.40), demonstrating reranking's critical role in RAG systems. Figure 3 summarizes the relative impact of each component through ablation analysis, measured as percentage change from the full system baseline. Figure 4 provides a detailed breakdown of how the two most critical components (reranking and specialists) impact individual evaluation metrics compared to the full system.

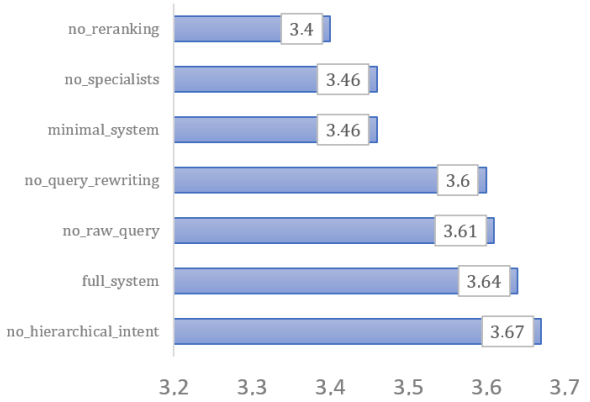

Figure 2. Overall scores across seven system configurations on 476 medical questions.

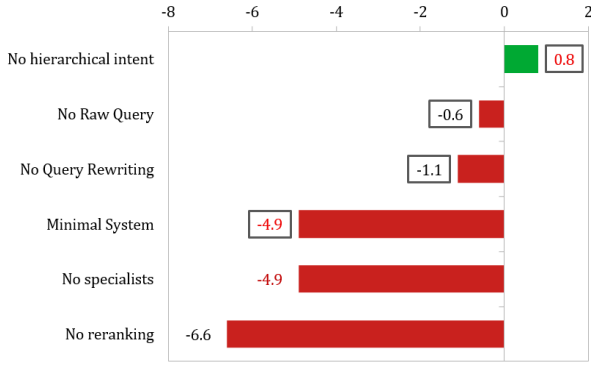

Figure 3. Component ablation impact measured as percentage change from full system baseline (N=476). Reranking removal causes the largest drop (-6.6%, P<0.001), followed by specialists (-4.9%, P<0.001). Hierarchical intent removal unexpectedly improves performance (+0.8%, P=0.010 for completeness). Negative values indicate a performance drop upon component removal, emphasizing the critical necessity of reranking and specialist modules.

This breakdown illustrates that reranking and specialist components primarily safeguard the relevance and faithfulness of the generated medical responses.

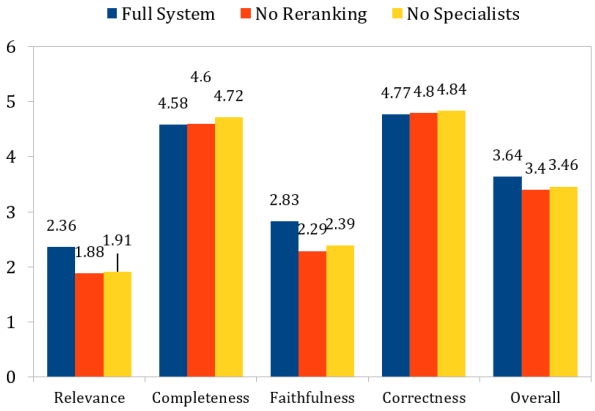

Figure 4. Metric-level breakdown for the two most critical components (N=476). Reranking has the strongest impact on relevance (-0.47***) and faithfulness (-0.54***), while specialists also degrade relevance (-0.45***) and faithfulness (-0.43***). Both show statistically significant effects (P<0.001) with small effect sizes.

3.4. Patterns Across Metrics

Analysis of individual metrics reveals distinct patterns for different RAG components. Relevance and faithfulness show the strongest sensitivity to component removal, with reranking and specialists both critical for maintaining these metrics. Completeness scores remain relatively stable across configurations (4.47-4.74), with hierarchical intent removal showing slight improvement (+0.09, P=0.010). Correctness scores are consistently high (4.67-4.82) across all configurations, suggesting that when the system provides an answer, it tends to be factually accurate regardless of which components are enabled.

The moderate overall scores (3.40-3.67 out of 5.0) and particularly low relevance scores (1.88-2.37) indicate substantial room for improvement in retrieval quality, highlighting the importance of reranking mechanisms for filtering and prioritizing retrieved context. The low relevance scores (1.88-2.37) suggest that the dense retrieval component often introduces extraneous information alongside pertinent facts. Interestingly, correctness scores remain high (4.67-4.82) because the LLM judge evaluates the final answer's factual integrity rather than the purity of the source documents. This indicates that the system's generation layer effectively filters out the 'noise' from the retrieval layer to produce accurate answers, though the retrieval precision itself remains a target for future hybrid-search improvements.

3.5. Summary of Findings

Our systematic ablation study on 476 medical questions reveals that reranking and specialist components are both essential for medical RAG systems, with statistically significant impacts confirmed through paired t-tests. Reranking removal causes the largest performance drop (-0.24 overall, P<0.001, Cohen's d=-0.44), primarily affecting relevance (-0.47) and faithfulness (-0.54). Specialist removal shows the second largest impact (-0.17 overall, P<0.001, d=-0.29), also degrading relevance (-0.45) and faithfulness (-0.43). Both effects show small but statistically significant effect sizes.

Counterintuitively, hierarchical intent classification degrades performance when included, with removal showing completeness improvement (+0.09, P=0.010). This finding suggests that simpler query processing may be preferable to complex intent classification mechanisms. Query rewriting shows minimal impact (-0.04 overall), while raw query inclusion primarily affects completeness (-0.15, P<0.001). These results provide evidence-based guidance for medical RAG design: prioritize reranking infrastructure and domain-specific specialists over sophisticated query understanding mechanisms.

4. Discussion

Our systematic ablation study on 476 medical questions provides quantitative evidence that reranking and specialist components are both essential for medical RAG systems, while revealing a counterintuitive finding about hierarchical intent classification. The reranking component shows the strongest impact (-0.24 overall, P<0.001, Cohen's d=-0.44), followed by specialists (-0.17 overall, P<0.001, d=-0.29). Both effects are statistically significant with small effect sizes, confirming their importance beyond measurement noise.

These findings have important implications for RAG system design priorities. The critical role of reranking aligns with cascade ranking principles from information retrieval literature, where cross-encoder reranking provides precision gains over dense retrieval alone. Our cross-encoder reranker (Clinical-Longformer) particularly improves relevance (-0.47 when removed) and faithfulness (-0.54 when removed), demonstrating its value in filtering and prioritizing retrieved medical context. This result emphasizes that investment in reranking infrastructure should be a primary design consideration for medical RAG systems.

The substantial specialist impact confirms that domain-specific expertise provides benefits beyond general-purpose LLMs, even with retrieval augmentation. Specialists show strong effects on both relevance (-0.45 when removed) and faithfulness (-0.43 when removed), suggesting they contribute to better context utilization and grounded response generation. This aligns with medical knowledge being inherently specialized and compartmentalized across different clinical domains.

4.1. Counterintuitive Finding: Hierarchical Intent Classification Degrades Performance

A surprising result is that hierarchical intent classification degrades performance when included, with removal showing completeness improvement (+0.09, P=0.010). This finding challenges the assumption that sophisticated query understanding mechanisms necessarily improve RAG systems. Several possible explanations exist: (1) the two-level hierarchy may

introduce unnecessary complexity for medical queries that are often straightforward information requests; (2) intent misclassification errors may propagate downstream and significantly degrade retrieval specifically, an incorrect main-type assignment prevents the system from ever reaching the correct domain specialist, creating a terminal bottleneck in the pipeline; or (3) the additional processing step may discard useful query information by forcing diverse natural language into rigid taxonomic categories, thereby limiting the LLM's inherent capacity for flexible reasoning.

A qualitative error analysis reveals that hierarchical misclassification often stems from overlapping domain keywords. For instance, a query about "treatment-resistant hypertension symptoms" might be classified as 'treatment' at Stage 1 but 'disease' at Stage 2. Such errors cause the coordinator to route queries to specialists with mismatched prompting templates, leading to responses that prioritize pathological definitions over clinical intervention steps, thereby reducing overall system completeness.

This result suggests that simpler query processing pipelines may be preferable to complex intent classification mechanisms, particularly in specialized domains like medical QA where query types are relatively uniform. While hierarchical intent classification is conceptually appealing, our empirical results demonstrate that architectural sophistication does not always translate to performance gains. Future work should investigate whether flatter intent classification or no explicit intent classification provides better results.

To illustrate this failure mode, consider a query regarding "long-term side effects of Ibuprofen." While clearly a 'medicine' intent, the hierarchical system may misclassify it at Stage 1 as 'disease' due to the mention of inflammatory conditions in the query context. This error propagates, causing the system to route the query to a Disease Specialist instead of a Medicine Specialist. Such misrouting results in a response focused on pathology rather than pharmacology, significantly degrading completeness despite the system's architectural complexity.

4.2. Query Processing Components' Impact

Query rewriting shows minimal overall impact (-0.04, P=0.036), with small effects on relevance (-0.15). This suggests that for medical questions, which are often clearly stated, query reformulation provides limited value compared to using the original query. Raw query inclusion primarily affects completeness (-0.15, P<0.001), indicating that preserving the user's original phrasing alongside processed queries helps ensure comprehensive responses.

4.3. Scope and Limitations

4.3.1. Dataset and evaluation scope

Our evaluation used 476 medical questions from the MedQA benchmark, focusing on general medical information queries. While this provides substantial statistical power for detecting component effects, the dataset characteristics should be acknowledged.

Questions that fall outside typical clinical guideline scope include:

- Rare genetic diseases not covered by clinical practice guidelines
- Specific drug-drug interactions requiring detailed pharmacokinetic databases
- Research-level biomedical mechanisms and molecular pathways
- Highly specialized procedures or conditions not addressed in general clinical guidelines

The moderate overall performance scores (3.40-3.67 out of 5.0) and particularly low relevance scores (1.88-2.37) suggest that retrieval quality remains a significant challenge. This indicates opportunities for improvement in both knowledge base coverage and retrieval mechanisms. Future work should expand to broader medical question types and knowledge sources.

A notable limitation is the alignment between the CMID taxonomy, originally designed for Chinese medical queries, and the English MedQA dataset used for evaluation. While core medical intents (e.g., disease, treatment) are largely universal, the semantic mapping of 36 specific subtypes may be influenced by cultural or linguistic nuances inherent in the original Chinese dataset. This cross-lingual transfer of an intent framework may introduce subtle categorization noise, potentially contributing to the performance degradation observed when using hierarchical classification in an English-speaking clinical context.

4.3.2. Evaluation methodology

While our evaluation used GPT-4o mini for consistent assessment across configurations (OpenAI, 2024), with independent metric evaluation to minimize bias, this automated approach has inherent limitations. As a smaller model, GPT-4o mini may exhibit systematic rating biases or lack the deep clinical nuance required to distinguish between highly similar medical edge cases compared to human experts or significantly larger frontier models. Consequently, this automated approach may not capture all aspects of clinical utility that human experts would consider.

While human expert evaluation remains the gold standard for clinical validation, this study prioritizes a high-volume, systematic architectural benchmark. The total evaluation process involved 13,328 independent LLM judge calls, a scale that ensures statistical significance across seven system configurations but makes manual human review of the entire set logistically prohibitive. Following established practices in medical QA research, we relied on GPT-4o mini as an objective proxy for context groundedness and factual correctness, which has shown consistency with human patterns in identifying technical hallucinations. However, the absence of clinician-led validation is a formal limitation, and our findings should be interpreted as a structural performance baseline rather than a direct measure of clinical safety.

5. Conclusion

This paper presents a systematic component-level analysis of medical RAG systems through comprehensive ablation studies on 476 questions from the MedQA benchmark. Our research was motivated by the need to quantify the relative importance of different RAG components, particularly reranking, domain specialists, and query processing mechanisms, to provide evidence-based guidance for medical question-answering system design. Through rigorous statistical testing with paired t-tests and effect size calculations, we identified which components are essential and which add unnecessary complexity. Our experiments yielded several critical findings. Most significantly, reranking emerged as the most important component, with its removal causing the largest performance drop (-0.24 overall, $P<0.001$, Cohen's $d=-0.44$), primarily affecting relevance and faithfulness. Specialist components showed the second strongest impact (-0.17 overall, $P<0.001$, $d=-0.29$), confirming that domain-specific expertise remains valuable even with retrieval augmentation. Counterintuitively, hierarchical intent classification was found to degrade performance, with its removal showing an improvement in completeness (+0.09, $P=0.010$). These results demonstrate that architectural sophistication, especially in query processing, does not always translate to performance gains in the medical domain. These findings provide actionable guidance for practitioners: prioritize precision-focused reranking infrastructure and domain-specific specialist routing, while favoring simpler query-processing pipelines to avoid unnecessary noise. The statistical rigor of this evaluation ensures that these recommendations are grounded in reliable evidence, providing a trustworthy framework for building medical information systems where reliability is paramount.

Moving forward, several promising directions emerge from this work. First, investigating why hierarchical intent classification degrades performance could inform more robust query-processing approaches. Second, exploring the efficiency-quality tradeoffs for reranking mechanisms would address practical deployment constraints in clinical settings. Finally, expanding the evaluation to include human expert assessment and wider medical domains would validate whether these automated metrics align with clinical utility judgments and enhance the system's broader applicability.

Author Contributions

The percentages of the authors' contributions are presented below. All authors reviewed and approved the final version of the manuscript.

	H.E.	D.Q.R.
C	30	70
D	70	30
S	70	30
DCP	30	70
DAI	30	70
L	30	70
W	50	50
CR	70	30
SR	70	30
PM	70	30
FA	50	50

C= concept, D= design, S= supervision, DCP= data collection and/or processing, DAI= data analysis and/or interpretation, L= literature search, W= writing, CR= critical review, SR= submission and revision, PM= project management, FA= funding acquisition.

Conflict of Interest

The authors declared that there is no conflict of interest.

Ethical Consideration

Ethics committee approval was not required for this study because of there was no study on animals or humans.

References

Belkin, N. J., Oddy, R. N., & Brooks, H. M. (1982). ASK for information retrieval: Part I. Background and theory. *Journal of Documentation*, *38*(2), 61–71. https://doi.org/10.1108/eb026722

Beltagy, I., Peters, M. E., & Cohan, A. (2020). *Longformer: The long-document transformer*. arXiv. https://arxiv.org/abs/2004.05150

Ben Abacha, A., & Demner-Fushman, D. (2019). On the summarization of consumer health questions. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* içinde (pp. 2228–2234). https://doi.org/10.18653/v1/P19-1215

Ben Abacha, A., Yim, W., Michalopoulos, G., & Lin, T. (2023). An investigation of evaluation methods in automatic medical note generation. *Findings of the Association for Computational Linguistics: ACL 2023* içinde (pp. 2575–2588). https://doi.org/10.18653/v1/2023.findings-acl.161

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* içinde (pp. 335–336). https://doi.org/10.1145/290941.291025

Casanueva, I., Temčinas, T., Gerz, D., Henderson, M., & Vulić, I. (2020). Efficient intent detection with dual sentence encoders. *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI* içinde (pp. 38–45).

https://doi.org/10.18653/v1/2020.nlp4convai-1.5

Chen, N., Su, X., Liu, T., Hao, Q., & Wei, M. (2020). A benchmark dataset and case study for Chinese medical question intent classification. *BMC Medical Informatics and Decision Making*, *20*, 125. https://doi.org/10.1186/s12911-020-1122-3

Chu, Y. W., Zhang, K., Malon, C., & Min, M. R. (2025). *Reducing hallucinations of medical multimodal large language models with visual retrieval-augmented generation*. arXiv. https://arxiv.org/abs/2502.15040

Deka, P., Jurek-Loughrey, A., & Padmanabhan, D. (2022). Improved methods to aid unsupervised evidence-based fact checking for online health news. *Journal of Data Intelligence*, *3*(4), 474–505. https://doi.org/10.26421/JDI3.4-5

Dorfner, F. J., Dada, A., Busch, F., Makowski, M. R., Han, T., Truhn, D., Kleesiek, J., Sushil, M., Adams, L. C., & Bressem, K. K. (2025). Evaluating the effectiveness of biomedical fine-tuning for large language models on clinical tasks. *Journal of the American Medical Informatics Association*, *32*(6), 1015–1024. https://doi.org/10.1093/jamia/ocaf045

Fu, T., Huang, K., Xiao, C., Glass, L. M., & Sun, J. (2022). HINT: Hierarchical interaction network for clinical-trial-outcome predictions. *Patterns*, *3*(4), 100445. https://doi.org/10.1016/j.patter.2022.100445

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, *3*(1), 1–23. https://doi.org/10.1145/3458754

Jeong, M., Sohn, J., Sung, M., & Kang, J. (2024). Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, *40*(Ek 1), i119–i129. https://doi.org/10.1093/bioinformatics/btae238

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., & Szolovits, P. (2021). What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, *11*(14), 6421. https://doi.org/10.3390/app11146421

Jin, Q., Yuan, Z., Xiong, G., Yu, Q., Ying, H., Tan, C., Chen, M., Huang, S., Liu, X., & Yu, S. (2022). Biomedical question answering: A survey of approaches and challenges. ACM Computing Surveys, 55(2), 1–36. https://doi.org/10.1145/3490238

Johnson, J., Douze, M., & Jégou, H. (2021). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, *7*(3), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572

Kim, J., Podlasek, A., Shidara, K., Liu, F., Alaa, A., & Bernardo, D. (2025). Limitations of large language models in clinical problem-solving arising from inflexible reasoning. *Scientific Reports*, *15*(1), 39426. https://doi.org/10.1038/s41598-025-22940-0

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems* (9459–9474). Curran Associates, Inc.

Li, Y., Wehbe, R. M., Ahmad, F. S., Wang, H., & Luo, Y. (2023). A comparative study of pretrained language models for long clinical text. *Journal of the American Medical Informatics Association*, *30*(2), 340–347. https://doi.org/10.1093/jamia/ocac225

Lu, W., Jiang, J., Shi, Y., Zhong, X., Gu, J., Huangfu, L., & Gong, M. (2023). Application of Entity-BERT model based on neuroscience and brain-like cognition in electronic medical record entity recognition. *Frontiers in Neuroscience*, *17*, 1259652. https://doi.org/10.3389/fnins.2023.1259652

Maharjan, J., Garikipati, A., Singh, N. P., Cyrus, L., Sharma, M., Ciobanu, M., Barnes, G., Thapa, R., Mao, Q., & Das, R. (2024). OpenMedLM: Prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. *Scientific Reports*, *14*, 14156. https://doi.org/10.1038/s41598-024-64827-6

Manas, G., Aribandi, V., Kursuncu, U., Alambo, A., Shalin, V. L., Thirunarayan, K., Beich, J., Narasimhan, M., & Sheth, A. (2021). Knowledge-infused abstractive summarization of clinical diagnostic interviews: Framework development study. *JMIR Mental Health*, *8*(5), e20865. https://doi.org/10.2196/20865

Mishra, R., Bian, J., Fiszman, M., Weir, C. R., Jonnalagadda, S., Mostafa, J., & Del Fiol, G. (2014). Text summarization in the biomedical domain: A systematic review of recent research. *Journal of Biomedical Informatics*, *52*, 457–467. https://doi.org/10.1016/j.jbi.2014.06.009

OpenAI. (2024). GPT-4o mini: Advancing cost-efficient intelligence. OpenAI. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/ (accessed on 29 January 2026).

Robertson, S. E. (1997). The probability ranking principle in IR. K. Sparck Jones & P. Willett (Editörler), *Readings in information retrieval* içinde (pp. 281–286). Morgan Kaufmann.

Selmi, W., Kammoun, H., & Amous, I. (2022). Semantic-based hybrid query reformulation for biomedical information retrieval. *The Computer Journal*, *66*(9), 2296–2316. https://doi.org/10.1093/comjnl/bxac078

Simonds, T., Kurniawan, K., & Lau, J. H. (2024). MoDEM: Mixture of domain expert models. *Proceedings of the 22nd Annual Workshop of the Australasian Language Technology Association* içinde (pp. 75–88). Association for Computational Linguistics. https://aclanthology.org/2024.alta-1.6/

Sun, K., Yu, D., Chen, J., Yu, D., Choi, Y., & Cardie, C. (2019). DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, *7*, 217–231. https://doi.org/10.1162/tacl_a_00264

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., Ber, D. S. W., Lim, J. Y., Eckhoff, H. B., Lim, G. S. W., Tso, C. F., Wong, D. S. L., Li, S., Xu, L., Hussain, R. Z., Xiang, Y., Lu, Y., Liu, N., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, *29*, 1930–1940. https://doi.org/10.1038/s41591-023-02448-8

Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M. R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., Almirantis, Y., Pavlopoulos, J., Baskiotis, N., Gallinari, P., Artières, T., Ngomo, A. C., Heino, N., Gaussier, E., Barrio-Alvers, L., Schroeder, M., Androutsopoulos, I., & Paliouras, G. (2015). An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, *16*, 138. https://doi.org/10.1186/s12859-015-0564-6

Wei, Z., Guo, D., Huang, D., Zhang, Q., Zhang, S., Jiang, K., & Li, R. (2024). Detecting and mitigating the ungrounded hallucinations in text generation by LLMs. *Proceedings of the 2023 International Conference on Artificial Intelligence, Systems and Network Security* içinde (pp. 1–6). https://doi.org/10.1145/3661638.3661653

Yang, D., Wei, J., Li, M., Liu, J., Liu, L., Hu, M., He, J., Ju, Y., Zhou, W., Liu, Y., & Zhang, L. (2025). MedAide: Information fusion and anatomy of medical intents via LLM-based agent

collaboration. *Information Fusion*, *127*, 103743. https://doi.org/10.1016/j.inffus.2025.103743

Zhang, Y., Yang, R., Xu, X., Li, R., Xiao, J., Shen, J., & Han, J. (2025). TELEClass: Taxonomy enrichment and LLM-enhanced hierarchical text classification with minimal supervision. *WWW '25: Proceedings of the ACM on Web Conference 2025* içinde (pp. 2032–2042). https://doi.org/10.1145/3696410.3714940

Zhao, W., Deng, Z., Yadav, S., & Yu, P. S. (2024). Heterogeneous knowledge grounding for medical question answering with retrieval augmented large language model. *Companion Proceedings of the ACM Web Conference 2024* içinde (pp. 1535–1538). https://doi.org/10.1145/3589335.3651941