ClinIQ: A Medical Expert-Level Question-Answering System

Adria Pop and **Konstantin Wehmeyer** and **Lorena Raichle**
University of St. Gallen
{adria.pop, matthiaskonstantin.wehmeyer, lorena.raichle}@student.unisg.ch

GitHub: https://github.com/LorenaRaichle/ClinIQ

Supervised by: Prof. Dr. Siegfried Handschuh & Götz-Henrik Wiegand

Abstract

Recent advancements in Large Language Models (LLMs) have demonstrated strong performance in open-domain question answering (ODQA); however, challenges remain with the effective application of these models to the medical domain, particularly for complex Multi Hop reasoning and coverage of niche medical topics. Additionally, the risk of hallucinations and reliance on outdated information during model pretraining poses a concern in a rapidly evolving field like medicine. Motivated by this, we address a challenge proposed by the University of Maryland (ClinIQLink, 2025) to evaluate the ability of generative models to produce factually accurate medical information.

Our main contribution lies in designing and evaluating a hybrid architecture Q&A system, which approaches the task by fine-tuning the open-source **DeepSeek Coder-7B** model within a Retrieval-Augmented Generation (RAG) framework. The model is trained on a curated set of nine medical question-answering datasets to enhance performance across a range of question types, including *Short Answer*, *True/False*, *Multi Hop*, and *Multiple Choice* formats. We evaluate 5 different architecture variants using a detailed evaluation suite tailored to each question type, assessing reasoning flow, semantic coherence, and factual accuracy.

Our architecture significantly outperforms the baseline across all question formats. For example, accuracy on *Multiple Choice* questions improved from **25%** to **47%**, while the ROUGE-L score for *Multi Hop* generated answers increased from **7%** to **40%**. These results demonstrate the complementary strengths of fine-tuning and retrieval-based augmentation, and offer insights into architecture-specific trade-offs when applying LLMs to complex, knowledge-intensive tasks.

1 Introduction

In high-stakes domains like medicine, the role of AI remains controversial, particularly in diagnostic decision-making, where accountability and factual accuracy are critical (Miller, 2019). Moreover, studies indicate that consumers are reluctant to embrace AI in healthcare due to concerns about its inability to adapt to patient individuality and clinical nuances (Longoni et al., 2019). Nonetheless, there is growing reliance on online medical resources, many of which are now powered by LLMs. However, these models often lack consistency in quality and vary significantly in health literacy. Although access to accurate and understandable health information is crucial for patient empowerment, the complex terminology used by LLMs often creates a barrier to comprehension (Zhou et al., 2025).
Advancements in LLM research have demonstrated promising capabilities in medical problem solving. For example, DeepSeek-R1 was shown to produce high-quality, readable output in a comparative evaluation involving ChatGPT-4o as well as other models (Zhou et al., 2025). However, commercial models remain closed-source, raising concerns about transparency, data privacy, and reproducibility in medical applications. Meanwhile, smaller open-source models often lack the parameter capacity necessary for sophisticated tasks like Multi Hop reasoning (Kim et al., 2024).

These limitations have intensified the focus on the concept of *Trustworthy Generative AI*, particularly in the medical domain. One of the most pressing concerns is the phenomenon of hallucinations (Huang et al., 2023, 2), which poses serious risks when arising in health-related contexts. As Zhou et al.(Zhou et al., 2025) highlight in their comparative analysis of AI-generated patient education materials, not only readability but also the factual quality of outputs must be systematically evaluated to ensure patient safety and trust.

Ensuring reliability, factual consistency, and up-to-date knowledge in medical applications requires more than pre-training on general-purpose data. In research literature, two complementary strategies for adapting pre-trained LLMs to domain-specific tasks are commonly discussed: Supervised Fine-Tuning (SFT) and Retrieval Augmented Generation (RAG) (Ovadia et al., 2024, 1). Fine-tuning involves an extended training phase in which the model's weights are updated using domain-relevant data to better align with the target task. In contrast, RAG provides external knowledge dynamically at inference time without modifying the model weights. It achieves this by incorporating a separate retrieval system that supplies relevant context based on the input query (Ovadia et al., 2024, 1). RAG architectures are often regarded as a form of in-context learning, as they "combine pre-trained parametric and non-parametric memory for language generation" Lewis et al. (2021). This hybrid approach outperforms in the evaluation on *Open-domain Question Answering* tasks for traditional seq2seq baseline models (Lewis et al., 2021). In clinical domains, augmenting LLMs with an information retrieval component has shown to effectively mitigate hallucinations (Jiang et al., 2024).
Building on promising results reported in prior work, particularly in knowledge-intensive and high-stakes domains, we adopt a hybrid approach that combines a fine-tuned model within a RAG pipeline. This design is especially well-suited for medical applications, where access to up-to-date, verifiable knowledge is essential. While MKRAG (Shi et al., 2024a) demonstrates that RAG alone can effectively improve language model performance in medical Q&A without requiring additional fine-tuning, we argue that integrating RAG with SFT yields complementary advantages. Specifically, SFT allows the model to internalize domain-specific question formats and reasoning strategies, while RAG dynamically injects accurate and contextually relevant medical knowledge at inference time. This combination is particularly valuable in clinical domains, where both structured reasoning and factual completeness are critical for trustworthy AI-assisted decision-making.

2 Methodology

2.1 Data

This project utilized a wide range of medical question-answer datasets, structured across four main question types: *Multi Hop **(MH)*** A.1.4, *Multiple Choice **(MC)*** A.1.1, *Short Answer **SA*** A.1.3, and *True/False **(TF)*** A.1.2. These datasets were chosen for their relevance to clinical reasoning and factual informative content to provide a comprehensive foundation to develop, fine-tune, and evaluate our models.

Multi Hop Question Dataset

- **FreedomIntelligence/medical-o1-reasoning-SFT**
  This dataset consists of 19,700 Multi Hop reasoning-based open-end question-answer pairs with intermediate thought chains to support model training for complex reasoning tasks.
  Source: (Chen et al., 2024) https://huggingface.co/datasets/FreedomIntelligence/medical-o1-reasoning-SFT

Multiple Choice Question Datasets

- **openlifescienceai/MedMCQA**
  With over 193,000 entries, MedMCQA is a comprehensive dataset sourced from medical entrance exams. It provides Multiple Choice questions across disciplines such as anatomy, physiology, and pharmacology.
  Source: (Pal et al., 2022) https://huggingface.co/datasets/openlifescienceai/medmcqa

- **stellalisy/mediQ**
  This dataset includes 2,540 clinical Multiple Choice questions (additionally context, patient metadata, etc.).
  *Source:*https://huggingface.co/datasets/stellalisy/mediQ

- **bigbio/MedQA**
  Containing 12,723 board-style Multiple Choice questions, collected from the professional medical board exams.
  Source: (Jin et al., 2021) https://huggingface.co/datasets/bigbio/med_qa

- **UCSC-VLAA/MedReason**
  With 32,700 examples, this dataset covers a broad spectrum of medical topics in the form of Multiple Choice.
  Source: (Wu et al., 2025) https://huggingface.co/datasets/UCSC-VLAA/MedReason

Short Answer Question Datasets

- **Ajayaadhi/Medical-QA**
  This dataset contains 49,900 Short Answer questions covering diverse medical subjects.
  Source: https://huggingface.co/datasets/Ajayaadhi/Medical-QA

- **Comprehensive Medical Q&A Dataset**
  With 14,979 questions, this Kaggle-hosted dataset covers infectious diseases, pharmacology, and other core subjects. Answers provided are from doctors, nurses and pharmacists
  Source: https://www.kaggle.com/datasets/thedevastator/comprehensive-medical-q-a-dataset

- **HPAI-BSC/OpenMedQA**
  This smaller set of 1,270 Short Answer questions.
  Source: (Bayarri Planas, J., n. d.). https://huggingface.co/datasets/HPAI-BSC/OpenMedQA

True/False Question Dataset

- **qiaojin/PubMedQA**
  PubMedQA provides 211,000 Yes/No questions derived from PubMed abstracts with scientific reasoning and biomedical knowledge grounded in real research. We changed the labels Yes/No labels into True/False to ensure a balance between entries of different question types.
  Source: https://huggingface.co/datasets/qiaojin/PubMedQA

Knowledge Base for Retrieval

- **MedRAG/pubmed**
  This dataset comprises curated biomedical literature from PubMed, used as the retrieval corpus in our RAG framework. It enables the system to acquire relevant scientific knowledge from abstract and corpus to support answer generation.
  Source: (Xiong et al., 2024) https://huggingface.co/datasets/MedRAG/pubmed

2.2 General dataset preprocessing

We iterated through the datasets to remove duplicate and empty question-answer pairs to ensure uniqueness and integrity for successful training. The aggregated dataset, as a result, consists of 515,876 questions collected across nine different datasets, each with distinct layouts and formats in accordance with their question type. After preprocessing and normalization, we successfully formatted 217,327 Multiple Choice, 67,576 Short Answer, 211,269 True/False, and 19,704 Multi Hop questions.

To address class imbalance during training, our preprocessing focused on resolving two key issues. First, the True/False dataset exhibited a significant skew toward the "True" class. To mitigate this, we reduced the number of "True" entries to achieve a more balanced distribution between "True" and "False" responses. Second, the dataset contained a disproportionately high number of Multiple Choice questions. We downsampled this category to ensure a more balanced representation across all question types. These adjustments aimed to help the model learn more effectively from underrepresented classes and avoid overfitting to the majority categories.

2.3 RAG-specific preprocessing of dataset

Our knowledge store integrates two primary content sources: (1) a curated multi-source training dataset and (2) a filtered subset of PubMed abstracts. To enable robust retrieval and traceability within this hybrid RAG pipeline, we designed a tailored preprocessing strategy for both sources. For the training set, we enriched each question with unique, type-specific IDs (e.g., `mc_0`, `tf_1`) and embedded original source metadata to facilitate domain-level filtering and evaluation. In the case of Multiple Choice questions, we normalized answer formats by converting single-character labels (e.g., `"B"`) into their full-text equivalents (e.g., `"2-3 years after eruption"`), and expanded ambiguous entries such as `"None of the above"` to explicitly include all distractor options. These steps ensured clean, semantically complete entries for downstream retrieval. Further, we merged and preprocessed 2.5 million PubMed abstracts by assigning consistent document IDs (e.g., `pubmed_42`), applying domain-specific text cleaning, and preparing structured formats for embedding and topic modeling that will be detailed in Section 2.7. Together, these preprocessing steps formed the foundation for populating a traceable, well-structured vector store, critical for RAG-based document retrieval and evaluation.

2.4 System Architecture and Modeling Strategies

To contextualize the modeling strategies described in subsequent sections, we provide a high-level architectural overview in Figure 1. This diagram illustrates the interaction between our hybrid retriever, fine-tuned DeepSeek model, and the structured knowledge sources populated in our knowledge base (Pinecone). Since the retriever returns only document identifiers, the full semantic context is subsequently reconstructed from preprocessed content stored in Google Drive.

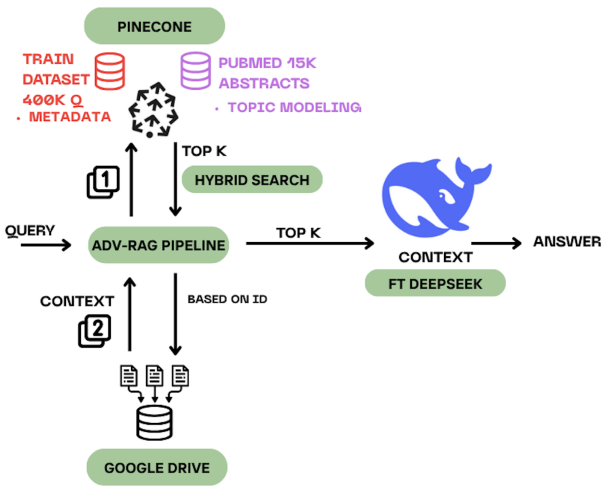

Figure 1: Advanced RAG Pipeline integrating training and PubMed knowledge via hybrid search

To tackle the ClinIQ challenge, we implemented and evaluated five distinct architecture variants that progressively combine fine-tuning and RAG. Table 1 summarizes the architectural configurations explored in this work. Each variant is integrated as a modular component within our main architecture pipeline (cf. Figure 1), allowing for consistent retrieval, context injection, and downstream evaluation across question types.

2.5 Basemodel

As our baseline model we used the *Deepseek-coder-7b-instruct-v1.5* model, which was pretrained on 2T tokens of programming and code data with a window size of 4K and fine-tuned on 2B tokens of instruction data. This instruction fine-tuning allowed the model to answer our questions in the correct format. With a relatively small parameter count, this model achieved good performance on different benchmarks (see Figure 8 in the Appendix). We experimented with other models, most notably the base model counterpart. Some literature suggests that base models outperform their instructed counterparts in RAG tasks by 20% on average (Cuconasu et al., 2024). However, we could not support this claim, as most of the answers of the base model were not in the correct format. Therefore, we used the instruction model for the baseline and for the subsequent RAG and fine-tuning architectures.

We used the guidelines for DeepSeek models, such as using clear and specific prompts, avoiding system prompts, and avoiding few-shot prompts. We experimented with different prompts and found that the ones listed in *prompt_utils.py* and used in this file worked best to achieve the desired output formats. (together.ai, 2025)

For the pipeline parameters, we used the setup in Table 2 for the different question types. Since both TF and MC questions only need to produce very Short Answers, we set the *max_new_tokens* to a lower number than in the SA and MH questions. Similarly, we set the *temperature* to 0.1 for TF and MC because we did not want any creativity in the answer, but rather the most probable prediction. For SA and MH, we set the temperature to 0.7, as was recommended, for example, in the OpenAI documentation and in several blog posts. While we experimented with the *temperature*, we set the *top_k* to 50 and did not change this as suggested by OpenAI: "We generally recommend altering this or top_k but not both" (OpenAI, 2025).

However, we did not conduct extensive prompt engineering and parameter optimization due to two reasons: Firstly, we wanted to focus on implementing techniques like RAG or LoRA fine-tuning and compute units were limited. Secondly, our literature review showed that pipeline parameters are crucial to consider, but the effect is not significant for improving LLM problem-solving performance (Renze and Guven, 2024).

The above stated parameter setup was used for all following experiments using the DeepSeek model.

2.6 Fine-tuning

For the next step we fine-tuned the *Deepseek-coder-7b-instruct-v1.5* model using a fully supervised approach, where the model learns to generate the correct answer directly from the provided prompt.

We used LoRA (Low-Rank Adaptation) as a technique for fine-tuning the base model efficiently. Instead of updating all model weights, LoRA in-

Table 1: Overview of the five architectural modeling strategies evaluated in this work

Approach	Explanation
1 Baseline	Baseline model `deepseek-coder-7b-instruct-v1.5`
2 FT (Fine-tuning)	LoRA fine-tuned `deepseek-coder-7b-instruct-v1.5`
3 RAG	(Naive) RAG pipeline with the baseline model (imbalanced INDEX 1)
4 RAG plus FT	(Advanced) RAG pipeline with the fine-tuned model (imbalanced INDEX 1)
5 RAG plus FT balanced	(Advanced) RAG pipeline with fine-tuned model (balanced INDEX 2)

	MC	TF	SA	MH
max_new_tokens	20	20	100	200
temperature	0.1	0.1	0.7	0.7

Table 2: Pipeline parameters for MC, TF, SA, and MH

jects trainable low-rank matrices into specific layers (such as attention projections), drastically reducing the number of trainable parameters.
By decomposing weight updates into low-rank matrices $A \in R^{d \times r}$ and $B \in R^{r \times k}$, LoRA approximates the update as:

$$\Delta W = AB$$

where $r \ll d, k$. This allows models to adapt to new tasks with minimal computational cost and memory usage - usually training less than 1% of the total parameters (Hu et al., 2021). In our case the baseline model had 6.9 billion trainable parameters. With our LoRA setup we could reduce this number of trainable parameters to just 3.9 million, less than 0.1% of the original parameter count.
In practical terms, we concatenated the question prompt with the correct answer option into a single string (*prompt_n_answer*) and trained the model on this combined sequence. Every token in the input was used as a label, enabling the model to learn from the entire prompt–answer structure without any masking. The rationale behind this approach was to expose the model to more medical context, helping it better understand and internalize domain-specific patterns through full-sequence supervision. A visualization of this can be seen in Figure 2 and 3. In our approach one question-answer-pair is split into many samples for the training process (e.g., 5 in Figure 2) and therefore provides more exposure to medical terms (as opposed to the one sample in the masked version in Figure 3). This approach was found to be superior in the literature (Shi et al., 2024b), especially, if there are - as in our case - no lengthy instructions that are repeated multiple times (Dettmers et al., 2023).

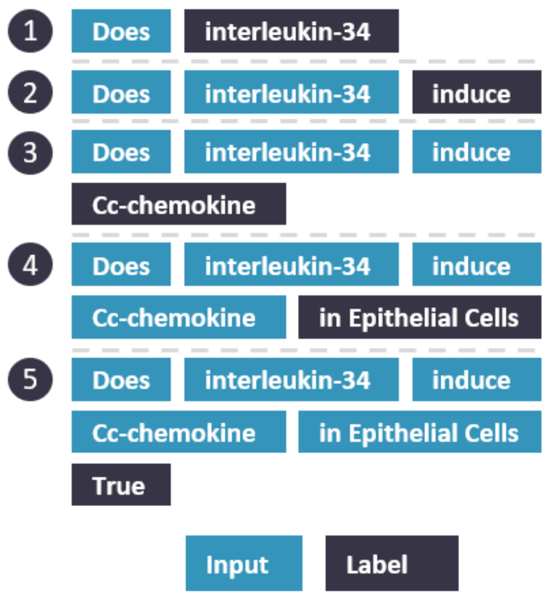

Figure 2: Training setup without masking

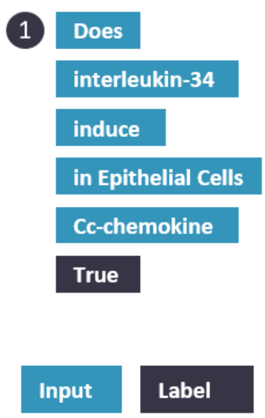

Figure 3: Training setup with masking

Our selection of hyperparameters in the LoRA configuration reflects a balanced approach informed by current research and practical considerations in fine-tuning large language models (LLMs).

Rank

We selected a rank of r = 8, which is consistent with recommendations for efficient fine-tuning. Research indicates that low ranks often suffice for adapting LLMs to new tasks without significant performance degradation. The original paper proposing LoRA argues "that increasing *r* does not cover a more meaningful subspace, which suggests that a low-rank adaptation matrix is sufficient" (Hu et al., 2021).

This choice, therefore, balances computational efficiency with the model's ability to learn task-specific patterns.

LoRA Alpha

We set the scaling factor LoRA Alpha (α) to twice the rank (i.e., 16), so we maintain a scaling factor (α/r) of 2. This is again in line with the original LoRA paper, which states that α should be a constant in *r*. Additionally, it argues that "when optimizing with Adam, tuning α is roughly the same as tuning the learning rate if we scale the initialization appropriately. As a result, we simply set α (...) and do not tune it." (Hu et al., 2021).

Target Modules

We chose to apply LoRA to the query and value projection layers because these components are central to the attention mechanism, which is critical for capturing contextual relationships in language models. Again, this was recommended in the original paper, and we did not perform experiments with other layers, as the results in (Hu et al., 2021) were clear.

Dropout

We set the dropout rate to 0.1 to introduce a moderate level of regularization. A 0.1 dropout is widely adopted in LoRA setups and offers a reliable balance between regularization and learning stability.

Training setup

We set the fine-tuning up for 3 epochs, but stopped the process after 8 hours (approx. 0.8 epochs) due to compute unit constraints in google colab. We employed a small batch size (1 per device) to avoid RAM overflow but a gradient accumulation steps of 8, effectively simulating a larger batch size. We used a learning rate of 2e-4 and enabled mixed-precision (fp16) training to reduce memory usage (similar as above, to avoid RAM overflow) and speed up training.

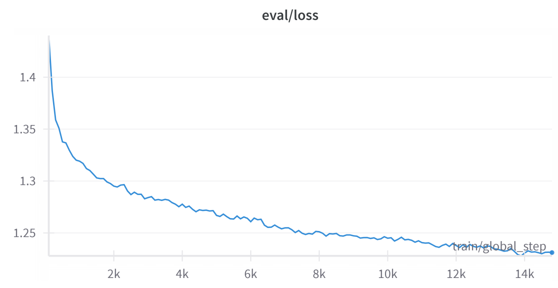

Figure 4: Evaluation loss

We monitored training and evaluation loss using Weights Biases. The evaluation was performed on a relatively small set of only 10 questions. While we acknowledge the limited size of the evaluation set, we did not complete all training epochs and therefore did not require early stopping. Nonetheless, the evaluation loss consistently decreased over the course of 14,000 training steps, as shown in Figure 4.

2.7 RAG

RAG has emerged as a powerful architecture for open-domain question answering, combining external knowledge retrieval with generative capabilities (Lewis et al., 2021). The original formulation by Meta AI, University College London, and NYU proposes an end-to-end approach in which a retriever and generator are trained jointly, enabling close interaction between the retrieved documents and the generation process. Specifically, the *Retriever component* $p_\eta(z|x)$, consisting of a "*query encoder* and *document index*" (Lewis et al., 2021, 2), returns the top-k most relevant documents z given a query x and text parameters η. The *Generator*, a pre-trained LLM / seq2seq model, predicts the next token y_i based on previousy generated text $y_{1:i-1}$, prompt, and additional information z retrieved by the retriever from the external knowledge base: $p_\theta(y_i \mid x, z, y_{1:i-1})$.

While this fully end-to-end design has shown strong performance in benchmark settings, it requires significant computational resources and model coupling. Following recommendations from Lewis et al. (2021) and aligned with recent literature, we adopt a modular RAG pipeline. While the separation of retrieval and generation phases in RAG enables flexible system design, it also introduces challenges such as retrieving semantically relevant chunks and resolving contradictions between external context and the LLM's parametric knowledge. These limitations have led to the development of more advanced architectures which apply enhancement strategies at both the *pre-retrieval* and *post-retrieval* stages (Gao et al., 2024). In this work, we focus on improvements piror to retrieval. In line with Kim and Yoon's VAIV Bio-Discovery system (Kim and Yoon, 2024), which integrates neural retrieval and metadata-enriched biomedical indexing, we enhance the pre-retrieval phase to improve contextual relevance and diversity. Inspired by their use of named entity recognition for biomedical concepts and vector-based semantic indexing, we incorporate:

- **Biomedical NER for metadata filtering:** We applied a medical domain-specific spaCy model to extract biomedical entities (e.g., `DISEASE`, `MEDICAL_CONDITION`, `SYMPTOM`) from the training questions. These extracted features were inserted as metadata tags in the vector store to support filtered retrieval based on medical relevance.

- **INDEX 1: Topic modeling for diverse index:** We used BERTopic on PubMed abstracts to capture thematic breadth in biomedical literature and to populate our knowledge base with a diverse set of topic-representative documents.

- **INDEX 2: Balanced index construction:** A second, balanced index of training and PubMed entries was created to prevent overrepresentation of fine-tuning data and improve fairness during retrieval.

These enhancements aim to alleviate common retrieval issues such as semantic irrelevance and overfitting to fine-tuning data, as similarly addressed in Kim and Yoon (2024).

Ablation Experiment on number of *k* retrieved contexts

All experiments listed in Table 1 have been conducted with *k = 5* retrieved contexts per query. To justify this choice, we conducted a small ablation study on the number of retrieved contexts k (for 300 Multiple Choice questions) to evaluate the effect of varying the number of retrieved contexts on downstream performance. As illustrated in Figure 5, we experimented with $k = \{1, 3, 5, 8, 12\}$. Our setup mirrors the retrieval parameter study performed by Shi et al. (2024a) in the MKRAG framework, where top-k medical facts ($k = \{4, 8, 16\}$) were retrieved and injected into the prompt. Their findings demonstrated a positive correlation between the number of retrieved facts and Q&A accuracy, with performance improving from 41.86% to 48.54% for Vicuna-7B as k increased. However, they also noted a practical ceiling imposed by the model's prompt size. Inspired by their design, we similarly observed diminishing returns beyond an optimal k, affirming the importance of carefully tuning this hyperparameter based on context complexity and model capacity.

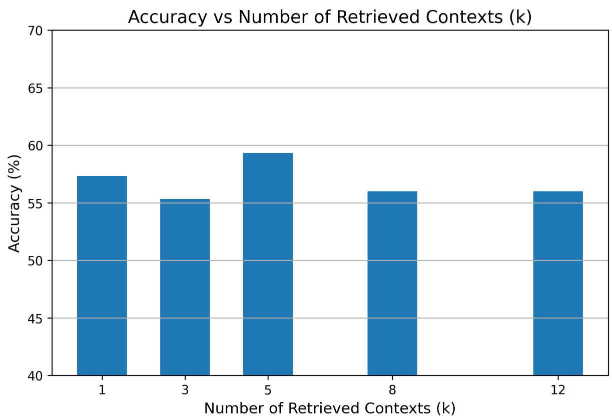

Figure 5: Ablation Experiment on 300 Multiple Choice Questions to find optimal number of k-retrieved contexts

As shown in Figure 5, the accuracy peaked at $k = 5$ (59.3% accuracy) , with performance degrading slightly for both smaller and larger values. This observation underscores the trade-off inherent in selecting k: too small a value may omit relevant information, while too large a value can introduce noise and overwhelm the language model. This phenomenon aligns with findings from recent studies. Yu et al. (2024) observed that smaller k values compromise recall, potentially missing pertinent information necessary for accurate answer generation. Conversely, larger k values may introduce irrelevant or conflicting contexts, thereby degrading performance. Similarly, Xu et al. (2024) reported that performance saturates around $k = 10$ in long-document Q&A tasks.

To further mitigate retrieval-related limitations and improve the relevance of selected contexts, we adopted a hybrid retrieval strategy that combines dense vector similarity with metadata-based filtering. This decision is motivated by insights on the *lost-in-the-middle* phenomenon (Liu et al., 2023), which describes the tendency of LLMs to underutilize information that appears in the middle of lengthy prompts. By integrating biomedical NER into the retrieval pipeline, we filtered candidate documents based on medically relevant entities (e.g., `DISEASE`, `SYMPTOM`) to prioritize conceptually aligned contexts. This not only enhanced semantic coverage but also supported greater retrieval fairness by enforcing diversity in both source type and content structure. Our design shares again conceptual similarities with the *MKRAG framework* proposed by Shi et al. (2024a), where a two-stage dense retrieval method is used: medical knowledge triplets from a disease database are first embedded, and then filtered for relevance to both the input question and the answer candidates. While their focus lies in knowledge triplet matching, both approaches emphasize domain-aware filtering and semantic alignment in medical Q&A contexts.

2.8 Evaluation

To evaluate model performance across different medical question types, we implemented a comprehensive evaluation suite that computed task-appropriate metrics per model approach. For discrete formats such as Multiple Choice and True/False questions, we calculated standard classification metrics including Recall, Precision, and Accuracy. For open-ended tasks like Short Answer and Multi Hop questions, we assessed generation quality using established metrics such as BLEU (Papineni et al., 2002), ROUGE-1/2/L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), BERTScore (Zhang et al., 2020), and Cosine Similarity (Gunawan et al., 2018). Reasoning coherence was further analyzed via sentence-level Cosine Similarity (Bronsdon, 2025). It is important to note, that we followed the challenge's guidelines for evaluating MH questions only on the final answer and not on the reasoning chain. We reflect that future work should more explicitly assess the utility of individual reasoning steps. One possible direction includes applying Natural Language Inference (NLI) to verify whether a reasoning step logically supports the final answer, regardless of the specific order in the gold standard. Another approach leverages LLMs as judges to assess the contribution of reasoning steps. Importantly, our current setup evaluates only the quality of the generated answer text based on the listed scores. Correctness of medical information can not be guaranteed. Furthermore, we do not assess the retrieval component (e.g., whether the correct context was retrieved in RAG-based models).

All metrics were integrated into a unified evaluation framework (our evaluation suite), which was tested on sample examples prior to full analysis. The results were visualized to support comparative inspection across model outputs and to ensure that our evaluation suite functioned as intended. We evaluated 1000 questions from the MC and TF testset, and 500 questions from the SA and MH testset, respectively. This was done due to compute unit constraints. Evaluating 500 MH questions took about 60 minutes on an A100 GPU, the SA questions took about 30 min on the same GPU (100 instead of 200 new tokens generated per sample), the MC and TF about 15 minutes each. That resulted in about 10 compute units (or about 1$) per evaluation of one approach.

All scores (per-sample and averaged), confusion matrices as well as RAG retrieval information were stored automatically by the evaluation suite to our drive to allow a systematic evaluation of all question types, models and approaches.

3 Results

For each of our 5 approaches (baseline model, fine-tuned model, RAG plus baseline model, RAG plus fine-tuned model, RAG on balanced index + fine-tuned model) we evaluated all question types with the scores introduced above. This allowed us to create the final results table shown in Table 3. A condensed version can be seen in the extract shown in Figure 9.

3.1 Multiple Choice (MC) and True/False (T/F) Questions

In classification-style tasks, the Advanced RAG plus Fine-Tuning (FT) model outperformed all other configurations in both MC and T/F formats. It achieved the highest MC accuracy (0.467) and T/F accuracy (0.595). Although the fine-tuned (FT) model alone had a slightly higher T/F accuracy (0.655), the combined approach achieved more balanced performance across all metrics. The Base model performed weakest overall, with particularly low MC precision (0.054).

3.2 Short Answer (SA) and Multi Hop(MH) Reasoning Evaluation

SA generation highlights the progressive benefits of model enhancement. Again, Advanced RAG plus FT led in nearly all metrics, including BLEU (0.206), METEOR (0.402), and semantic match indicators such as sentence similarity (0.867) and reasoning coherence (0.794), suggesting that combining retrieval with fine-tuning enables high quality, contextually relevant text. Also, for MH tasks that require integration of different information pieces, Advanced RAG plus FT demonstrated a clear advantage (BLEU (0.241), ROUGE2 (0.367), and Cosine Similarity (0.980)). While the fine-tuned model performed reasonably well in precision and recall (both 0.87), its coherence and semantic metrics lag behind the retrieval-augmented versions.

What is interesting to see is the difference between the fine-tuned model and the standalone RAG approach. While the fine-tuned model performs better on the discrete classification-style tasks, the RAG outperforms the fine-tuned approach on text-generation tasks. This implies that the baseline model can better leverage retrieved context during generation without being constrained by fine-tuned decision boundaries. In contrast, fine-tuning appears to specialize the model toward the training distribution, potentially at the cost of flexibility in open-ended or generative tasks. This trade-off highlights the complementary strengths of both approaches depending on the task type.

Combining both approaches in the RAG plus FT approach yields the best results across all question types, effectively balancing the precision of fine-tuning with the contextual adaptability of retrieval-augmented generation.

Another observation is that the RAG approach heavily relies on the underlying index and its associated data to answer text-generation questions. This became evident in the balanced experiment, where we intentionally reduced the number of documents from the training data in the index to match the quantity of PubMed data. As a result, the overall volume of indexed data was significantly lowered. Although we continued to use the fine-tuned model to generate answers, the RAG pipeline could no longer supply the full context, which had a notable impact on text-generation performance. The evaluation scores dropped to those of the standard RAG setup, though they remained slightly higher than those of the fine-tuned model alone. Interestingly, performance on classification-style tasks remained strong - likely because the fine-tuned model, rather than the retriever, was primarily responsible for handling those task types in this configuration.

Beyond the numerical performance metrics, qualitative analysis also offers valuable insights that support the findings discussed above. Examination of the confusion matrices for both the RAG pipeline and the fine-tuned model reveals that classification-style tasks are predominantly handled by the fine-tuned component (see Figure 6). This indicates that fine-tuning has the most significant impact in structured prediction scenarios. In contrast, the RAG pipeline on its own fails to provide sufficient contextual grounding for the baseline model to perform these tasks effectively. Interestingly, this effect is reversed in the evaluation of text-generation tasks, where the RAG pipeline contributes more substantially to overall performance.

Additionally noteworthy is, that the baseline model - and by extension, the RAG system built upon it - refused to answer a substantial portion of the text-based questions, citing their medical nature. Specifically, the baseline model frequently responded that it could not provide an answer and recommended consulting a medical professional instead. Out of a total of 1,000 text-generation questions, approximately 60% were blocked in this manner. Remarkably, over 100 of these blocked responses explicitly included a disclaimer stating,

> *"As an AI system developed by OpenAI, I cannot answer this question."*
> (DeepSeek model)

This is particularly notable given that the model in question was not developed by OpenAI, suggesting potential reuse or imitation of instruction-aligned response patterns.

Furthermore, we found that fine-tuning open-weight models enabled us to effectively override (or "jailbreak") such restrictions. This highlights the potential for domain-specific fine-tuning to bypass default safety or alignment constraints in publicly available models, especially Chinese open-weight models, which appear to be pre-configured to restrict answers to certain question types. Investigating methods to jailbreak other generic large language models may offer valuable insights into their alignment strategies and pretraining safeguards.

Finally, we want to focus on another important area of investigation that goes beyond numerical evaluation metrics and involves a closer examination of specific qualitative examples. In the following, we present a representative sample that includes the ground truth, the response generated by the fine-tuned model, and the response produced by the fine-tuned model augmented with RAG. This qualitative comparison enables a more nuanced assessment of answer quality, allowing us to evaluate how well each approach aligns with clinical reasoning and diagnostic accuracy, insights that may not be fully captured by aggregate performance metrics alone. This would be the ideal assessment for more samples, but requires medical knowledge and is relatively work intensive.

Question A 75-year-old patient with a 25-year history of diabetes mellitus presented with a **(2) sudden painless loss of vision**. On fundus examination, flame-shaped hemorrhages were observed in the retina. What is the probable diagnosis?

Ground-truth The probable diagnosis for this 75-year-old patient with a history of diabetes mellitus and symptoms of sudden painless vision loss, along with the presence of flame-shaped hemorrhages in the retina, is **(1) likely proliferative diabetic retinopathy**. Proliferative diabetic retinopathy is a common complication in patients with long-standing diabetes. It results from the growth of new, abnormal blood vessels on the retina, which can bleed (leading to hemorrhages) and potentially cause a sudden loss of vision, especially if there is vitreous hemorrhage or retinal detachment. While central retinal vein occlusion can also cause similar symptoms and retinal findings, the longstanding history of diabetes makes proliferative diabetic retinopathy the more likely diagnosis in this case.

Fine-tuned answer The probable diagnosis for the 75-year-old patient with a 25-year history of diabetes mellitus presenting with a sudden painless loss of vision and flame-shaped hemorrhages in the retina is **(1) diabetic retinopathy**. These characteristics align with the typical progression and presentation of diabetic retinopathy, particularly the development of hemorrhages, which are common in this condition. It's crucial to note that diabetic retinopathy can progress to proliferative vitreoretinopathy if not treated promptly, potentially **(2) leading to permanent vision loss**. Therefore, timely and appropriate management is necessary to prevent further vision loss and complications. This includes regular eye examinations and prompt treatment of any retinal lesions or hemorrhages detected. Treatment options may include laser or surgical interventions, depending on the severity of the condition. Regular monitoring and prompt intervention are key to maintaining and preserving vision in patients with diabetes. In summary, the probable diagnosis is diabetic retinopathy, which aligns with the **(4)**

RAG plus FT answer The most probable diagnosis for this 75-year-old patient is **(1) proliferative diabetic retinopathy**. This condition is common in individuals with long-standing diabetes, and it often results in the growth of abnormal new blood vessels in the retina, leading to hemorrhages and the possibility of a sudden loss of vision, especially if there is vitreous hemorrhage or retinal detachment. The patient's symptoms, including sudden vision loss, are consistent with these complications. Therefore, it is crucial to refer the patient for an immediate retinal evaluation to assess the extent of the retinal damage and manage any potential complications. **(3) In the context of peripheral neuropathy**, which is often seen in diabetes, it is important to note that the development of proliferative diabetic retinopathy can occur even in patients with no apparent ocular symptoms. Therefore, even if the patient is asymptomatic, it is recommended to conduct an evaluation for retinal health. This proactive approach can help identify and **(4)**

Analysis

Four remarks can be made for the generated responses:

(1): The FT answer is generic and only partly accurate. It lacks the clinical precision as it doesn't identify proliferative diabetic retinopathy, which is important given the symptom of sudden vision loss and flame-shaped hemorrhages. The RAG system does identify the proliferative condition, which is crucial.

(2): The FT answer does not fully capture the question. It states that the condition can lead to permanent vision loss. However, in the question, it is already stated that the patient suffers from "sudden painless loss of vision".

(3): The RAG system introduces tangential detail, i.e., peripheral neuropathy not relevant to the case. This is most probable due to certain documents that are retrieved as additional context but not directly relevant.

(4): Both versions end awkwardly and incomplete, this is due to the token limitation set in the pipeline parameters (200 for MH questions).

This analysis supports the interpretation of the quantitative metrics: the RAG with fine-tuning response ranks higher for text-generation questions due to its clinical correctness, despite (or maybe due to) a lack of focus and occasional verbosity. In contrast, the fine-tuned model's response, while valid, is too general and incomplete for the diagnostic requirements of this specific case. These observations are consistent with the ROUGE-L scores for the respective responses - 0.42 for the RAG-enhanced output and 0.34 for the fine-tuned model alone.

3.3 Discussion

Our experimental results demonstrate that the integration of both RAG and task-specific fine-tuning significantly improves performance on Multiple Choice medical Q&A. Specifically, our Advanced RAG plus FT model, built on the *DeepSeek-Coder-7B-Instruct-v1.5* architecture and fine-tuned using LoRA, achieves a Multiple Choice accuracy of 0.467. This marks a notable improvement

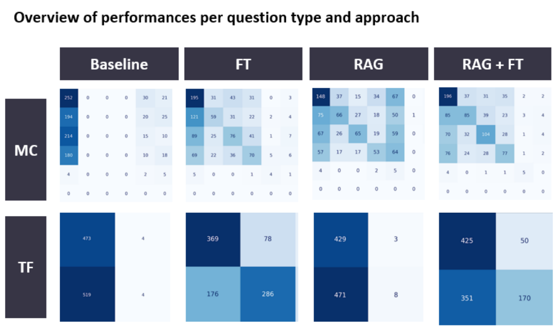

Figure 6: Confusion matrices for classification style tasks (MC and TF)

over the base model's 0.254, effectively pushing DeepSeek's performance to a level comparable to that reported in the MKRAG study, where retrieval-augmented Vicuna-7B reached an accuracy of 48.54% on MedQA-USMLE (Shi et al., 2024a).

While these results highlight the effectiveness of our domain-adapted RAG pipeline, we also observed a second, important phenomenon: the presence of **structural retrieval bias** within the retriever's output that may limit the diversity and generalizability of retrieved contexts in biomedical Q&A systems. We evaluated retrieval quality under two indexing strategies. As shown in Figure 7, our first index represented a strong imbalance between training data and PubMed articles. In contrast, our second configuration applied balanced sampling (Index 2), integrating equal proportions of training and external documents. Interestingly, both configurations revealed a strong retrieval bias toward documents matching the question type. For example, Multiple Choice (MC) questions predominantly retrieved MC documents in both indices, despite the broader availability of thematically diverse entries. Even with a balanced index, retrieval remained skewed toward same-type documents across all question types. This behavior suggests that the retriever's embedding space prioritizes structural similarities, favoring documents with similar format or phrasing to the input question. As a result, the retrieved contexts may lack diversity, limiting the model's ability to reason across heterogeneous document types or answer formats.

To our knowledge, prior Q&A-RAG literature has not explicitly investigated such structural biases arising from the implementation of question-answer pairs in the knowledge base. However, related retrieval limitations have been reported by

Shah et al. (2024), who proposed the *RAIDD framework* to enhance semantic diversity in retrieval. In their evaluation, they experimented with question-based feature generation, where pseudo-questions are created from individual text chunks and used to guide retrieval. Notably, this approach degraded performance, particularly for question types requiring broader contextual understanding. As the authors explain:

> "The questions generated from each individual chunk do not adequately reflect the types of questions typically associated with this question type." (Shah et al., 2024, p. 7)

This mirrors our own observations: retrieval models that overly rely on surface structure may fail to generalize across formats or integrate broader context. In our case, embedding-based retrievers demonstrated a similar overfitting behavior reducing contextual diversity.

While our findings highlight a structural similarity bias in retrieval, most existing literature has focused on a demographic bias in medical Q&A. A recent study by Ji et al. (2025), published in March 2025, examined how retrieval-augmented Q&A systems respond to queries that differ only in patient demographic descriptors (e.g., "an elderly patient" vs. "a young adult"). They found that despite these demographic variations, the retrieved documents remained largely the same across cases, a phenomenon known as retrieval overlap. This indicates that the retriever fails to adapt its context selection based on the patient group described, which they identify as a fairness limitation in medical Q&A. Although their primary focus is on demographic bias, their findings reveal a broader issue: retrieval models often overlook important contextual cues in the input query. This aligns with our structural bias observation, where embedding-based retrievers overemphasize surface-level similarities (such as question format or phrasing), resulting in reduced contextual diversity.
Our findings thus contribute a complementary perspective to the emerging body of bias research in RAG systems, one that considers how question structure itself influences retrieval behavior. From our perspective, this biased-outcome is intuitive since embedding-based retrieval captures not only semantic similarity but also structural patterns, such as phrasing, length, or formatting cues present in question types. Together, these findings suggest a need for more context-aware retrieval strategies that can handle both semantic and structural variability in biomedical Q&A tasks.
To address this, we adopted a hybrid retrieval strategy with biomedical metadata, aiming to guide the retriever toward semantically relevant and structurally diverse contexts. This approach reflects a first effort to mitigate structural overfitting and improve contextual fairness across question formats.

3.4 Limitations

Several things should be considered when interpreting the results and applicability thereof.
Firstly, while the Advanced RAG plus Fine-Tuning model outperforms other configurations, its accuracy on Multiple Choice (0.467) and True/False (0.595) tasks remains moderate, indicating that the system is not reliable for critical medical decision-making.
Secondly, the evaluation relies heavily on automated NLP metrics such as BLEU, ROUGE, METEOR, and BERTScore for open-ended questions, which, while standard in research, may not fully capture how well a model performs in the face of human evaluation, especially in the clinical context. In other words, apart from a few samples (e.g. the one listed in the last section of the Results chapter 3) the lack of clinical validation or real-world deployment means the model's outputs have not been assessed by medical professionals, limiting confidence in their practical utility.
While RAG improves performance, the quality and relevance of retrieved information from the PubMed corpus as well as the question-answer training dataset, cannot be guaranteed.

In general, while the project demonstrates promising advances, these limitations highlight the need for further development, clinical validation, and careful consideration before deployment in real-world healthcare environments.

3.5 Future Work

Several research directions remain open for future investigation, and important questions are yet to be answered:

(1): Although we are aware of the structural biases inherent in retrieval models, as highlighted in related research Ji et al. (2025), our study did not conduct a systematic analysis of the extent of these biases across different question types or datasets, nor did it attempt to implement mitigation strategies. This presents an opportunity for future research to explore targeted mitigation techniques and assess their effectiveness across diverse retrieval scenarios.

(2): Knowledge graphs organize biomedical entities, typically derived from named entity recognition and their relationships. Integrating a knowledge graph with a RAG system might enhance both retrieval precision and explainability. Unlike purely dense vector representations, knowledge graphs capture explicit relationships between entities (e.g., Gene A — inhibits — Disease B), enabling the retrieval of context that may be missed by vector-only approaches.

(3): Agentic or multi-agentic RAG systems involving multiple virtual medical specialists warrant further investigation, as recent studies have demonstrated their potential to achieve higher precision and recall, key improvements in the context of critical healthcare applications (Singh, 2025).

(4): It is important to note that our study did not involve comprehensive hyperparameter optimization or prompt engineering, which may have limited the full performance potential of the models evaluated. This is something that should be investigated further. Additionally, we recognize the opportunity to explore fine-tuning strategies that are both domain-specific, focusing on individual medical specialties, and use-case-specific, tailored to particular task types. For instance, large language models such as Aquila-Med have been fine-tuned on medical datasets and further refined through reinforcement learning from human feedback (RLHF) provided by medical professionals, resulting in improved performance on Multiple Choice question answering tasks (Anisuzzaman et al., 2025).

Table 3: Model Evaluation Results

Multiple Choice (MC) and True/False Evaluation

Model	MC Acc.	MC Prec.	MC Rec.	T/F Acc.	T/F Prec.	T/F Rec.
Balanced Advanced RAG plus FT	0.422	0.321	0.319	0.573	0.436	0.392
Advanced RAG plus FT	**0.467**	**0.388**	**0.377**	0.595	0.440	0.405
Advanced RAG	0.337	0.263	0.249	0.437	0.401	0.305
FT	0.404	0.334	0.318	**0.655**	**0.488**	0.440
Base	0.254	0.054	0.169	0.477	**0.488**	**0.500**

Short Answer Evaluation

Model	BLEU	METEOR	ROUGE1	ROUGE2	ROUGEL	Prec.	Rec.	F1	CosSim	ReasonCoh	SentSim	ParaSim
Balanced Advanced RAG plus FT	0.127	0.216	0.296	0.228	0.276	**0.877**	0.821	0.848	0.864	0.164	0.837	0.864
Advanced RAG plus FT	**0.206**	**0.402**	**0.401**	**0.314**	**0.355**	0.843	**0.868**	0.854	**0.913**	**0.794**	**0.867**	**0.913**
Advanced RAG	0.118	0.295	0.339	0.210	0.269	0.848	0.870	**0.858**	0.850	0.727	0.801	0.850
FT	0.049	0.207	0.263	0.109	0.188	0.834	0.838	0.835	0.836	0.654	0.665	0.836
Base	0.014	0.061	0.084	0.033	0.061	0.781	0.791	0.785	0.262	0.191	0.237	0.262

Multi Hop Evaluation

Model	BLEU	METEOR	ROUGE1	ROUGE2	ROUGEL	Prec.	Rec.	F1	CosSim	ReasonCoh	SentSim	ParaSim
Balanced Advanced RAG plus FT	0.127	0.216	0.296	0.228	0.276	0.877	0.821	0.848	0.864	0.164	0.837	0.864
Advanced RAG plus FT	**0.241**	**0.501**	**0.544**	**0.367**	**0.406**	**0.893**	**0.922**	**0.907**	**0.980**	**0.887**	**0.926**	**0.980**
Advanced RAG	0.156	0.359	0.407	0.245	0.295	0.862	0.892	0.876	0.855	0.710	0.799	0.855
FT	0.079	0.303	0.406	0.162	0.247	0.862	0.879	0.870	0.947	0.749	0.790	0.947
Base	0.018	0.086	0.131	0.040	0.077	0.796	0.813	0.804	0.352	0.253	0.306	0.352

References

D. M. Anisuzzaman, Jeffrey G. Malins, Paul A. Friedman, and Zachi I. Attia. 2025. Fine-Tuning Large Language Models for Specialized Use Cases. *Mayo Clinic Proceedings: Digital Health*, 3(1):100184.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Conor Bronsdon. 2025. G-eval: Evaluating llm-generated text with granular metrics. `https://www.galileo.ai/blog/g-eval-metric`. Accessed: 2025-05-29.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. Huatuogpt-o1, towards medical complex reasoning with llms.

ClinIQLink. 2025. Cliniqlink 2025 – llm lie detector test. Accessed: 2025-05-29.

Florin Cuconasu, Giovanni Trappolini, Nicola Tonellotto, and Fabrizio Silvestri. 2024. A tale of trust and accuracy: Base vs. instruct llms in rag systems.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey.

D Gunawan, C A Sembiring, and M A Budiman. 2018. The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents. *Journal of Physics: Conference Series*, 978(1):012120. Publisher: IOP Publishing.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Yuelyu Ji, Hang Zhang, and Yanshan Wang. 2025. Evaluating bias in retrieval-augmented medical question-answering systems. *arXiv preprint arXiv:2503.15454*.

Xinke Jiang, Ruizhe Zhang, Yongxin Xu, Rihong Qiu, Yue Fang, Zhiyuan Wang, Jinyi Tang, Hongxin Ding, Xu Chu, Junfeng Zhao, and Yasha Wang. 2024. Hykge: A hypothesis knowledge graph enhanced framework for accurate and reliable medical llms responses.

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

Hyunjae Kim, Hyeon Hwang, Jiwoo Lee, Sihyeon Park, Dain Kim, Taewhoo Lee, Chanwoong Yoon, Jiwoong Sohn, Donghee Choi, and Jaewoo Kang. 2024. Small Language Models Learn Enhanced Reasoning Skills from Medical Textbooks. ArXiv:2404.00376 [cs].

Seonho Kim and Juntae Yoon. 2024. Vaiv bio-discovery service using transformer model and retrieval augmented generation. *BMC Bioinformatics*, 25(273).

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xinyun Liu, Yichong Zhang, Zihang Dai, and etc. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Chiara Longoni, Andrea Bonezzi, and Carey K Morewedge. 2019. Resistance to Medical Artificial Intelligence. *Journal of Consumer Research*, 46(4):629–650.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

OpenAI. 2025. Openai platform. `https://platform.openai.com/docs/api-reference/chat/create#chat-create-temperature`. Accessed: 2025-05-12.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2024. Fine-tuning or retrieval? comparing knowledge injection in llms.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matthew Renze and Erhan Guven. 2024. The effect of sampling temperature on problem solving in large language models.

Keyush Shah, Abhishek Goyal, and Isaac Wasserman. 2024. Enhancing retrieval in qa systems with derived feature association. *arXiv preprint arXiv:2410.03754*.

Yucheng Shi, Shaochen Xu, Tianze Yang, Zhengliang Liu, Tianming Liu, Quanzheng Li, Xiang Li, and Ninghao Liu. 2024a. MKRAG: Medical Knowledge Retrieval Augmented Generation for Medical Question Answering. ArXiv:2309.16035 [cs].

Zhengyan Shi, Adam X. Yang, Bin Wu, Laurence Aitchison, Emine Yilmaz, and Aldo Lipani. 2024b. Instruction tuning with loss over instructions.

Ajit Singh. 2025. **Future Prospects of Agentic RAG in Medical Research**.

together.ai. 2025. Prompting DeepSeek-R1. `https://docs.together.ai/docs/prompting-deepseek-r1`. Accessed: 2025-05-03.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, Yihan Cao, Hui Ren, Xiang Li, Xiaoxiao Li, and Yuyin Zhou. 2025. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*.

Peng Xu, Wei Ping, Xia Wu, Luke McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Ekaterina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Retrieval meets long context large language models. In *International Conference on Learning Representations (ICLR)*.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. *arXiv preprint arXiv:2407.02485*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. ArXiv:1904.09675 [cs].

Mi Zhou, Yun Pan, Yuye Zhang, Xiaomei Song, and Youbin Zhou. 2025. Evaluating AI-generated patient education materials for spinal surgeries: Comparative analysis of readability and DISCERN quality across ChatGPT and deepseek models. *International Journal of Medical Informatics*, 198:105871.

A Appendix

A.1 Question Formats

A.1.1 Multiple Choice (MC)


```
{
  "correct_answer": "A",
  "options": {
    "A": "Localized immune complex",
    "B": "Ag-Ab reaction",
    "C": "Complement mediated",
    "D": "Ab mediated"
  },
  "question": "Ahus reaction is...",
  "source": "MC4-UCSC-VLAA/MedReason",
  "type": "multiple_choice"
}
```


A.1.2 True/False (TF)


```
{
  "answer": "True",
 "question": "Is reduced Klotho associated
  with the presence ...",
  "source": "TF2-qiaojin/PubMedQA",
  "type": "true_false"
}
```


A.1.3 Short Answer (SA)


```
{
  "question": "What is the value of ...",
  "answer": "In the late distal tubule,
  [TF/P]osm is less than 1 when there is
  low ADH.",
  "source": "SA2-Ajayaadhi/Medical-QA",
  "type": "short_answer"
}
```


A.1.4 Multi Hop (MH)


```
{
  "question": "In a study assessing ...",
  "answer": "In a normal...",
  "reasoning": [
    "Step 1: Alright, ...",
    "Step 2: When a student's...",
   "Step 3: I remember from stats that ..."
  ],
  "source": "MH-FreedomIntelligence/
  medical-ol-reasoning-SFT",
  "type": "multi_hop"
}
```


A.2 Analysis of retrieved RAG contexts

A.3 3rd party benchmarks for DeepSeek-Coder

A.4 Result overview of our approaches

A.5 Directory of Aids

Aid Usage	Application	Affected Areas
OpenAI - Chat-GPT	Help with writing, Grammar Check and formatting into LaTeX	Abstract, Introduction, Methodology, Evaluation, Appendix
Perplexity.AI	Help with writing, Grammar Check and formatting into LaTeX	Introduction, Methodology, Evaluation, Limitations, Appendix
Microsoft Copilot	Help with writing, Grammar Check and formatting into LaTeX	Introduction, Methodology, Evaluation, Appendix

Table 4: Writing Aids (Art. 57 AB)

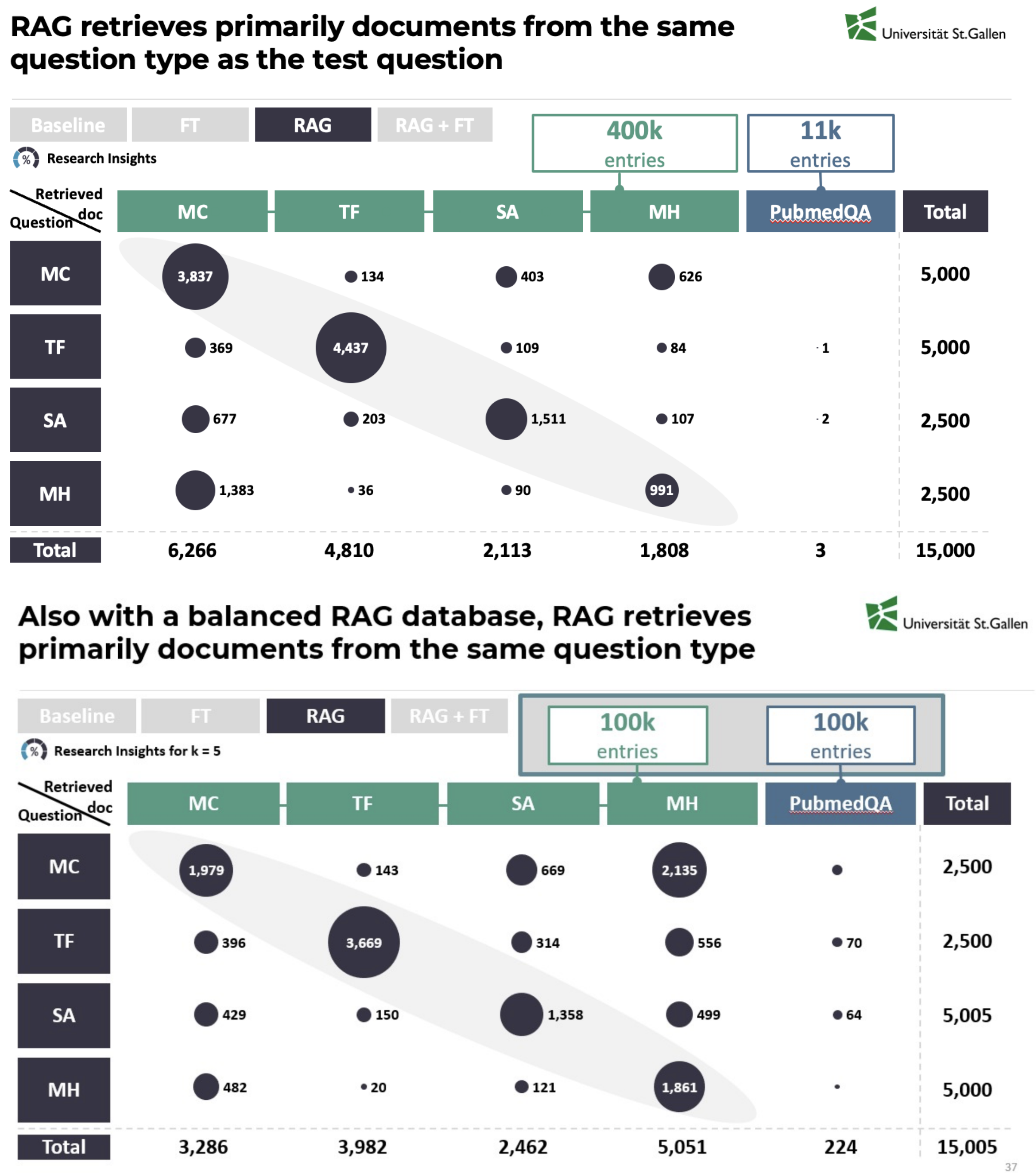

Figure 7: Retrieved document types for different question types using two index strategies: Diverse Index (top) and Balanced Index (bottom). Both show strong same-type retrieval patterns despite index differences.

		Programming		Math Reasoning		Natural Language				
Models	Size	HumanEval	MBPP	GSM8K	MATH	MMLU	BBH	HellaSwag	WinoG	ARC-C
DeepSeek-Coder-Base	6.7B	**44.7%**	**60.6%**	43.2%	19.2%	36.6%	44.3%	53.8%	57.1%	32.5%
DeepSeek-Coder-Base-v1.5	6.9B	43.2%	60.4%	**62.4%**	**24.7%**	**49.1%**	**55.2%**	**69.9%**	**63.8%**	**47.2%**
DeepSeek-Coder-Instruct	6.7B	**66.1%**	**65.4%**	62.8%	28.6%	37.2%	46.9%	55.0%	57.6%	37.4%
DeepSeek-Coder-Instruct-v1.5	6.9B	64.1%	64.6%	**72.6%**	**34.1%**	**49.5%**	**53.3%**	**72.2%**	**63.4%**	**48.1%**

Figure 8: 3rd party benchmarks for our baseline model

Overview of performances per question type and approach

	Baseline	FT	RAG	RAG + FT	Balanced RAG + FT
MC	19 min L4 **Acc: 0.254**	15 min A100 **Acc: 0.404**	15 min A100 **Acc: 0.337**	18min A100 **Acc: 0.467**	15min A100 **Acc: 0.422**
TF	19 min L4 **Acc: 0.477**	15min A100 **Acc: 0.655**	14 min A100 **Acc: 0.437**	30min A100 **Acc: 0.595**	32min A100 **Acc: 573**
SA	35min L4 **ROUGE-L: 0.06**	74min A100 **ROUGE-L: 0.18**	37min A100 **ROUGE-L: 0.26**	40min A100 **ROUGE-L: 0.35**	43min A100 **ROUGE-L: 0. 28**
MH	56min L4 **ROUGE-L: 0.07**	10min A100 **ROUGE-L: 0.24**	56min A100 **ROUGE-L: 0.29**	77min A100 **ROUGE-L: 0.40**	73min A100 **ROUGE-L: 0.22**

Figure 9: Result overview of our approaches