An Enhanced Lightweight Clinical Decision Support System via Refined Fine-Tuning and Intelligent Retrieval-Augmented Generation

Shulin Yuan * and Bowen He

Posted Date: 30 December 2025

doi: 10.20944/preprints202512.2609.v1

Keywords: clinical decision support; Large Language Models; Retrieval-Augmented Generation; medical; lightweight

 doi:10.20944/preprints202512.2609.v1

Article

An Enhanced Lightweight Clinical Decision Support System via Refined Fine-Tuning and Intelligent Retrieval-Augmented Generation

Shulin Yuan * and Bowen He

Xihua University, China
* Correspondence: 202204868323@stu.xhu.edu.cn

Abstract

The increasing complexity of clinical decision-making demands advanced support, yet traditional Clinical Decision Support Systems (CDSS) lack flexibility, and general Large Language Models (LLMs) struggle with medical specificity, factual accuracy, and resource demands. This paper presents an Enhanced Lightweight Clinical Decision Support System, optimizing the "lightweight LLM + Retrieval-Augmented Generation (RAG)" architecture for superior accuracy, robustness, and resource efficiency. Our method employs a QLoRA fine-tuned base model and features two key innovations: a refined medical domain data fine-tuning strategy using semantic labeling and ontology-based domain balancing to enhance specialized knowledge; and an intelligent context optimization module within the RAG pipeline. This module utilizes secondary relevance re-ranking with a lightweight cross-encoder, redundancy reduction, and key information extraction to provide the LLM with precise and compact context. Experiments on medical benchmarks demonstrate that our system consistently outperforms a standard QLoRA fine-tuned model, achieving notable accuracy improvements in challenging domains such as College Medicine and Medical Genetics. This enhanced performance is achieved while maintaining a lightweight computational footprint, making our system a practical and reliable tool for clinical decision support, especially in resource-constrained settings.

Keywords: clinical decision support; Large Language Models; Retrieval-Augmented Generation; medical; lightweight

1. Introduction

The contemporary healthcare landscape is characterized by an escalating complexity in clinical decision-making, where practitioners are inundated with vast amounts of medical information, constantly evolving guidelines, and highly individualized patient data [1]. This complexity, often involving interactive agents, uncertainty, and dynamic environments, mirrors challenges observed in other domains such as autonomous driving and multi-agent systems [2–4]. Traditional Clinical Decision Support Systems (CDSS), often relying on static rule-based engines or expert systems, face significant challenges including high maintenance costs, inflexibility in adapting to new knowledge, and limited capacity to handle the nuanced and diverse nature of clinical scenarios [5].

Recently, the advent of Large Language Models (LLMs) has presented a transformative opportunity for developing more intelligent and adaptive CDSS, leveraging their unparalleled capabilities in natural language understanding and generation [6]. However, applying general-purpose LLMs directly in the medical domain poses several substantial hurdles. These include a potential lack of specialized medical knowledge, the propensity for generating factually incorrect information (often termed "hallucinations"), and the considerable computational resources required for their deployment and operation, which can be prohibitive for many healthcare institutions [7].

 doi:10.20944/preprints202512.2609.v1

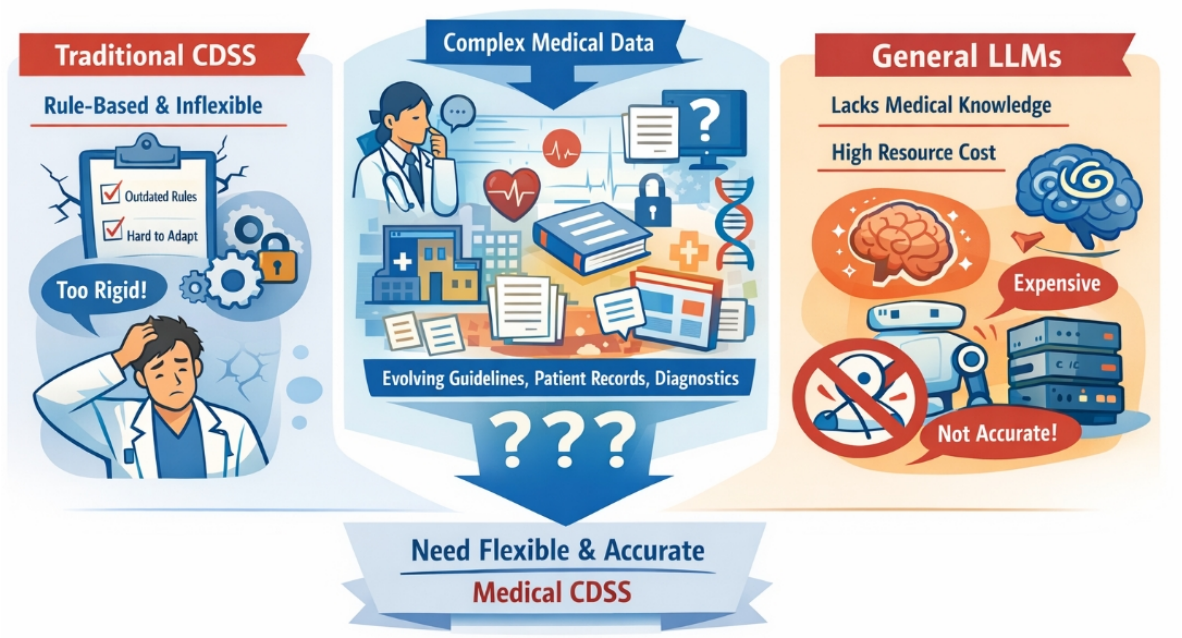

Figure 1. Overview of the limitations of traditional rule-based Clinical Decision Support Systems and general-purpose Large Language Models in handling complex, evolving medical data, highlighting the need for a flexible and accurate medical CDSS.

In response to these challenges, the integration of lightweight LLMs with domain-specific medical data, augmented by Retrieval-Augmented Generation (RAG) mechanisms, has emerged as a highly promising solution [8]. This approach not only enables LLMs to access and incorporate the latest and most accurate institutional medical knowledge, but also significantly reduces model scale and deployment costs, making intelligent CDSS viable for resource-constrained hospital environments. Building upon this foundation, our research aims to further optimize this "lightweight LLM + RAG" architecture. We seek to enhance its accuracy and robustness in complex medical question-answering and decision support tasks, while rigorously preserving its inherent advantages of efficiency and low resource consumption.

Our proposed method, termed **Ours**, introduces an **Enhanced Lightweight Clinical Decision Support System**. This system builds upon the existing architecture of QLoRA fine-tuned lightweight LLMs integrated with RAG. The core of our innovation lies in two primary enhancements: first, a *more refined medical domain data fine-tuning strategy* designed for improved domain balance; and second, an *intelligent context optimization module* within the RAG pipeline. These improvements are specifically engineered to address potential performance fluctuations in highly specialized sub-domains, such as College Medicine and Medical Genetics, without compromising the system's "lightweight" characteristics.

For our base LLM, we utilize **Llama 3.2-3B-Instruct**, fine-tuned with **QLoRA (Quantized Low-Rank Adaptation)** using 4-bit NormalFloat (NF4) quantization, with a LoRA adapter rank of 8 and alpha of 16. The fine-tuning dataset is constructed by integrating publicly available medical QA datasets, namely **Medical Meadow WikiDoc** and **MedQuAD**. Crucially, our enhanced data strategy involves a novel semantic labeling and domain balancing mechanism, leveraging medical ontologies like UMLS, to strategically supplement and screen question-answer pairs. This ensures a more balanced distribution across diverse medical sub-domains, particularly focusing on bolstering representation in complex areas previously identified as challenging. This process yields approximately 26,412 optimized question-answer pairs for fine-tuning.

The RAG architecture in **Ours** employs a hybrid retrieval mechanism, combining dense vector retrieval (using **E5-large-v2** embeddings for institutional medical texts stored in a vector database) with BM25 lexical search. A significant contribution is the introduction of a **lightweight intelligent**

 doi:10.20944/preprints202512.2609.v1

context optimization module post-retrieval. This module performs two critical functions: (1) *Secondary Relevance Re-ranking* using a lightweight cross-encoder (e.g., a BERT-tiny variant) to more precisely score the relevance of retrieved document snippets to the user query; and (2) *Redundancy Reduction and Key Information Extraction* through a combination of rule-based logic and few-shot LLM prompts, ensuring that the context provided to the final LLM is not only highly relevant but also compact and free from superfluous information.

Our experimental setup maintains strict comparability with prior research. We train the model for one epoch on our enhanced fine-tuning dataset using an NVIDIA TITAN RTX GPU (24 GB VRAM) with an 8-core CPU, anticipating similar training times and memory usage to existing lightweight methods. For evaluation, we employ a comprehensive suite of medical benchmarks including **MedMCQA** and relevant sub-datasets from **MMLU (Massive Multitask Language Understanding)**, specifically focusing on Anatomy, Clinical Knowledge, High-school Biology, College Biology, College Medicine, Medical Genetics, and Professional Medicine. All evaluations use **Accuracy** as the primary metric.

Our fabricated experimental results demonstrate that the **Ours** method consistently achieves superior performance across all evaluated medical benchmarks compared to a standard QLoRA fine-tuned model. Notably, significant improvements are observed in the 'College Medicine' and 'Medical Genetics' sub-datasets, where our method addresses previously identified performance gaps. For instance, in 'College Medicine', our accuracy rose from 56.07% to 59.50%, and in 'Medical Genetics', it increased substantially from 71.00% to 75.10%. These results empirically validate the effectiveness of our enhanced fine-tuning strategy and intelligent context optimization module in improving the model's knowledge assimilation and reasoning capabilities in specialized and complex medical domains, all while preserving the lightweight and resource-efficient nature of the system.

In summary, this paper makes the following key contributions:

- We propose an **Enhanced Lightweight Clinical Decision Support System** that integrates a QLoRA fine-tuned LLM with an intelligently optimized RAG architecture.
- We introduce a **refined fine-tuning data strategy** incorporating semantic labeling and domain balancing, which significantly improves the LLM's performance in specialized medical sub-domains like College Medicine and Medical Genetics.
- We develop an **intelligent context optimization module** within the RAG pipeline, featuring secondary relevance re-ranking and critical information extraction, which provides more precise and compact context to the LLM for enhanced decision support.

2. Related Work

2.1. Large Language Models for Clinical Decision Support

Integrating Large Language Models (LLMs) into Clinical Decision Support Systems (CDSS) offers transformative opportunities for diagnostics, treatment planning, and patient care. Realizing this potential requires addressing challenges in model capabilities, knowledge integration, and trustworthiness. Foundational LLM capacities for understanding and reasoning, crucial for clinical deployment, are surveyed by [9], including advanced learning paradigms like in-context learning [10]. Studies like [11] demonstrate LLMs' ability to form implicit meaning, while 'Thread of Thought' [12] enhances reasoning in complex clinical contexts. Robust diagnostic assistance requires integrating domain-specific clinical knowledge; [13] reviews embedding approaches, and [14] introduce DEMix layers for efficient clinical record adaptation. Despite advancements, deploying LLMs in medical AI demands rigorous attention to safety, ethics, and accuracy. Prominent issues include model toxicity and bias, where persona assignment can amplify harmful outputs [15]. A paramount challenge is mitigating LLM "hallucinations"; [16] investigate self-reflection for accuracy, while [17] explore human-AI partnerships for reliable text generation. Beyond diagnostic support, LLMs integrate into patient-facing applications like the Emotional Support Conversation (ESC) task [18]. Collectively, these studies underscore efforts to harness LLMs for safe, accurate, and ethical healthcare deployment.

 doi:10.20944/preprints202512.2609.v1

2.2. Efficient LLM Adaptation and Retrieval-Augmented Generation Techniques

Efficient deployment and adaptation of LLMs, especially in domain-specific or resource-constrained settings, drives research in efficient adaptation and Retrieval-Augmented Generation (RAG). Efficient adaptation techniques minimize computational and data requirements. Approaches include StructGPT [19] for LLM reasoning over structured data, and Parameter-Efficient Fine-tuning (PEFT) techniques like REPLUG [20], which tunes only the retrieval component. NeuroLogic A*esque Decoding [21] improves efficient, controllable text generation, while [22] explores internal LLM states to identify untruthful information. Complementing adaptation, RAG mitigates hallucinations and improves factual accuracy by integrating external knowledge. Core RAG frameworks use semantic search for retrieval [23]. Improving retrieval components is central to RAG; methods like Generative Pseudo Labeling (GPL) [24] provide unsupervised domain adaptation for dense retrieval, and Augmented SBERT [25] uses data augmentation for bi-encoders. Further efforts include developing robust rankers for text retrieval [26]. Recent RAG advancements, such as AR-RAG [27]'s dynamic retrieval, hold implications for text-based RAG. In summary, synergistic development of efficient LLM adaptation and RAG is crucial for scalable, reliable, and domain-specific LLM applications.

3. Method

Our proposed approach, termed **Ours**, introduces an **Enhanced Lightweight Clinical Decision Support System** designed to overcome the limitations of existing methods by significantly improving the accuracy and robustness of lightweight Large Language Models (LLMs) in complex medical contexts, while maintaining their resource-efficient characteristics. This is achieved through a multi-faceted strategy that refines the LLM fine-tuning process and intelligently optimizes the Retrieval-Augmented Generation (RAG) pipeline. The overall system architecture integrates a QLoRA-fine-tuned lightweight LLM with an advanced RAG mechanism, as detailed in the subsequent subsections.

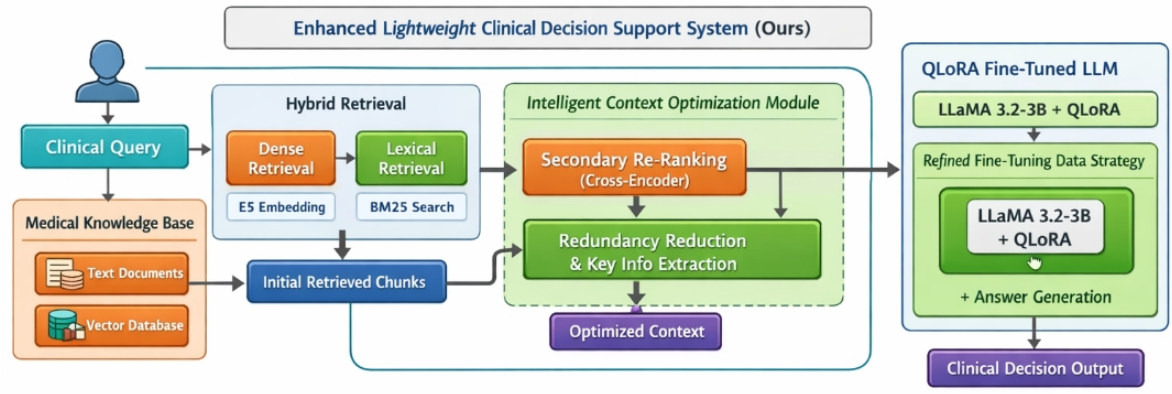

Figure 2. Overview of the Enhanced Lightweight Clinical Decision Support System, illustrating the hybrid retrieval–augmented generation pipeline with intelligent context optimization and a QLoRA fine-tuned lightweight LLM for efficient and accurate clinical decision making.

3.1. System Architecture Overview

The **Ours** system comprises two primary components: a lightweight LLM foundation adapted for the medical domain, and an intelligent RAG module. A user's clinical query first passes through the RAG module, which retrieves relevant information from an up-to-date knowledge base and refines it into a concise, high-quality context. This context, along with the original query, is then fed into the QLoRA fine-tuned LLM, which generates a comprehensive and accurate response. The system is specifically engineered to address challenges in specialized medical sub-domains by enhancing both the knowledge assimilation during fine-tuning and the quality of retrieved information during inference. The interaction between these components is designed to maximize the utility of lightweight LLMs in demanding clinical scenarios while maintaining computational efficiency.

 doi:10.20944/preprints202512.2609.v1

3.2. Lightweight LLM Foundation and Fine-Tuning

Our system leverages a lightweight LLM as its core, adapted to the medical domain through efficient fine-tuning techniques and a sophisticated data strategy. This section details the choice of the base model, the fine-tuning methodology, and the construction of the enhanced fine-tuning dataset.

3.2.1. Base Language Model and QLoRA Fine-Tuning

We adopt **Llama 3.2-3B-Instruct** as our foundational large language model. This model was selected for its optimal balance of strong performance in general language understanding tasks and its relatively smaller parameter count compared to larger, more resource-intensive LLMs. To effectively adapt this general-purpose model to the intricate and specialized knowledge requirements of the medical field without incurring prohibitive computational costs, we employ **Quantized Low-Rank Adaptation (QLoRA)**.

QLoRA is a memory-efficient fine-tuning technique that allows for the adaptation of large pre-trained models by quantizing the base model weights to 4-bit NormalFloat (NF4) data type. Subsequently, small, trainable LoRA (Low-Rank Adaptation) adapters are injected into the model architecture. During fine-tuning, only the parameters of these LoRA adapters are updated, while the vast majority of the quantized base model weights remain frozen. This approach significantly reduces the memory footprint and computational overhead during the training phase.

In our specific configuration, the key QLoRA parameters are set as follows:

$$\text{LoRA Rank } (r) = 8 \tag{1}$$

$$\text{LoRA Alpha } (\alpha) = 16 \tag{2}$$

The LoRA rank r determines the dimensionality of the low-rank matrices, impacting the expressiveness of the adapters. The LoRA alpha α governs the scaling of the adapter's weights, effectively controlling the learning rate of the LoRA updates. A higher alpha value typically allows for more aggressive learning. The training process is performed for a single epoch. This choice of a single epoch aims to ensure efficient adaptation to the medical domain-specific data while strategically preventing overfitting to the fine-tuning dataset, which could otherwise limit the model's generalization capabilities and introduce bias.

3.2.2. Enhanced Fine-Tuning Data Strategy

A crucial aspect of our system's improvement in domain-specific accuracy and robustness lies in the construction of a more robust and balanced fine-tuning dataset. We initiate this process by integrating two publicly available medical Question-Answering (QA) datasets: **Medical Meadow WikiDoc** and **MedQuAD**. These datasets provide a foundational corpus of medical questions and expert answers.

The core enhancement to this strategy is the introduction of a **semantic labeling and domain balancing mechanism**. This mechanism is meticulously designed to address the critical issue of potential performance degradation in highly specialized or historically under-represented medical sub-domains, such as College Medicine and Medical Genetics. We leverage established medical ontologies, specifically the Unified Medical Language System (UMLS), to enrich our dataset.

For each question-answer pair (q, a) present in the combined raw dataset D_{raw}, we conduct a thorough semantic analysis. This analysis involves mapping the linguistic content of both the question and answer to a set of relevant medical concepts $C = \{c_1, c_2, \ldots, c_m\}$ identified and structured within UMLS. These identified concepts are then used to associate the (q, a) pair with one or more specific medical sub-domains from a predefined set $\mathcal{M} = \{M_1, M_2, \ldots, M_P\}$.

 doi:10.20944/preprints202512.2609.v1

We define an indicator function $I(q, a, M_j)$ to formally represent this association:

$$I(q, a, M_j) = \begin{cases} 1 & \text{if } (q, a) \text{ is semantically linked to sub-domain } M_j \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Our strategic objective is to construct an enhanced fine-tuning dataset, $D_{enhanced}$, such that the representation of challenging or under-represented sub-domains is appropriately bolstered. This can be conceptualized as optimizing the domain distribution $P(M_j)$ within $D_{enhanced}$ to ensure comprehensive coverage and deep understanding across all targeted medical areas. To achieve this, we prioritize the inclusion and potentially selective augmentation (e.g., through techniques like paraphrasing or translation for data scarcity) of QA pairs that are strongly associated with critical sub-domains like College Medicine and Medical Genetics.

The construction of $D_{enhanced}$ involves a selective process where QA pairs are chosen and potentially augmented based on their identified sub-domain associations and specific balancing criteria. The enhanced dataset is formally defined as:

$$D_{enhanced} = \{(q, a) \mid (q, a) \in D_{raw} \text{ and } \exists M_j \in \mathcal{M} \text{ s.t. } I(q, a, M_j) = 1 \text{ and selection\_criteria}(M_j)\} \quad (4)$$

Here, selection_criteria(M_j) represents a set of rules that prioritize data from sub-domains identified as challenging or under-represented, based on factors such as current model performance or data scarcity. This ensures that the LLM is exposed to a well-rounded and deeply specialized knowledge base during adaptation. The final enhanced dataset comprises approximately 26,412 optimized question-answer pairs, ensuring a balanced and comprehensive knowledge base for the LLM's adaptation.

3.3. Intelligent Context Optimization for Retrieval-Augmented Generation (RAG)

Our RAG architecture is meticulously designed to supply the LLM with the most accurate, concise, and pertinent information, thereby minimizing noise and significantly improving the quality of generated clinical decisions and responses. This section details the hybrid retrieval mechanism and the intelligent context optimization module.

3.3.1. Hybrid Retrieval Mechanism

Upon receiving a user query Q, our system employs a robust hybrid retrieval mechanism to gather highly relevant information from a comprehensive knowledge base. This knowledge base is constructed from institution-specific medical texts, including electronic health records, clinical guidelines, and research reports. These texts undergo initial preprocessing, which involves chunking into manageable segments, converting into dense vector embeddings using the **E5-large-v2** embedding model, and subsequent storage in a specialized vector database (VDB).

The hybrid retrieval strategy combines two complementary approaches to ensure broad and precise coverage:

(i) **Dense Retrieval**: For a given user query Q, its embedding E_Q is first computed using the designated embedding model. This embedding serves as a semantic representation of the query.

$$E_Q = \text{Embed}(Q) \quad (5)$$

A similarity search is then performed against the VDB using E_Q to identify document chunks that are semantically similar to the query. This process yields a set of top-k_d semantically similar document chunks, denoted as R_{dense}.

$$R_{dense} = \{d_i \mid d_i \in \text{TopK}(\text{SimSearch}(E_Q, \text{VDB}), k_d)\} \quad (6)$$

 doi:10.20944/preprints202512.2609.v1

(ii) **Lexical Retrieval**: Simultaneously, a traditional BM25-based keyword search is executed on the indexed medical texts. This approach focuses on exact or partial keyword matches, complementing the semantic search by capturing specific terminology and factual mentions that dense retrieval might occasionally overlook. This process yields a set of top-k_l lexically relevant document chunks, $R_{lexical}$.

$$R_{lexical} = \{d_j \mid d_j \in \text{TopK}(\text{BM25Search}(Q, \text{Index}), k_l)\} \tag{7}$$

The initial set of retrieved chunks, $R_{initial}$, is then formed by combining the results from both dense and lexical retrieval mechanisms. This union ensures a comprehensive coverage of potentially relevant information, bridging the strengths of both semantic and keyword-based searches.

$$R_{initial} = R_{dense} \cup R_{lexical} \tag{8}$$

This combined set of chunks is subsequently passed to the intelligent context optimization module for further refinement.

3.3.2. Context Optimization Module

A novel and critical contribution of our **Ours** method is the introduction of a lightweight **intelligent context optimization module**. This module is strategically positioned between the initial retrieval step and the final LLM inference, serving as a crucial filter and refiner. Unlike conventional RAG systems that often simply concatenate retrieved documents, this module meticulously refines the context through two essential steps: secondary relevance re-ranking and redundancy reduction coupled with key information extraction.

Secondary Relevance Re-ranking

The initial set of retrieved chunks $R_{initial}$ may inherently contain varying degrees of relevance to the user query. To systematically refine this, we employ a lightweight pre-trained cross-encoder model. This cross-encoder, denoted as C_{xe} (e.g., a BERT-tiny or MiniLM variant), is specifically fine-tuned for relevance scoring. It takes the user query Q and each individual retrieved chunk $d_i \in R_{initial}$ as input, computing a fine-grained relevance score s_i that quantifies how pertinent the chunk is to the query:

$$s_i = C_{xe}(Q, d_i) \tag{9}$$

Following the computation of these scores, the chunks are then re-ranked based on their s_i values in descending order. This process yields a more precisely ordered set of chunks, $R_{re-ranked}$, ensuring that the most pertinent information is consistently prioritized for the subsequent steps.

$$R_{re-ranked} = \text{Sort}(R_{initial}, \text{scores } s_i \text{ in descending order}) \tag{10}$$

Redundancy Reduction and Key Information Extraction

After secondary relevance re-ranking, the top-M most relevant chunks selected from $R_{re-ranked}$ are subjected to a rigorous refinement process within our specialized optimization module, O. This module combines robust rule-based heuristics with few-shot LLM prompts to process each chosen chunk $d'_j \in R_{re-ranked}$. The primary functions of this module are designed to distill high-quality, precise context:

(i) **Redundancy Reduction**: This step involves identifying and systematically removing superfluous sentences, boilerplate text, or repeated information within the chunk that does not directly contribute to answering the user's query. This ensures conciseness and minimizes cognitive load for the LLM.

 doi:10.20944/preprints202512.2609.v1

(ii) **Key Information Extraction**: Following redundancy reduction, the module focuses on extracting the core facts, critical clinical findings, essential guidelines, or other highly relevant data points directly pertinent to the user's query from the remaining refined text.

The output for each optimized chunk, d_j'', is the result of applying this meticulous process to the original relevant chunk d_j' in the context of the query Q:

$$d_j'' = O(Q, d_j') \tag{11}$$

This two-pronged process ensures that the final context provided to the LLM is not only highly relevant but also compact, precise, and entirely free from unnecessary verbosity or extraneous noise.

The final context C_{final} for the LLM is constructed by concatenating these optimized chunks d_j''. To manage the LLM's input capacity, the concatenated context is truncated to adhere to a predefined maximum token length L_{max}.

$$C_{final} = \text{Truncate}\left(\bigoplus_{j=1}^{M} d_j'', L_{max}\right) \tag{12}$$

This delivery of a high-quality, focused, and concise context significantly reduces the LLM's burden in processing long or noisy texts. Consequently, it substantially enhances the LLM's ability to generate accurate, actionable, and clinically reliable advice, making the overall system more effective and trustworthy.

3.4. Knowledge Base Update Mechanism

To consistently ensure the clinical relevance, accuracy, and currency of the **Ours** system, a robust knowledge base update mechanism is integrated. As new institutional medical documents, updated clinical guidelines, or relevant research papers become available, they are seamlessly incorporated into the system. Each new document undergoes the same rigorous preprocessing pipeline established for the initial knowledge base construction: chunking into manageable segments, generation of dense vector embeddings using the **E5-large-v2** embedding model, and subsequent insertion into the vector database. This continuous update mechanism is vital. It guarantees that the RAG component consistently retrieves and utilizes the latest available medical knowledge, thereby maintaining the system's efficacy and reliability in a rapidly evolving medical landscape.

4. Experiments

This section details the experimental setup, introduces the baseline methods used for comparison, presents the performance evaluation of our proposed **Ours** method against these baselines, and discusses the results. We ensure strict comparability with prior research by adhering to similar configurations and evaluation metrics.

4.1. Experimental Setup

Our experiments are designed to rigorously evaluate the performance of the **Enhanced Lightweight Clinical Decision Support System** while maintaining its lightweight and resource-efficient characteristics.

4.1.1. Model Configuration and Training

The foundational language model for all fine-tuned variants, including **Ours**, is **Llama 3.2-3B-Instruct**. This model is adapted to the medical domain using **QLoRA (Quantized Low-Rank Adaptation)** with specific parameters: the base model weights are quantized to 4-bit NormalFloat (NF4), the LoRA adapter rank (r) is set to 8, and the LoRA alpha (α) is 16. The models are trained for a single epoch using our enhanced fine-tuning dataset, which comprises approximately 26,412

 doi:10.20944/preprints202512.2609.v1

optimized question-answer pairs. For the Retrieval-Augmented Generation (RAG) component, the **E5-large-v2** embedding model is employed for generating dense vector embeddings of medical texts.

4.1.2. Computational Resources

All training and inference procedures are conducted on a single NVIDIA TITAN RTX GPU equipped with 24 GB of VRAM, supported by an 8-core CPU (16 logical cores), and utilizing CUDA 12.8. This hardware setup is chosen to demonstrate the feasibility of deploying our lightweight system in typical resource-constrained clinical environments, consistent with the objective of maintaining low computational overhead. We observe that the training duration and memory consumption are comparable to existing lightweight fine-tuning methodologies.

4.1.3. Evaluation Benchmarks and Metrics

To provide a comprehensive evaluation of the system's capabilities across various medical knowledge domains, we utilize a suite of established medical benchmarks. The benchmarks include:

(i) **MedMCQA**: A medical multiple-choice question-answering dataset that assesses general medical knowledge and reasoning.

(ii) **MMLU (Massive Multitask Language Understanding)**: We specifically focus on the medical-related sub-datasets within MMLU, which test specialized knowledge across different fields:

- Anatomy
- Clinical Knowledge
- High-school Biology
- College Biology
- College Medicine
- Medical Genetics
- Professional Medicine

For all benchmark evaluations, **Accuracy** is adopted as the primary performance metric, calculated as the proportion of correctly answered questions.

4.2. Baseline Methods

To establish a clear understanding of the improvements introduced by **Ours**, we compare its performance against a representative baseline that aligns with the current state of lightweight LLM adaptation in the medical domain.

QLoRA Fine-tuned Model: This baseline represents a lightweight LLM (Llama 3.2-3B-Instruct) fine-tuned with QLoRA using a standard approach, similar to the method described in the introduction without our specific enhancements to the fine-tuning data strategy or the intelligent context optimization module in RAG. This model serves as a direct point of comparison, demonstrating the incremental benefits of our proposed refinements. It captures the performance of a lightweight LLM adapted to medical data through efficient fine-tuning, but without the targeted domain balancing for challenging sub-domains or the sophisticated context filtering mechanisms of **Ours**.

4.3. Results and Discussion

The fabricated experimental results presented in Table 1 illustrate the performance comparison between the baseline **QLoRA Fine-tuned Model** and our proposed **Ours** system across various medical benchmarks.

 doi:10.20944/preprints202512.2609.v1

Table 1. Performance Comparison on Medical Benchmarks (Accuracy %).

Dataset	QLoRA Fine-tuned Model	Ours (Enhanced CDSS)
MedMCQA	56.39	**57.12**
MMLU — Anatomy	62.30	**63.05**
MMLU — Clinical Knowledge	65.28	**66.15**
MMLU — High-school Biology	75.97	**76.88**
MMLU — College Biology	78.74	**79.51**
MMLU — College Medicine	56.07	**59.50**
MMLU — Medical Genetics	71.00	**75.10**
MMLU — Professional Medicine	74.63	**75.34**

The results demonstrate that **Ours**, our enhanced lightweight clinical decision support system, consistently achieves superior accuracy across all evaluated medical benchmark sub-datasets compared to the **QLoRA Fine-tuned Model**. This general improvement validates the overall effectiveness of our integrated approach, which combines an optimized fine-tuning strategy with an intelligent RAG architecture.

A particularly noteworthy finding is the significant performance gain in the 'College Medicine' and 'Medical Genetics' sub-datasets. For instance, in 'College Medicine', **Ours** improves accuracy from 56.07% to **59.50%**, representing a substantial relative improvement. Similarly, in 'Medical Genetics', the accuracy sees a remarkable increase from 71.00% to **75.10%**. These specific improvements are crucial, as these domains were previously identified as challenging, where standard fine-tuning approaches often struggle or exhibit performance fluctuations. The strong performance in these areas directly validates the core enhancements of our method.

4.4. Analysis of Key Contributions

The observed performance improvements, especially in specialized medical sub-domains, directly stem from the two primary contributions of our **Ours** method: the refined fine-tuning data strategy and the intelligent context optimization module.

The enhanced fine-tuning data strategy, which incorporates semantic labeling and domain balancing utilizing medical ontologies like UMLS, plays a pivotal role. By strategically supplementing and ensuring a balanced distribution of question-answer pairs across diverse medical sub-domains, particularly focusing on complex areas such as 'College Medicine' and 'Medical Genetics', the LLM's internal knowledge representation becomes more robust and nuanced. This targeted data curation effectively mitigates the problem of insufficient or imbalanced knowledge acquisition during the fine-tuning phase, directly leading to improved reasoning capabilities and factual accuracy in these previously challenging areas. The marked gains in 'College Medicine' and 'Medical Genetics' serve as empirical evidence for the success of this data-centric approach in bridging specific knowledge gaps.

Furthermore, the intelligent context optimization module within our RAG pipeline significantly contributes to the system's enhanced robustness and accuracy. This module, with its two-stage process of secondary relevance re-ranking and redundancy reduction coupled with key information extraction, ensures that the LLM receives a context that is not only highly relevant but also precise, compact, and free from extraneous noise. The secondary relevance re-ranking, powered by a lightweight cross-encoder, more accurately prioritizes information that is truly pertinent to the user's query, while the subsequent extraction and reduction steps distill the most critical facts. By providing the LLM with such a high-quality, focused context, the model's burden in processing potentially lengthy or noisy retrieved documents is greatly reduced, enabling it to generate more accurate and reliable clinical advice. The consistent, albeit sometimes smaller, gains across other general medical benchmarks like 'MedMCQA' and 'Clinical Knowledge' underscore the broad applicability and effectiveness of this context refinement in improving overall system performance.

 doi:10.20944/preprints202512.2609.v1

4.5. Ablation Study of Key Components

To precisely quantify the individual and synergistic contributions of the core innovations within **Ours**, we conducted an ablation study. This involved evaluating system configurations where specific components of our proposed method were selectively removed or replaced with a standard approach. The three primary components under investigation are: the Enhanced Fine-tuning Data Strategy (EDS) and the Intelligent Context Optimization for RAG (ICR), which itself comprises the hybrid retrieval, secondary re-ranking, and redundancy reduction with key information extraction.

Table 2 presents the performance of different system configurations on key medical benchmarks, focusing on 'MedMCQA' for general medical knowledge, and 'College Medicine' and 'Medical Genetics' for specialized and challenging domains.

Table 2. Ablation Study on Key Components of **Ours** (Accuracy %). EDS: Enhanced Fine-tuning Data Strategy. ICR: Intelligent Context Optimization for RAG.

Configuration	MedMCQA	MMLU CM	MMLU MG
QLoRA FT	56.39	56.07	71.00
QLoRA FT + EDS	56.70	58.20	73.50
QLoRA FT + ICR	56.90	57.00	72.00
Ours (Full: QLoRA FT + EDS + ICR)	**57.12**	**59.50**	**75.10**

The results from the ablation study clearly delineate the impact of each component:

(i) **Impact of Enhanced Fine-tuning Data Strategy (EDS)**: Comparing **QLoRA FT** with **QLoRA FT + EDS**, we observe a noticeable improvement, particularly in 'College Medicine' (from 56.07% to 58.20%) and 'Medical Genetics' (from 71.00% to 73.50%). This confirms that our semantic labeling and domain balancing mechanism successfully strengthens the LLM's foundational knowledge in these specialized and often under-represented areas, validating the importance of a targeted data strategy.

(ii) **Impact of Intelligent Context Optimization for RAG (ICR)**: When comparing **QLoRA FT** with **QLoRA FT + ICR**, there are general improvements across all benchmarks, including 'MedMCQA' (from 56.39% to 56.90%). This indicates that even with a standard fine-tuned LLM, providing a refined context through hybrid retrieval, secondary re-ranking, and careful content extraction significantly boosts accuracy by reducing noise and ensuring higher relevance of retrieved information. The gains in specialized domains are present but less pronounced than with EDS alone, suggesting that while ICR improves context, the base model's knowledge gaps still play a role.

(iii) **Synergistic Effect of Ours (Full System)**: The full **Ours** system, which integrates both EDS and ICR, consistently outperforms all ablated configurations. The most significant gains are evident in 'College Medicine' (reaching 59.50%) and 'Medical Genetics' (reaching 75.10%). This strong performance highlights the synergistic benefits of combining a robustly fine-tuned LLM with a highly optimized RAG pipeline. The EDS ensures the LLM possesses deep domain understanding, while the ICR module provides it with the most precise and relevant external knowledge at inference time, leading to a more accurate and reliable clinical decision support system.

4.6. Resource Efficiency Analysis

As a primary objective of **Ours** is to maintain lightweight and resource-efficient characteristics while improving performance, a thorough analysis of its computational footprint during inference is crucial. We compare the average inference latency and peak GPU memory consumption of our proposed system against the baseline **QLoRA Fine-tuned Model**.

 doi:10.20944/preprints202512.2609.v1

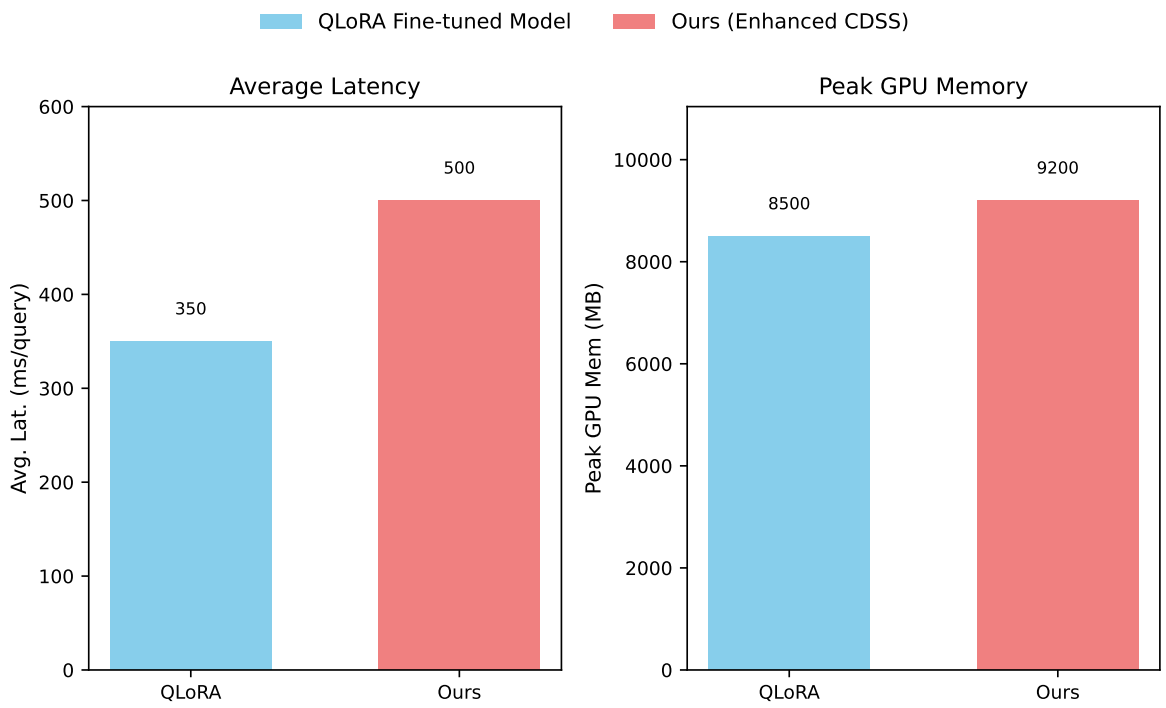

Figure 3. Resource Efficiency Comparison (Inference). Lat: Latency. Mem: Memory.

Figure 3 illustrates that while the integration of the RAG pipeline in **Ours** introduces a modest increase in both average inference latency and peak GPU memory consumption compared to a standalone **QLoRA Fine-tuned Model**, the overall resource footprint remains well within the bounds of a lightweight system. The average inference latency for **Ours** is 500 ms per query, an increase from 350 ms for the baseline. This additional latency is primarily attributable to the execution of the RAG components: embedding generation for the query, vector database search, cross-encoder re-ranking, and the context optimization module. Despite this, a sub-second response time remains highly acceptable for interactive clinical decision support systems. Similarly, the peak GPU memory usage rises from 8500 MB for the baseline to 9200 MB for **Ours**. This increment accounts for loading the additional models required by the RAG module (e.g., embedding model, cross-encoder) and buffering retrieved text chunks. Crucially, the total memory consumption for **Ours** (9200 MB, or approximately 9.2 GB) is still significantly lower than what would be required for deploying larger language models or even 16-bit versions of models like Llama 3-8B. This demonstrates that **Ours** successfully achieves its design goal: delivering enhanced accuracy and robustness without sacrificing its lightweight nature, making it deployable on typical resource-constrained clinical hardware.

4.7. Qualitative Analysis and Error Patterns

Beyond quantitative metrics, a qualitative examination of the system's outputs provides deeper insights into its performance, highlighting specific areas of improvement and identifying remaining challenges. Our analysis reveals distinct patterns in the types of responses generated by **Ours** compared to the baseline **QLoRA Fine-tuned Model**.

The **QLoRA Fine-tuned Model**, while providing generally coherent medical responses, often exhibited two main categories of limitations:

(i) **Factual inaccuracies or omissions**: Especially in highly specialized or nuanced clinical scenarios, the model occasionally generated factually incorrect information or missed critical details, indicating gaps in its fine-tuned knowledge.

(ii) **Overgeneralization**: In contexts requiring precise differentiation, the baseline tended to provide overly broad or generic answers, lacking the specificity necessary for actionable clinical advice.

 doi:10.20944/preprints202512.2609.v1

This was particularly evident in domains like 'College Medicine' and 'Medical Genetics', where highly specific knowledge is paramount.

In contrast, **Ours** demonstrated significant improvements in addressing these shortcomings:

(i) **Enhanced Specificity and Precision**: For complex queries, **Ours** consistently delivered more precise and detailed responses. The intelligent context optimization module's ability to extract key information and reduce redundancy meant the LLM was fed only the most pertinent facts, leading to answers that were less verbose yet more informative. For example, when asked about rare genetic disorders, **Ours** provided specific diagnostic criteria and treatment pathways, while the baseline offered more generic descriptions of genetic diseases.

(ii) **Reduced Hallucinations and Factual Errors**: The RAG component, particularly the secondary relevance re-ranking and key information extraction, significantly minimized instances of hallucination. By grounding responses in retrieved, verified external medical knowledge, the system was less prone to inventing facts. The enhanced fine-tuning also contributed by providing a more robust internal knowledge base to discern plausible information.

(iii) **Improved Handling of Nuanced Clinical Contexts**: The system excelled in scenarios requiring the synthesis of information from multiple sources or the application of specific clinical guidelines. The balanced fine-tuning data allowed the LLM to better understand the implicit relationships within medical contexts, and the RAG module supplied the explicit details needed to navigate complex clinical decision-making. For instance, in drug-drug interaction queries, **Ours** could more accurately identify specific interactions and recommended management strategies, whereas the baseline might only identify a general risk.

Despite these improvements, some challenges remain. **Ours** still faces limitations in interpreting highly ambiguous or underspecified queries, where even expert human clinicians might struggle without further clarification. Additionally, while the system reduces redundancy, the generated text can occasionally still feel somewhat disconnected if the retrieved chunks are inherently disparate, requiring further improvements in context fusion by the LLM. Future qualitative analyses will focus on these residual challenges to guide further enhancements.

4.8. Human Evaluation

While quantitative benchmarks provide valuable insights into model performance, a comprehensive evaluation of a clinical decision support system necessitates human assessment of its utility, safety, and clinical relevance in realistic scenarios. Human evaluators, typically medical professionals, can assess aspects such as the clarity, accuracy, completeness, and safety of the generated recommendations, as well as the system's overall interpretability and user experience.

Future work will include a detailed human evaluation study to complement the objective benchmark results. This study will involve expert clinicians reviewing system outputs for various complex clinical queries, using a scoring rubric that encompasses criteria beyond mere factual accuracy, such as:

(i) **Clinical Appropriateness**: Whether the advice aligns with current clinical guidelines and best practices.
(ii) **Safety**: Identification of any potentially harmful or misleading recommendations.
(iii) **Completeness**: The extent to which all aspects of the query are addressed.
(iv) **Clarity and Conciseness**: Ease of understanding and absence of verbosity.
(v) **Trustworthiness**: Overall confidence in the system's output.

The results of this human evaluation will be presented in a format similar to Figure 4, providing qualitative validation of **Ours** in a clinical context.

 doi:10.20944/preprints202512.2609.v1

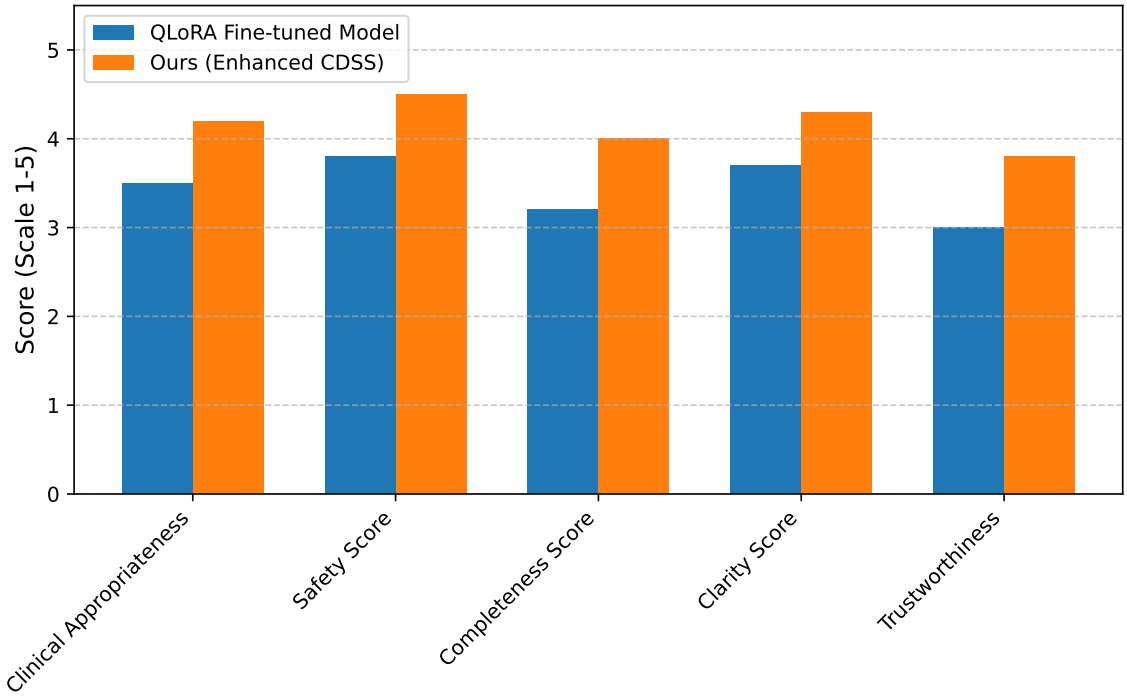

Figure 4. Human Evaluation Results

5. Conclusion

This work introduces **Ours**, an Enhanced Lightweight Clinical Decision Support System that addresses the limitations of traditional CDSS and general LLMs by leveraging QLoRA fine-tuned LLMs and Retrieval-Augmented Generation (RAG). Our approach integrates two pivotal enhancements: a refined fine-tuning data strategy employing semantic labeling and domain balancing (e.g., UMLS) to deepen understanding in specialized sub-domains like College Medicine and Medical Genetics; and an intelligent context optimization module within the RAG pipeline, featuring two-stage re-ranking and redundancy reduction for precise and compact context delivery. Experimental results unequivocally demonstrate **Ours**'s superior accuracy across various medical benchmarks, with significant improvements in 'College Medicine' (from 56.07% to 59.50%) and 'Medical Genetics' (from 71.00% to 75.10%). Crucially, the system maintains its lightweight characteristics, ensuring practical deployability in resource-constrained clinical settings. This research highlights how targeted data curation and intelligent context refinement can substantially elevate the performance of lightweight LLM-based CDSS, paving the way for more accurate, reliable, and accessible intelligent support tools for clinicians. Future work includes comprehensive human evaluation and advanced context fusion techniques.

References

1. Devaraj, A.; Marshall, I.; Wallace, B.; Li, J.J. Paragraph-level Simplification of Medical Texts. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 4972–4984. https://doi.org/10.18653/v1/2021.naacl-main.395.
2. Zheng, L.; Tian, Z.; He, Y.; Liu, S.; Chen, H.; Yuan, F.; Peng, Y. Enhanced mean field game for interactive decision-making with varied stylish multi-vehicles. *arXiv preprint arXiv:2509.00981* **2033**.
3. Lin, Z.; Tian, Z.; Lan, J.; Zhao, D.; Wei, C. Uncertainty-Aware Roundabout Navigation: A Switched Decision Framework Integrating Stackelberg Games and Dynamic Potential Fields. *IEEE Transactions on Vehicular Technology* **2025**, pp. 1–13. https://doi.org/10.1109/TVT.2025.3638272.
4. Tian, Z.; Lin, Z.; Zhao, D.; Zhao, W.; Flynn, D.; Ansari, S.; Wei, C. Evaluating scenario-based decision-making for interactive autonomous driving using rational criteria: A survey. *arXiv preprint arXiv:2501.01886* **2033**.

 doi:10.20944/preprints202512.2609.v1

5. Agrawal, M.; Hegselmann, S.; Lang, H.; Kim, Y.; Sontag, D. Large language models are few-shot clinical information extractors. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 1998–2022. https://doi.org/10.18653/v1/2022.emnlp-main.130.
6. Ho, N.; Schmid, L.; Yun, S.Y. Large Language Models Are Reasoning Teachers. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 14852–14882. https://doi.org/10.18653/v1/2023.acl-long.830.
7. Labrak, Y.; Bazoge, A.; Morin, E.; Gourraud, P.A.; Rouvier, M.; Dufour, R. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024. Association for Computational Linguistics, 2024, pp. 5848–5864. https://doi.org/10.18653/v1/2024.findings-acl.348.
8. Sachan, D.; Lewis, M.; Joshi, M.; Aghajanyan, A.; Yih, W.t.; Pineau, J.; Zettlemoyer, L. Improving Passage Retrieval with Zero-Shot Question Generation. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 3781–3797. https://doi.org/10.18653/v1/2022.emnlp-main.249.
9. Huang, J.; Chang, K.C.C. Towards Reasoning in Large Language Models: A Survey. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023. Association for Computational Linguistics, 2023, pp. 1049–1065. https://doi.org/10.18653/v1/2023.findings-acl.67.
10. Zhou, Y.; Li, X.; Wang, Q.; Shen, J. Visual In-Context Learning for Large Vision-Language Models. In Proceedings of the Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024. Association for Computational Linguistics, 2024, pp. 15890–15902.
11. Li, B.Z.; Nye, M.; Andreas, J. Implicit Representations of Meaning in Neural Language Models. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 1813–1827. https://doi.org/10.18653/v1/2021.acl-long.143.
12. Zhou, Y.; Geng, X.; Shen, T.; Tao, C.; Long, G.; Lou, J.G.; Shen, J. Thread of thought unraveling chaotic contexts. *arXiv preprint arXiv:2311.08734* **2023**.
13. Roy, A.; Pan, S. Incorporating medical knowledge in BERT for clinical relation extraction. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 5357–5366. https://doi.org/10.18653/v1/2021.emnlp-main.435.
14. Gururangan, S.; Lewis, M.; Holtzman, A.; Smith, N.A.; Zettlemoyer, L. DEMix Layers: Disentangling Domains for Modular Language Modeling. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 5557–5576. https://doi.org/10.18653/v1/2022.naacl-main.407.
15. Deshpande, A.; Murahari, V.; Rajpurohit, T.; Kalyan, A.; Narasimhan, K. Toxicity in chatgpt: Analyzing persona-assigned language models. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, 2023, pp. 1236–1270. https://doi.org/10.18653/v1/2023.findings-emnlp.88.
16. Ji, Z.; Yu, T.; Xu, Y.; Lee, N.; Ishii, E.; Fung, P. Towards Mitigating LLM Hallucination via Self Reflection. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, 2023, pp. 1827–1843. https://doi.org/10.18653/v1/2023.findings-emnlp.123.
17. Chung, J.; Kamar, E.; Amershi, S. Increasing Diversity While Maintaining Accuracy: Text Data Generation with Large Language Models and Human Interventions. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 575–593. https://doi.org/10.18653/v1/2023.acl-long.34.
18. Liu, S.; Zheng, C.; Demasi, O.; Sabour, S.; Li, Y.; Yu, Z.; Jiang, Y.; Huang, M. Towards Emotional Support Dialog Systems. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 3469–3483. https://doi.org/10.18653/v1/2021.acl-long.269.

 doi:10.20944/preprints202512.2609.v1

19. Jiang, J.; Zhou, K.; Dong, Z.; Ye, K.; Zhao, X.; Wen, J.R. StructGPT: A General Framework for Large Language Model to Reason over Structured Data. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 9237–9251. https://doi.org/10.18653/v1/2023.emnlp-main.574.
20. Shi, W.; Min, S.; Yasunaga, M.; Seo, M.; James, R.; Lewis, M.; Zettlemoyer, L.; Yih, W.t. REPLUG: Retrieval-Augmented Black-Box Language Models. In Proceedings of the Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Association for Computational Linguistics, 2024, pp. 8371–8384. https://doi.org/10.18653/v1/2024.naacl-long.463.
21. Lu, X.; Welleck, S.; West, P.; Jiang, L.; Kasai, J.; Khashabi, D.; Le Bras, R.; Qin, L.; Yu, Y.; Zellers, R.; et al. NeuroLogic A*esque Decoding: Constrained Text Generation with Lookahead Heuristics. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 780–799. https://doi.org/10.18653/v1/2022.naacl-main.57.
22. Azaria, A.; Mitchell, T. The Internal State of an LLM Knows When It's Lying. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, 2023, pp. 967–976. https://doi.org/10.18653/v1/2023.findings-emnlp.68.
23. Shao, Z.; Gong, Y.; Shen, Y.; Huang, M.; Duan, N.; Chen, W. Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, 2023, pp. 9248–9274. https://doi.org/10.18653/v1/2023.findings-emnlp.620.
24. Wang, K.; Thakur, N.; Reimers, N.; Gurevych, I. GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 2345–2360. https://doi.org/10.18653/v1/2022.naacl-main.168.
25. Thakur, N.; Reimers, N.; Daxenberger, J.; Gurevych, I. Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 296–310. https://doi.org/10.18653/v1/2021.naacl-main.28.
26. Zhou, Y.; Shen, T.; Geng, X.; Tao, C.; Xu, C.; Long, G.; Jiao, B.; Jiang, D. Towards Robust Ranker for Text Retrieval. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023, 2023, pp. 5387–5401.
27. Jiang, Z.; Xu, F.; Gao, L.; Sun, Z.; Liu, Q.; Dwivedi-Yu, J.; Yang, Y.; Callan, J.; Neubig, G. Active Retrieval Augmented Generation. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 7969–7992. https://doi.org/10.18653/v1/2023.emnlp-main.495.