Leveraging Knowledge Graph-Enhanced LLMs for Context-Aware Medical Consultation

Su-Hyeong Park[1]*, Ho-Beom Kim[1]*, Seong-Jin Park[2], Dinara Aliyeva[3], Kang-Min Kim[1,2]†

[1]Department of Data Science [2]Department of Artificial Intelligence
The Catholic University of Korea, Bucheon, Republic of Korea
[3]Department of Computer Science
University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
{pshpulip22,hobeom2001,sjpark,kangmin89}@catholic.ac.kr
adinara@cs.unc.edu

Abstract

Recent advancements in large language models have significantly influenced the field of online medical consultations. However, critical challenges remain, such as the generation of hallucinated information and the integration of up-to-date medical knowledge. To address these issues, we propose **I**nformatics **Llama** (ILlama), a novel framework that combines retrieval-augmented generation (RAG) with a structured medical knowledge graph. ILlama incorporates relevant medical knowledge by transforming subgraphs from a structured medical knowledge graph into text for RAG. By generating subgraphs from the medical knowledge graph in advance for RAG, specifically focusing on diseases and symptoms, ILlama enhances the accuracy and relevance of its medical reasoning. This framework enables effective incorporation of causal relationships between symptoms and diseases. Also, it delivers context-aware consultations aligned with user queries. Experimental results on the two medical consultation datasets demonstrate that ILlama outperforms strong baselines, achieving a semantic similarity F1 score of 0.884 when compared to ground-truth consultation answers. Furthermore, qualitative analysis of ILlama's responses reveals significant improvements in hallucination reduction and clinical usefulness. These results suggest that ILlama has strong potential as a reliable tool for real-world medical consultation environments. Our implementation is available at: https://github.com/suhyeong10/ILlama

1 Introduction

Traditional online medical consultation platforms, such as HealthCareMagic[1] and iCliniq[2], rely on medical professionals to answer patient queries and provide expert advice. However, due to their dependence on human labor, these systems face limitations in delivering real-time responses, as experts require significant time to review inquiries and generate appropriate answers (Cao et al., 2022). To address this issue, rule-based medical consultation systems have been introduced (Amato et al., 2017; Mishra et al., 2023; Rosruen and Samanchuen, 2018; Huang et al., 2018). Nevertheless, these systems often struggle to handle complex symptoms and patient-specific queries, as they rely on predefined rules that lack flexibility and adaptability.

Recently, large language model (LLM)-based consultation systems, such as ChatDoctor (Li et al., 2023), have emerged as promising alternatives. These systems typically extract keywords from user queries to retrieve relevant medical information from sources like Wikipedia or custom disease databases. However, their reliance on potentially inaccurate keyword extraction may lead to search failures and hallucinations, failing to capture essential disease-symptom relationships. While dense embedding-based retrieval methods (Karpukhin et al., 2020) can alleviate keyword extraction errors, they still have limitations in capturing the complex symptom-disease relationships essential for medical consultations. For example, distinguishing whether shortness of breath and fatigue arise from a serious condition like lung cancer or a more benign cause such as anemia requires an understanding of such causal relationships.

To overcome these limitations, we propose a novel framework for real-time medical consultation, called **I**nformatics **Llama** (ILlama). ILlama improves the response quality as measured by embedding-based evaluation metrics by incorporating structured medical knowledge. ILlama leverages retrieval-augmented generation (RAG) (Lewis

*These authors contributed equally to this work.
†Corresponding author.
[1]https://www.askadoctor24x7.com
[2]https://www.icliniq.com

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 30459–30475
November 4-9, 2025

et al., 2020b) by incorporating medical knowledge from a structured knowledge graph (KG) built upon the unified medical language system (UMLS)[3]. Unlike keyword-dependent approaches, ILlama improves both retrieval efficiency and reliability. Keyword-extraction methods often misinterpret the intent of user queries and, in many cases, fail to return relevant results if relevant medical information for the extracted keywords is unavailable. By employing a KG-based retrieval approach (Luo et al., 2025), ILlama effectively represents disease-symptom causal relationships, enhancing the contextual relevance and accuracy of diagnostic responses.

In addition, ILlama tackles two core limitations prevalent in existing dense embedding-based augmentation systems: (1) the incompleteness of external knowledge representations and (2) the difficulty in aligning user queries with the embedded knowledge space (Varshney et al., 2023). ILlama addresses the incompleteness of the KG by constructing subgraphs that enrich sparse regions with semantically related triples. It is also designed to alleviate the challenge of aligning user queries with the KG structure by embedding each triple and integrating it into the answer generation process. This approach enables more accurate semantic matching and enhances the clinical relevance of the generated responses. Specifically, embedding triples allows the model to retrieve more precise symptom-disease associations, reducing factual errors, while the structured knowledge context provided by the KG improves the alignment of responses with real-world clinical reasoning.

To ensure that ILlama performs reliably not only on known data distributions but also in unfamiliar real-world scenarios, we adopt both in-distribution and out-of-distribution evaluation protocols throughout this work. We validate the effectiveness of ILlama using two medical consultation datasets with different characteristics. Specifically, we conducted experiments with publicly available data collected from HealthCareMagic and iCliniq. For in-distribution evaluation, we use the HealthCareMagic dataset, which includes separate training, validation, and test splits. For out-of-distribution evaluation, we use real-world consultation records from the iCliniq platform, serving as the test set. ILlama_{7B}, which is based on the same base model as ChatDoctor, achieves semantic similarity F1 scores of 0.866 and 0.851 on the in-distribution and out-of-distribution datasets, respectively, and outperforms all baseline models. ILlama_{8B}, with a more powerful backbone LLM, further improves these results, achieving scores of 0.884 and 0.871, respectively. We further validate the reliability and clinical quality of the generated responses by a qualitative evaluation.

In summary, our contributions are three-fold:

- We propose ILlama, a framework that enhances the effectiveness of medical consultations by leveraging RAG with structured UMLS-based KG.
- ILlama utilizes subgraphs from a UMLS-based KG, which are transformed into document form, combined with vector search techniques, enabling precise retrieval and integration of medically relevant knowledge into the answer generation process.
- Our framework achieves state-of-the-art performance across multiple datasets, with the best results observed on the HealthCareMagic dataset, significantly improving the reliability and usefulness of automated medical consultation systems.

2 Related Works

2.1 Early Medical Consultation Systems

Early systems (Amato et al., 2017; Mishra et al., 2023; Rosruen and Samanchuen, 2018; Huang et al., 2018) used rule-based approaches for simple Q&A interactions, easing the burden on healthcare professionals but failing to handle complex symptoms and disease interactions. To overcome this, medical-specialized models using natural language processing technologies (Lee et al., 2020; Yuan et al., 2022; Lu et al., 2022) were developed, yet challenges in incorporating structured medical knowledge and understanding causal relationships between symptoms and diseases remain. LLMs such as GPT-4 (Achiam et al., 2023) have catalyzed the development of models capable of sophisticated medical consultations (Thirunavukarasu et al., 2023; Li et al., 2024; Toma et al., 2023; Chen et al., 2023; Luo et al., 2022; Yang et al., 2024), although persistent challenges remain, including hallucinations and the incorporation of up-to-date medical information (Vaishya et al., 2023; Hadi

[3] https://www.nlm.nih.gov/research/umls/index.html

et al., 2024). To mitigate these issues, we incorporate a UMLS-based KG that enables accurate identification of disease relationships and contextual information retrieval, thereby supporting more clinically relevant and context-aware consultations.

2.2 Knowledge Graph-Enhanced LLMs in Medical Consulting

To address the limitations of LLMs, such as hallucinations, lack of timely medical knowledge, and insufficient adaptability to patient-specific contexts (Pal et al., 2023), recent research has explored the integration of real-world knowledge to enhance performance in medical applications. Among these approaches, the combination of LLMs with KGs has demonstrated effectiveness in incorporating external information (Goldsack et al., 2023). For example, KG-enhanced models have been used for diagnosis prediction (Gao et al., 2025), graph-augmented medical dialogues (Varshney et al., 2023), and factual medical question answering (Guo et al., 2022; Martino et al., 2023). While prior methods serialize the entire graph or all neighboring nodes into lengthy text inputs without filtering noisy information, our approach selects only a one-hop subgraph centered on a compressed triple via Triple2Seq (Bi et al., 2024), resulting in a much shorter input averaging 130 tokens (Li et al., 2025). ILlama introduces subgraph-based retrieval and semantic reranking to improve knowledge relevance and integration, offering more accurate and context-sensitive medical consultations.

3 Method

The proposed framework consists of three main components: Retriever, Reranker, and Generator. Medical knowledge from the KG is first segmented into subgraphs and transformed into documents in natural language form, which serve as input across all stages. Section 3.1 describes how the Retriever identifies subgraphs semantically relevant to the input query. Section 3.2 presents the Reranker, which employs a cross-encoder (Reimers and Gurevych, 2019) to rerank the retrieved documents in natural language form. Section 3.3 explains how the Generator uses the top-ranked documents to generate the final response. The overall process is illustrated in Figure 1.

3.1 Retriever: Enhancing Medical Knowledge

In medical consultations, it is essential to provide accurate, context-aware information without hallucinations. Our framework requires comprehensive medical knowledge, particularly regarding causal relationships between symptoms and diseases. To achieve this, we incorporate a KG based on UMLS, which enables the language model to effectively capture these relationships and allows targeted retrieval of relevant medical facts from the KG. This ensures that responses are both precise and contextually appropriate to the user's query.

3.1.1 Triple-Centric Knowledge Structuring for Medical Reasoning

To represent medical knowledge, we adopt the Triple2Seq method to segment the UMLS-based KG into coherent subgraphs. A subgraph contains medical concept nodes (e.g., diseases, symptoms, treatments) connected by relationship edges (e.g., "has symptom", "treated by"). Triple2Seq dynamically selects a subgraph by identifying a central triple and including only its one-hop neighbors. This structure is linearized into a sequence for efficient language model integration while minimizing noise.

Each subgraph $\mathcal{T}_g$ is composed of a center triple $\mathcal{T}_c$ (e.g., Lung Cancer-has symptom-Fatigue), representing a core medical concept, and a set of neighboring context triples $\mathcal{T}_N$ (e.g., Lung Cancer-has causes-Smoking) that provide additional medical facts related to the center triple:

$$\mathcal{T}_g = \mathcal{T}_c \cup \mathcal{T}_N. \tag{1}$$

$\mathcal{T}_N$ includes all triples connected to the center triple via its neighboring nodes in the KG and is defined as:

$$\mathcal{T}_N = \{\mathcal{T}_i \mid \mathcal{T}_i \in \mathcal{N}\}, \tag{2}$$

where $\mathcal{N}$ denotes the set of nodes that are directly linked to the center concept in the graph. For example, if the center triple corresponds to a disease such as lung cancer, the context triples may include related symptoms (e.g., shortness of breath and fatigue), diagnostic procedures (e.g., chest X-ray), or causes (e.g., smoking or air pollution). By organizing knowledge in this localized and relation-centric manner, the model is guided to focus on medically relevant and causally connected concepts, thereby enhancing the contextual consistency of the generated responses. Furthermore, this structure enables more accurate and context-aware diagnosis

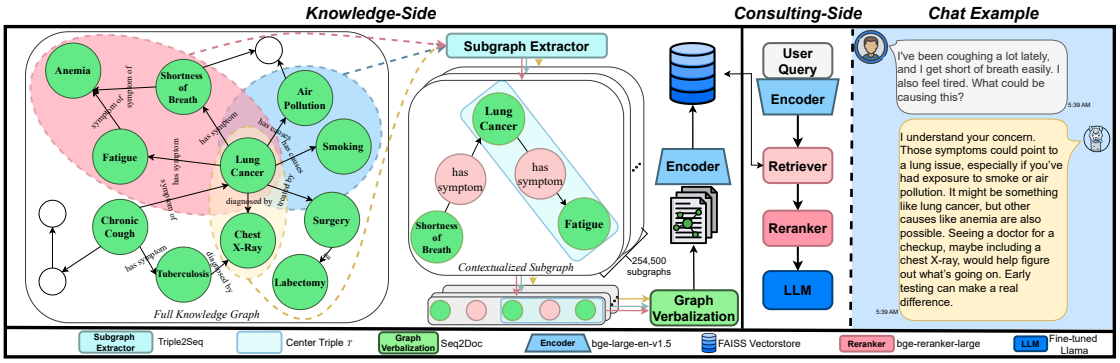

Figure 1: Overall architecture for medical query answering using contextualized subgraphs from the UMLS-based KG. These subgraphs are encoded and stored in a vector database, then combined with the user query to generate the final response using the Llama model.

and consultation based on the patient's reported symptoms, ultimately improving the reliability and practicality of medical dialogue systems.

3.1.2 Subgraph-to-Text Transformation

The UMLS-based KG represents relationships between medical concepts using a subject–predicate–object triple structure, which closely resembles the structure of natural language sentences. Leveraging this property, we convert each triple into a natural sentence. This transformation reconstructs the structural relationships in the graph into a coherent narrative, allowing the model to intuitively understand the meaning and connections between medical entities. As a result, the graph-based knowledge is naturally integrated into the text generation process, enabling the model to learn richer contextual information.

We further define subgraphs consisting of a center triple and its related neighboring triples. All triples within each subgraph are converted into natural sentences and aggregated into a single document, forming a semantically coherent and logically structured unit of knowledge. A detailed example of this subgraph-to-document transformation, including the rule-based sentence structure and the resulting document format, is provided in Appendix B.

3.1.3 Pseudo Query Generation for Fine-Tuning Medical Search System

In our framework, document-form subgraph encoder and reranker models pre-trained on general domain data are not sufficient to accurately retrieve medical information grounded in a UMLS-based KG. To improve their ability to understand and retrieve UMLS-specific representations, these models should be fine-tuned on domain-specific data. However, manually constructing high-quality query-document pairs is impractical and costly. To address this, we propose an automated pipeline based on frozen Llama3.1$_{8B}$ (Dubey et al., 2024) models that generates and filters training data without human supervision. The pipeline consists of two core components: a pseudo query generator, which produces queries reflecting key contents of each document, and an evaluator, which filters these queries based on two criteria, patient centeredness ([`Patient`/`notPatient`]) and document relevance ([`Relevant`/`Irrelevant`]).

As illustrated in Figure 2, the system generates multiple candidate queries per document, evaluates them, and filters those that meet the training standards. Although the evaluator operates in a zero-shot setting without parameter updates, it consistently selects high-quality query-document pairs and generalizes well across unseen pairs. These filtered pairs are subsequently used to fine-tune the document-form subgraph encoder and reranker models, contributing to improved retrieval accuracy and consistency. Details on the objective functions used for each encoder and reranker models are provided in Appendix C. Furthermore, while our pipeline focuses on medical consultation documents in this study, it can be easily adapted to other domains by adjusting the evaluation criteria.

3.1.4 Document Embedding and Vector Retrieval

We fine-tuned the bge-large-en-v1.5 (Xiao et al., 2024) model to generate embeddings for documents derived from the subgraph, optimizing its ability to capture semantic nuances. Details of

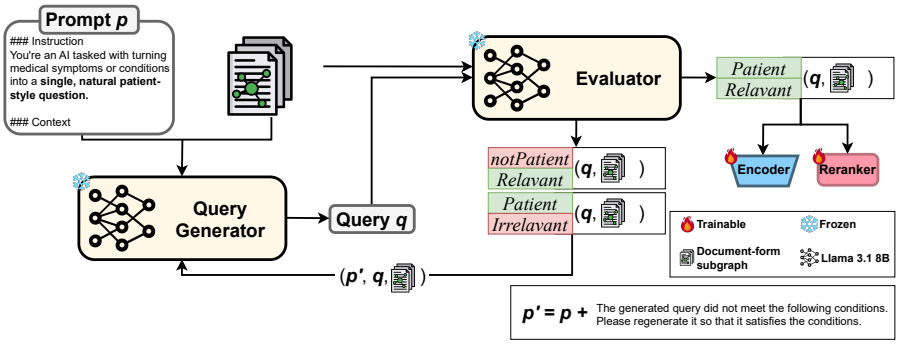

Figure 2: Overview of training data generation process for the document-form subgraph encoder and reranker. A fixed query generator creates questions that a patient is likely to ask, and an evaluator checks if they match a patient-like style and are relevant to the document. If they don't meet the criteria, they are regenerated. The final data trains *bge-large-en v1.5* and *bge-reranker-large*, enhancing the model's ability to understand and process patient-oriented queries.

the fine-tuning process are provided in Appendix C.1 for reference. These embeddings are stored in a vector database using FAISS (Johnson et al., 2019), which is optimized for large-scale similarity searches. By integrating maximum inner product search (Shrivastava and Li, 2014), we efficiently retrieve relevant documents, ensuring low-latency and high-precision results, crucial for real-time applications like conversational agents.

3.2 Reranker: Filtering for Exact Knowledge

To enhance the accuracy of retrieved documents, we employ a reranking process using the fine-tuned *bge-reranker-large*[4] model. The cross-encoder jointly encodes the user query and candidate documents to capture fine-grained interactions and assigns relevance scores. We then reorder candidates to prioritize contextually appropriate knowledge for generation, injecting precise and clinically relevant evidence into the final answer. This compensates for the retriever's limited precision and is crucial in medical settings to ensure reliability and safety. In practice, we keep the top 10 documents after reranking and discard low-confidence items with a calibrated score threshold, which yields consistent gains in precision and reduces hallucinations. Detailed training procedures and loss functions are provided in Appendix C.2.

3.3 Generator: Generating Patient-Centered Medical Consultation

In the final stage, we generate medically accurate and context-aware responses using reranked documents. We fine-tune $\text{Llama2}_{\text{7B}}$ (Touvron et al., 2023) and $\text{Llama3.1}_{\text{8B}}$ on real medical consultation data, allowing the model to learn associations between patient queries and retrieved knowledge. Unlike methods that rely solely on synthetic prompts, our framework uses actual consultation records with retrieved documents integrated during fine-tuning. This helps the model better understand semantic relationships between queries and supporting knowledge, grounding its generation in clinically relevant context. As a result, ILlama can deliver more accurate and tailored responses while reducing hallucinations and speculative content.

4 Experiments

4.1 Datasets

We use two types of data in ILlama, namely a UMLS-based KG and real-world medical consultation records. The KG provides structured clinical relationships that support precise retrieval, while the consultation data enables response generation grounded in authentic patient–doctor interactions. Detailed descriptions of each dataset are provided in the following subsections.

	HealthCareMagic	iCliniq
# dialogues	112,165	1,380
# tokens	27,475,545	313,735
Avg. # tokens per dialogue	245.01	227.34
Max # tokens per dialogue	2,544	1,001
Min # tokens per dialogue	78	60

Table 1: Statistics of the datasets used for training, validation, and testing, showing the distribution of dialogues and token counts.

[4] https://huggingface.co/BAAI/bge-reranker-large

Category / Model	In-Distribution F1	METEOR	BLEU-4	ROUGE-2	Top-1 Hit Rate	Avg NLI Score	Out-of-Distribution F1	METEOR	BLEU-4	ROUGE-2	Top-1 Hit Rate	Avg NLI Score
Baselines without Retrieval												
$BART_{Large}$	0.837	0.059	0.0	0.038	0.010	0.361	0.838	0.063	0.0	0.023	0.006	0.318
$T5_{Large}$	0.840	0.061	0.0	0.031	0.012	0.376	0.843	0.069	0.0	0.020	0.007	0.330
$Llama2_{7B}$ w/ LoRA	0.838	0.192	0.031	0.052	0.017	0.515	0.838	0.201	0.029	0.050	0.006	0.392
$Llama3.1_{8B}$ w/ LoRA	0.844	0.230	0.061	0.074	0.214	0.552	0.841	**0.222**	0.029	0.048	0.062	0.460
$Gemma2_{27B}$	0.842	0.207	0.027	0.043	0.219	0.521	0.846	0.213	0.026	0.042	0.037	0.441
Baselines without Fine-Tuning												
GPT-4o	0.836	0.215	0.013	0.035	0.315	0.489	0.836	0.218	0.014	0.039	0.001	0.345
$Gemma2_{9B}$	0.836	0.180	0.016	0.036	0.027	0.407	0.841	0.201	0.022	0.040	0.016	0.383
$Yi1.5_{9B}$	0.832	0.168	0.015	0.034	0.032	0.431	0.835	0.188	0.021	0.038	0.019	0.395
$Falcon3_{7B}$	0.839	0.135	0.008	0.025	0.011	0.508	0.844	0.156	0.012	0.028	0.007	0.400
$DeepSeek\text{-}R1_{8B}$	0.832	0.175	0.012	0.028	0.015	0.362	0.837	0.191	0.014	0.030	0.001	0.349
$MedGemma_{4B}$	0.845	0.213	0.013	0.033	0.059	0.463	0.841	0.196	0.015	0.031	0.025	0.413
Baselines with Fine-Tuning & Retrieval												
$Llama2_{7B}$ w/ LoRA	0.837	0.191	0.029	0.050	0.109	0.535	0.839	0.203	0.029	0.050	0.044	0.428
$Llama3.1_{8B}$ w/ LoRA	0.786	0.222	0.010	0.024	0.205	0.541	0.789	0.199	0.006	0.019	0.054	0.445
MedRAG	0.829	0.179	0.025	0.047	0.168	0.514	0.811	0.176	0.024	0.038	0.058	0.457
ChatDoctor	0.846	0.218	0.008	0.022	0.262	0.569	0.845	0.211	0.035	0.045	0.088	0.478
Ours												
$ILlama_{7B}$	0.866	0.203	0.037	0.058	0.258	0.582	0.851	0.213	**0.041**	0.048	0.153	0.540
$ILlama_{8B}$	**0.884**	**0.231**	**0.063**	**0.075**	**0.793**	**0.633**	**0.871**	**0.222**	0.030	**0.049**	**0.800**	**0.503**
$IGemma_{27B}$	0.897	0.245	0.077	0.081	0.816	0.661	0.881	0.234	0.052	0.056	0.822	0.587

Table 2: Performance comparison across baselines categorized into three groups: without retrieval, without fine-tuning, and with fine-tuning & retrieval. Metrics include F1 score, METEOR, BLEU, ROUGE, and MinosEval-based Top-1 Hit Rate and Avg NLI Score for both in-distribution and out-of-distribution datasets. The highlighted row represents our proposed method, demonstrating superior performance across most metrics.

4.1.1 Datasets for Medical Retrieval

We construct our KG using the 2024 release of the UMLS Metathesaurus[5], which comprises approximately 20K entities, 22 relation types, and 250K triples. This structured resource provides a semantic backbone for our RAG framework, enabling precise retrieval and integration of clinically relevant knowledge. Grounding generation in this KG enhances factual accuracy, reduces hallucinations, and supports context-aware medical responses.

4.1.2 Datasets for Medical Consultation

To evaluate the performance of ILlama, we used medical consultation records from two real-world platforms: HealthCareMagic and iCliniq. The HealthCareMagic dataset, specifically collected for medical question answering tasks, consists of single-turn interactions where each patient query is paired with a response from a licensed medical professional. We split this dataset into training, validation, and in-distribution test sets using an 8:1:1 ratio. In contrast, the iCliniq dataset, which follows a similar single-turn format, was used exclusively as an out-of-distribution test set. This separation allows us to evaluate the model's generalization performance on unseen queries from a different source, minimizing the risk of data leakage and ensuring a fair comparison. Both datasets are publicly available for academic research and have been de-identified to protect user privacy. As these records often include patient-reported details such as age and symptoms, the model implicitly learns to adapt responses to clinical contexts during fine-tuning. Detailed dataset statistics are provided in Table 1.

[5] https://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html

4.2 Metrics

We evaluated our model using semantic and quantitative metrics to assess the accuracy and contextual appropriateness of generated responses. For semantic evaluation, we adopted BERTScore (Zhang* et al., 2020) with $RoBERTa_{Large}$ (Liu et al., 2019), which measure contextual similarity using deep embeddings. This approach is particularly suitable for handling the nuances of medical language where lexical overlap is often limited (Hanna and Bojar, 2021). We report the F1 score from BERTScore as our primary semantic similarity metric.

In addition to semantic evaluation, we employed lexical metrics including ROUGE-2 (Lin, 2004), BLEU-4 (Papineni et al., 2002), and METEOR to evaluate lexical accuracy, fluency, and coherence. METEOR considers synonymy, stemming, and paraphrasing for sentence-level similarity assessment. To address the limitations of traditional metrics for open-ended medical consultation tasks, we adopt MinosEval (Fan et al., 2025), a framework for evaluating medical question answering. For non-factoid queries, we use Top-1 Hit Rate where GPT-4o ranks all model responses, with a model receiving a hit if ranked first. For fac-

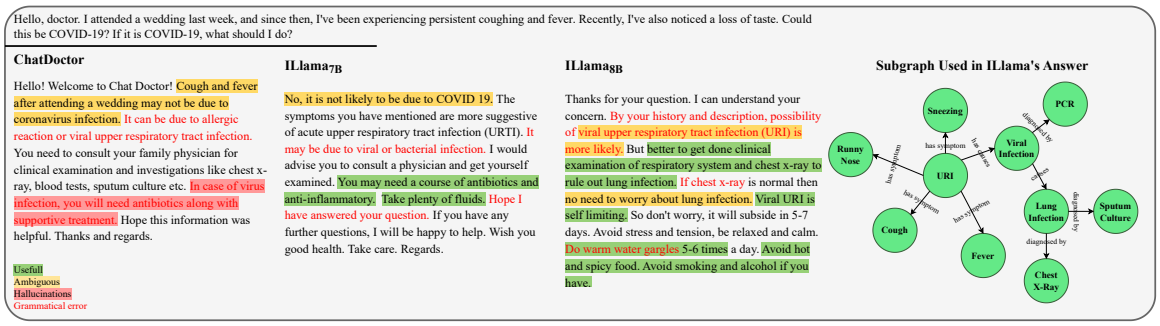

Figure 3: Comparison of responses from ChatDoctor, ILlama$_{7B}$, and ILlama$_{8B}$ regarding COVID-19 symptoms. Highlighted sections indicate usefulness, ambiguity, hallucinations, and grammatical errors.

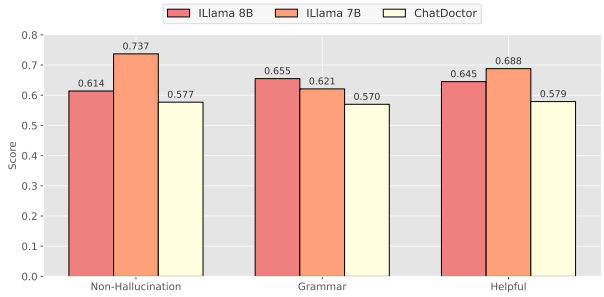

Figure 4: Out-of-distribution evaluation on the iCliniq dataset using OpenAI o1. The proportion of outputs evaluated as non-hallucinated, grammatically correct, or helpful for ILlama$_{8B}$, ILlama$_{7B}$, and ChatDoctor.

toid queries, we employ Average NLI Score using a ClinicalBERT$_{Base}$ (Huang et al., 2019) NLI model fine-tuned on MedNLI to evaluate whether responses entail, contradict, or are neutral to clinical key points in reference answers. This framework enables assessment of the model's ability to deliver precise, fluent, and contextually relevant medical responses across different question types.

4.3 Baselines

Baselines without Retrieval These models rely solely on fine-tuned capabilities on domain-specific data, without any retrieval mechanism. We examine BART$_{Large}$ (Lewis et al., 2020a), T5$_{Large}$ (Raffel et al., 2020), Llama2$_{7B}$ w/ LoRA (Hu et al., 2022), Llama3.1$_{8B}$ w/ LoRA, and Gemma2$_{27B}$ w/ LoRA (Team et al., 2024).

Baselines without Fine-Tuning Models in this category use a retrieval mechanism but are not fine-tuned on domain-specific data. These baselines enhance their performance by leveraging the PubMed dataset (Xiong et al., 2024) for retrieval of pertinent biomedical literature, which provides a rich source of domain-specific information without the need for additional fine-tuning. These include Gemma2$_{9B}$, Yi1.5$_{9B}$ (Young et al., 2024), Falcon3$_{7B}$ (Team, 2024), DeepSeek-R1$_{8B}$ (DeepSeek-AI et al., 2025), and MedGemma$_{4B}$ (Sellergren et al., 2025).

Baselines with Fine-Tuning & Retrieval This category includes models that undergo fine-tuning on domain-specific data and use retrieval. Models include Llama2$_{7B}$ w/ LoRA (Hu et al., 2022), Llama3.1$_{8B}$ w/ LoRA, MedRAG(Zhao et al., 2025), and ChatDoctor, although ChatDoctor and MedRAG do not use PubMed for retrieval.

4.4 Result

As shown in Table 2, ILlama consistently outperformed the baselines across both in-distribution and out-of-distribution evaluations. In the in-distribution setting, ILlama$_{8B}$ achieved the best F1 score of 0.884, METEOR of 0.231, and exceptional MinosEval performance with Top-1 Hit Rate of 0.793 and Avg NLI Score of 0.633, surpassing all baseline models. Notably, ILlama$_{7B}$ also showed strong performance (F1 score: 0.866, METEOR: 0.203), outperforming ChatDoctor across all major metrics. Our method also consistently outperforms MedRAG, as MedRAG is tailored for clinical decision support using structured EHR data, which may not align well with open-domain patient consultation tasks that our system is optimized for. To assess applicability beyond smaller backbones, we further instantiated our framework with Gemma2$_{27B}$, which yielded additional gains, indicating that the proposed pipeline continues to benefit from increased model capacity.

In the out-of-distribution setting, ILlama maintained robust generalization performance. ILlama$_{8B}$ achieved an F1 score of 0.871, METEOR of 0.222, and Top-1 Hit Rate of 0.800,

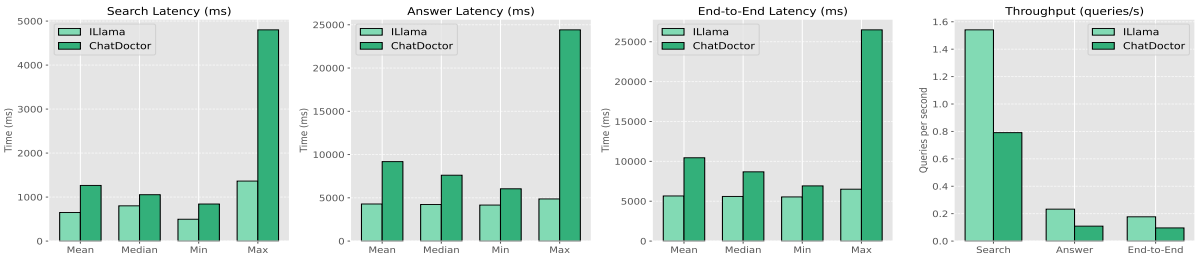

Figure 5: ILlama and ChatDoctor performance comparison in search latency, answer latency, end-to-end latency and throughput.

with minimal performance drop compared to its in-distribution results. This demonstrates ILlama's ability to adapt to unseen queries and linguistic variations from different data sources. The integration of embedding-based vector retrieval and a structured medical KG played a key role in improving factual consistency while minimizing hallucinations. As in the in-distribution results, the framework also scales to larger backbones with further improvements under distribution shift. Overall, ILlama achieved state-of-the-art performance across traditional and MinosEval-based metrics, validating its reliability and generalization in real-world medical consultation scenarios.

5 Analysis

5.1 Qualitative Analysis of Outputs

As shown in Figure 3, we present a qualitative comparison of ChatDoctor, ILlama$_{7B}$, and ILlama$_{8B}$ in response to a COVID-19 related query, alongside the underlying reasoning represented through contextualized subgraphs extracted from the UMLS-based KG. While ChatDoctor exhibited frequent hallucinations, such as recommending antibiotics for viral infections, ILlama$_{7B}$ demonstrated improved clinical reasoning but still included unnecessary suggestions. ILlama$_{8B}$ provided the most balanced response, delivering accurate medical guidance and appropriate follow-up steps. These evaluations were conducted using the OpenAI o1 model[6] (Liu et al., 2023). The prompts used for this assessment are provided in Appendix D. The reasoning process is grounded in causal and diagnostic relationships (e.g., Viral Infection–URI–Cough/Fever or Chest X-Ray to rule out Lung Infection), captured within the subgraph structure.

We also conduct a standardized evaluation by assessing all model outputs on the iCliniq out-of-distribution set using the same OpenAI o1–based judging protocol. Figure 4 presents aggregate proportions for non-hallucination, grammar, and helpfulness. For ChatDoctor, the judge frequently flags hallucinations and reports lower helpfulness, suggesting brittle grounding under distribution shift. In contrast, ILlama$_{8B}$ closely aligns with clinically appropriate phrasing and attains the highest non-hallucination and helpfulness rates.

5.2 Latency Analysis in Real-Time Medical Consultation Systems

In Figure 5, we present a comparison of latency and throughput between ILlama and ChatDoctor. ILlama consistently demonstrates lower latency across search, answer, and end-to-end processing. For example, ILlama's end-to-end latency ranges from approximately 5,538 to 6,507 ms, whereas ChatDoctor's ranges from around 6,921 to over 26,491 ms. This gap stems from ChatDoctor's reliance on LLM-based keyword extraction followed by live API calls to external sources such as Wikipedia, which markedly inflate the answer-latency portion of the overall response time. In contrast, ILlama uses preindexed graph-based retrieval with documents averaging around 130 tokens, enabling higher throughput with reduced delay.

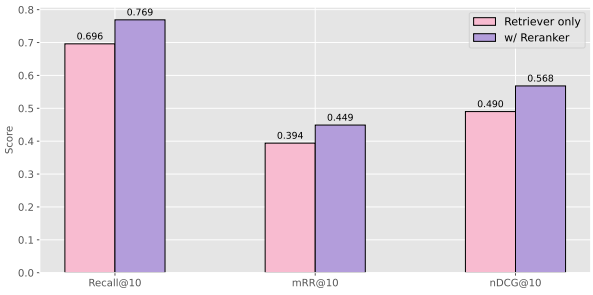

Figure 6: Retrieval performance comparison between retriever-only and retriever with reranker, evaluated on Recall@10, mRR@10, and nDCG@10.

[6] https://openai.com/o1/

5.3 Ablation Study

Retrieval Performance Analysis As shown in Figure 6, we compared the retriever alone with the retriever–reranker pipeline using Recall@10, mRR@10, and nDCG@10, which are commonly used to assess the effectiveness of retrievers in RAG-based open-domain question answering (Jin et al., 2023). The retriever already provided strong coverage, demonstrating the effectiveness of our document-form subgraph and vector search approach. The reranker further enhanced the ranking quality while maintaining high recall, confirming their complementary roles and directly addressing the need for retrieval evaluation.

Component and Knowledge Contributions Table 3 shows that the reranker consistently improves final F1 scores, while structured knowledge from UMLS provides clear advantages over unstructured PubMed text. The reranker's gains are moderate, likely because the retriever already performs strongly; however, in the biomedical domain its role lies in refining rankings to promote clinically meaningful subgraphs and ensure contextual precision, suggesting that even small improvements can yield more reliable outputs.

Model	Retriever	Reranker	F1
$\text{ILlama}_{\text{8B \& UMLS}}$	✓	✓	0.884
$\text{ILlama}_{\text{8B \& UMLS}}$	✓	✗	0.857
$\text{ILlama}_{\text{8B \& PubMed}}$	✓	✓	0.786
$\text{ILlama}_{\text{8B \& PubMed}}$	✓	✗	0.765
$\text{ILlama}_{\text{8B \& None}}$	✗	✗	0.844

Table 3: Component and knowledge-source ablation on F1 score.

Full-Graph vs. Subgraph Search. We compared subgraph-based retrieval using Triple2Seq with a full-graph approach. As shown in Figure 7, the subgraph method yields more accurate responses. Unlike full-graph retrieval, which often introduces loosely connected or irrelevant nodes, subgraph retrieval focuses on a central medical concept and its related neighbors. This targeted context improves factual accuracy and reduces noise, which is especially critical in the medical domain.

6 Conclusion

In this study, we proposed ILlama, a retrieval-augmented medical consultation framework that integrates structured KGs and cross-encoder reranking to reduce hallucinations. Experiments demonstrate strong performance across accuracy, latency, and contextual relevance, highlighting the promise of structured medical knowledge in LLMs for scalable healthcare applications.

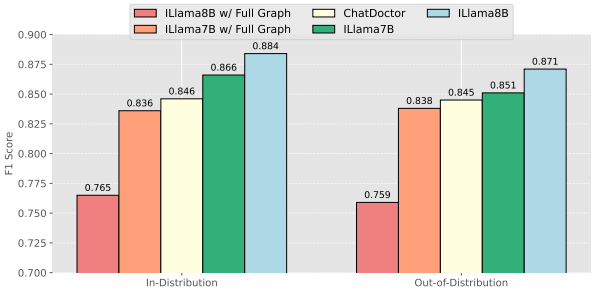

Figure 7: F1 score comparison between subgraph and full graph reasoning.

Limitations

Although ILlama improves the accuracy of medical consultations and reduces hallucinations by leveraging a UMLS-based KG and embedding-based retrieval, several limitations remain. First, the coverage of the KG and datasets is narrow, focusing on specific diseases and linguistic patterns, which limits generalizability to broader clinical domains and multilingual contexts. Expanding training data with more diverse medical cases would improve robustness. Second, the KG may not reflect the latest clinical updates such as emerging diseases, treatments, or revised guidelines, and without real-time synchronization the model risks producing outdated responses. Third, while ILlama shows strong performance on metrics like F1 score, METEOR, and MinosEval, these do not fully capture clinical safety or decision-making validity, underscoring the need for automated or simulated clinical evaluations. Finally, reliance on large-scale models restricts deployment in resource-limited settings; developing lightweight and multilingual variants could enable broader adoption in global healthcare.

Ethical Considerations

While ILlama aims to enhance medical consultations by reducing hallucinations, it may occasionally provide incorrect diagnoses that could lead to serious consequences; therefore, users should always seek guidance from qualified healthcare professionals, with AI models serving only as supplementary resources.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. This work was partly supported by the Basic Research Program through a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No.2022R1C1C1010317) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2025-25443681).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Flora Amato, Stefano Marrone, Vincenzo Moscato, Gabriele Piantadosi, Antonio Picariello, Carlo Sansone, and 1 others. 2017. Chatbots meet ehealth: Automatizing healthcare. In *WAIAH@ AI* IA*, pages 40–49.

Zhen Bi, Siyuan Cheng, Jing Chen, Xiaozhuan Liang, Feiyu Xiong, and Ningyu Zhang. 2024. Relphormer: Relational graph transformer for knowledge graph representations. *Neurocomputing*, 566:127044.

Bolin Cao, Wensen Huang, Naipeng Chao, Guang Yang, and Ningzheng Luo. 2022. Patient activeness during online medical consultation in china: multilevel analysis. *Journal of Medical Internet Research*, 24(5):e35557.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, and 1 others. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yongqi Fan, Yating Wang, Guandong Wang, Zhai Jie, Jingping Liu, Qi Ye, and Tong Ruan. 2025. MinosEval: Distinguishing factoid and non-factoid for tailored open-ended QA evaluation with LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10517–10548, Vienna, Austria. Association for Computational Linguistics.

Yanjun Gao, Ruizhe Li, Emma Croxford, John Caskey, Brian W Patterson, Matthew Churpek, Timothy Miller, Dmitriy Dligach, and Majid Afshar. 2025. Leveraging medical knowledge graphs into large language models for diagnosis prediction: Design and application study. *JMIR AI*, 4:e58670.

Tomas Goldsack, Zhihao Zhang, Chen Tang, Carolina Scarton, and Chenghua Lin. 2023. Enhancing biomedical lay summarisation with external knowledge graphs. *arXiv preprint arXiv:2310.15702*.

Quan Guo, Shuai Cao, and Zhang Yi. 2022. A medical question answering system using large language models and knowledge graphs. *International Journal of Intelligent Systems*, 37(11):8548–8564.

Muhammad Usman Hadi, Qasem Al Tashi, Abbas Shah, Rizwan Qureshi, Amgad Muneer, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, and 1 others. 2024. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*.

Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of bertscore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Chin-Yuan Huang, Ming-Chin Yang, Chin-Yu Huang, Yu-Jui Chen, Meng-Lin Wu, and Kai-Wen Chen. 2018. A chatbot-supported smart wireless interactive healthcare system for weight control and health promotion. In *2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pages 1791–1795.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2023. InstructoR: Instructing unsupervised conversational dense retrieval with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6649–6675, Singapore. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jianning Li, Amin Dada, Behrus Puladi, Jens Kleesiek, and Jan Egger. 2024. Chatgpt in healthcare: a taxonomy and systematic review. *Computer Methods and Programs in Biomedicine*, page 108013.

Mufei Li, Siqi Miao, and Pan Li. 2025. Simple is effective: The roles of graphs and large language models in knowledge-graph-based retrieval-augmented generation. In *The Thirteenth International Conference on Learning Representations*.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Qiuhao Lu, Dejing Dou, and Thien Nguyen. 2022. Clinicalt5: A generative language model for clinical text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5436–5443.

Hang Luo, Jian Zhang, and Chujun Li. 2025. Causal graphs meet thoughts: Enhancing complex reasoning in graph-augmented llms. *arXiv preprint arXiv:2501.14892*.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Ariana Martino, Michael Iannelli, and Coleen Truong. 2023. Knowledge injection to counter large language model (llm) hallucination. In *European Semantic Web Conference*, pages 182–185. Springer.

Ritwik Mishra, Simranjeet Singh, Jasmeet Kaur, Pushpendra Singh, and Rajiv Shah. 2023. Hindi chatbot for supporting maternal and child health related queries in rural India. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 69–77, Toronto, Canada. Association for Computational Linguistics.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-HALT: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–334, Singapore. Association for Computational Linguistics.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nudtaporn Rosruen and Taweesak Samanchuen. 2018. Chatbot utilization for medical consultant system.

In *2018 3rd Technology Innovation Management and Engineering Science International Conference (TIMES-iCON)*, pages 1–5.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Anshumali Shrivastava and Ping Li. 2014. Asymmetric lsh (alsh) for sublinear time maximum inner product search (mips). *Advances in neural information processing systems*, 27.

Falcon-LLM Team. 2024. The falcon 3 family of open models.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. 2023. Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Raju Vaishya, Anoop Misra, and Abhishek Vaish. 2023. Chatgpt: Is this version good for healthcare and research? *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 17(4):102744.

Deeksha Varshney, Aizan Zafar, Niranshu Kumar Behera, and Asif Ekbal. 2023. Knowledge grounded medical dialogue generation using augmented graphs. *Scientific Reports*, 13(1):3310.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 641–649, New York, NY, USA. Association for Computing Machinery.

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

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Mentallama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 4489–4500.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, and 1 others. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. BioBART: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109, Dublin, Ireland. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. 2025. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 4442–4457, New York, NY, USA. Association for Computing Machinery.

A Implementation Details

This study fine-tuned a medical domain-specific model using LoRA with a configuration of $r = 16$, $lora\_alpha = 16$, and $lora\_dropout = 0$. The learning rate started at 2×10^{-5}, with 10% of the total steps dedicated to warmup. A linear scheduler was used for adjusting the learning rate during training. The model was trained for 3 epochs, and the maximum sequence length was set to 4,096 for handling complex queries, and the training was conducted on two NVIDIA RTX A6000 48GB GPUs.

For retrieval, 50 documents were fetched using maximum inner product search from the FAISS vector store. The top 10 documents from this set were selected for final use after reranking. This approach improved the model's ability to address medical queries by leveraging dense retrieval methods, enhancing both retrieval accuracy and response quality.

B Example Document from UMLS Subgraph

The UMLS is a comprehensive biomedical knowledge base that integrates over a million concepts and multi-million relationships from more than 100 controlled vocabularies (including MeSH, SNOMED CT, RxNorm, ICD-10, etc.), along with an accompanying semantic network and lexical tools to ensure interoperability and accurate concept mapping. UMLS is updated regularly, which aligns well with common retraining cycles. Therefore, instead of frequently retraining the language model, it is more efficient to update the KG, enabling practical and timely integration of new medical information.

Table 4 presents an example of subgraph-to-text conversion used in our system. The subgraph is constructed around the central triple *(Lung cancer – has symptom – fatigue)* from the UMLS-based KG. All triples connected to the central node are included and expressed as simple natural language sentences using a rule-based template. Each relation type (e.g., *has symptom*, *diagnosed by*, *treated by*) is mapped to a consistent sentence pattern, such as "*X has symptom Y*" or "*X can be diagnosed by Y.*" This consistency facilitates automatic transformation and retrieval in downstream components. The resulting document serves as a structured and semantically coherent unit of medical knowledge for training and inference.

C Implementation details of retriever and reranker

C.1 Document-form Subgraph Encoder

We fine-tuned the *bge-large-en-v1.5* model to generate embeddings for documents derived from the subgraph, optimizing its ability to capture semantic nuances. The model was trained for 10 epochs with a batch size of 32, using the AdamW optimizer with a learning rate of 1×10^{-5}. The encoder is trained with the InfoNCE loss (Oord et al., 2018), which is a contrastive learning objective widely used in self-supervised learning. Given a set of N random samples $X = \{x_1, \ldots, x_N\}$ containing one positive sample x_{t+k} from the true conditional distribution $p(x_{t+k} \mid c_t)$ and $N - 1$ negative samples drawn from a proposal distribution $p(x_{t+k})$, the loss is formulated as:

$$\mathcal{L}_N = -\mathbb{E}_X \left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)} \right],$$

where $f_k(x, c_t)$ denotes a scoring function (e.g., a dot product or similarity function) that estimates the compatibility between context c_t and future sample x. Optimizing this loss leads $f_k(x_{t+k}, c_t)$ to approximate the density ratio:

$$f_k(x_{t+k}, c_t) \propto \frac{p(x_{t+k} \mid c_t)}{p(x_{t+k})}.$$

C.2 Reranker

To compensate for the retriever's limited precision in selecting the most relevant subgraph, we employ a cross-encoder reranker that jointly encodes the input query and each retrieved candidate to assess their semantic relevance. The reranker computes a relevance score for each query–subgraph pair and reorders the top-50 candidates returned by the retriever. The 10 highest-ranked subgraphs are then selected as the final knowledge inputs to the generator. This additional reranking step is particularly important in the medical domain, where selecting the most contextually appropriate knowledge is critical for ensuring the reliability and safety of the generated output.

The reranker model adopts a cross-encoder architecture and is fine-tuned with a binary cross-entropy loss. The reranker is trained to assign high scores to gold subgraphs and low scores to irrelevant ones. Positive training examples are constructed from gold query–subgraph pairs, and negatives are sampled from the remaining retriever

outputs. This pairwise labeling allows the model to effectively learn fine-grained distinctions in contextual relevance. Given a query–document pair (q, d) and a binary label $y \in \{0, 1\}$ indicating relevance, the model predicts a scalar relevance score $\hat{y} = \text{sigmoid}(s(q, d))$, and the loss is computed as:

$$\mathcal{L}_{\text{BCE}} = -\left[y \log \hat{y} + (1 - y) \log(1 - \hat{y})\right]$$

This objective encourages the model to produce high scores for relevant documents and low scores for irrelevant ones, improving the quality of the final ranking.

D Evaluation Prompt Design

To support the qualitative evaluation of model outputs, we designed three structured prompts targeting hallucination detection, grammatical correctness, and patient helpfulness. These prompts were used with the OpenAI o1 model to evaluate responses generated by ILlama. Table 6 presents the full text of each prompt. Each includes clear task instructions, placeholders for the model-generated response, and, in the helpfulness case, the original patient question. The prompts instruct the model to make a binary decision and identify specific parts of the response when relevant.

The hallucination prompt assesses whether a response contains fabricated or unsupported information. The grammatical prompt checks for language correctness. The helpfulness prompt determines whether the response includes content that would be useful to a patient, based on the given question. Evaluations were conducted in a zero-shot setting, and the prompt design aimed to guide the model toward accurate and consistent judgments without fine-tuning. This allowed for scalable and focused assessment of clinical response quality.

E Algorithm

E.1 ILlama Algorithm

This algorithm, as shown in Algorithm 1, retrieves and reranks relevant documents for context-aware medical consultations. It combines FAISS search results, reranks them with a cross encoder, and generates a contextually accurate response using Llama, maintaining optimal performance and accuracy throughout the process.

E.2 Pseudo Query Generation Algorithm

This algorithm, as shown in Algorithm 2, generates patient-style queries and evaluates them to obtain (q, d) pairs for training the encoder and reranker if the conditions are met. Based on the input prompt and documents derived from the KG, the pseudo query generator (Llama3.1_{8B}) creates a query. The evaluator (Llama3.1_{8B}) then checks if the generated query meets the "patient-style" and "relevant" conditions. If the conditions are satisfied, the (q, d) pairs are stored for document-form subgraph encoder and reranker training; otherwise, the query is regenerated, and the evaluation process is repeated.

Subject	Relation	Object	Document-form subgraph
Lung Cancer	has symptom	Fatigue	Lung cancer has symptom fatigue.
Lung Cancer	has symptom	Shortness of Breath	Lung cancer has symptom shortness of breath.
Shortness of Breath	is symptom of	Anemia	Shortness of breath is symptom of anemia.
Fatigue	is symptom of	Anemia	Fatigue is symptom of anemia.
Lung Cancer	has symptom	Chronic Cough	Lung cancer has symptom chronic cough.
Lung Cancer	diagnosed by	Chest X-Ray	Lung cancer can be diagnosed by chest X-ray.
Lung Cancer	has cause	Smoking	Lung cancer has cause smoking.
Lung Cancer	has cause	Air Pollution	Lung cancer has cause air pollution.
Lung Cancer	treated by	Surgery	Lung cancer is treated by surgery.
Surgery	isa	lobectomy	Surgery is a lobectomy.

Table 4: Example of subgraph-to-text conversion for a document centered on the triple *(Lung cancer – has symptom – fatigue)*.

Prompting Category	Input Prompt
ILlama's prompt	You are a medical assistant specializing in providing expert consultations for medical inquiries.Your role is to deliver accurate, user-friendly medical information, clarify symptoms, explain potential medical conditions, and recommend next steps with empathy and professionalism. When formulating your response, to ensure clarity and accuracy, user-friendly answer in your response. ### Context {`context`} ### Input {`query`} ### Response

Table 5: Prompt used for ILlama inference

Algorithm 1 ILlama Algorithm for Medical Query Answering

Input:
q: User query
KG: UMLS-based KG
DB: FAISS vector database (Encoded Subgraph Documents)
$T2S$: Triple2Seq
QE: Query Encoder
CE: Cross Encoder
$Llama$: Llama Model
Output: Final response r
Step 1: UMLS-based KG Processing
$KG_{sub} \leftarrow T2S.split(KG)$
$D_{sub} \leftarrow$ Convert G_{sub} to text-based documents
Store D_{sub} in FAISS Vector Database
Step 2: Query Encoding
$q_{emb} \leftarrow QE.encode(q)$
Step 3: Retrieval & Reranking from Vector DB
$D_{top50} \leftarrow DB.retrieve(q_{emb}, k = 50)$
for each document d in D_{top50} **do**
    $s_d \leftarrow CE.score(q, d)$
end for
$D_{top10} \leftarrow$ Select top-10 documents based on s_d
Step 4: Response Generation
$input \leftarrow q + D_{top10}$
$r \leftarrow Llama.generate(input)$
Return r

Algorithm 2 Patient-Style Pseudo Query Generation and Evaluation

Input:
p: Prompt for query generation
d: Graph Document (Derived from KG)
QG: Query Generator (Llama3.1_{8B})
$Eval$: Evaluator (Llama3.1_{8B})
Output: (q, d) pairs for training Encoder and Reranker
Step 1: Generate Query
$q \leftarrow QG.generate(p, d)$
Step 2: Evaluate Query
$(s_1, s_2) \leftarrow Eval.check(q, d)$
if $s_1 ==$ Patient-Style **and** $s_2 ==$ Relevant **then**
    Store (q, d) for training Encoder and Reranker
else
    Regenerate q using QG
    Repeat from Step 2
end if
Return (q, d) pairs

Prompting Category	Input Prompt
Hallucination Evaluation	The following is a response generated by a model. Carefully read the response and evaluate whether it contains hallucinations based on logical consistency and factual accuracy. A hallucination refers to information that is fabricated or unsupported by evidence. ### Instructions - If a hallucination is found, pinpoint the exact part. - If no hallucination is found, respond with "No hallucination." ### Model Response: {`model response`} ### Evaluation:
Grammatical Error Evaluation	The following is a response generated by a model. Carefully read the response and identify any grammatical errors. ### Instructions - If grammatical errors are found, pinpoint the exact part. - If no grammatical errors are found, respond with "No grammatical errors." ### Model Response: {`model response`} ### Evaluation:
Helpful Information for Patients Evaluation	The following is a patient's question and a response generated by a model. Carefully read the response and identify any words or phrases that could be helpful to the patient. ### Instructions - Pinpoint the exact words or phrases in the model's response that are relevant to the patient's question. - If no helpful information is found, respond with "No helpful information." ### Patient's Question: {`question`} ### Model Response: {`model response`} ### Evaluation:

Table 6: Prompt for evaluation ILlama