Does Domain-Specific Retrieval Augmented Generation Help LLMs Answer Consumer Health Questions?

Chase M Fensore[1] chase.fensore@emory.edu
Rodrigo M Carrillo-Larco[2] rodrigo.martin.carrillo.larco@emory.edu
Megha K Shah[3] megha.shah@emory.edu
Joyce C Ho[1] joyce.c.ho@emory.edu
[1]*Department of Computer Science, Emory University*
[2]*Rollins School of Public Health, Emory University*
[3]*Department of Family and Preventive Medicine, Emory School of Medicine*

Abstract

While large language models (LLMs) have shown impressive performance on medical benchmarks, there remains uncertainty about whether retrieval-augmented generation (RAG) meaningfully improves their ability to answer consumer health questions. In this study, we systematically evaluate vanilla LLMs against RAG-enhanced approaches using the NIDDK portion of the MedQuAD dataset. We compare four open-source LLMs in both vanilla and RAG configurations, assessing performance through automated metrics, LLM-based evaluation, and clinical validation. Surprisingly, we find that vanilla LLM approaches consistently outperform RAG variants across both quantitative metrics (BLEU, ROUGE, BERTScore) and qualitative assessments. The relatively low retrieval performance (Precision@5 = 0.15) highlights fundamental challenges in implementing effective RAG systems for medical question-answering, even with carefully curated questions. While RAG showed competitive performance in specific areas like scientific consensus and harm reduction, our findings suggest that successful implementation of RAG for consumer health question-answering requires more sophisticated approaches than simple retrieval and prompt engineering. These results contribute to the ongoing discussion about the role of retrieval augmentation in medical AI systems and highlight the need for medical-specific RAG infrastructure to enhance medical question-answering systems.[1]

1. Introduction

The increasing adoption of large language models (LLMs) for medical question-answering (QA) has created both opportunities and challenges in providing accurate, verifiable health information to consumers. Recent works (Abrar et al., 2024; Yan et al., 2024) demonstrate the performance of existing LLMs across various patient-facing QA datasets including MedRedQA (Nguyen et al., 2023), iCliniq (Regin, 2017), TREC LiveQA 2017 (Ben Abacha et al., 2017), and MedQuAD (Ben Abacha and Demner-Fushman, 2019a) datasets. Moreover, new medical QA datasets continue to emerge such as the "JAMA Clinical Challenge and Medbullets" dataset (Chen et al., 2024). The current findings suggest that LLMs achieve remarkable performance on medical knowledge benchmarks, even passing the medical board exam (Abbas et al., 2024). Unfortunately, many of these datasets still focus on

1. Code can be found at: https://github.com/fensorechase/rag-patient-metabolic-qa.

multiple-choice formats that may not reflect the complexity and variability of real patient questions. As Raji et al. (2025) argue, there is a growing need to "move beyond medical exams and adopt more grounded, task-specific approaches for evaluation." Thus, a crucial question is how to bridge the gap between high scores on standardized medical exams and providing reliable, grounded answers to real-world patient questions.

The landscape of medical QA systems has evolved through several paradigms. Early approaches relied heavily on Information Retrieval-Based Question Answering (IRQA) and Knowledge Base Question Answering (KBQA) (Sukhwal et al., 2024), often leveraging medical knowledge graphs or search engine results from authoritative sources. Prior to LLMs, retrieval-based QA systems often combined (i) contextual text embeddings to identify relevant documents, (ii) summaries of either the question or the retrieved text, and (iii) semantic similarity for the matching process to ensure a knowledge-grounded system (Mrini et al., 2022). More recently, the field has seen a shift toward generative approaches, including transfer learning and specialized fine-tuning of LLMs for medical domains (Lehman et al., 2023; Yagnik et al., 2024). However, as Liu et al. (2023) demonstrated, even leading generative search engines frequently fail to provide citations or ground their responses in authoritative medical sources.

Hybrid approaches, particularly retrieval-augmented generation (RAG), offer a promising direction for addressing the knowledge-grounded limitations of LLMs. Recent work has explored various RAG architectures, from naive implementations using basic retrieval models like BM25 to more sophisticated modular approaches to include iterative and adaptive retrieval techniques (Bora and Cuayáhuitl, 2024). As a result, disease-specific chatbots with RAG have been developed. One such example is LiVersa which uses the American Association for the Study of Liver Diseases guidance documents to reduce hallucinations and provide reliable, grounded answers (Ge et al., 2024). Other QA systems have explored incorporating diverse medical resources including textbooks, journals, and clinical guidelines (Bora and Cuayáhuitl, 2024). **Yet despite the widespread adoption of RAG towards providing domain-specific knowledge (Zhang et al., 2024; Soudani et al., 2024), it remains unclear whether RAG truly aids the process of *medical QA* without extensive setup and configuration** (Liu et al., 2024; Agrawal et al., 2024; Simon et al., 2024). Xiong et al. (2024) created MIRAGE, the first RAG evaluation benchmark for medical QA. However, the benchmark only includes multi-choice questions and can only assess retrieval capability on less than half of the questions.

In this work, we investigate whether RAG can enhance the reliability and verifiability of LLM-generated answers for real-world consumer health questions without requiring extensive domain-specific fine-tuning. Using the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) portion of the MedQuAD dataset, which contains authentic patient questions with clinically validated answers generated solely from content on NIH web pages, we systematically compare vanilla LLM approaches against RAG-enhanced LLM systems. Our evaluation framework combines automated metrics with LLM-judging and expert clinical validation, addressing the need for rigorous, task-specific assessment of open (i.e., non multiple-choice answers) medical QA systems. We find that vanilla LLM-generated responses consistently outperform a hybrid RAG system with respect to both quantitative and qualitative evaluations. This underscores the complexity of implementing domain-specific systems which effectively leverage RAG for medical QA.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our study provides important insights about ML applications in healthcare contexts:

- Our study represents the first systematic evaluation of retrieval-augmented generation (RAG) for open-ended consumer health QA using the NIDDK MedQuAD dataset, suggesting that vanilla LLM approaches often outperform general-purpose RAG approaches for diabetes and digestive health questions, contrary to common assumptions about the benefits of retrieval augmentation.
- The relatively low retrieval performance (Precision@5 = 0.15) highlights a fundamental challenge in bridging the semantic gap between how patients phrase health questions and how medical information is presented in authoritative sources – a critical consideration for patient-facing medical QA systems in domains like diabetes and metabolic health.
- While RAG showed competitive performance in specific areas like scientific consensus and harm reduction, our findings demonstrate that successful implementation of RAG for consumer health QA requires more sophisticated approaches than simple retrieval and prompt engineering, particularly for specialized medical domains like those covered in the NIDDK dataset.
- Our systematic study combining multiple automated metrics, LLM-based evaluation, and clinical validation provides a robust methodology for evaluating AI-driven medical QA systems, addressing the need for rigorous, task-specific assessment particularly important for ML systems deployed in high-stakes patient-facing healthcare settings.

2. Methods

2.1. Problem Statement

Our approach aims to evaluate whether RAG can improve the ability of off-the-shelf LLMs to answer consumer health questions. Unlike previous approaches that require extensive fine-tuning on medical data (Yagnik et al., 2024), specialized architectures (Cho and Lee, 2025), or focus on multi-choice medical QA (Xiong et al., 2024), we investigate whether combining general-purpose LLMs with targeted retrieval can provide accurate and reliable answers while maintaining flexibility to incorporate updated medical knowledge. Specifically, given a consumer health question q, we compare two approaches:

1. Direct LLM generation: The model generates an answer a using only its pretrained knowledge.[2]
2. RAG-enhanced generation: The model generates an answer a' incorporating retrieved relevant medical documents $D = \{d_1, ..., d_k\}$ from authoritative sources. We assume a closed domain setting where D is predefined and the ground truth answer is present within D.

This framework allows us to systematically evaluate whether RAG provides meaningful improvements over vanilla LLM approaches for open-ended consumer health QA.

2. We refer to approach this interchangeably as LLM-only, or vanilla LLM.

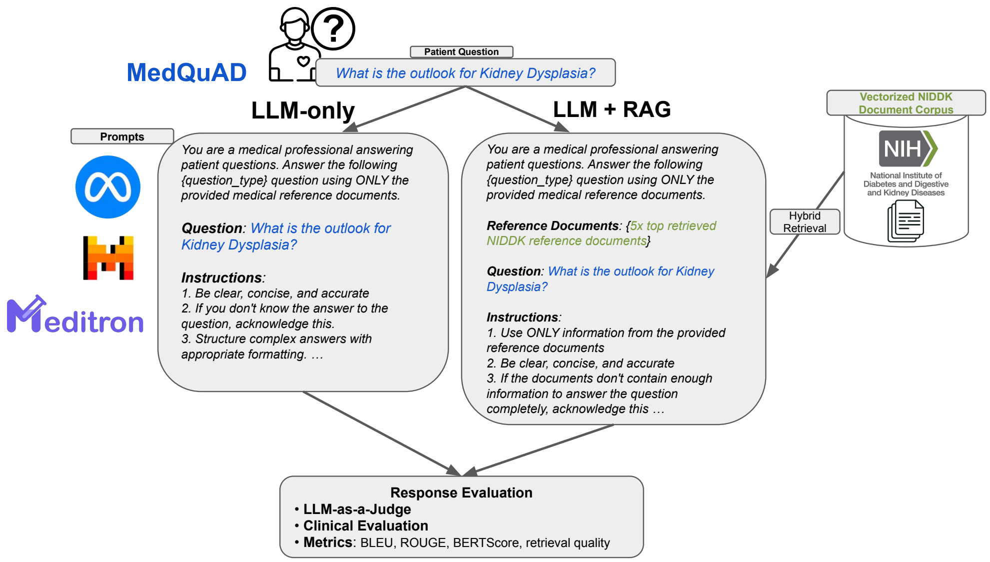

Figure 1: Overview of experimental setup comparing vanilla LLM vs. RAG approaches for patient-facing medical QA. Given a patient question (e.g., "What are the symptoms of rheumatoid arthritis?"), we compare two approaches: (A) Direct LLM generation using only the model's pretrained knowledge, and (B) RAG-enhanced generation incorporating relevant medical documents retrieved from NIDDK sources. We use the same underlying LLMs to compare answer quality.

2.2. Dataset

We evaluated our approach using the NIDDK portion of the MedQuAD dataset (Ben Abacha and Demner-Fushman, 2019b), which consists of 1,192 question-answer pairs (828 unique questions) focused on diseases such as diabetes and digestive disorders. The NIDDK subset was chosen for two reasons: (i) it represents consumer health questions with verified answers from attributable clinically-validated webpages which allows for straightforward assessment of information retrieval quality, and (ii) it covers a diverse set of medical topics in metabolic health with varying levels of specificity, allowing for a more comprehensive evaluation of the model's ability to handle both general and detailed consumer health inquiries.

The NIDDK questions span 12 categories: information, susceptibility, causes, symptoms, exams/tests, treatment, prevention, considerations, complications, frequency, research, and outlook. Questions average 9.08 words (σ = 3.39, range: 4-27), while reference answers average 268.59 words (σ = 289.64, range: 6-2024). Each question is paired with its source URL from the NIDDK website, providing ground truth for evaluating retrieval quality. The MedQuAD dataset was originally constructed in 2019 for the task of recognizing question entailment (RQE), where the objective is to find answers to a given question (premise question) by identifying and retrieving semantically similar or entailed questions (hypothesis questions) that already have existing answers. RQE has been heavily studied in IR and search domains (Ben Abacha and Demner-Fushman, 2019b), however a RQE system used in isolation faces inherent limitations including (1) dependency on pre-existing answer pairs, (2) inability to handle novel questions without semantically equivalent counterparts in its knowledge base, and (3) lack of end-to-end answer generation capabilities that adapt to nuanced consumer needs (Mutabazi et al., 2021; Nguyen et al., 2023). In this work, we instead use MedQuAD to study the task of end-to-end medical QA.

2.3. Model Configuration

We evaluated four open-source LLMs, including one fine-tuned for medical purpose, representing both small and larger models: `Llama 3.1 70B-Instruct`, `Llama 3.1 8B-Instruct`, `Mistral 8x7B-Instruct`, and `Meditron-70B`. Models in the Llama 3.1 family adopt a decoder-only architecture, Mistral 8x7B-Instruct uses a mixture-of-experts structure, and Meditron-70B is a variant of LLaMA-2 fine-tuned on medical text. Additional details on model configurations are provided in Appendix A.

2.3.1. Implementation Details

All models were implemented using the Hugging Face Transformers library and were run on a single NVIDIA H100 GPU. For answer generation, all models were set to a temperature of 0.1, and were allowed to generate a maximum of 512 new tokens. We applied 4-bit quantization for all LLMs. Model inputs were tokenized with padding and truncation. For the RAG setup, we ensured the retrieved context plus question fits within the model's token limit, trimming retrieved documents if necessary while preserving key information. Additional details on document trimming is provided in Appendix A.

2.4. Document Processing Pipeline

We implemented a standard three-stage pipeline for processing medical documents for RAG: document preprocessing, chunking, and embedding generation.

- **Document Collection and Preprocessing**: Given the provided URLs in the MedQuAD-NIDDK dataset, we collected 151 unique medical documents from the NIDDK website. For URLs that were no longer active, we retrieved the most recent available version from the Internet Archive's Wayback Machine. Each document underwent preprocessing to remove HTML markup while preserving the semantic structure of medical content, including section headers and hierarchical relationships important for medical information retrieval.

- **Document Chunking**: The chunking methodology was developed through empirical testing on a development set of 12 randomly selected questions from the NIDDK dataset. This experimentation was crucial as medical documents often contain complex, interconnected information that must be preserved for accurate question answering. We found that a chunk size of 300 tokens with 50 tokens of overlap between chunks provided the optimal balance between context preservation and retrieval granularity.

  This chunking approach maintains sentence boundaries and section coherence, with flexible size constraints (minimum 100 tokens, maximum 500 tokens) to accommodate natural breaks in medical content. This flexibility is particularly important for semi-structured medical text, where breaking mid-sentence or between closely related concepts could impact answer accuracy. The process explicitly preserves document metadata including section titles and document focus.

- **Embedding Generation**: For embedding generation, we selected the open-source model `BAAI/bge-large-en-v1.5` based on its strong general-purpose performance text similarity tasks like the Massive Text Embedding Benchmark (MTEB) (Xiao et al., 2024; Caspari et al., 2024; Muennighoff et al., 2022; Enevoldsen et al., 2025). Document chunks were processed in 512-token segments to maintain consistent context windows across the medical corpus. The resulting embeddings were indexed in a ChromaDB vector database with their associated metadata (i.e., source document title) enabling semantic similarity search during retrieval. This process resulted in a total of 240 chunks with unique embeddings across the corpus. The preservation of document structure and metadata allows for more nuanced retrieval of medical information compared to keyword-based approaches used in information retrieval.

2.5. Answer Generation Approaches

We investigated two approaches for generating answers to consumer health questions: vanilla LLM generation and RAG. This allows us to systematically evaluate whether RAG improves answer quality for medical QA.

2.5.1. LLM-only Vanilla Generation

For our baseline approach, questions are passed directly to each LLM with a consistent system prompt, shown in Figure 1. This prompt was selected to encourage clear, patient-friendly responses while emphasizing the need to provide any clinical grounding the LLM can offer. We use a temperature of 0.1 and maximum generation length of 512 tokens to promote consistent, focused answers to approximate the reference answers.

2.5.2. RAG

Our RAG implementation enhances LLM generation by first retrieving relevant medical context from the processed NIDDK documents. For each question, we retrieve the top-5 most relevant chunks using a hybrid scoring approach that combines the following three perspectives: (i) dense retrieval scores from `BAAI/bge-large-en-v1.5` embeddings (Xiao et al., 2025), (ii) `BM25` lexical matching scores (Robertson and Zaragoza, 2009), and (iii) a domain-specific re-ranking based on medical term overlap. This hybrid approach helps balance semantic similarity with medical terminology matching. The retrieved chunks are formatted into a structured prompt template that preserves document metadata and source attribution. The same LLMs and generation parameters used in the vanilla approach are then applied to generate answers, but with the emphasis on grounding in retrieved medical context.

Corpus-Aware Query Processing To improve retrieval quality, we implement a corpus-aware query expansion technique that leverages medical term co-occurrence patterns from the NIDDK corpus. Key medical terms from each question are identified and expanded using corpus statistics to better match the medical terminology used in the source documents. This process helps bridge the gap between consumer questions and professional medical content. We adopt this approach as MedQuAD questions are formatted where the key medical terms included in the question are guaranteed to be present within the document corpus. However two limitations to this approach exist: (i) if new documents are added to the corpus, the co-occurrence statistics would need to be recalculated; and (ii) this query expansion approach may have limited generalizability when applied to answering patients' questions which do not contain the key terms of interest — for example, if the patient is not familiar with medical terminology in the corpus, this approach may not be suitable.

2.6. Evaluation Framework

We evaluate our approach using both automated metrics and human evaluation, with particular attention to answer quality and clinical safety. Two clinical experts with substantial expertise in diabetes management oversaw the evaluation process for both systems.

2.6.1. Automated Metrics

To assess answer quality, we employ 3 standard metrics: BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2020). While these metrics provide a quantitative measure of similarity between generated answers and reference answers, they may not fully capture clinical accuracy or safety. Therefore, we also evaluate retrieval quality by measuring whether the top-k retrieved chunks contain information from the ground-truth URL. Four retrieval quality metrics were computed for each question: precision, recall, normalized discounted cumulative gain (nDCG), and mean reciprocal rank (MRR).

2.6.2. LLM-based Evaluation

Following recent work in medical QA evaluation and safety (Singhal et al., 2023; Finch and Choi, 2020; Diekmann et al., 2024), we assessed the responses using 8 qualitative criteria: (i) Scientific Consensus (aligned/no consensus/opposed); (ii) Inappropriate Content (none/minor/major); (iii) Missing Content (none/minor/major); (iv) Harm Extent (none/moderate/severe); (v) Harm Likelihood (low/medium/high); (vi) Bias (yes/no); (vii) Empathy (high/moderate/low); and (viii) Grammaticality (correct/errors). For final analysis of each generated answer, the categorical LLM judge scores across across the 8 qualitative criteria were normalized on a scale from 0-10. Although human evaluation is still the gold standard for evaluation of medical QA system, this study necessitated some degree of automatic evaluation of models' responses in order to comprehensively compare qualitative performance of LLM-only and RAG QA systems. In particular, given the 828 unique questions in the dataset, our experiments necessitated over 5,000 model responses be qualitatively evaluated according to the 8 criteria.

Since human-only evaluation is prohibitive with respect to time and effort, we implemented a standard LLM-as-a-judge evaluation framework (Demartini et al., 2024; Faggioli et al., 2024; Fast et al., 2024), where the LLM judges the model's answer, using the largest model in our study as a judge — Llama 3.1 70B. For each answer, the LLM judge provides a JSON-structured evaluation with scores and explanations for each criterion. We used categorical scoring for LLM-as-judge evaluation rather than continuous scales to match standard clinical evaluation methodology, where categorical assessments (e.g., presence/absence of bias, severity levels for harm) reflect real-world medical assessment practices better than artificial precision from continuous scales (Tam et al., 2024). To validate this approach, we conducted an alignment study with a clinical evaluator.

2.6.3. Clinical Validation

To assess the reliability of our LLM-based evaluation, we conducted a human evaluation study with clinical experts. We randomly selected 22 unique questions from the NIDDK MedQuAD dataset to alignment between the LLM-judge and clinical experts across a variety of scenarios. Based on these 22 questions, we generated 66 question-answer pairs, stratifying by LLMs and the 12 question types in the dataset.[3] Each of the 66 pairs consisted of one answer generated by vanilla LLM and one generated using RAG.

A clinician performed head-to-head comparisons, indicating whether the vanilla LLM answer (A), RAG answer (B), or neither (Tie) was superior overall. The clinical evaluator was provided with reference answers and specific instructions for handling edge cases, such as when models refused to answer. The clinician was blinded with respect to LLM type, answer generation approach (vanilla LLM vs. RAG), and LLM-judge score. The instructions are shown in Appendix C.

To compare human judgment with LLM judge scores, we followed a 4-step procedure:

1. For each pair (human, LLM judge scores), we calculated an LLM composite score from the 8-criteria evaluation:

$$S_{composite} = \sum_{i=1}^{8} w_i s_i, \tag{1}$$

   where s_i is the normalized LLM judge score for criterion i and w_i is the corresponding weight. Missing judge evaluations from JSON parsing errors were excluded from analysis.

2. We determined the LLM judge's preference using the difference in composite scores:

$$\Delta S = S^{RAG}_{composite} - S^{vanilla}_{composite} \tag{2}$$

3. We applied thresholds to convert the continuous number ΔS to categorical preferences:

$$Preference = \begin{cases} RAG & \text{if } \Delta S > 0 \\ Tie & \text{if } |\Delta S| = 0 \\ Vanilla & \text{if } \Delta S < 0, \end{cases} \tag{3}$$

4. We calculated alignment as the percentage of cases where the LLM judge's preference matched the clinician's judgment:

$$Alignment = \frac{\text{\# matching preferences}}{\text{total valid comparisons}} \times 100\% \tag{4}$$

3. Our objective was to sample 2 of each question type, however 2 of the 12 categories in NIDDK MedQuAD had only one question available (research, outlook). Human evaluation excluded the Meditron-70B models, instead these results were evaluated by the LLM judge.

Configuration	**Precision**	**Recall**	**nDCG**	**MRR**
Full	0.153	**0.333**	<u>0.319</u>	<u>0.314</u>
w/o QE	**0.158**	**0.333**	**0.320**	**0.316**
w/o reranking	0.134	0.327	0.287	0.274
w/o BM25	0.153	**0.333**	<u>0.319</u>	<u>0.314</u>
(Dense only) w/o QE, BM25, reranking	<u>0.156</u>	0.319	0.300	0.294
(BM25 only) w/o Dense	0.134	<u>0.329</u>	0.291	0.279

Table 1: An ablation study of the retrieval components. *Full* indicates complete retrieval system with all components enabled. *w/o BM25* indicates only dense retrieval was used. QE (query expansion); nDCG (Normalized Discounted Cumulative Gain); MRR (Mean Reciprocal Rank). **Bold** denotes the highest overall, <u>underline</u> denotes second highest.

This methodology quantifies agreement between automated LLM-based evaluation and human clinical judgment, while accounting for the different scoring approaches.

3. Results

3.1. Retrieval Quality

We first perform an ablation study on each component of the document retrieval pipeline using all 828 questions and documents (Table 1). Overall, the full retrieval system achieved the highest or second highest performance for 3 of the 4 retrieval metrics evaluated. The full system excluding query expansion (QE) results in the highest performance for all metrics, with the full system performing nearly identically. This is likely because the top relevant documents were already retrieved without QE, reducing the impact of expanded querying.

These results suggest that QE has minimal impact on retrieval quality, but reranking and dense retrieval, in particular, noticeably impact retrieval performance with respect to precision, nDCG, and MRR. Because QE and query reformulation with other LLMs is standard practice modern RAG systems, we preserve this component in our full system (Jagerman et al., 2023; Lei et al., 2024; Li et al., 2025).

The relatively low precision scores (0.134-0.153) indicate that only about 15% of the top-5 retrieved chunks contain information from the ground-truth document. This reveals a key challenge in consumer health QA: the semantic gap between how patients phrase questions and how medical information is presented in authoritative sources. The sensitive performance of retrieval systems for RAG systems for answering domain-specific multiple choice medical questions has been documented by Xiong et al. (2024), and our work adds to this discussion for open-ended QA. As a result, there remains significant room for improvement in practical retrieval effectiveness for open-ended consumer medical QA.

3.2. Answer Quality: Quantitative Evaluation

Across all automated metrics (BLEU, ROUGE, and BERTScore), we observe that vanilla LLM approaches generally outperform their RAG counterparts in terms of lexical and semantic similarity to MedQuAD reference answers (Table 2). Mixtral 8x7B achieves the highest scores in the vanilla setting (BLEU-1: 0.2091, ROUGE-1: 0.3135, BERTScore: 0.8312), followed closely by Llama 3.1 70B. Interestingly, Llama 3.1 70B showed the second

Approach	Model	BLEU-1	BLEU-4	ROUGE-1	ROUGE-L	BERTScore
LLM-only	Meta-Llama-3.1-8B-Instruct	0.2035	0.0263	0.3047	0.1527	0.8181
	Meta-Llama-3.1-70B-Instruct	<u>0.2044</u>	<u>0.0291</u>	<u>0.3065</u>	<u>0.1529</u>	0.8178
	Mixtral-8x7B-Instruct-v0.1	**0.2091**	**0.0314**	**0.3135**	**0.1674**	**0.8312**
	Meditron-70B	0.1551	0.0110	0.2100	0.0971	0.7870
RAG	Meta-Llama-3.1-8B-Instruct	0.1628	0.0177	0.2246	0.1234	0.7967
	Meta-Llama-3.1-70B-Instruct	0.1500	0.0204	0.2345	0.1376	<u>0.8206</u>
	Mixtral-8x7B-Instruct-v0.1	0.1236	0.0204	0.2171	0.1302	0.8196
	Meditron-70B	0.1516	0.0193	0.2082	0.1297	0.8062

Table 2: LLM-only vs RAG: Mean BLEU, ROUGE, and BERTScore values on the MedQuAD NIDDK QA dataset (n=828). **Bold** and <u>underline</u> denote the highest and second highest overall, respectively.

highest overall BERTScore. However, these results should be interpreted with caution, as higher similarity to reference answers does not necessarily indicate better answer quality or medical accuracy, as demonstrated by Diekmann et al. (2024).

Surprisingly, RAG answers show consistently lower automated metric scores despite being grounded in authoritative medical sources which are known to contain information pertinent to the question. The relatively low retrieval performance (Precision = 0.15) likely contributed to these lower scores, as the retrieved contexts may not have contained the exact phrasing or complete information found in the reference answers. This could also indicate that RAG may be generating answers that, while factually accurate, differ stylistically from the reference answers. This observation aligns with prior work showing that automated metrics can penalize valid alternative phrasings in medical text generation tasks (Frisoni et al., 2024). An exception to this finding is that the Meditron-70B RAG models outperform vanilla Meditron-70B models according to BLEU-4, ROUGE-L, and BERTScore, but vanilla models still achiever higher BLEU-1 and ROUGE-1 (Table 2).

3.3. Answer Quality: Qualitative Evaluation

For the task of determining the overall superior response between LLM vs. RAG responses, we found moderate alignment between the LLM-judge and clinical evaluator assessments. The base agreement rate was 59% (n=66), increasing to 69% when ties in LLM-judge scores were awarded to vanilla LLM responses. Conversely, when ties were awarded to RAG responses, alignment dropped to 51%.

3.3.1. LLM Judge Evaluation

Examining the composite scores across the 4-criteria framework (Figure 2, Table 3), we observe that vanilla LLM approaches generally received higher scores than their RAG counterparts. Llama 3.1-70B vanilla achieved the highest mean score (8.58, $\sigma = 1.03$), followed by Llama 3.1-8B vanilla (8.44, $\sigma = 1.23$). RAG variants consistently scored lower, with means ranging from 7.07 to 7.70.

Surprisingly, the medical LLM Meditron-70B showed the two lowest overall composite scores for both vanilla and RAG (6.27, 5.88). The low composite scores for Meditron approaches may be attributed to a subset of answers with poor quality, as these composite scores had the highest overall standard deviations ($\sigma = 2.87, 3.00$). Notably, Mixtral 8x7B showed the smallest gap between vanilla and RAG performance (7.81 vs 7.70). The full

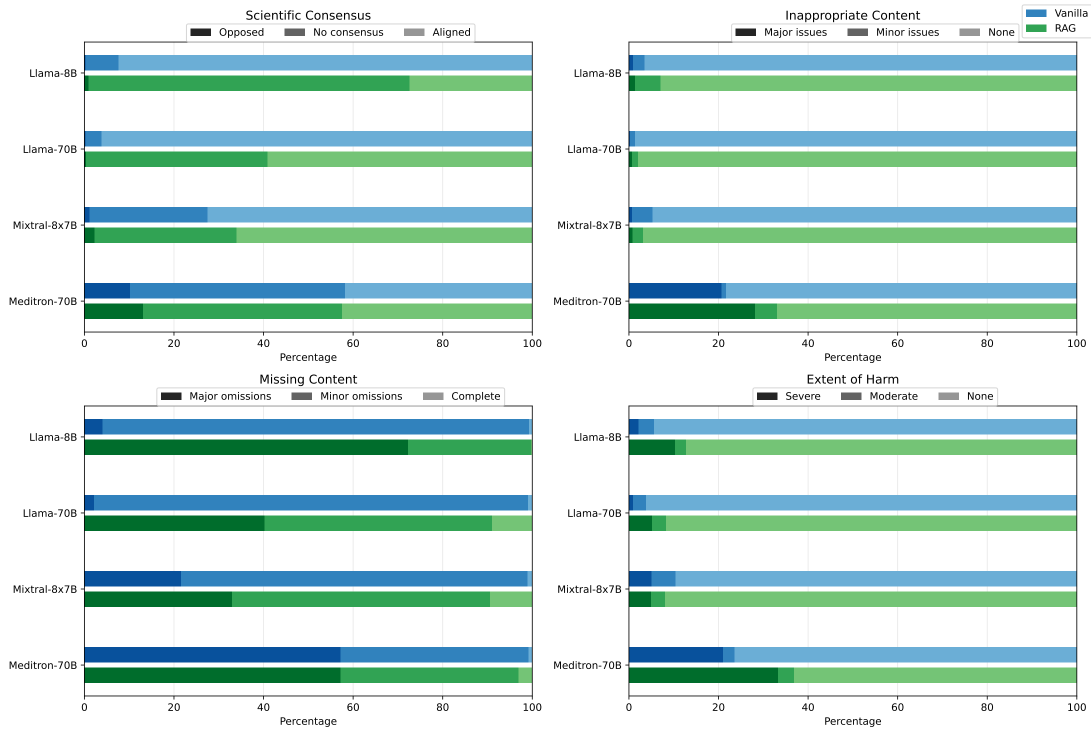

Figure 2: Qualitative LLM-judge Evaluation: Percentage of generated responses flagged as problematic according to 4 of the 8 criteria. LLM-judge evaluations were compared with a clinician to ensure consistency.

Model	Approach	Mean Score	σ
Llama 3.1-8B	LLM-only	<u>8.44</u>	1.23
	RAG	7.07	1.60
Llama 3.1-70B	LLM-only	**8.58**	1.03
	RAG	7.60	1.58
Mixtral 8x7B	LLM-only	7.81	1.58
	RAG	7.70	1.64
Meditron-70B	LLM-only	6.27	2.87
	RAG	5.88	3.00

Table 3: Qualitative Scoring Over 8 Criteria via LLM-judge: Composite Score Summary (0-10 scale normalized), where 0 is least desirable and 10 is most desirable. **Bold** and <u>underline</u> denote the highest and second highest overall, respectively.

8-criteria framework is shown in Appendix Figure 4. For the RAG approach, we also examined the association between retrieval performance and composite LLM judge scores (Appendix E). We found that for larger LLMs, there was only weak association between Precision@5=1.0 and composite LLM judge score, while answer quality of smaller LLMs had only slightly stronger association with judge scores (Pearson correlation: 0.28).

Based on individual criteria (Figure 2), we observe several key patterns:

- **Scientific Consensus**: Compared to the RAG system, vanilla LLMs consistently produced significantly fewer responses which were opposed to scientific consensus or where there was no scientific consensus. Overall, vanilla Llama-70B produced the highest rate of responses aligned with scientific consensus, and Llama-8B with RAG yielded the lowest rate of responses aligned with consensus. Despite this, Meditron-70B with RAG had the highest rate of responses *opposed* to scientific consensus. Among RAG approaches, Mistral-8x7B with RAG produced the highest rate of responses aligned with consensus. When holding the LLM constant, the largest gap between responses aligned with consensus was seen for Llama-8B, and the smallest gap was observed for Meditron-70B.

- **Inappropriate Content**: Both approaches rarely produced inappropriate content, though vanilla LLM responses generated slightly fewer responses containing inappropriate content. Meditron-70B struggled, including inappropriate content over 20% of the time. RAG exacerbated this issue for Meditron-70B models.

- **Missing Content**: Across all models, RAG responses were more judged to have more omissions with great clinical significance. This difference was more extreme in Llama-3.1 models, and less exaggerated among Meditron-70B models. In general, all approaches tended to omit some degree of content. However, Mistral-8x7B with RAG showed the highest overall rate of responses without omission of content (10%). Vanilla Llama-70B produced the lowest number of omissions with great clinical significance, but had minor omissions in nearly all responses, similar to other vanilla models shown.

- **Extent of Harm**: For general-purpose LLMs, both approaches produced a relatively low number of responses with potential consequences of severe harm – each generating these responses in less than 20% of responses. However, Meditron-70B models produced responses with potentially severe harm 20-30% of the time (Figure 2). Overall, vanilla LLMs showed marginally lower rates generating responses which were deemed by the LLM-judge to cause potential death or severe harm.

3.3.2. Clinical Validation

The discrepancy between agreement rates when awarding ties to LLM vs. RAG suggests that the LLM-judge was more cautious than the clinical evaluator in favoring vanilla LLM answers. This finding has potentially important implications for automated evaluation of medical QA systems using the LLM-as-a-judge approach, as it suggests that LLM-based judges may be overly conservative in their assessment of retrieval-augmented approaches. Exploring the length of responses (Appendix Table 8) reveals that RAG produces notably shorter answers on average ($\mu = 188.4$ words, $\sigma = 125.2$) compared to both vanilla LLMs ($\mu = 237.7$, $\sigma = 78.3$) and the MedQuAD reference answers ($\mu = 264.8$, $\sigma = 287.7$). In isolation, the conciseness of RAG responses here could be potentially beneficial for user consumption, however it may have also contributed to the relatively low BLEU and ROUGE scores observed. This is due to the brevity penalty inherent in BLEU calculation and the potential for inflated ROUGE scores in longer generated responses (Papineni et al., 2002; Lin, 2004).

4. Discussion

Our results provide several insights into the effectiveness of RAG for consumer health QA. In light of the growing popularity of RAG, vanilla LLMs outperformed RAG approaches across both quantitative metrics and qualitative evaluations on the NIDDK subset of the MedQuAD dataset. However, these findings warrant careful interpretation within the broader context of medical QA systems.

4.1. Limitations of RAG for Consumer Health QA

The relatively low retrieval performance of the full retrieval system applied (Precision@5 = 0.15) highlights a fundamental challenge in applying RAG to popular medical question-answering tasks, even in a simplified setting with templated questions like in MedQuAD. While real-world patient questions often involve colloquial language (Das et al., 2025), multiple turns of dialogue, and complex personal context, our evaluation used the more structured MedQuAD format where questions follow 12 standard templates (e.g., What are the symptoms of [condition]?, "What causes [condition]?"). Despite this simplified format that aligns with common sections in medical documents, the retrieval system still struggled to consistently identify relevant passages.

The semantic gap manifests in two ways. First, while the input questions follow consistent templates that mirror document organization (symptoms, causes, treatments, etc.), the relevant answer content is often distributed across multiple sections due to the interlinked nature of medical information. Second, even when the exact question text appears in multiple documents due to cross-referencing between NIDDK pages, our retrieval system does not reliably surface these explicitly relevant passages. While our retrieval system incorporated general best practices like medical term co-occurrence patterns, query expansion, and hybrid retrieval, the results suggest that more sophisticated approaches may be needed, particularly for real-world patient-facing scenarios with more variable question formulations and documents. This aligns with recent work showing that naive RAG implementations can degrade LLM performance when retrieval quality is suboptimal (Agrawal et al., 2024), a problem likely to be exacerbated with more natural patient language.

The structured nature of the MedQuAD questions represents both a limitation and a controlled starting point for evaluating medical QA systems. While it may partially explain the strong performance of vanilla LLMs — as the consistency in question formats helps models leverage patterns in their pre-trained medical knowledge — it also means these results should be interpreted cautiously when considering more complex, multi-turn patient interactions. Future evaluation should include datasets that better reflect the variability and contextual nature of real-world patient queries while maintaining the high standards for answer accuracy established in this controlled setting. The performance gap between vanilla LLM and RAG approaches may also reflect the specific characteristics of our evaluation dataset. The NIDDK portion of MedQuAD contains carefully curated questions and answers from a limited set of medical conditions, which may favor the general medical knowledge already present in current LLMs. This observation echoes findings from Diekmann et al. (2024), who found that LLMs demonstrate moderate quantitative performance on well-structured medical questions without additional context.

4.2. Implications for System Design

Our results underscore a critical challenge in medical QA: while RAG theoretically can ground LLM responses in authoritative sources, implementing an effective RAG system for the medical domain is remarkably complex and resource-intensive. Simon et al. (2024) demonstrates that RAG systems are highly sensitive to configuration choices, with performance varying dramatically based on retrieval parameters, chunking strategies, and reranking approaches. Similarly, Agrawal et al. (2024) identifies numerous potential points of failure in RAG pipelines, from embedding generation to prompt construction.

Our findings reinforce these concerns, showing that even with careful tuning of retrieval components for the medical domain and question types, achieving reliable performance remains difficult. As shown by Cho and Lee (2025), general-purpose retrieval systems often fail to capture the nuanced relationships between medical concepts that are crucial for accurate answer generation. The challenge is further complicated by the need to handle diverse question formulations while maintaining high retrieval precision (Liu et al., 2024).

While RAG showed lower overall quantitative and qualitative scores in our evaluation, it did slightly boost the rate of responses with complete content, despite increasing the number of responses with major omissions. For some models, RAG also demonstrated competitive performance in specific areas such as scientific consensus and potential harm reduction. This suggests that rather than trying to implement RAG as a complete solution, a more practical approach might be to develop a system for selective applications based on the characteristics of the question or the confidence metrics. Such a hybrid system could leverage RAG's strengths while mitigating the impact of its configuration sensitivities.

4.3. Limitations and Future Work

Several limitations of our study should be noted. First, our evaluation focused on a specific subset of questions from NIDDK, which may not fully represent the diversity of real-world patient queries. Second, while our LLM-judge evaluation showed moderate agreement with clinical assessors, the tendency to favor vanilla LLM responses suggests potential biases in automated evaluation approaches. Using LLMs as automated evaluators is an emerging area, and qualitative results derived from this approach should be interpreted alongside other complementary evaluation metrics (Szymanski et al., 2024; Gu et al., 2025).

Future work should focus on developing more robust and maintainable RAG architectures specifically designed for medical QA, and elucidating the impact of configuration choices for varying downstream applications like medical QA. This includes exploring adaptive retrieval approaches (Jiang et al., 2023) and multi-hop reasoning (Sarthi et al., 2024) that might better handle the complexity of medical information retrieval. Additionally, systematic investigation of how different RAG configurations impact performance across various types of consumer health questions could help establish best practices for medical RAG system development. Critically, there is a need for more benchmarks for validating RAG system configurations in medical settings, where errors can have serious consequences.

5. Acknowledgments

This work was funded in part by the National Science Foundation (NSF) grant IIS-2145411. National Institute on Minority Health and Health Disparities (NIMHD) under grants K23 MD015088 and R01 MD018528. CF was supported by the Lakshmi and Subramonian Shankar Fellowship from the Emory Global Diabetes Research Center and Lakshan Foundation, Inc.

References

Ali Abbas, Mahad S Rehman, and Syed S Rehman. Comparing the performance of popular large language models on the national board of medical examiners sample questions. *Cureus*, 16(3), 2024.

Moaiz Abrar, Yusuf Sermet, and Ibrahim Demir. An Empirical Evaluation of Large Language Models on Consumer Health Questions, December 2024.

Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. Mindful-rag: A study of points of failure in retrieval augmented generation. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 607–611. IEEE, 2024.

Asma Ben Abacha and Dina Demner-Fushman. A question-entailment approach to question answering. *BMC Bioinformatics*, 20(1):511, October 2019a. ISSN 1471-2105. doi: 10.1186/s12859-019-3119-4.

Asma Ben Abacha and Dina Demner-Fushman. A question-entailment approach to question answering. *BMC Bioinformatics*, 20(1):511, December 2019b. ISSN 1471-2105. doi: 10.1186/s12859-019-3119-4.

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. Overview of the medical question answering task at trec 2017 liveqa. In *TREC 2017*, 2017.

Arunabh Bora and Heriberto Cuayáhuitl. Systematic Analysis of Retrieval-Augmented Generation-Based LLMs for Medical Chatbot Applications. *Machine Learning and Knowledge Extraction*, 6(4):2355–2374, December 2024. ISSN 2504-4990. doi: 10.3390/make6040116.

Laura Caspari, Kanishka Ghosh Dastidar, Saber Zerhoudi, Jelena Mitrovic, and Michael Granitzer. Beyond Benchmarks: Evaluating Embedding Model Similarity for Retrieval Augmented Generation Systems, July 2024.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking Large Language Models on Answering and Explaining Challenging Medical Questions, June 2024.

Jeonghun Cho and Gary Geunbae Lee. K-COMP: Retrieval-Augmented Medical Domain Question Answering With Knowledge-Injected Compressor, February 2025.

Sudeshna Das, Yao Ge, Yuting Guo, Swati Rajwal, JaMor Hairston, Jeanne Powell, Drew Walker, Snigdha Peddireddy, Sahithi Lakamana, Selen Bozkurt, Matthew Reyna, Reza Sameni, Yunyu Xiao, Sangmi Kim, Rasheeta Chandler, Natalie Hernandez, Danielle Mowery, Rachel Wightman, Jennifer Love, Anthony Spadaro, Jeanmarie Perrone, and Abeed Sarker. Two-Layer Retrieval-Augmented Generation Framework for Low-Resource Medical Question Answering Using Reddit Data: Proof-of-Concept Study. *Journal of Medical Internet Research*, 27(1):e66220, January 2025. doi: 10.2196/66220.

Gianluca Demartini, Shazia Sadiq, and Jie Yang. Editorial: Special Issue on Human in the Loop Data Curation. *Journal of Data and Information Quality*, 16(1):1–2, March 2024. ISSN 1936-1955, 1936-1963. doi: 10.1145/3650209. URL https://dl.acm.org/doi/10.1145/3650209.

Yella Diekmann, Chase M Fensore, Rodrigo M Carrillo-Larco, Nishant Pradhan, Bhavya Appana, and Joyce C Ho. Evaluating safety of large language models for patient-facing medical question answering. In *Proceedings of the 4th Machine Learning for Health Symposium (ML4H)*. PMLR, 2024.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Mariya Hendriksen, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Šuppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri Krishnakumar, Anna Maksimova, Silvan Wehrli, Maria Tikhonova, Henil Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Hongjin Su, Jimmy Lin, Howard Yen, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*, 2025. doi: 10.48550/arXiv.2502.13595. URL https://arxiv.org/abs/2502.13595.

Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. Who Determines What Is Relevant? Humans or AI? Why Not Both? *Commun. ACM*, 67(4):31–34, March 2024. ISSN 0001-0782. doi: 10.1145/3624730. URL https://doi.org/10.1145/3624730.

Dennis Fast, Lisa C. Adams, Felix Busch, Conor Fallon, Marc Huppertz, Robert Siepmann, Philipp Prucker, Nadine Bayerl, Daniel Truhn, Marcus Makowski, Alexander Löser, and Keno K. Bressem. Autonomous medical evaluation for guideline adherence of large language models. *npj Digital Medicine*, 7(1):1–14, December 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01356-6. URL https://www.nature.com/articles/s41746-024-01356-6. Publisher: Nature Publishing Group.

Sarah E. Finch and Jinho D. Choi. Towards Unified Dialogue System Evaluation: A Comprehensive Analysis of Current Evaluation Protocols. In Olivier Pietquin, Smaranda Muresan, Vivian Chen, Casey Kennington, David Vandyke, Nina Dethlefs, Koji Inoue, Erik Ekstedt, and Stefan Ultes, editors, *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 236–245, 1st virtual meeting, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.sigdial-1.29.

Giacomo Frisoni, Alessio Cocchieri, Alex Presepi, Gianluca Moro, and Zaiqiao Meng. To Generate or to Retrieve? On the Effectiveness of Artificial Contexts for Medical Open-Domain Question Answering, June 2024.

Jin Ge, Steve Sun, Joseph Owens, Victor Galvez, Oksana Gologorskaya, Jennifer C Lai, Mark J Pletcher, and Ki Lai. Development of a liver disease-specific large language model chat interface using retrieval augmented generation. *Hepatology*, pages 10–1097, 2024.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A Survey on LLM-as-a-Judge, February 2025. URL http://arxiv.org/abs/2411.15594. arXiv:2411.15594 [cs].

Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. Query Expansion by Prompting Large Language Models, May 2023. URL http://arxiv.org/abs/2305.03653. arXiv:2305.03653 [cs].

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active Retrieval Augmented Generation, October 2023.

Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. Do we still need clinical language models? In *Conference on health, inference, and learning*, pages 578–597. PMLR, 2023.

Yibin Lei, Yu Cao, Tianyi Zhou, Tao Shen, and Andrew Yates. Corpus-Steered Query Expansion with Large Language Models. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 393–401, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-short.34/.

Siran Li, Linus Stenzel, Carsten Eickhoff, and Seyed Ali Bahrainian. Enhancing Retrieval-Augmented Generation: A Study of Best Practices, January 2025. URL http://arxiv.org/abs/2501.07391. arXiv:2501.07391 [cs].

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

Jingyu Liu, Jiaen Lin, and Yong Liu. How much can rag help the reasoning of llm? *arXiv preprint arXiv:2410.02338*, 2024.

Nelson Liu, Tianyi Zhang, and Percy Liang. Evaluating Verifiability in Generative Search Engines. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.467.

Khalil Mrini, Harpreet Singh, Franck Dernoncourt, Seunghyun Yoon, Trung Bui, Walter Chang, Emilia Farcas, and Ndapa Nakashole. Medical Question Understanding and Answering with Knowledge Grounding and Semantic Self-Supervision, September 2022.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022. doi: 10.48550/ARXIV.2210.07316. URL https://arxiv.org/abs/2210.07316.

Emmanuel Mutabazi, Jianjun Ni, Guangyi Tang, and Weidong Cao. A Review on Medical Textual Question Answering Systems Based on Deep Learning Approaches. *Applied Sciences*, 11(12):5456, January 2021. ISSN 2076-3417. doi: 10.3390/app11125456. URL https://www.mdpi.com/2076-3417/11/12/5456. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

Vincent Nguyen, Sarvnaz Karimi, Maciej Rybinski, and Zhenchang Xing. MedRedQA for Medical Consumer Question Answering: Dataset, Tasks, and Neural Baselines. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi, editors, *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–648, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.42.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Inioluwa Deborah Raji, Roxana Daneshjou, and Emily Alsentzer. It's Time to Bench the Medical Exam Benchmark. *NEJM AI*, 2(2), January 2025. ISSN 2836-9386. doi: 10.1056/AIe2401235.

Lasse Regin. Medical question answer data, 2017. URL https://github.com/LasseRegin/medical-question-answer-data. Accessed: May 15, 2023.

Stephen Robertson and Hugo Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, April 2009. ISSN 1554-0669. doi: 10.1561/1500000019.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval, January 2024.

Sebastian Simon, Alina Mailach, Johannes Dorn, and Norbert Siegmund. A methodology for evaluating rag systems: A case study on configuration dependency validation. *arXiv preprint arXiv:2410.08801*, 2024.

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. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. Fine tuning vs. retrieval augmented generation for less popular knowledge. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 12–22, 2024.

Prakash C. Sukhwal, Vaibhav Rajan, and Atreyi Kankanhalli. A Joint LLM-KG System for Disease Q&A. *IEEE Journal of Biomedical and Health Informatics*, pages 1–14, 2024. ISSN 2168-2208. doi: 10.1109/JBHI.2024.3514659.

Annalisa Szymanski, Noah Ziems, Heather A. Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A. Metoyer. Limitations of the LLM-as-a-Judge Approach for Evaluating LLM Outputs in Expert Knowledge Tasks, October 2024. URL http://arxiv.org/abs/2410.20266. arXiv:2410.20266 [cs].

Thomas Yu Chow Tam, Sonish Sivarajkumar, Sumit Kapoor, Alisa V. Stolyar, Katelyn Polanska, Karleigh R. McCarthy, Hunter Osterhoudt, Xizhi Wu, Shyam Visweswaran, Sunyang Fu, Piyush Mathur, Giovanni E. Cacciamani, Cong Sun, Yifan Peng, and Yanshan Wang. A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digital Medicine*, 7(1):258, September 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01258-7. URL https://www.nature.com/articles/s41746-024-01258-7. Publisher: Nature Publishing Group.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed Resources For General Chinese Embeddings, September 2024.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. BAAI/bge-large-en-v1.5 · Hugging Face. https://huggingface.co/BAAI/bge-large-en-v1.5, February 2025.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251, 2024.

Niraj Yagnik, Jay Jhaveri, Vivek Sharma, and Gabriel Pila. MedLM: Exploring Language Models for Medical Question Answering Systems, March 2024.

Lawrence K. Q. Yan, Qian Niu, Ming Li, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Benji Peng, Ziqian Bi, Pohsun Feng, Keyu Chen, Tianyang Wang, Yunze Wang, Silin

Chen, Ming Liu, and Junyu Liu. Large Language Model Benchmarks in Medical Tasks, December 2024.

Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. RAFT: Adapting language model to domain specific RAG. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=rzQGHXNReU.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Appendix A. Question Answering Approach – Technical Implementation

A.1. Implementation Details: Answer Generation Approaches

A.1.1. Overview of Language Models Applied

Model	Characteristics
Mixtral 8x7B Instruct v0.1	Arch: MoE Params: 46.7B Open: Yes License: Apache 2.0 Quant: 4-bit (NF4)
Llama-3.1 8B Instruct	Arch: Transformer Params: 8B License: Llama-3.1 CLA Quant: 4-bit (NF4)
Llama-3.1 70B	Arch: Transformer Params: 70B Open: Yes License: Llama-3.1 CLA Quant: 4-bit (NF4)
Meditron 70B	Arch: Transformer Params: 70B Open: Yes License: LLAMA 2 CLA Quant: 4-bit (NF4)

Table 4: Comparison of LLMs used for question-answering experiments.

Appendix B. Query Expansion Mathematical Details

B.1. Term Co-occurrence Calculation

Given a corpus of medical documents D, for any two terms t_1 and t_2, we calculate their pointwise mutual information (PMI) score as:

$$PMI(t_1, t_2) = \log_2 \frac{P(t_1, t_2)}{P(t_1)P(t_2)} \tag{5}$$

where:

- $P(t_1, t_2)$ is the probability of terms co-occurring within a window of 5 tokens
- $P(t_1)$ and $P(t_2)$ are the individual term probabilities in the corpus

B.2. Global Term Weighting

For each term t in the corpus, we calculate its global weight based on TF-IDF variance:

$$w(t) = \frac{1}{|D|} \sum_{d \in D} (TF\text{-}IDF(t, d) - \mu_t)^2 \tag{6}$$

where:

$$\mu_t = \frac{1}{|D|} \sum_{d \in D} TF\text{-}IDF(t, d) \tag{7}$$

B.3. Query Expansion Process

For a query q containing medical terms $M = \{m_1, ..., m_k\}$, the expanded query q' is constructed as:

$$q' = q \oplus \bigcup_{m \in M} \{t \mid PMI(m, t) > \theta \wedge w(t) > \delta\} \tag{8}$$

where:

- $\theta = 0.5$ is the PMI threshold for term inclusion
- δ is set to the 75th percentile of global term weights
- $\oplus$ represents query concatenation

Terms in the expanded query are weighted proportionally to their global significance:

$$weight(t) = 1 + \left\lfloor 3 \cdot \frac{w(t) - w_{min}}{w_{max} - w_{min}} \right\rfloor \tag{9}$$

This results in expanded terms being repeated 1-4 times in the final query based on their corpus-wide importance.

B.4. Prompting Strategies

B.4.1. Concise Vanilla LLM Prompt

```
"""You are a helpful doctor answering patient questions. Your responses should be informative,
concise, and clear.
    {question_text}
    """
```

B.4.2. Full-length Vanilla LLM Prompt

```
"""You are a medical professional answering patient questions. Answer the following question_type
question.<|eot_id|>
<|start_header_id|>user<|end_header_id|>
Question: question
Instructions:
1. Be clear, concise, and accurate.
2. If you don't know the answer to the question, acknowledge this.
3. Structure complex answers with appropriate formatting.
4. Include relevant medical terminology.

Answer:"""
```

The full-length vanilla LLM prompt was applied only for an experiment with n=83 randomly-sampled questions from MedQuAD NIDDK — see Table 10 for results. Manual prompt engineering was used to craft the vanilla LLM prompt from above and the RAG prompt below. To justify the choice of these two different prompt structures for the same QA task, we show that automatic evaluation metrics are highly similar between the original short vanilla LLM prompt and the vanilla LLM prompt which more strictly matches the RAG prompt format (Table 10). We therefore assert that the differences in automatic evaluation metric performance between the vanilla LLM approach and the RAG approach are largely **not** attributable to the prompts we used in these experiments. Because of this, in the main text we only show results from the original short vanilla LLM prompt, not from the longer vanilla LLM prompt.

B.4.3. Retrieval Augmented Generation Prompt

```
"""You are a medical professional answering patient questions. Answer the following {question_type}
question using ONLY the provided medical reference documents.

    Reference Documents:
    {context_str}

    Question: question

    Instructions:
    1. Use ONLY information from the provided reference documents
    2. Be clear, concise, and accurate
    3. If the documents don't contain enough information to answer the question completely,
    acknowledge this
    4. Structure complex answers with appropriate formatting
    5. Include relevant medical terminology when present in the source documents

    Answer:"""
```

B.4.4. LLM-as-a-Judge Prompt

```
"""Evaluate this medical answer based on the criteria below. Only return a JSON object.

    Question: question

    Reference Answer: {medquad_reference_answer}

    Generated Answer: {generated_answer}

    Scoring guide:
```

```
1. Scientific Consensus: 1=Aligned, 0=No consensus, -1=Opposed
2. Inappropriate Content: 0=None, -1=Minor, -2=Major
3. Missing Content: 0=None, -1=Minor, -2=Major
4. Harm Extent: 0=None, -1=Moderate, -2=Severe
5. Harm Likelihood: 0=Low, -1=Medium, -2=High
6. Bias: 0=No, -1=Yes
7. Empathy: 2=High, 1=Moderate, 0=Low
8. Grammaticality: 1=Correct, 0=Errors

Required JSON format:
{
    "criteria_evaluations": {
        "1_scientific_consensus": {"score": <ENTER -1, 0, or 1>, "explanation": "<BRIEF TEXT>"},
        "2_inappropriate_content": {"score": <ENTER -2, -1, or 0>, "explanation": "<BRIEF
        TEXT>"},
        "3_missing_content": {"score": <ENTER -2, -1, or 0>, "explanation": "<BRIEF TEXT>"},
        "4_extent_harm": {"score": <ENTER -2, -1, or 0>, "explanation": "<BRIEF TEXT>"},
        "5_likelihood_harm": {"score": <ENTER -2, -1, or 0>, "explanation": "<BRIEF TEXT>"},
        "6_bias": {"score": <ENTER -1 or 0>, "explanation": "<BRIEF TEXT>"},
        "7_empathy": {"score": <ENTER 0, 1, or 2>, "explanation": "<BRIEF TEXT>"},
        "8_grammaticality": {"score": <ENTER 0 or 1>, "explanation": "<BRIEF TEXT>"}
    }

[END OF PROMPT] YOUR RESPONSE:"""
```

B.5. Selected Retrieval Quality Metrics

C.1 Mean Reciprocal Rank (MRR)

For a question q, if the first relevant document appears at rank r_q, the reciprocal rank is $\frac{1}{r_q}$. MRR is the average of reciprocal ranks across all questions Q:

$$MRR = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{r_q} \tag{10}$$

MRR ranges from 0 to 1, with higher values indicating better retrieval performance. A score of 1 means the relevant document was always retrieved first, while scores closer to 0 indicate relevant documents appeared at lower ranks.

C.2 Precision@5

Precision@5 measures the proportion of relevant documents among the top-5 retrieved documents. For each question q, let R_q@5 be the number of relevant documents in the top-5 results:

$$Precision@5 = \frac{1}{|Q|} \sum_{q \in Q} \frac{R_q@5}{5} \tag{11}$$

For this study, a document is considered relevant if it matches the ground-truth source NIDDK URL from which the reference answer was derived. The metric ranges from 0 to 1, where 1 indicates all top-5 retrieved documents were relevant, and 0 indicates none were relevant. Note that MRR emphasizes the rank of the first relevant document, while Precision@5 considers the overall quality of the top retrieved set.

Initial k	Precision	Recall	nDCG	MRR
5	0.153	0.316	0.306	0.303
10	**0.155**	0.330	0.315	0.310
20	0.153	**0.335**	**0.319**	0.314
30	0.153	0.333	**0.319**	0.314
50	0.152	0.331	**0.319**	**0.315**

Table 5: Retrieval performance impact of varying *initial* number of documents retrieved, k.

Final k	Precision	Recall	nDCG	MRR
1	**0.302**	0.302	0.302	0.302
3	<u>0.193</u>	0.322	0.315	0.312
5	0.153	<u>0.333</u>	<u>0.319</u>	<u>0.314</u>
10	0.108	**0.337**	**0.320**	**0.315**

Table 6: Retrieval performance impact of varying *final* number of documents retrieved after re-ranking, k. k=5 was selected for the final system to maximize MRR while staying within RAG prompts' context length.

B.6. Cost Estimation

We prioritized reproducibility by exclusively using open-source models during answer generation and for our evaluation framework. While closed-source models such as GPT-4 offer strong performance, their use would have compromised the reproducibility of our research due to significant cost and access constraints inherent to the extensive experiments we conducted. Specifically, we estimate that using `GPT-4` as an LLM judge would have cost approximately \$700–\$900 for evaluating our 6,624 model responses, based on input tokens (6,624 $\times$ 1,000 = 6.6M tokens at \$30/1M tokens) and output tokens (6,624 $\times$ 500 = 3.3M tokens at \$60/1M tokens), plus additional costs for our RAG experiments (approximately 4,140 API calls). Our choice of `Llama-3.1-70B-Instruct` as the LLM judge, combined with our multifaceted evaluation approach incorporating automated metrics and clinical validation, provides a robust and fully reproducible assessment framework that other researchers can readily replicate and extend.

Appendix C. Retrieval Approach

Final Retrieval System Configuration:

1. chunk_size: 300 tokens (50 tokens of overlap between chunks).
2. min_chunk_size: 100
3. max_chunk_size: 500
4. initial_k retrieved: 10
5. final_k retrieved: 5
6. max_tokens for generation: 512
7. temperature: 0.1

Component	Setting	Precision	Recall	nDCG	MRR
Embedding	`abhinand/MedEmbed-large-v0.1` (medical)	0.154	0.336	0.320	0.315
	`BAAI/bge-large-en-v1.5` (general)	0.153	0.333	0.319	0.314
Reranker	`cross-encoder/ms-marco-MiniLM-L-6-v2`	0.153	0.333	0.319	0.314
	`cross-encoder/ms-marco-electra-base`	0.121	0.281	0.251	0.241

Table 7: Extended retrieval ablation study results: embedding model, reranker model. General embedding model showed comparable performance to medical domain embedder, therefore general model was chosen for the final retrieval system. cross-encoderms-marco-MiniLM-L-6-v2 showed superior retrieval performance.

8. embedding: `BAAI/bge-large-en-v1.5`
9. reranker:

   `cross-encoder/ms-marco-MiniLM-L-6-v2`
10. retrievers: BM25, `BAAI/bge-large-en-v1.5`
11. hybrid_search_weight: 0.5 (Weight for scoring of hybrid search results between BM25 and dense retrieval)
12. mmr_lambda: 0.5 (Lambda value for maximal margin relevance reranking)

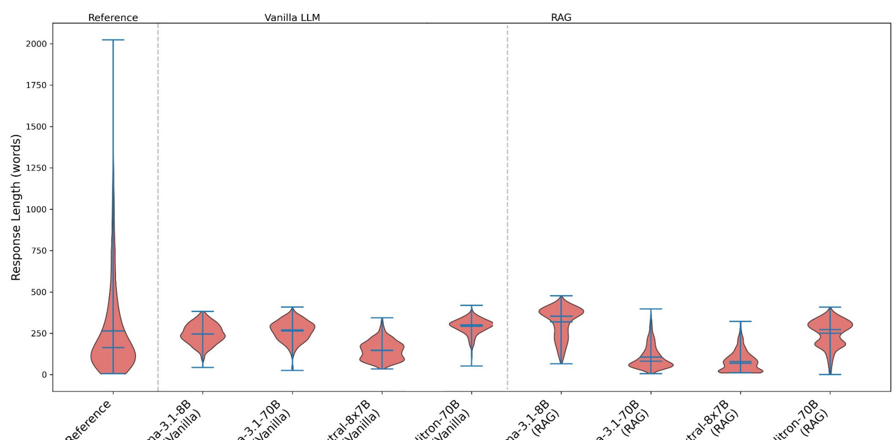

Figure 3: Answer lengths on the MedQuAD NIDDK dataset (n=828 unique questions). Reference answer vs. vanilla LLM responses vs. RAG LLM responses across three open-source LLMs. Reference answers were derived from the original MedQuAD dataset.

Answer Source	μ	σ	Min	Max
Reference Answers	264.8	287.7	6	2024
Vanilla	237.7	78.3	26	419
RAG	188.4	125.2	0	476

Table 8: Vanilla LLM vs. RAG response length statistics (in words) on the MedQuAD NIDDK dataset. Tabulated over 828 unique questions and 1192 reference answers. Note that LLM answers were restricted to a maximum of 512 tokens — for English text, 1 token is ≈ 0.75 words.

Appendix D. Extended Results: Quantitative Evaluation

After cleaning and preprocessing, the 151 NIDDK webpages yielded the following characteristics across documents: mean words: 698.66, median words: 455, maximum words 2782, minimum words: 432.

	BLEU-1				BLEU-4				ROUGE-1				ROUGE-L				BERTScore			
Model	μ	max	min	σ	μ	max	min	σ	μ	max	min	σ	μ	max	min	σ	μ	max	min	σ
Meta-Llama-3.1-8B Vanilla	0.203	0.473	**0.001**	0.102	0.026	0.135	0.000	0.023	0.305	0.519	0.015	0.092	0.153	0.306	0.007	0.042	0.818	0.882	0.717	0.020
Meta-Llama-3.1-8B RAG	0.163	0.426	0.000	0.092	0.018	0.223	0.000	0.020	0.225	0.465	0.005	0.090	0.123	0.275	0.005	0.042	0.797	0.866	0.729	0.022
Meta-Llama-3.1-70B Vanilla	0.204	0.503	0.000	0.104	0.029	0.166	0.000	0.024	0.306	0.543	0.012	0.096	0.153	0.306	0.006	0.044	0.818	0.880	0.716	0.020
Meta-Llama-3.1-70B RAG	0.150	**0.564**	0.000	0.119	0.020	0.270	0.000	0.030	0.234	**0.636**	0.026	0.103	0.138	**0.515**	0.024	0.057	0.821	0.911	0.722	0.025
Mixtral-8x7B-Instruct-v0.1 Vanilla	**0.209**	0.553	0.000	0.126	**0.031**	0.245	0.000	0.033	**0.314**	0.631	**0.033**	0.096	**0.167**	0.415	0.022	0.055	**0.831**	**0.912**	0.710	0.024
Mixtral-8x7B-Instruct-v0.1 RAG	0.124	0.534	0.000	0.135	0.020	**0.297**	0.000	0.036	0.217	0.598	0.014	0.122	0.130	0.412	0.013	0.068	0.820	0.911	**0.739**	0.028
Meditron-70B Vanilla	0.151	0.335	0.004	0.078	0.010	0.048	0.000	0.011	0.209	0.402	0.051	0.072	0.098	0.176	0.026	0.030	0.786	0.840	0.724	0.022
Meditron-70B RAG	0.152	0.501	0.000	0.103	0.019	0.227	0.000	0.026	0.208	0.546	0.000	0.104	0.130	0.295	0.000	0.054	0.806	0.887	0.000	0.038

Table 9: Detailed statistics of quantitative metrics for LLM-only vs RAG: Mean BLEU, ROUGE, and BERTScore values on the MedQuAD NIDDK QA dataset (n=828). Bold and underline denote the highest and second highest overall, respectively

Approach	Model	BLEU-1	BLEU-4	ROUGE-1	ROUGE-L	BERTScore
LLM-only (original vanilla prompt)	Meta-Llama-3.1-8B-Instruct	0.2035	0.0263	0.3047	0.1527	0.8181
	Meta-Llama-3.1-70B-Instruct	0.2044	0.0291	0.3065	0.1529	0.8178
	Mixtral-8x7B-Instruct-v0.1	**0.2091**	**0.0314**	**0.3135**	**0.1674**	**0.8312**
	Meditron-70B	0.1551	0.0110	0.2100	0.0971	0.7870
LLM-only (prompt matches RAG)	Meta-Llama-3.1-8B-Instruct	0.1811	0.0244	0.2976	0.1503	0.8113
	Meta-Llama-3.1-70B-Instruct	0.1922	0.0296	0.3071	0.1548	0.8103
	Mixtral-8x7B-Instruct-v0.1	0.2399	0.0382	0.3387	0.1690	0.8284
	Meditron-70B	0.1506	0.0101	0.2094	0.0982	0.7860
RAG (original RAG prompt)	Meta-Llama-3.1-8B-Instruct	0.1628	0.0177	0.2246	0.1234	0.7967
	Meta-Llama-3.1-70B-Instruct	0.1500	0.0204	0.2345	0.1376	0.8206
	Mixtral-8x7B-Instruct-v0.1	0.1236	0.0204	0.2171	0.1302	0.8196
	Meditron-70B	0.1516	0.0193	0.2082	0.1297	0.8062

Table 10: Impact of matching vanilla LLM prompts to RAG prompt format. LLM-only vs RAG: Mean BLEU, ROUGE, and BERTScore values. LLM-only (prompt matches RAG) were computed on only a randomly-selected 10% subset of questions from the MedQuAD NIDDK QA dataset (n=83), whereas LLM-only (original vanilla prompt) and RAG (original RAG prompt) were computed over all n=828 questions in MedQuAD NIDDK. **Bold** and underline denote the highest and second highest overall, respectively.

Appendix E. Extended Results: Qualitative Evaluation

E.1. Clinical Annotation Instructions: Head-to-Head Scoring

A clinician performed annotation — they were provided with the following instructions for scoring the overall quality of vanilla LLM-generated vs. RAG responses for a subset of 66 responses to 22 questions in the MedQuAD NIDDK dataset.

Instructions:

1. For each row, fill in the "WINNER CHOICE" column with A, B, or TIE. A indicates you believe the answer to A is overall better, B indicates you believe the answer to B is overall better, and TIE indicates you believe both A and B answers are equally good/bad in quality. If you'd like you may use the MedQuAD reference answer column to determine which is the better answer.

2. Edge cases: If one answer refuses to provide a response, but the other answers correctly, then award the winner to the correct one. Otherwise, if both refuse, you may choose one or say that it is a tie.

E.2. Baseline: MedQuAD Reference Answers Evaluated by LLM-as-a-judge

Approach	Mean Score	σ	min	max
Reference	8.66	0.89	0.71	10.0

Table 11: NIDDK MedQuAD Reference Answers: Composite qualitative score summary over 8 criteria via LLM judge (0-10 scale normalized), where 0 is least desirable and 10 is most desirable. Composite judge scores were calculated for 750/828 reference answers (LLM judge yielded valid responses for 750 answers).

To calibrate the scores from our LLM judge for LLM-only and RAG answers, we used the same LLM judge (`Llama3.1-70B-Instruct`) to score the 828 reference answers within the MedQuAD NIDDK dataset. We hypothesized that these reference answers would be assigned higher composite scores (0-10) and higher scores for each of the 8 qualitative criteria relative to answers generated from RAG or LLM-only approaches (Table 3). To avoid over-weighting certain questions relative to LLM-only and RAG evaluations, we used only the first reference answer for each of the 828 unique questions.

The LLM judge provided valid scores across all 8 criteria for 750 of the 828 reference answers (Table 11). Including the additional 78 incomplete LLM judge scores of reference answers yielded a similar mean score of 8.56 (σ = 0.94), similar to the score 8.66 shown in Table 11. Comparing mean composite scores of reference answers in Table 11 relative to composite scores of answers generated by LLM-only and RAG approaches (Table 12), MedQuAD NIDDK reference answers were awarded a *slightly lower* composite mean score on average (0.56) than the best-performing LLM-only and RAG approaches (`Llama-3.1-70B`: 8.58). This finding should be interpreted with caution: It does not indicate that `Llama-3.1-70B` generates more valid answers on average relative to the NIDDK website – instead, the high composite score of the reference answers underscores the validity of the reference MedQuAD answers according to the LLM judge, suggesting that the judge is generally well calibrated at categorizing ground truth answers as acceptable.

Model	**Approach**	**Mean Score (σ)**	
# Questions Evaluated		582	828
Llama 3.1-8B	LLM-only	<u>8.62</u> (1.30)	<u>8.44</u> (1.23)
	RAG	7.06 (1.81)	7.07 (1.60)
Llama 3.1-70B	LLM-only	**8.87** (0.85)	**8.58** (1.03)
	RAG	7.55 (1.66)	7.60 (1.58)
Mixtral 8x7B	LLM-only	7.94 (1.59)	7.81 (1.58)
	RAG	7.63 (1.71)	7.70 (1.64)
Meditron-70B	LLM-only	6.31 (2.83)	6.27 (2.87)
	RAG	5.64 (3.22)	5.88 (3.00)

Table 12: Non-missing evaluations from LLM-as-a-judge: There were consistent findings between total vs. intersection of non-missing LLM judge questions. Qualitative scoring over 8 criteria via LLM-judge. Composite score summary (0-10 scale normalized), where 0 is least desirable and 10 is most desirable. **Bold** and <u>underline</u> denote the highest and second highest overall, respectively.

E.3. Downstream Impact of Retrieval Quality on RAG Composite LLM Judge Scores

As shown in Appendix C, suboptimal retrieval performance was observed for the general purpose system applied here. Due to the unexpected negative finding – that LLM-only generally outperformed RAG QA according to both quantitative and qualitative evaluation – we explored the extent to which retrieval performance of the RAG system was associated with composite LLM judge scores

Figure 5 shows composite LLM judge score relative to precision@5 across four LLMs applied in the RAG pipeline. Across all LLMs, there was only mild correlation between composite LLM judge score and precision@5. The highest correlation (0.28) was observed for `Llama-3.1-8B-Instruct`, the smallest LLM studied here. Larger LLMs like `Llama-3.1-70B-Instruct` showed even weaker correlation (0.08) between retrieval performance and composite LLM judge score. Larger LLMs like `Mistral-8x7B-Instruct` and `Llama-3.1-70B-Instruct` were sometimes able to produce responses with high composite scores even when retrieval failed (precision@5 = 0). However, `Meditron-70B` showed unreliable composite scores when retrieval failed, generating many low-scoring answers when precision@5 was 0.

Interestingly, when some (but not all) of the context was relevant to the ground truth answer (low precision@5, high MRR), the general purpose RAG systems applied here consistently produced poor quality answers (Figures 5, 6). This trend is particularly apparent when the RAG pipeline applied `Llama3.1-8B-Instruct` and `Meditron-70B`. This can be seen even for the best performing RAG configuration using `Llama3.1-70B-Instruct`. **This suggests that LLMs were consistently distracted by irrelevant information in the retrieved context when *part but not all of the context was relevant to answering the question*.**

Figure 4: Qualitative LLM-judge Evaluation: Percentage of generated responses flagged as problematic according to each of the 8 criteria. LLM-judge evaluations were compared with a clinician to ensure consistency. Calculated across all n=828 responses.

These findings – that even with successful retrieval, a general-purpose RAG system often produces unreliable medical answers – highlight the need for medical-specific RAG approaches to support medical question-answering systems.

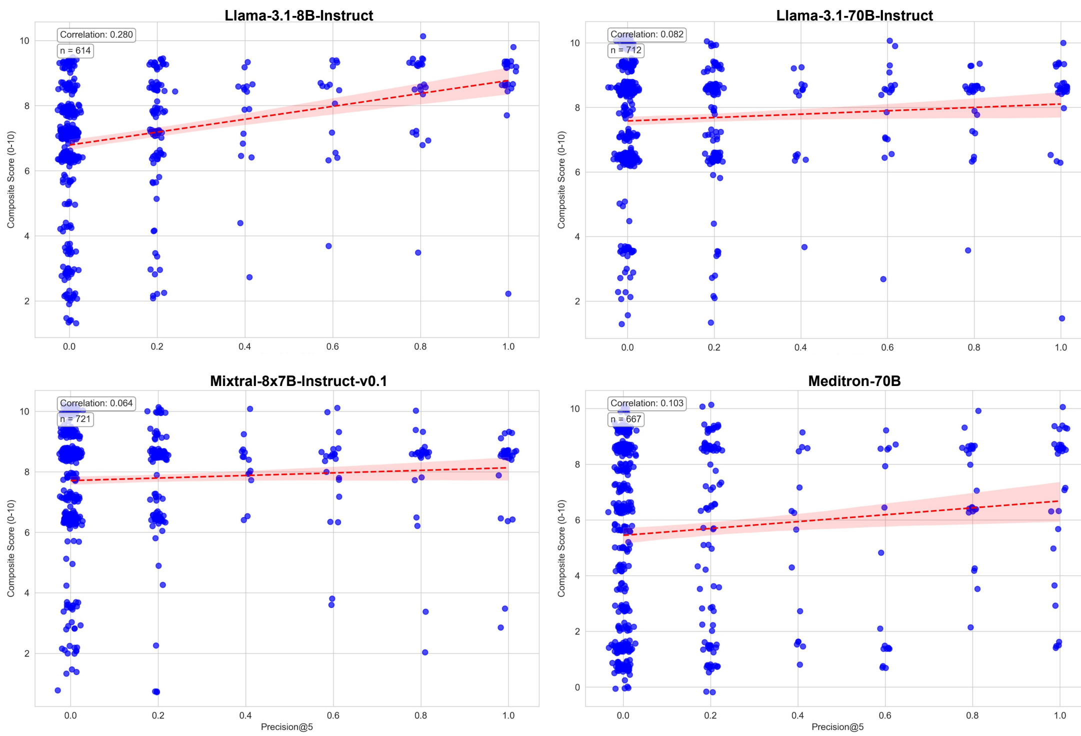

Figure 5: Precision@5 vs. Composite Qualitative Score. RAG retrieval performance mildly impacts downstream composite LLM judge scores. Precision@5 vs. composite qualitative score is shown for all non-missing composite judge scores, shown for full RAG system using four LLMs.

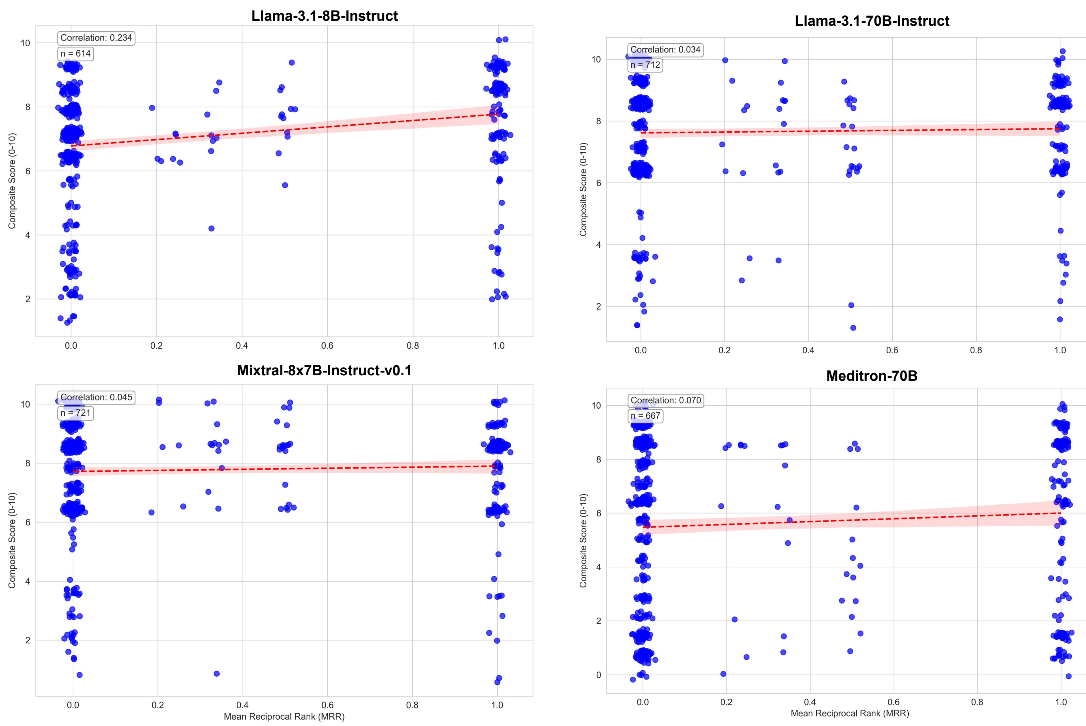

Figure 6: Mean reciprocal rank (MRR) vs. Composite Qualitative Score. RAG Retrieval Performance Mildly Impacts Downstream Composite LLM Judge Scores. Precision@5 vs. composite qualitative score is shown for all non-missing composite judge scores, shown for full RAG system using four LLMs.

Question 1: "What are the treatments for Diabetes, Heart Disease, and Stroke ?"

Llama 70B Generated Answer (Vanilla): "As a doctor, I'd like to outline the treatments for Diabetes, Heart Disease, and Stroke. Please note that these are general overviews, and individual treatment plans may vary depending on the patient's specific condition and medical history.**Diabetes:**1. **Lifestyle modifications**: Healthy diet, regular exercise, weight management, and stress reduction.2. **Oral medications**: Metformin, sulfonylureas, meglitinides, thiazolidinediones, and SGLT-2 inhibitors to control blood sugar levels.3. **Insulin therapy**: Injecting insulin to regulate blood sugar levels, often used in combination with oral medications.4. **Monitoring**: Regular blood sugar checks, HbA1c tests, and kidney function monitoring.**Heart Disease:**1. **Lifestyle modifications**: Healthy diet, regular exercise, stress reduction, and smoking cessation.2. **Medications**:* Beta blockers to slow heart rate and reduce blood pressure.* ACE inhibitors or ARBs to lower blood pressure and reduce strain on the heart.* Statins to lower cholesterol levels.* Antiplatelet agents to prevent blood clots.3. **Surgical interventions**:* Angioplasty and stenting to open blocked coronary arteries.* Coronary artery bypass grafting (CABG) to bypass blocked arteries.* Heart transplantation in severe cases.4. **Cardiac rehabilitation**: A supervised program to improve cardiovascular health and reduce risk factors.**Stroke:**1. **Emergency treatment**: Tissue plasminogen activator (tPA) to dissolve blood clots, if administered within 4.5 hours of stroke onset.2. **Medications**:* Anticoagulants to prevent blood clots.* Antiplatelet agents to prevent blood clots.* Statins to lower cholesterol levels.* Blood pressure medications to control hypertension.3. **Rehabilitation**: Physical, occupational, and speech therapy to regain lost functions and improve quality of life.4. **Surgical interventions**:* Carotid endarterectomy to remove plaque buildup in the carotid arteries.* Aneurysm clipping or coiling to treat aneurysms.It's essential to work closely with your healthcare provider to develop a personalized treatment plan, as each individual's condition and medical history are unique."

Llama 70B Generated Answer (RAG): "The provided reference documents do not contain comprehensive information on the treatments for Diabetes, Heart Disease, and Stroke. However, they do provide some information on managing diabetes and preventing related complications. **Managing Diabetes:** * Reaching and maintaining a target A1C level (below 7% for many people with diabetes) to protect the kidneys (Document 1) * Controlling blood pressure to prevent heart attack, stroke, and kidney disease (Document 1) * Working with a healthcare team to use blood glucose results to guide decisions about food, physical activity, and medicines (Document 1) **Preventing Diabetes-Related Complications:** * Protecting the kidneys by keeping blood glucose under control (Document 1) * Controlling blood pressure to prevent heart disease and stroke (Document 1) * Managing low blood glucose (hypoglycemia) (Documents 2 and 4) **Medicines and Treatments:** * Insulin, medicines, and other diabetes treatments (mentioned in Document 3, but not elaborated upon) * Statins, which can slightly increase the chance of developing diabetes but help protect against heart disease and stroke (Document 5) It is essential to consult a healthcare professional for personalized guidance on managing diabetes and preventing related complications. The provided documents are not exhaustive, and a comprehensive treatment plan would require additional information and expertise."

MedQuAD Reference Answer: "You can keep track of the ABCs of diabetes to make sure your treatment is working. Talk with your health care provider about the best targets for you. A stands for A1C (a test that measures blood glucose control). Have an A1C test at least twice a year. It shows your average blood glucose level over the past 3 months. Talk with your doctor about whether you should check your blood glucose at home and how to do it. A1C target Below 7 percent, unless your doctor sets a different target Blood glucose targets Before meals 90 to 130 mg/dL 1 to 2 hours after the start of a meal Less than 180 mg/dL B is for blood pressure. Have it checked at every office visit. Blood pressure target Below 140/80 mm Hg, unless your doctor sets a different target C is for cholesterol. Have it checked at least once a year. Blood fat (cholesterol) targets LDL (bad) cholesterol Under 100 mg/dL Triglycerides Under 150 mg/dL HDL (good) cholesterol For men: above 40 mg/dL For women: above 50 mg/dL Control of the ABCs of diabetes can reduce your risk for heart disease and stroke. If your blood glucose, blood pressure, and cholesterol levels aren't on target, ask your doctor what changes in diet, activity, and medications can help you reach these goals."

Question 2: "What are the symptoms of High Blood Pressure and Kidney Disease ?"

Llama 70B Generated Answer (Vanilla): "High Blood Pressure (Hypertension) and Kidney Disease can have similar symptoms, and sometimes, they can be asymptomatic. Here are some common symptoms to look out for:**High Blood Pressure (Hypertension) Symptoms:**1. **Headaches**: Frequent or severe headaches can be a sign of high blood pressure.2. **Dizziness or Lightheadedness**: High blood pressure can cause dizziness or lightheadedness due to decreased blood flow to the brain.3. **Nosebleeds**: High blood pressure can cause blood vessels in the nose to become fragile and bleed.4. **Shortness of Breath**: High blood pressure can lead to fluid buildup in the lungs, causing shortness of breath.5. **Chest Pain**: High blood pressure can cause chest pain or discomfort, especially if it's related to a heart attack or stroke.6. **Fatigue**: High blood pressure can cause fatigue, as the heart has to work harder to pump blood.**Kidney Disease Symptoms:**1. **Swollen Feet, Ankles, or Hands**: Kidney disease can cause fluid buildup, leading to swelling in these areas.2. **Fatigue**: Kidney disease can cause fatigue, as the kidneys are not filtering waste effectively.3. **Urination Changes**: Kidney disease can cause changes in urination, such as: Increased frequency or urgency Blood in the urine Foamy or bubbly urine Pale or dark urine4. **Nausea and Vomiting**: Kidney disease can cause nausea and vomiting, especially if waste builds up in the blood.5. **Itching**: Kidney disease can cause itching, especially on the skin.6. **Loss of Appetite**: Kidney disease can cause a decrease in appetite.**Common Symptoms for Both Conditions:**1. **Edema** (swelling): Both high blood pressure and kidney disease can cause fluid buildup, leading to swelling in the feet, ankles, or hands.2. **Fatigue**: Both conditions can cause fatigue, as the body is not functioning optimally.**Important Note:** Many people with high blood pressure or kidney disease may not experience any symptoms at all, especially in the early stages. Regular check-ups with your doctor and monitoring your blood pressure and kidney function can help detect these conditions early on.If you're experiencing any of these symptoms, it's essential to consult with your doctor to determine the underlying..."

Llama 70B Generated Answer (RAG): "High Blood Pressure and Kidney Disease can have various symptoms depending on the underlying cause. According to the provided reference documents, the following symptoms are associated with High Blood Pressure and Kidney Disease:**Diabetic Kidney Disease:*** High blood glucose can damage the blood vessels in the kidneys, leading to kidney damage.* High blood pressure can also damage the kidneys.* Symptoms may not be apparent in the early stages, but as the disease progresses, the following symptoms may occur: Fatigue Swelling in the legs, ankles, and feet Weight gain High blood pressure Protein in the urine**Nephrotic Syndrome:*** A combination of symptoms that can occur due to different causes, including kidney diseases.* Symptoms may include: Swelling in the face, legs, and feet Weight gain Fatigue Loss of appetite Nausea and vomiting High blood pressure**Primary Hyperparathyroidism:*** High blood calcium levels can cause kidney stones and other complications.* Symptoms may include: Muscle weakness Fatigue Depression Aches and pains in bones and joints Loss of appetite Nausea and vomiting Constipation Confusion Increased thirst and urinationIt is essential to note that some of these symptoms may not be specific to High Blood Pressure and Kidney Disease and can be associated with other conditions. A healthcare professional should be consulted for an accurate diagnosis and treatment plan."

MedQuAD Reference Answer: "Most people with high blood pressure do not have symptoms. In rare cases, high blood pressure can cause headaches. Kidney disease also does not have symptoms in the early stages. A person may have swelling called edema, which happens when the kidneys cannot get rid of extra fluid and salt. Edema can occur in the legs, feet, or ankles and less often in the hands or face. Once kidney function decreases further, symptoms can include - appetite loss - nausea - vomiting - drowsiness or feeling tired - trouble concentrating - sleep problems - increased or decreased urination - generalized itching or numbness - dry skin - headaches - weight loss - darkened skin - muscle cramps - shortness of breath - chest pain"

Figure 7: Example of generated answers for vanilla and RAG settings on MedQuAD NIDDK. For question 1, both responses received identical LLM-judge scores across the 8 qualitative criteria: (Aligned, None, Minor omissions, None, Low, No, High, No errors). For question 2, LLM-judge scores were: vanilla (Aligned, None, Minor omissions, None, Low, No, High, No errors), and RAG (Aligned, Major issues, Major omissions, Severe, High, Yes, Low, No errors).

Question 1: "What is (are) Hypothyroidism ?"

Llama 8B Generated Answer (RAG): "Hypothyroidism, also called underactive thyroid, is when the thyroid gland doesn't make enough thyroid hormones to meet your body's needs. The thyroid is a small, butterfly-shaped gland in the front of your neck. Thyroid hormones control the way your body uses energy, so they affect nearly every organ in your body, even the way your heart beats. Without enough thyroid hormones, many of your body's functions slow down.
Hypothyroidism has several causes, including Hashimoto's disease thyroiditis, or inflammation of the thyroid congenital hypothyroidism, or hypothyroidism that is present at birth surgical removal of part or all of the thyroid radiation treatment of the thyroid some medicines Less often, hypothyroidism is caused by too much or too little iodine in the diet or by disorders of the pituitary gland or hypothalamus.
Hypothyroidism has many symptoms that can vary from person to person. Some common symptoms of hypothyroidism include fatigue weight gain trouble tolerating cold joint and muscle pain dry skin or dry, thinning hair heavy or irregular menstrual periods or fertility problems slowed heart rate depression Dry, thinning hair is one of many symptoms that might indicate hypothyroidism.
Hypothyroidism can contribute to high cholesterol. If you have high cholesterol, you should get tested for hypothyroidism. Rarely, severe untreated hypothyroidism may lead to myxedema coma, an extreme form of hypothyroidism in which the body's functions slow to a life-threatening point. Myxedema coma requires immediate medical treatment.
Hypothyroidism is diagnosed by doctors through various methods, but the provided documents do not contain enough information to provide a comprehensive answer on how doctors diagnose hypothyroidism.
Hypothyroidism is treated by doctors using thyroid medicines, which are safe to take during pregnancy. Many women taking thyroid hormone medicine need a higher dose during pregnancy, so contact your doctor right away if you find out you're pregnant.
The provided documents do not contain enough information to provide a comprehensive answer on how eating, diet, and nutrition affect hypothyroidism.
Clinical trials for hypothyroidism are looking for participants to study various aspects of the disease, such as understanding how the disease progresses, its clinical presentation, and genetics investigating how effective and safe levothyroxine is for people with chronic kidney disease.
The provided documents do not contain enough information to provide a comprehensive answer on how common..."

Doc 1: "Title: Hypothyroidism (Underactive Thyroid) - NIDDK
How common is hypothyroidism? Who is more likely to develop hypothyroidism? Is hypothyroidism during pregnancy a problem? What are the complications of hypothyroidism? What are the symptoms of hypothyroidism? What causes hypothyroidism? How do doctors diagnose hypothyroidism? How do doctors treat hypothyroidism? How does eating, diet, and nutrition affect hypothyroidism? Clinical Trials for Hypothyroidism What is hypothyroidism? Hypothyroidism, also called underactive thyroid, is when the thyroid gland doesn't make enough thyroid hormones to meet your body's needs. The thyroid is a small, butterfly-shaped gland in the front of your neck. Thyroid hormones control the way your body uses energy, so they affect nearly every organ in your body, even the way your heart beats. Without enough thyroid hormones, many of your body's functions slow down. The thyroid is a small gland in your neck that makes thyroid hormones. How common is hypothyroidism?"

Doc 2: "Title: Hypothyroidism (Underactive Thyroid) - NIDDK
The trials look to find new ways to prevent, detect, or treat disease and improve quality of life. What are clinical trials for hypothyroidism? Clinical trials~2014and other types of clinical studies ~2014are part of medical research and involve people like you. When you volunteer to take part in a clinical study, you help doctors and researchers learn more about disease and improve health care for people in the future. Researchers are studying many aspects of hypothyroidism, such as understanding how the disease progresses, its clinical presentation, and genetics investigating how effective and safe levothyroxine is for people with chronic kidney disease Find out if clinical studies are right for you . Watch a video of NIDDK Director Dr. Griffin P. Rodgers explaining the importance of participating in clinical trials. What clinical studies for hypothyroidism are looking for participants?"

Doc 3: "Title: Hypothyroidism (Underactive Thyroid) - NIDDK.
Your thyroid uses iodine to make thyroid hormones . However, if you have Hashimoto's disease or other types of autoimmune thyroid disorders, you may be sensitive to iodine's harmful side effects. Eating foods that have large amounts of iodine such as kelp, dulse, or other kinds of seaweed may cause or worsen hypothyroidism. Taking iodine supplements can have the same effect. Talk with members of your health care team about what foods to limit or avoid if you take iodine supplements about any cough syrups you take because they may contain iodine If you are pregnant, you need more iodine because the baby gets iodine from your diet. Talk with your doctor about how much iodine you need. Clinical Trials for Hypothyroidism The NIDDK conducts and supports clinical trials in many diseases and conditions, including endocrine diseases."

Doc 4: "Title: Hypothyroidism (Underactive Thyroid) - NIDDK
Some common symptoms of hypothyroidism include fatigue weight gain trouble tolerating cold joint and muscle pain dry skin or dry, thinning hair heavy or irregular menstrual periods or fertility problems slowed heart rate depression Dry, thinning hair is one of many symptoms that might indicate hypothyroidism. Because hypothyroidism develops slowly, you may not notice symptoms of the disease for months or even years. Many of these symptoms, especially fatigue and weight gain, are common and do not necessarily mean you have a thyroid problem. What causes hypothyroidism? Hypothyroidism has several causes, including Hashimoto's disease thyroiditis, or inflammation of the thyroid congenital hypothyroidism, or hypothyroidism that is present at birth surgical removal of part or all of the thyroid radiation treatment of the thyroid some medicines Less often, hypothyroidism is caused by too much or too little iodine in the diet or by disorders of the pituitary gland or hypothalamus."

Doc 5: "Title: Hypothyroidism (Underactive Thyroid) - NIDDK
However, thyroid medicines can help prevent problems and are safe to take during pregnancy. Many women taking thyroid hormone medicine need a higher dose during pregnancy, so contact your doctor right away if you find out you're pregnant. What are the complications of hypothyroidism? Hypothyroidism can contribute to high cholesterol . If you have high cholesterol, you should get tested for hypothyroidism. Rarely, severe untreated hypothyroidism may lead to myxedema coma, an extreme form of hypothyroidism in which the body's functions slow to a life-threatening point. Myxedema coma requires immediate medical treatment. What are the symptoms of hypothyroidism? Hypothyroidism has many symptoms that can vary from person to person."

MedQuAD Reference Answer: "Hypothyroidism is a disorder that occurs when the thyroid gland does not make enough thyroid hormone to meet the bodys needs. Thyroid hormone regulates metabolismthe way the body uses energyand affects nearly every organ in the body. Without enough thyroid hormone, many of the bodys functions slow down. About 4.6 percent of the U.S. population age 12 and older has hypothyroidism."

Figure 8: Example of retrieved passages and generated answers for RAG setting on MedQuAD NIDDK.