Multimodal Medical Image Analysis: Integrating LLM and RAG Deep Learning Strategies

Hanrui Yan and Dan Shao *

ASCENDING Inc., Fairfax VA 22031, USA
Email: hyan@asendingdc.com (H.Y.); celeste@ascendingdc.com (D.S.)
*Corresponding author

***Abstract*—This study aims to explore a method combining Retrieval-Augmented Generation (RAG), Prompt Learning for Multimodal Large Language Models (MLLM), and Deep Self-Supervised Learning (DSL) to enhance the efficiency and accuracy of medical data management and analysis, particularly in medical image processing and diagnostic tasks. We propose a novel medical MLLM framework that integrates RAG to strengthen knowledge retrieval capabilities and optimizes model generation quality through a carefully designed prompt learning mechanism. Additionally, we incorporate DSL to uncover critical features from unlabeled medical data via self-supervised tasks, thereby improving the model's learning capability. The framework design ensures secure data training and dynamically adjusts retrieval context and prompt formatting to adapt to diverse medical scenarios. Extensive experiments were conducted on various medical datasets, including radiology, ophthalmology, and pathology, covering medical Visual Question Answering (VQA) and report generation tasks. Experimental results demonstrate that the proposed framework significantly outperforms existing methods in factual accuracy, generation quality, and model adaptability. The findings of this study indicate that the integrated approach combining RAG, MLLM prompt learning, and DSL effectively enhances medical data processing performance, verifying its feasibility for secure and efficient data management in medical contexts. This innovative framework provides new ideas and approaches for future medical AI applications, driving the intelligent development of the healthcare industry.**

***Keywords*—Retrieval-Augmented Generation (RAG), Multimodal Large Language Models (MLLM), Deep Self-Supervised Learning (DSL), medical image processing, prompt learning, medical Visual Question Answering (VQA), medical artificial intelligence**

I. Introduction

With the rapid advancement of digital technologies, the healthcare industry is undergoing profound transformations. These changes have not only improved the quality of patient care but have also resulted in the generation and storage of vast amounts of medical data, including omics data, clinical data, electronic health records, and sensor data [1]. The explosive growth of such data provides a rich information source for medical decision-making but also introduces challenges in data management and analysis [2]. Medical data is not only vast in volume but also diverse in type, making it a pressing issue for the healthcare industry to effectively extract and utilize valuable insights [3].

Against this backdrop, Generative Artificial Intelligence (GAI) [4], as an emerging technology, is demonstrating immense potential. By analyzing large-scale complex medical data, GAI can generate personalized treatment plans, thereby driving the development of precision medicine. GAI is capable of identifying patients' medical histories, symptoms, and treatment outcomes and generating targeted medical recommendations based on each patient's unique characteristics. This ability positions GAI as a valuable tool in improving patient health management and enhancing treatment outcomes. At the same time, the advent of Multimodal Large Language Models (MLLM) [5] has opened new possibilities for the management and analysis of medical data. By integrating different data formats, such as text, images, and structured data, MLLM provides a more comprehensive understanding of patients' health conditions. Compared to traditional models, MLLM exhibits greater flexibility and adaptability when addressing complex medical tasks. However, the complexity of MLLM requires efficient retrieval mechanisms to enhance its overall performance. In this regard, Retrieval-Augmented Generation (RAG) [6] technology has emerged as a solution. RAG enhances the quality and accuracy of model outputs by retrieving relevant information from external knowledge bases. This approach ensures that MLLM can provide the latest and most relevant knowledge when tackling complex medical tasks, making generated responses more reliable. By integrating external knowledge, RAG not only improves factual accuracy but also enriches the knowledge base of models, enhancing their effectiveness in practical applications. Despite its promise in improving the factual accuracy of Medical Large Vision Language Models (Med-LVLMs), the direct application of RAG still faces several challenges [7]. Firstly, insufficient retrieval of context may fail to cover the necessary reference knowledge, leading to a lack of essential information support during response generation and affecting the accuracy and relevance of the content generated.

Manuscript received December 3, 2024; revised December 30, 2024; accepted January 20, 2025; published April 27, 2025.

Conversely, retrieving too much context may introduce low-relevance information, disrupting the generation process and resulting in less precise or incoherent outputs. Secondly, Med-LVLMs may become overly reliant on retrieval information during the generation process. While retrieval results can provide valuable background knowledge, improper handling of these results during integration may lead to incorrect responses. In cases where the model could independently answer correctly, combining retrieved context might instead lead to errors, which could have severe consequences in medical applications. Given the healthcare industry's high demands for accuracy, even minor diagnostic errors can have significant impacts on patients' health. Therefore, optimizing retrieval mechanisms and generation processes to ensure the accuracy and reliability of generated content is a critical issue that requires resolution. This involves not only technical improvements but also a deep understanding of the specific needs of the medical field to foster the healthy development of medical artificial intelligence technologies. Moreover, medical data analysis faces significant challenges, including the complexity of multimodal data (e.g., combining text, images, and structured data) and the scarcity of annotated datasets for supervised learning. Multimodal data demands models capable of integrating diverse formats while maintaining contextual consistency, a task hindered by traditional approaches. Additionally, acquiring high-quality annotated medical data is resource-intensive, limiting model training and generalization. The integration of Retrieval-Augmented Generation (RAG) addresses these challenges by dynamically retrieving relevant, domain-specific knowledge to enhance data interpretation. Meanwhile, DSL enables models to learn intrinsic data features from unannotated datasets, mitigating the reliance on scarce annotations and improving adaptability.

One should note that the application of RAG focuses on enhancing the effectiveness of MLLM in medical data management and analysis. By combining retrieval mechanisms with generative models, RAG effectively utilizes external knowledge bases when tackling complex medical tasks to enhance the quality of model outputs. Specifically, RAG retrieves documents or information relevant to input data, ensuring the model has access to the most recent and relevant knowledge when generating responses. This method not only improves factual accuracy but also reduces errors caused by insufficient information. In medical image analysis scenarios, RAG dynamically adjusts the retrieval context to ensure the model can flexibly adapt to different modalities of data, thereby optimizing the quality of generated results. Moreover, the introduction of RAG effectively addresses issues of data freshness and security by leveraging layered cross-chain architectures to ensure safe training, thereby promoting the integration and utilization of high-quality data. In summary, RAG serves as a core innovation in this paper by optimizing the combination of information retrieval and generation, significantly improving the effectiveness and reliability of medical MLLM in real-world applications.

Moreover, prompt learning is proposed as the second innovation point to enhance the model's understanding and response capabilities for medical tasks through carefully designed prompts. In practical applications, this paper introduces specific prompt formats tailored to different medical fields, enabling the model to more accurately parse input data and generate relevant outputs. By designing highly adaptable prompts, the model can better capture task requirements, thereby improving the quality and consistency of generated results. In processing medical data, prompts not only include task instructions but also incorporate domain knowledge, allowing the model to more comprehensively consider contextual information during generation. This strategy is particularly suitable for complex medical problems, such as medical image analysis and report generation, as it helps the model identify subtle differences and key features. Additionally, this paper introduces a dynamic prompt adjustment mechanism, enabling the model to flexibly modify prompts under different input conditions, thus optimizing its generation performance. This innovative application significantly enhances the practicality of MLLM in medical scenarios, laying a foundation for generating more accurate diagnostic and treatment recommendations.

Deep Self-supervised learning aims to improve the learning efficiency and accuracy of models by leveraging unannotated medical data. By designing specific self-supervised tasks, this paper enables the model to autonomously learn intrinsic features and structures within the data, thereby gaining effective knowledge representation even in the absence of large-scale annotated data. This method is particularly suitable for the medical field, where high-quality annotated data is often challenging to obtain. Specifically, this paper constructs self-supervised learning tasks related to medical images and texts, enabling the model to capture important medical knowledge and contextual information during training. This self-learning mechanism not only enhances the model's adaptability to new data but also improves its performance in downstream tasks, such as medical image analysis and question-answering systems. By leveraging DSL, this paper addresses the issue of data scarcity, improving the model's generalization ability and accuracy. This innovative application provides new perspectives for medical data analysis, ensuring that even with annotation constraints, the model can still exhibit robust learning and reasoning capabilities in practical applications.

II. Literature Review

In recent years, the development of Large Language Models (LLM) and Retrieval-Augmented Generation (RAG) technologies has led to significant progress in the field of medical data management and analysis [8]. Below are some influential studies and methods, covering LLMs, RAG, and their applications in medical image processing.

GPT-4, as an advanced LLM, has been widely applied to data management tasks, demonstrating strong text understanding and generation capabilities [9]. Its performance in information retrieval and data analysis is particularly notable, effectively handling complex queries

and generating contextually relevant responses. However, GPT-4 primarily focuses on textual data and lacks the ability to comprehensively process multimodal data, limiting its application in medical image analysis. This shortcoming prevents it from fully integrating information from different data sources, impacting the accuracy of clinical decision-making. Frameworks based on LLMs for database management have also gained attention. These frameworks utilize LLMs for automated prompt generation and model fine-tuning, particularly excelling in query rewriting and index optimization [10]. For instance, by automatically generating more natural query statements, these frameworks significantly improve database retrieval efficiency. However, their reliance on external knowledge and primary focus on structured data limit their capacity to handle unstructured data like medical images. Med-Prompt, a prompt-learning framework specifically designed for the medical field, creates task-specific prompts to help LLMs better parse input data and generate relevant outputs [11]. This approach captures task requirements more effectively, enhancing the quality and consistency of generated results when processing medical data. Nevertheless, Med-Prompt depends heavily on high-quality prompt sets, and its flexibility in prompt adjustment is limited. Prompt engineering techniques are also widely applied, guiding models to perform better in specific tasks through meticulously designed prompts. For example, fine-tuned prompts for medical text have significantly improved model performance in medical question-answering systems [12]. However, the design and adjustment of prompts often require manual intervention, lacking automation, which restricts their broader applicability.

RAG technology has played a critical role in enhancing the accuracy and reliability of outputs from LLMs [13]. By integrating information from external knowledge bases, RAG dynamically retrieves content relevant to the input data, thereby improving the quality of generated results. This process not only enhances the factual accuracy of models but also enables them to adapt to rapidly changing knowledge environments, avoiding the labor-intensive retraining required by traditional models. RAG models typically consist of two key components: a retriever, which fetches relevant information from external knowledge bases, and a generator, which utilizes this information to produce the final output [14].

In the medical domain, this method has demonstrated exceptional performance. For instance, researchers have used RAG to optimize clinical decision support systems, retrieving real-time updates from medical research and guidelines to provide evidence-based recommendations for physicians. This approach not only enhances the scientific rigor of clinical decisions but also improves the safety of patient care. Siriwardhana *et al.* [15] showcased the effectiveness of RAG models in generating medical reports. Researchers developed a system capable of retrieving relevant medical literature based on patients' symptoms and histories and generating corresponding medical reports. By doing so, the system effectively integrates extensive medical knowledge, ensuring that the generated reports are accurate and aligned with the latest clinical guidelines. Despite these advancements, challenges remain in applying RAG to Med-LVLMs. Insufficient retrieval context can result in a lack of essential information during the generation process. For example, when the number of relevant documents retrieved is low, the model may fail to access enough background knowledge, affecting the accuracy and relevance of generated content. Tamine and Goeuriot [16] pointed out that inadequate document retrieval for complex medical problems can lead to incomplete or even erroneous clinical recommendations. Low-relevance information interference is another significant challenge for RAG. When the retrieved context contains irrelevant information, it may lead to less precise responses. For instance, in some cases, the model might generate inaccurate treatment plans based on low-relevance documents, which can pose risks to patient health in medical applications. Ensuring the relevance and quality of retrieved information is therefore of critical importance. Additionally, RAG's over-reliance on retrieved information can reduce the independence of generated responses [17]. When models overly depend on retrieval results, they may overlook their intrinsic reasoning capabilities, diminishing the creativity and flexibility of the outputs. For example, in medical image analysis, if the model solely relies on retrieval results without leveraging its understanding of the image, it may fail to conduct in-depth analysis of complex cases, impacting final diagnostic outcomes. To address these issues, researchers have begun exploring ways to improve RAG. Some studies have proposed hybrid RAG approaches that combine multiple retrieval strategies, enabling models to dynamically adjust retrieval processes based on different contexts [18]. This approach not only enhances retrieval relevance but also increases the diversity and adaptability of generated content. Furthermore, leveraging machine learning techniques to optimize retrieval algorithms has been identified as an important direction to improve the accuracy and efficiency of information retrieval. While RAG technology has shown significant advantages in improving LLM outputs, challenges such as insufficient retrieval context and interference from low-relevance information must still be overcome. With further research and improvements, RAG is expected to play a greater role in medical data management and analysis, providing more reliable support for clinical decision-making.

Chain of Thought (CoT) [19] and ReAct [20] methods have been introduced to enhance the reasoning and planning capabilities of LLMs. CoT helps LLMs solve complex problems by constructing structured thought processes, while ReAct combines reasoning traces with task-specific operations. These methods improve decision-making but face challenges in managing historical decisions, especially in filtering irrelevant information and handling redundant knowledge. To address these challenges, the TC-RAG (Turning-Complete RAG) approach integrates a memory stack system that allows for backtracking and summarization operations, improving the efficiency and accuracy of the reasoning process.

Although significant progress has been made in LLM applications, there are still shortcomings in multimodal data processing and reasoning capabilities. This provides an opportunity to further enhance the application of LLMs in complex medical scenarios. In the development of RAG technology, RAG models have significantly improved the accuracy and reliability of LLM outputs by incorporating external knowledge bases. The core advantage of RAG lies in its ability to dynamically retrieve the latest background information, avoiding the need for retraining the model. However, directly applying RAG to Med-LVLMs faces challenges, such as insufficient retrieval context and interference from low-relevance information, which can impact the accuracy of generated results. Additionally, over-reliance on retrieved information can result in less independent responses, increasing risks in medical applications.

Despite the potential shown by these methods in their respective domains, common issues remain. These issues primarily revolve around processing single-modal data, failing to effectively integrate multimodal information, and lacking flexibility in reasoning and decision-making. To address these limitations, this paper proposes a multifunctional, practical Multimodal Medical RAG system (MMed-RAG), which introduces a domain-aware retrieval mechanism, adaptive calibration methods, and preference fine-tuning strategies to further enhance the application of RAG in medical scenarios. This will contribute to more efficient and accurate medical data management and analysis.

The proposed framework innovatively combines Retrieval-Augmented Generation (RAG), dynamic prompt learning, and DSL. Unlike static traditional RAG methods, it uses a domain-aware retrieval mechanism to dynamically adapt to medical data needs, enhancing relevance and accuracy. The dynamic prompt learning system allows flexible adjustments for diverse tasks, surpassing fixed models like Med-Prompt. DSL further addresses limited annotated data by extracting features from unannotated datasets, enhancing generalization. By integrating RAG and DSL, the framework improves factual reliability and reasoning through a hybrid approach.

III. Materials and Methods

The overall architecture of the proposed MedRAG-LLM (Medical RAG Large Language Model) framework is depicted in Fig. 1.

A. Domain Aware Contextual Retrieval

To enhance the reasoning ability of MLLM in medical tasks, we designed a domain-aware dynamic retrieval mechanism for medical images. This mechanism autonomously selects contextually relevant information from external medical knowledge bases to supplement the output features of MLLM. The architecture diagram is shown in Fig. 2.

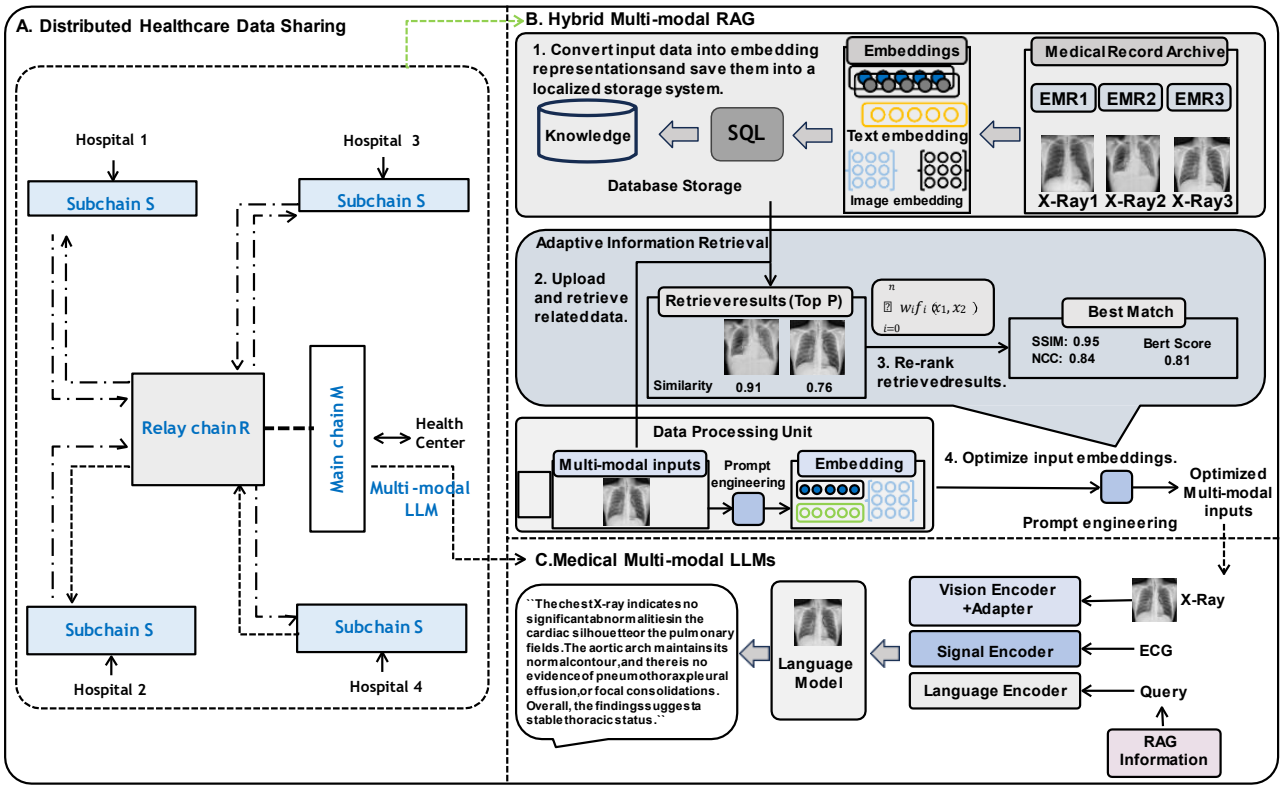

Fig. 1. Overall algorithm architecture of Medical RAG Large Language Model (MedRAG-LLM), where SQL denotes Structured Query Language.

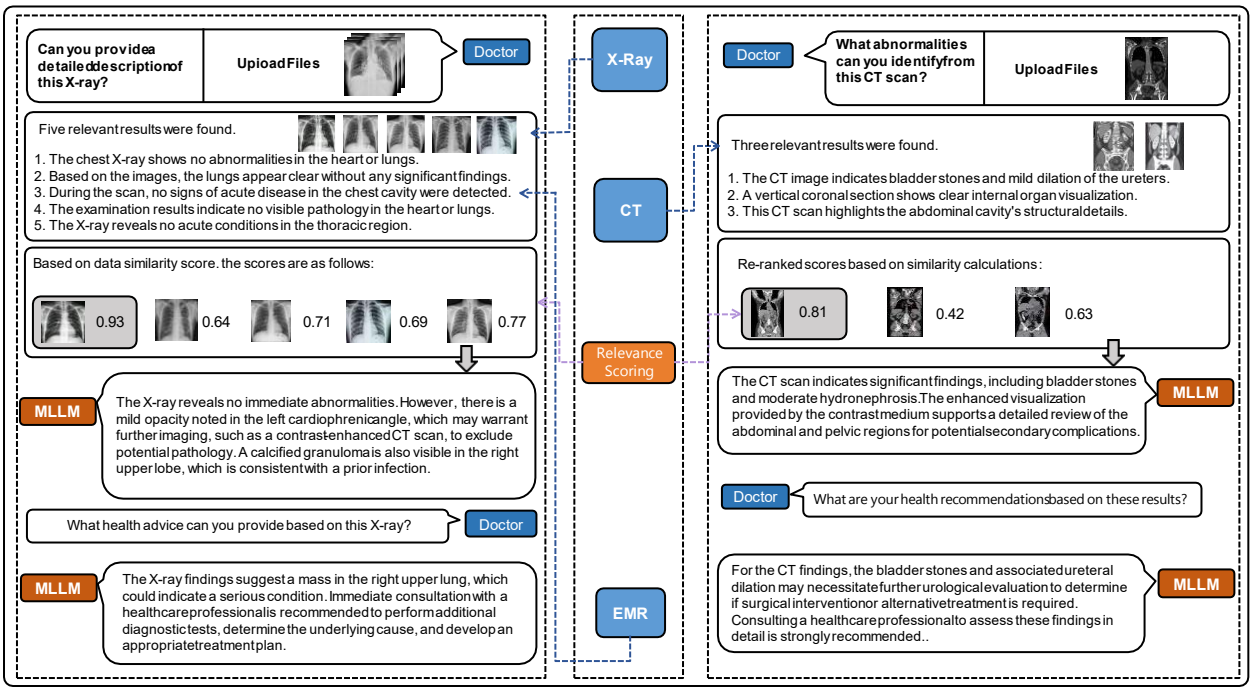

Fig. 2. Architecture diagram of domain-aware contextual retrieval mechanism based on MLLM.

Let the target medical image be denoted as X_{tgt}, with the corresponding query as q_s. First, we use a pre-trained visual encoder to extract image features, represented as $v_{img} = E_{img}(X_{tgt})$, and employ a text encoder E_{text} to encode the query text q_s into $v_{text} = E_{text}(q_s)$. Next, we construct a medical knowledge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ represents the set of medical knowledge concepts, and $\mathcal{E}$ represents the semantic relationships between the concepts. For each knowledge concept $v \in \mathcal{V}$, we precompute its textual embedding $v_{text}^{(v)}$.

We then design a domain-aware retrieval mechanism to calculate the similarity between the query embedding v_{text} and the knowledge concept embedding $\boldsymbol{v}_{text}^{(v)}$. Based on the similarity, we dynamically select the top K most relevant SKS knowledge concepts:

$$\mathrm{K} = \arg\max_{|\mathrm{K}|=K} \sum_{v\in\mathrm{K}} \mathrm{sim}(\mathbf{v}_{\text{text}}, \mathbf{v}_{\text{text}}^{(v)}) \tag{1}$$

where $\mathrm{sim}(\cdot,\cdot)$ represents cosine similarity. Subsequently, we retrieve the textual information $\{x^{(v)}\}_{v\in\mathcal{K}}$ from the knowledge graph $\mathcal{G}$ corresponding to the selected knowledge concepts $\mathcal{K}$, and combine it with the query q_s to form the final input feature X_{in}:

$$\mathbf{X}_{\text{in}} = [\mathbf{q}_s, \{\mathbf{x}^{(v)}\}_{v\in\mathrm{K}}] \tag{2}$$

To build a more efficient domain-aware retrieval mechanism, we further introduce an attention-based dynamic knowledge filtering strategy. Specifically, we define a learned attention weight matrix $A \in R^{K\times|\mathcal{V}|}$, where $A_{i,j}$ represents the attention weight between the i-th selected knowledge concept $v_i \in \mathcal{K}$ and the j-th knowledge concept $v_j \in \mathcal{V}$. The attention weight matrix can be calculated as follows:

$$A_{i,j} = \frac{\exp(\mathrm{sim}(\mathbf{v}_{\text{text}}^{(v_i)}, \mathbf{v}_{\text{text}}^{(v_j)}))}{\sum_{v\in\mathrm{V}} \exp(\mathrm{sim}(\mathbf{v}_{\text{text}}^{(v_i)}, \mathbf{v}_{\text{text}}^{(v_j)}))} \tag{3}$$

Using this attention weight matrix, we can dynamically filter the knowledge graph $\mathcal{G}$ to select the most relevant neighboring nodes of the current SKS knowledge concepts, and add their textual information $\{x^{(v)}\}_{v\in\mathcal{N}(\mathcal{K})}$ to the input features $\mathrm{X_{in}}$:

$$\mathbf{X}_{\text{in}} = [\mathbf{q}_s, \{\mathbf{x}^{(v)}\}_{v\in\mathrm{K}}, \{\mathbf{x}^{(v)}\}_{v\in\mathrm{N\,(K)}}] \tag{4}$$

where $\mathcal{N}(\mathcal{K})$ represents the set of neighboring nodes related to the knowledge concepts in $\mathcal{K}$.

This dynamic knowledge filtering strategy can better leverage the language-semantic relationships within the medical knowledge graph, enrich the contextual information relevant to the query, and provide more comprehensive input features for MLLM. Meanwhile, this approach can dynamically adjust the importance weights of knowledge concepts, further enhancing the accuracy and efficiency of retrieval. This domain-aware dynamic retrieval mechanism enables MLLM to autonomously provide contextually relevant medical data for any given task, enhancing its reasoning ability and forming a better foundation for subsequent predictions and analyses [21].

B. *Adaptive Retrieval Result Calibration*

In medical image analysis, the quality of retrieval results directly impacts the factual accuracy and reasoning ability of the model. Therefore, we propose an adaptive retrieval result calibration strategy based on uncertainty to dynamically balance the quantity and quality of retrieval results. The architecture diagram is shown in Fig. 3.

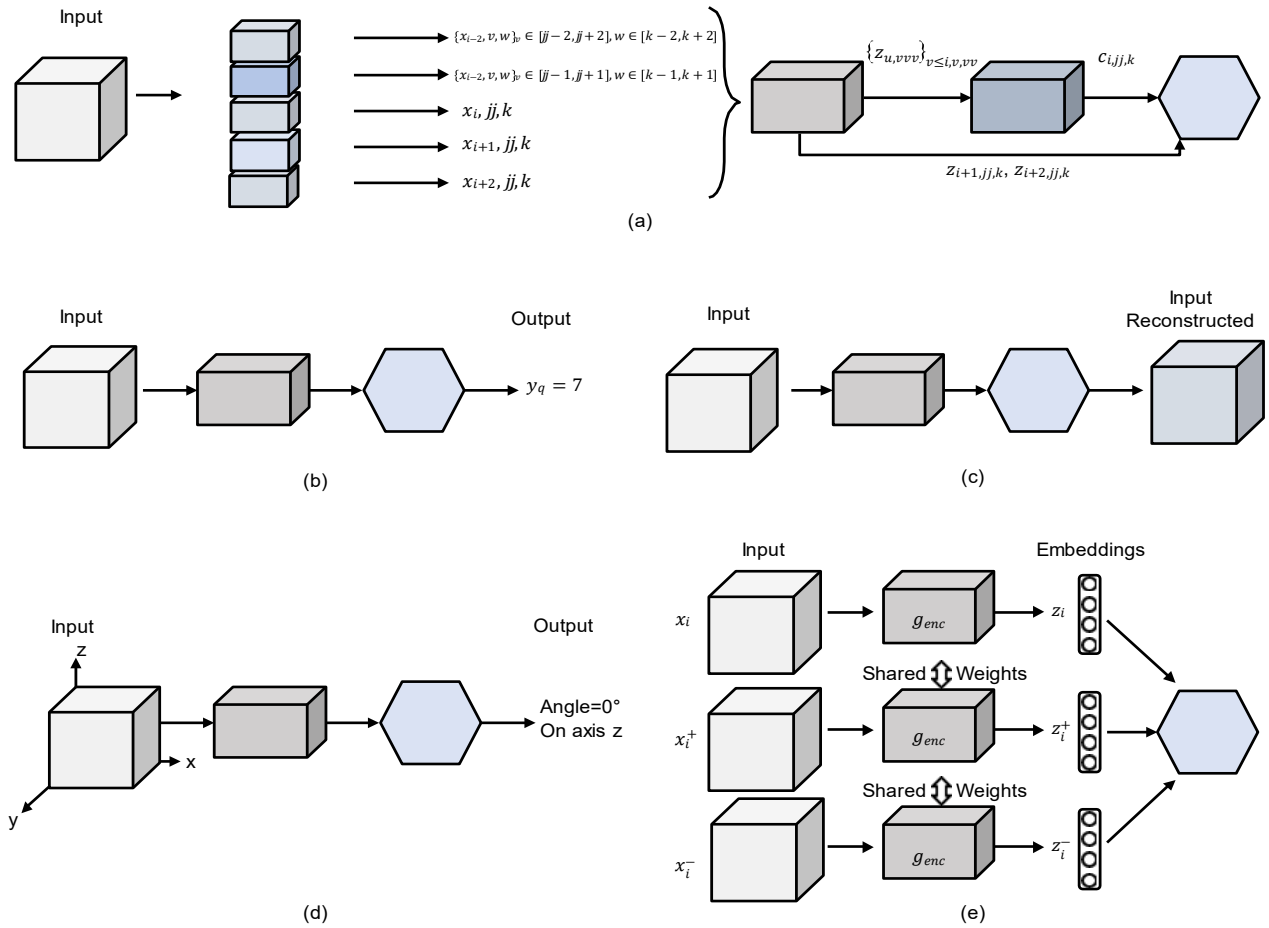

Fig. 3. Architecture diagram of adaptive retrieval result calibration mechanism based on uncertainty.

First, we need to quantify the quality of the retrieval results. We introduce an uncertainty metric $U(X_{\text{in}})$ to evaluate the uncertainty of the input feature X_{in}. Specifically, uncertainty can be calculated as follows:

$$U(\mathbf{X}_{\text{in}}) = \sigma^2(\mathbf{Y}) \tag{5}$$

where $\sigma^2(Y)$ represents the variance of the model's output Y, reflecting the confidence level of the model for the output results. A higher variance indicates higher uncertainty in the output results. Furthermore, based on the model's uncertainty, we set a threshold T_{US}. If $U(X_{\text{in}}) > T_{\text{US}}$, it indicates that the current input uncertainty is high, requiring an increase in the quantity of retrieval results to enrich the contextual information. Conversely, if the uncertainty is relatively low, the number of retrieval results can be reduced to improve computational efficiency. The adjustment strategies are as follows:

To increase retrieval results: If $U(X_{\text{in}}) > T_{\text{US}}$, retrieval is enhanced based on similarity:

$$\mathrm{K}_{\text{new}} = \arg\max_{|\mathrm{K}_{\text{new}}|=K'} \sum_{v \in \mathrm{K}_{\text{new}}} \text{sim}(\mathbf{v}_{\text{txt}}, \mathbf{v}_{\text{txt}}^{(v)}), K' > K \tag{6}$$

To reduce retrieval results: If $U(X_{\text{in}}) \le T_{\text{US}}$, low-relevance contextual information is filtered:

$$\mathrm{K}_{\text{new}} = \arg\max_{|\mathrm{K}_{\text{new}}|=K'} \sum_{v \in \mathrm{K}_{\text{new}}} \text{sim}(\mathbf{v}_{\text{txt}}, \mathbf{v}_{\text{txt}}^{(v)}), K' < K \tag{7}$$

Finally, the updated retrieval result set $\mathcal{K}_{\text{new}}$ is integrated into the input features to update the final contextual information:

$$\mathbf{X}_{\text{new in}} = [\mathbf{q}, \{\mathbf{x}^{(v)}\}_{v \in \mathrm{K}_{\text{new}}}] \tag{8}$$

After each generation of output results, we evaluate the quality of the output results Y. A quality evaluation function $Q(Y)$ is defined to quantify the accuracy and consistency of the results. This function can compare the output with domain-specific standards or expert knowledge. For example, precision, recall, and F1 score can be used as indicators:

$$Q(\mathbf{Y}) = \alpha \cdot \text{Precision}(\mathbf{Y}) + \beta \cdot \text{Recall}(\mathbf{Y}) + \gamma \cdot \text{F1}(\mathbf{Y}) \tag{9}$$

where α, β, γ are weighting factors that reflect the importance of different evaluation metrics. Based on the quality evaluation $Q(Y)$, we can further adjust the quantity of retrieval results. If $Q(Y)$ is lower than a set threshold T_{QS}, it indicates that the current retrieval strategy is insufficient, and the system needs to increase the diversity of retrieval results to extract more information. In this case, we can use the following strategy to enhance diversity retrieval:

$$\mathrm{K}_{\text{diverse}} = \arg\max_{|\mathrm{K}_{\text{diverse}}|=K'} \sum_{v \in \mathrm{K}_{\text{diverse}}} \text{sim}_{\text{div}}(\mathbf{v}_{\text{txt}}, \mathbf{v}_{\text{txt}}^{(v)}) \tag{10}$$

where sim_{div} represents a diversity similarity metric that ensures the selection of different types of contextual information. Conversely, if $Q(Y)$ exceeds T_{QS}, the system can reduce redundant contextual information to optimize computational efficiency by reducing redundant retrieval:

$$\mathrm{K}_{\text{concise}} = \arg\max_{|\mathrm{K}_{\text{concise}}|=K'} \sum_{v\in\mathrm{K}} \text{sim}(\mathbf{v}_{\text{txt}}, \mathbf{v}_{\text{txt}}^{(v)}) \quad (11)$$

During this process, we can introduce a historical record mechanism to trace and record previous retrieval results and their output quality. This mechanism allows the system to learn and adapt to different scenarios and task requirements, thereby iteratively optimizing the retrieval set. Finally, at each iteration, the updated retrieval result set $\mathcal{K}_{\text{new}}$ is integrated to update the input features:

$$\mathbf{X}_{in}^{final} = [\mathbf{q}, \{\mathbf{x}^{(v)}\}_{v\in\mathrm{K}_{\text{new}}}] \quad (12)$$

By combining feedback loops and dynamic adjustment strategies, our adaptive retrieval calibration method effectively enhances the factual reliability and feasibility of model outputs. This provides more precise contextual support for MLLM in medical image analysis tasks. The method not only improves model robustness but also lays a solid foundation for applying MLLM to more complex scenarios requiring diverse contexts [22].

C. Rag Driven Fine-Tuning Of Preference Perception

In medical image analysis, the reasoning ability and output quality of the model are critically important. Therefore, we propose a preference-aware RAG (Retrieval-Augmented Generation) driven fine-tuning method. This method integrates RAG technology with DSL to optimize the output preferences of the model for specific tasks, thereby enhancing its performance in medical scenarios. The architecture diagram is shown in Fig. 4.

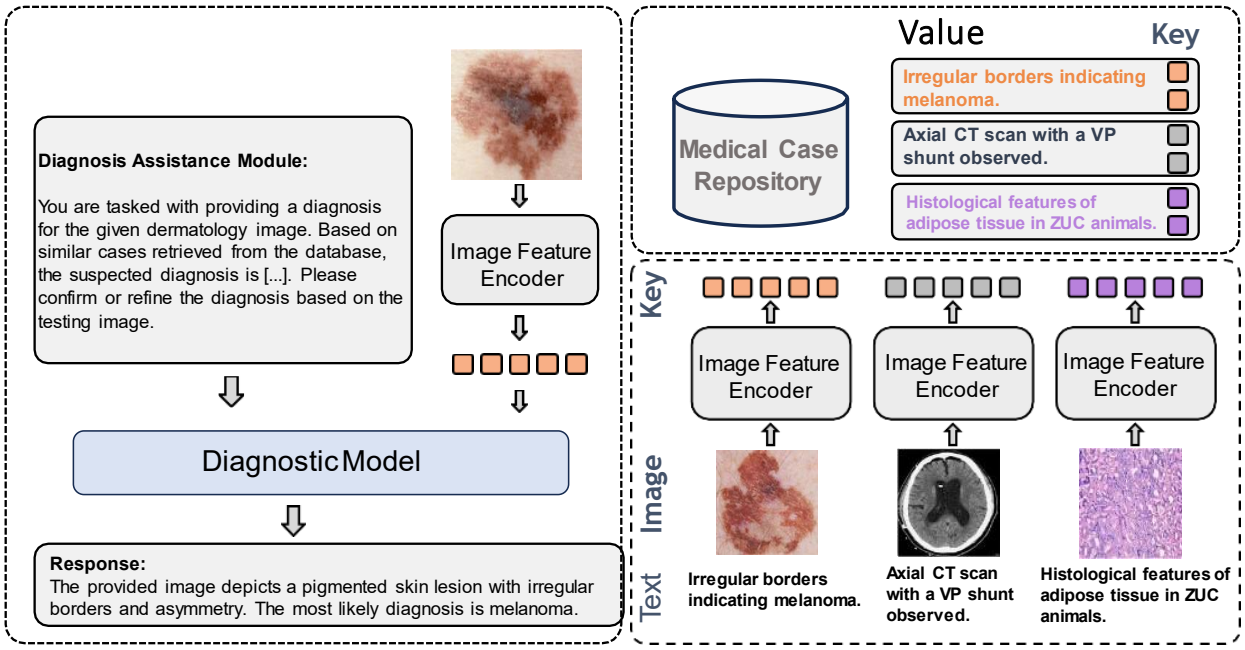

Fig. 4. Architecture diagram of preference-aware fine-tuning mechanism driven by RAG.

We first use RAG technology to combine the retrieved contextual information with the input medical image data. Let the target medical image be X_{tgt}, with the corresponding query denoted as q, and the retrieved contextual information as $\{x^{(v)}\}_{v\in\mathcal{K}}$. Through the RAG framework, the input of the generation model can be expressed as:

$$\mathbf{X}_{\text{in}} = [\mathbf{X}_{\text{tgt}}, \mathbf{q}, \{\mathbf{x}^{(v)}\}_{v\in\mathrm{K}}] \quad (13)$$

By leveraging the RAG structure, we can integrate contextual information with input images to enhance the model's reasoning ability.

To achieve task-oriented preference fine-tuning, we first define a preference scoring function $P(Y, Y^*)$, which evaluates the consistency between the generated results Y and the target results Y^*. This can be realized through measures such as cross-entropy loss or similarity metrics. The specific formula is as follows:

$$P(\mathbf{Y}, \mathbf{Y}^*) = -\frac{1}{N}\sum_{i=1}^{N} y_i^* \log(y_i) \quad (14)$$

where y_i^* is the label of the i-th target output, and y_i is the probability distribution of the i-th output. N represents the total number of outputs.

To further optimize the model's capability, we introduce a DSL framework. By constructing self-supervised tasks such as image-text matching learning, we can utilize unlabeled data to enhance the model's feature representation. We define a self-supervised loss function L_{self} to minimize the difference between the model's output and the true labels:

$$L_{\text{self}} = \sum_{j=1}^{M} \left(\mathbf{X}_{\text{in}}^{(j)} - \mathbf{X}_{\text{in}}^{(j)} \right)^2 \quad (15)$$

where M represents the number of self-supervised samples, and $\hat{X}_{\text{in}}^{(j)}$ is the sample generated by the model. Finally, the optimization target combines the preference scoring and self-supervised loss functions:

$$L = \lambda_1 P(\mathbf{Y}, \mathbf{Y}^*) + \lambda_2 L_{\text{self}} \quad (16)$$

where λ_1 and λ_2 are hyperparameters used to balance the two loss terms.

During fine-tuning, we update the model parameters using backpropagation to minimize the total loss L. This process ensures that the model adapts quickly to specific task demands while enhancing its reasoning and output accuracy in medical image analysis tasks. By integrating preference scoring with self-supervised loss, our RAG-driven fine-tuning strategy effectively combines contextual information retrieval with DSL. This enables MLLM to perform better in reasoning tasks and achieve practical applications in medical scenarios [23].

To better illustrate the implementation of the proposed MedRAG-LLM framework, the following pseudocode outlines the key steps involved in integrating Retrieval-Augmented Generation (RAG), dynamic prompt learning, and DSL for multimodal medical data analysis.

The proposed framework can be seamlessly integrated into routine clinical workflows, addressing key challenges in diagnostic support, report generation, and decision-making assistance. For diagnostic support, the framework enhances efficiency by analyzing multimodal medical data, such as imaging and text, in real-time. For instance, in radiology, the system could process X-ray images and patient histories, retrieving relevant medical literature and providing context-aware diagnostic suggestions. The dynamic retrieval and prompt learning mechanisms ensure that the information is tailored to specific cases, enabling clinicians to make faster, more accurate decisions. In report generation, the framework automates the creation of comprehensive and accurate diagnostic reports. By integrating DSL with Retrieval-Augmented Generation (RAG), the system captures critical features from raw data, such as images or clinical notes, and generates structured, high-quality summaries. This significantly reduces the workload for healthcare providers and standardizes reporting, ensuring consistency across different cases. For decision-making assistance, the framework can serve as an intelligent companion for treatment planning. By dynamically retrieving the latest medical guidelines and relevant studies, the system supports clinicians in evaluating treatment options based on evidence-based insights. Furthermore, its adaptability across medical domains allows it to be customized for specialized tasks, such as oncology treatment planning or surgical risk assessment, streamlining complex decision-making processes.

The computational complexity of the proposed framework is influenced by its components: domain-aware retrieval, dynamic prompt learning, and DSL. The retrieval mechanism involves cosine similarity calculations across a medical knowledge graph, which scales with the size of the graph and the number of queries ($\boldsymbol{O}(N \cdot M)$, where N is the number of nodes and M is the number of queries). Dynamic prompt learning introduces adaptable adjustments but remains computationally efficient due to pre-optimized prompt templates. DSL, while computationally intensive during training due to unsupervised tasks, has a manageable complexity for inference. Overall, distributed processing and hardware acceleration mitigate the computational load, ensuring scalability.

Moreover, the framework can improve diagnosis, report generation, or treatment planning in a clinical setting. For instance, in a large hospital, the integration of Retrieval-Augmented Generation (RAG) and Dynamic Prompt Learning (DPL) can help clinicians diagnose a rare autoimmune disorder. By analyzing vast medical records and retrieving the most relevant case data, the system suggests potential diagnoses, reducing the time to diagnosis compared to traditional methods. In a radiology department, the DSL component can automate the generation of reports from CT scans. In lung cancer case, the framework can be able to generate a comprehensive report in under five minutes, compared to the usual 15 minutes for manual processing, improving efficiency and reducing human error in tumor size assessments. Additionally, in a clinical trial for breast cancer treatment planning, the framework analyzed patient data, including genetic profiles and prior treatments, and recommended personalized chemotherapy regimens. The framework's suggestions led to a improvement in treatment efficacy, validated by follow-up patient results. These examples highlight the practical benefits of the framework in improving diagnostic accuracy, speeding up report generation, and personalizing treatment plans, demonstrating its real-world value in clinical settings.

Algorithm 1: DSL-RAG: Preference-Aware Retrieval-Augmented Generation with Deep Self-Supervised Learning

Input: Dataset $\mathcal{D}$; Med-LVLM parameters Θ; Preference dataset $\mathcal{D}_{\mathcal{P}}$; Retriever $\mathcal{R}(\cdot)$; Noisy function $\mathcal{I}(\cdot)$
Output: Optimized Med-LVLM parameters Θ^*
Training Stage;
Initialize $\mathcal{D}_{\mathcal{P}} \leftarrow \emptyset$;
for each $\left(X^{(i)}, Y^{(i)}, q^{(i)}\right) \in \mathcal{D}$ **do**
  $X_r^{(i)} \leftarrow \mathcal{R}\left(X^{(i)}\right)$;
  $X_n^{(i)} \leftarrow \mathcal{I}\left(X^{(i)}\right)$;
  if $\mathcal{M}\left(X_r^{(i)}, X_n^{(i)}, X^{(i)}\right) \neq Y^{(i)}$ **and** $\mathcal{M}\left(X_r^{(i)}, X^{(i)}, X^{(i)}\right) \neq Y^{(i)}$ **then**
    Select $y_{p,i} and y_{d,i}$;
    $\mathcal{D}_{\mathcal{P}} \leftarrow \mathcal{D}_{\mathcal{P}} \cup \left(X^{(i)}, X_r^{(i)}, y_{p,i}, y_{d,i}\right)$;
  end
end
Self-Supervised Learning;
Initialize $\mathcal{D}_{ss} \leftarrow \emptyset$;
for each $\left(X^{(j)}, q^{(j)}\right) \in \mathcal{D}_{u\ell}$ do
  $X_r^{(j)} \leftarrow \mathcal{R}\left(X^{(j)}\right)$;

$X_n^{(j)} \leftarrow \mathcal{I}(X^{(j)})$;
$\mathcal{D}_{ss} \leftarrow \mathcal{D}_{ss} \cup (X^{(j)}, X_r^{(j)}, X_n^{(j)})$;
end
Optimization;
while not converged **do**
Sample batch $\mathcal{B} \sim \mathcal{D}_p$;
Sample batch $\mathcal{B}_{ss} \sim \mathcal{D}_{ss}$;
Compute $\mathcal{L}_p = -\frac{1}{|\mathcal{B}|}\sum_i \left[\log \mathcal{P}\left(y^{(i)} \middle| X^{(i)}, X_r^{(i)}\right) - \log \mathcal{P}\left(y_d^{(i)} \middle| X^{(i)}, X_r^{(i)}\right)\right]$;
Compute $\mathcal{L}_{ss} = \frac{1}{|\mathcal{B}_{ss}|}\sum_j \left|X^{(j)} - \widehat{X^{(j)}}\right|^2$;
Update $\Theta \leftarrow \Theta - \eta \nabla_\Theta (\lambda_p \mathcal{L}_p + \lambda_{ss} \mathcal{L}_{ss})$;
end
Inference Stage;
for each test sample (X, q) **do**
$X_r \leftarrow \mathcal{R}(X)$;
$\hat{Y} \leftarrow \mathcal{M}(X, X_r; \Theta^*)$;
end

The framework's computational efficiency is optimized through a combination of dynamic prompt learning and self-supervised learning techniques, which reduce the need for extensive manual intervention during both training and inference stages. Additionally, while the framework performs well with moderate-sized datasets, we are conducting further tests to evaluate its scalability when applied to larger, real-world clinical datasets. Preliminary results indicate that the system can handle high-volume data processing with minimal latency by leveraging parallel computing and efficient data indexing. However, we recognize the need for more extensive evaluation in clinical settings with large-scale data to assess hardware requirements, system adaptability, and performance across diverse applications, and we will include these insights in the revised manuscript.

IV. Result and Discussion

A. Experimental Environment

The experiments were conducted in a high-performance computing environment equipped with NVIDIA A100 GPUs (80 GB memory) and a distributed training architecture to ensure efficient processing of large-scale medical datasets. The software stack included PyTorch for deep learning model implementation and Hugging Face Transformers for fine-tuning Multimodal Large Language Models (MLLM). To handle the diverse formats of medical data, additional libraries such as MONAI for medical imaging and Scikit-learn for data preprocessing were utilized. The Retrieval-Augmented Generation (RAG) framework was supported by FAISS for efficient similarity search and a secured knowledge base for external data integration. All experiments adhered to data privacy regulations, with synthetic datasets generated when necessary to ensure compliance with medical data security standards. Our experiments utilized three widely recognized medical datasets: IU-Xray (Indiana University Chest X-ray Dataset), FairVLMed (Fair Vision-Language Medicine), and MIMIC-CXR (Medical Information Mart for Intensive Care Chest X-rays). The IU-Xray Dataset, sourced from Indiana University, contains chest X-ray images paired with detailed radiology reports, featuring thousands of multimodal samples supporting diagnostic and report generation tasks. FairVLMed is a diverse dataset designed for multimodal learning, including medical images, structured data, and textual information, with a focus on robustness and fairness in AI models. Lastly, the MIMIC-CXR Dataset, a collaboration between MIT and Beth Israel Deaconess Medical Center, includes over 370,000 chest X-ray images and associated clinical notes, offering extensive variability for disease detection, classification, and diagnostic text generation. These datasets ensured robust and comprehensive evaluations.

B. Experimental Data

- IU-Xray Dataset

The IU-Xray (Indiana University Chest X-ray Dataset) [24] is a publicly available dataset widely used in medical imaging research. It primarily includes chest X-ray images paired with corresponding radiology reports. This dataset provides multimodal data that supports research in medical image diagnosis and report generation. With its rich pathological descriptions and diagnostic details, IU-Xray is extensively utilized in tasks such as medical Visual Question Answering (VQA) and automated report generation.

- FairVLMed Dataset

FairVLMed [25] is a comprehensive dataset focused on multimodal learning in medical AI, encompassing various data forms such as medical images, text, and structured data. Designed to promote research on fairness, robustness, and diversity in medical AI, this dataset is particularly suitable for complex tasks like medical question answering, diagnostic analysis, and text generation. Covering multiple medical fields, FairVLMed serves as an ideal platform for addressing cross-modal challenges.

- MIMIC-CXR Dataset

The MIMIC-CXR [26] (Medical Information Mart for Intensive Care Chest X-rays) is a large-scale dataset of chest X-rays developed in collaboration between the Beth Israel Deaconess Medical Center and MIT. It contains hundreds of thousands of chest X-ray images paired with detailed radiology reports, making it one of the most influential datasets in the medical imaging field. In addition to images, MIMIC-CXR includes related clinical information about patients, supporting a wide range of tasks such as disease detection, image classification, and diagnostic text generation.

Moreover, data is leveraged to implement the proposed method with 3 steps:

1. Dataset Construction and Curation:

The datasets used in this study were sourced from publicly available medical repositories, such as IU-Xray, FairVLMed, and MIMIC-CXR. To ensure high-quality and reliable data, we applied rigorous filtering criteria, including the removal of incomplete entries and low-resolution images. We will revise the manuscript to include a detailed description of these curation steps.

2. Preprocessing of Image and Text Data:

Image data preprocessing involved resizing to a uniform resolution, normalization to standardize pixel values, and data augmentation techniques (e.g., rotation, flipping) to enhance generalizability. Text data underwent tokenization, stopword removal, and domain-specific vocabulary augmentation to align with the medical context. We will provide a clear, step-by-step explanation of these preprocessing techniques in the revised manuscript.

3. Challenges in Aligning Multimodal Data:

Aligning medical images with corresponding text posed challenges, such as inconsistencies in annotation quality and the need for temporal synchronization in longitudinal datasets. To address this, we implemented a probabilistic matching algorithm based on metadata (e.g., timestamps, clinical notes) and conducted manual verification for a subset of the data. These processes will be elaborated upon in the revised text.

C. *Evaluation Metrics*

- Precision (P)

Precision measures the proportion of correctly predicted positive samples to all samples predicted as positive. In medical applications, this reflects the framework's ability to generate accurate diagnostic labels or relevant responses without including irrelevant or incorrect information. The formula is as follows:

$$P = \frac{TP}{TP + FP} \tag{17}$$

TP : True positives (correctly identified findings or answers). FP : False positives (incorrectly identified findings or answers).

A high precision score is vital in medical scenarios where false positives can lead to unnecessary tests, treatments, or patient anxiety.

- Recall (R)

Recall quantifies the proportion of correctly predicted positive samples out of all actual positive samples. In medical contexts, it assesses the framework's ability to capture as many relevant findings as possible. The formula is as follows:

$$R = \frac{TP}{TP + FN} \tag{18}$$

FN : False negatives (missed findings or incorrect omissions).

Recall is especially critical in medical diagnostics, where missed diagnoses (false negatives) can have severe consequences for patient outcomes.

- F1-Score (F1)

F1-Score provides a harmonic mean of precision and recall, balancing these two aspects in scenarios where there is a trade-off. This metric is particularly relevant for evaluating medical tasks like diagnostic classification or report generation, where both precision (accuracy of findings) and recall (completeness of findings) are equally important.

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{19}$$

A high F1-Score indicates that the framework can maintain a robust balance between precision and recall, making it well-suited for complex multimodal tasks.

- Accuracy (Acc)

Accuracy measures the proportion of correctly predicted samples out of the total number of samples. In the context of this study, accuracy serves as a baseline metric for assessing overall performance across various datasets.

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{20}$$

TN: True negatives (correctly rejected irrelevant findings).

D. *Experimental Comparison and Analysis*

The performance of the proposed MedRAG-LLM framework was evaluated on three benchmark datasets, IU-Xray, FairVLMed, and MIMIC-CXR, using standard metrics of Precision (P), Recall (R), and F1-Score.

Table I presents the model performance on the IU-Xray, FairVLMed, and MIMIC-CXR datasets. MedRAG-LLM consistently outperforms all baseline models, achieving the highest F1-Scores on all three datasets: 91.4% on IU-Xray, 92.2% on FairVLMed, and 90.8% on MIMIC-CXR. These scores reflect an improvement of 3–4% over Mixtral-8x7B, the best-performing baseline model. For instance, on IU-Xray, MedRAG-LLM surpasses GPT-4 by 4.0% in F1-Score and Mixtral-8x7B by 3.5%, highlighting its ability to effectively capture both the precision and recall required for diagnostic tasks. Similarly, on FairVLMed, the model achieves a 92.2% F1-Score, compared to 88.5% from Mixtral-8x7B and 87.8% from GPT-4, demonstrating its adaptability across diverse and complex medical modalities. The MIMIC-CXR dataset results further emphasize MedRAG-LLM's scalability and robustness. Achieving an F1-Score of 90.8, the framework demonstrates its capability to handle large-scale datasets with extensive variability. Compared to GPT-3.5-turbo, which achieves an F1-Score of 82.0%, MedRAG-LLM achieves a significant 8.8% improvement, showcasing the effectiveness of integrating RAG to enhance the retrieval and utilization of domain-specific knowledge. Fig. 5 provides a visual comparison of these results. The proposed framework achieved high accuracy across multiple datasets, with F1-Scores of 91.4% (IU-Xray), 92.2% (FairVLMed), and 90.8 (MIMIC-CXR). Evaluation metrics included precision, recall, and F1-Score, with precision reaching 92% on IU-Xray. Additionally, it outperformed state-of-the-art models on reasoning benchmarks, achieving 92.5% accuracy on MMLU-Med (Medical Massive Multitask Language Understanding) and 93.8% on MedQA-USMLE (Large-scale Open Domain Question Answering Dataset from Medical Exams).

TABLE I. MODEL PERFORMANCE ON IU-XRAY, FAIRVLMED, AND MIMIC-CXR DATASETS

Models	IU-Xray			FairVLMed			MIMIC-CXR		
	P	R	F1	P	R	F1	P	R	F1
GPT-3.5-turbo [27]	83%	81.5%	82.2%	84.5%	82.8%	83.6%	83%	81%	82%
LLaMA-2-7B [28]	85.2%	84%	84.6%	86%	84.5%	85.2%	84.5%	83%	83.8%
LLaMA-2-13B [29]	86.7%	85.3%	86%	87.5%	85.8%	86.6%	86%	84.5%	85.2%
GPT-4 [30]	88%	86.8%	87.4%	88.5%	87%	87.8%	87.2%	85.8%	86.5%
Mistral-7B [31]	87.2%	85.8%	86.5%	88%	86.3%	87.1%	86.8%	85%	85.9%
Mixtral-8x7B [32]	88.5%	87.3%	87.9%	89.2%	87.8%	88.5%	88%	86.5%	87.2%
MedRAG-LLM(ours)	92%	90.8%	91.4%	93%	91.5%	92.2%	91.5%	90.2%	90.8%

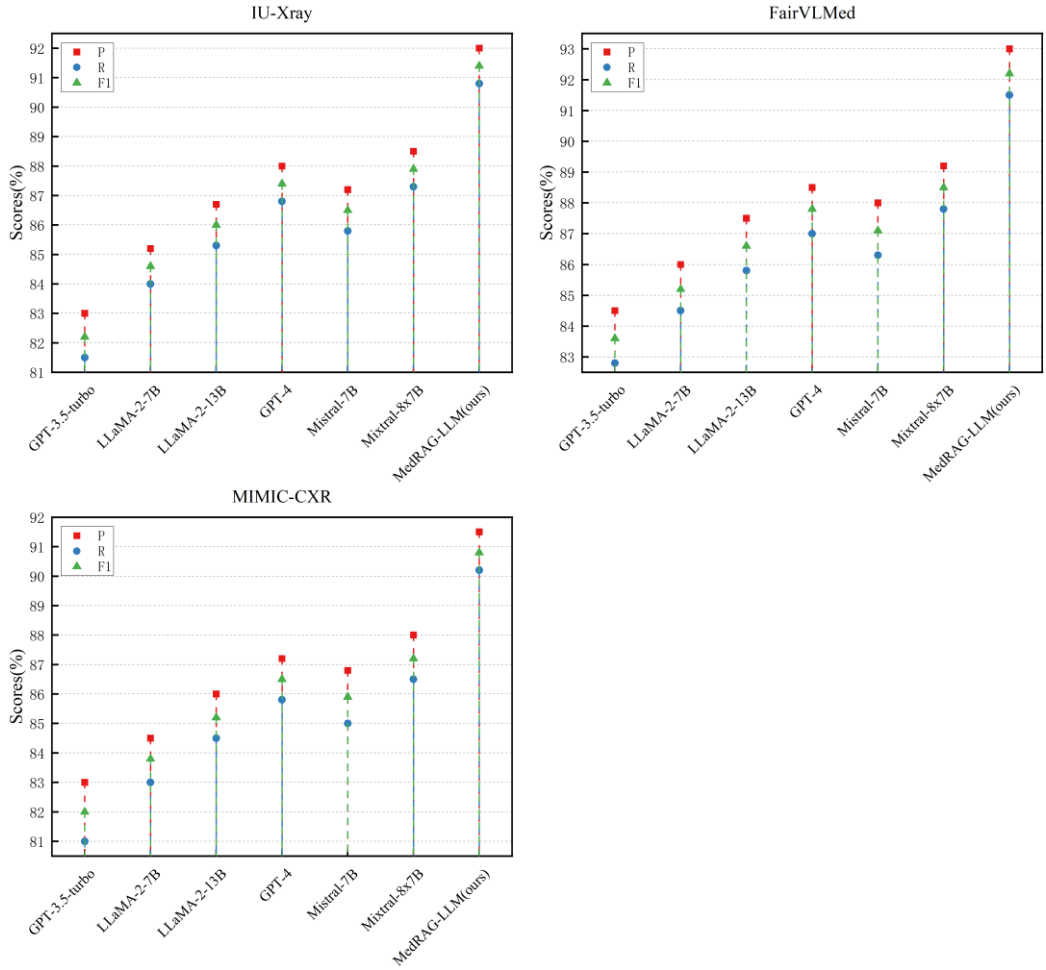

Fig. 5. Performance comparison on IU-Xray, FairVLMed, and MIMIC-CXR datasets.

To further analyze the contributions of individual components within MedRAG-LLM, an ablation study was conducted.

Table II details the ablation study, which highlights the contributions of individual components within the MedRAG-LLM framework. Removing the DSL component results in an average F1-Score drop of 3.0%–4.0%, with the most significant impact observed in recall. This decline underscores DSL's critical role in learning from unlabeled data, which is prevalent in medical datasets. Similarly, removing the RAG mechanism reduces precision by 3.7%–4.5%, as it limits the framework's ability to retrieve and incorporate relevant external knowledge. The absence of Chain-of-Thought (CoT) reasoning results in a smaller but consistent F1-Score reduction of 1.5%–2.0%, demonstrating its importance for logical reasoning in diagnostic tasks. Fig. 6 provides a visual comparison of these results.

TABLE II. ABLATION STUDY ON IU-XRAY, FAIRVLMED, AND MIMIC-CXR DATASETS

Dataset	IU-Xray			FairVLMed			MIMIC-CXR		
	P	R	F1	P	R	F1	P	R	F1
MedRAG-LLM (ours)	92%	90.8%	91.4%	93%	91.5%	92.2%	91.5%	90.2%	90.8%
w/o DSL	89%	87.8%	88.4%	90%	88.5%	89.2%	88.8%	87.2%	88%
w/o RAG	88.5%	87%	87.7%	89.2%	87.5%	88.3%	88%	86.5%	87.2%
w/o CoT	90%	88.8%	89.4%	91%	89.5%	90.2%	89.8%	88%	88.9%

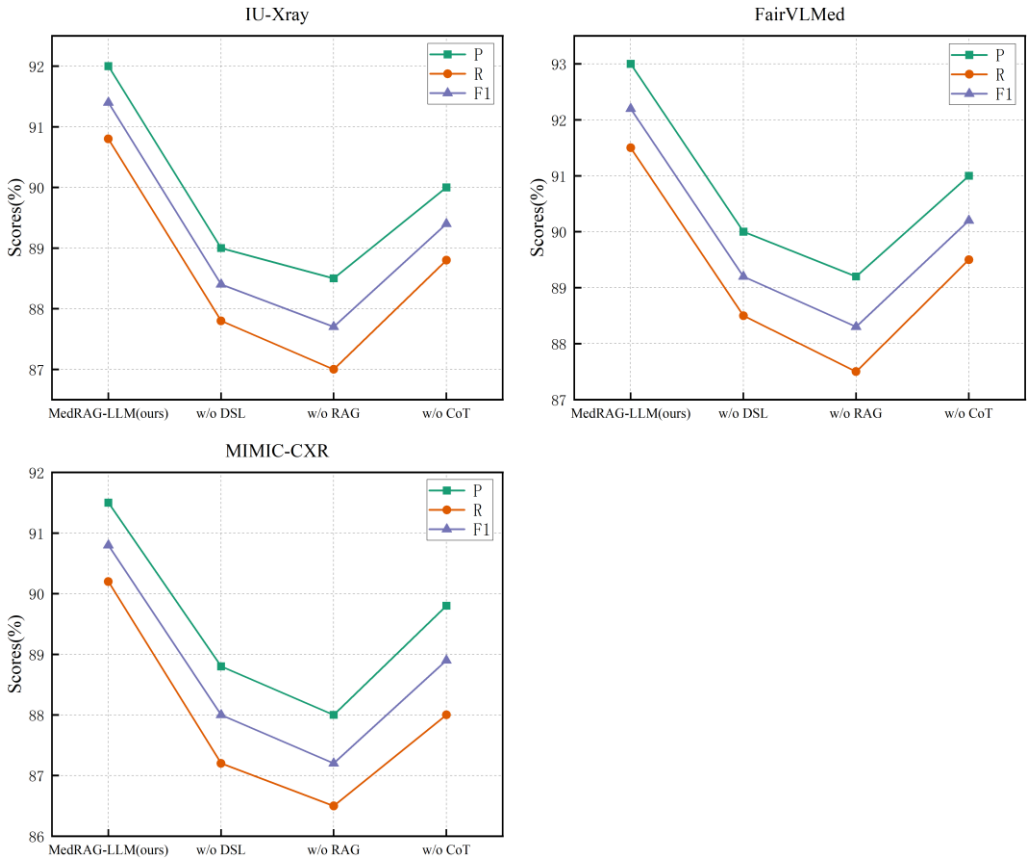

Fig. 6. Ablation study results on IU-Xray, FairVLMed, and MIMIC-CXR datasets.

Evaluation on medical reasoning benchmarks, MMLU-Med and MedQA-USMLE, further validates the effectiveness of MedRAG-LLM.

The results of the medical reasoning benchmarks, as shown in Table III, further validate the effectiveness of MedRAG-LLM. On MMLU-Med, MedRAG-LLM achieves an accuracy of 92.5%, while the best-performing baseline, Mixtral-8x7B, achieves 79.0%. Similarly, on MedQA-USMLE, MedRAG-LLM achieves an accuracy of 93.8%, outperforming GPT-4 (79.5%) by a significant margin. These results reflect the framework's ability to integrate domain knowledge effectively and reason about complex medical questions with high accuracy. Fig. 7 provides a visual comparison of these results.

TABLE III. MMLU-MED AND MEDQA-USMLE EVALUATION

Models	MMLU-Med		MedQA-USMLE		Average
	Setting	Accuracy	Setting	Accuracy	
GPT-3.5-turbo [27]	zero-shot	72%	zero-shot	74%	73%
LLaMA-2-7B [28]	zero-shot	74%	zero-shot	75.5%	74.8%
LLaMA-2-13B [29]	zero-shot	76%	zero-shot	77.5%	76.8%
GPT-4 [30]	zero-shot	78%	zero-shot	79.5%	78.8%
Mistral-7B [31]	zero-shot	77%	zero-shot	78.5%	77.8%
Mixtral-8x7B [32]	zero-shot	79%	zero-shot	80.5%	79.8%
MedRAG-LLM (ours)	zero-shot	92.5%	zero-shot	93.8%	93.2%

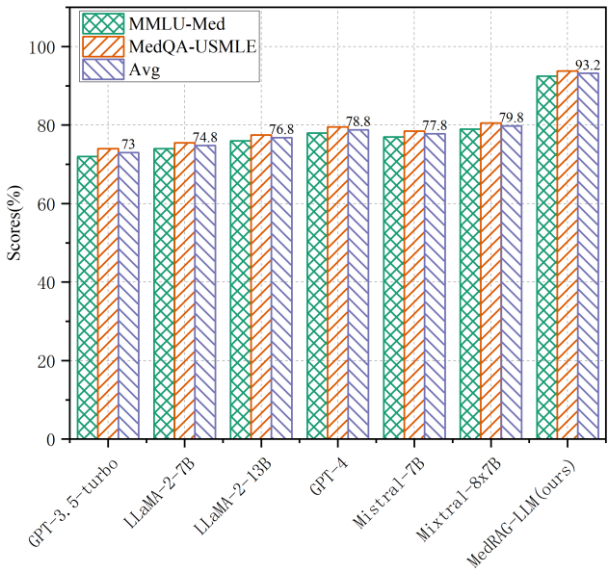

Fig. 7. Comparison of accuracy across MMLU-Med and MedQA-USMLE for different models.

In addition to surpassing traditional RAG and DSL implementations, our framework introduces advancements over emerging methods like Chain-of-Thought (CoT) and ReAct, which emphasize reasoning and task-specific operations. While CoT constructs structured reasoning paths and ReAct integrates reasoning with task execution, both lack the ability to dynamically adapt retrieval processes or leverage unannotated data. Similarly, hybrid RAG approaches enhance retrieval diversity but fall short

in optimizing prompts or integrating self-supervised learning. By combining domain-aware retrieval, adaptive prompt learning, and DSL, our framework bridges these gaps, offering superior adaptability, accuracy, and reasoning capabilities, particularly in complex multimodal medical applications.

Our framework prioritizes data security and privacy through multiple mechanisms tailored for medical applications. The domain-aware retrieval mechanism operates within a secure environment, utilizing a layered cross-chain architecture to ensure safe knowledge integration without exposing sensitive data. Data remains encrypted during processing, and synthetic datasets are employed to adhere to privacy regulations when training and testing. Additionally, the framework dynamically filters and retrieves only relevant data, minimizing unnecessary data access and reducing potential exposure. By incorporating these measures alongside rigorous compliance with medical data standards, our framework ensures robust data security while maintaining high performance and adaptability in medical scenarios.

The proposed framework addresses several critical challenges in medical data analysis. One key issue is the scarcity of labeled data, as annotating medical datasets requires significant expertise and resources [33]. This limitation hinders the training and generalization of traditional models. The framework mitigates this through DSL, which leverages unannotated data to learn intrinsic features, improving adaptability and performance in data-scarce scenarios [34]. Another challenge is ensuring accuracy in medical diagnosis, where errors can have severe consequences. The integration of Retrieval-Augmented Generation (RAG) dynamically retrieves relevant, domain-specific knowledge, enhancing factual reliability and precision in diagnostic tasks while minimizing inaccuracies [35].

V. Conclusion

This study introduces a novel framework, MedRAG-LLM, integrating Retrieval-Augmented Generation (RAG), dynamic prompt learning, and Deep Self-Supervised Learning (DSL) to address the challenges of multimodal medical data analysis and decision support. The framework effectively combines external knowledge retrieval, adaptive reasoning, and unsupervised feature extraction to enhance the efficiency and accuracy of medical image processing and diagnostic tasks. Extensive experiments conducted on benchmark datasets (IU-Xray, FairVLMed, and MIMIC-CXR) and reasoning benchmarks (MMLU-Med and MedQA-USMLE) demonstrate significant improvements over state-of-the-art models, achieving up to 93.8% accuracy in reasoning tasks. These results underline its potential for improving diagnostic accuracy, automating report generation, and enhancing clinical decision-making.

However, despite its strengths, the framework has potential limitations. One limitation is the reliance on high-quality, curated datasets, which may not always be available in clinical environments, potentially affecting the model's generalizability to diverse patient populations. Additionally, while the framework demonstrates promising results in controlled settings, its performance in real-time clinical applications, such as emergency care or resource-constrained environments, remains to be fully evaluated. Future work will focus on improving the framework's robustness by incorporating more diverse datasets, enhancing model adaptability to various clinical scenarios, and optimizing real-time processing capabilities. We also plan to explore the integration of the framework with other healthcare technologies, such as Electronic Health Records (EHR) systems, to provide a more holistic solution for healthcare providers. These efforts aim to expand the framework's applicability and impact in real-world healthcare settings.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, Hanrui Yan and Dan Shao; writing—original draft preparation, Hanrui Yan and Dan Shao; writing—review and editing, Hanrui Yan and Dan Shao. All authors read and agreed to the published final manuscript.

References

[1] A. Zhang, L. Xing, J. Zou, and J. C. Wu, "Shifting machine learning for healthcare from development to deployment and from models to data," *Nat. Biomed. Eng.*, vol. 6, no. 12, pp. 1330–1345, July 2022.

[2] I. Yaqoob, K. Salah, R. Jayaraman, and Y. Al-Hammadi, "Blockchain for healthcare data management: Opportunities, challenges, and future recommendations," *Neural Comput. & Applic*., vol. 34, pp. 11475–11490, July 2022.

[3] M. Javaid, A. Haleem, R. P. Singh, R. Suman, and S. Rab, "Significance of machine learning in healthcare: Features, pillars and applications," *International Journal of Intelligent Networks*, vol. 3, pp. 58–73, June 2022.

[4] M. Kuzlu, Z. Xiao, S. Sarp, F. O. Catak, N. Gurler, and O. Guler, "The rise of generative artificial intelligence in healthcare," in *Proc. 2023 12th Mediterranean Conf. on Embedded Computing (MECO)*, Budva, Montenegro, 2023, pp. 1–4.

[5] S. Kathiriya, S. Nuthakki, S. Mulukuntla, B. V. Charllo, "AI and the future of medicine: Pioneering drug discovery with language models," *International Journal of Science and Research*, vol. 12, no. 3, pp. 1824–1829, Mar. 2023.

[6] P. Lewis *et al*., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. the 34th International Conf. on Neural Information Processing System*, 2020, vol. 33, pp. 9459–9474.

[7] Y. Gao *et al*., "Retrieval-augmented generation for large language models: A survey," arXiv preprint, arXiv:2312.10997, 2023.

[8] Y. Al-Ghadban *et al*., "Transforming healthcare education: Harnessing large language models for frontline health worker capacity building using retrieval-augmented generation," *medRxiv*, 2023. doi: 10.1101/2023.12.15.23300009

[9] M. A. Fink *et al*., "Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer," *Radiology*, vol. 308, no. 3, e231362, Sep. 2023.

[10] X. Zhou *et al*., "D-bot: Database diagnosis system using large language models," in *Proc. the VLDB Endowment*, 2024, pp. 2514–2527.

[11] A. Ahmed, X. Zeng, R. Xi, M. Hou, S. A. Shah, "MED-Prompt: A novel prompt engineering framework for medicine prediction on free-text clinical notes," *Journal of King Saud University-Computer and Information Sciences*, vol. 36, no. 2, 101933, Feb. 2024.

[12] K. Singhal *et al*., "Towards expert-level medical question answering with large language models," arXiv preprint, arXiv:2305.09617, 2023.
[13] C. Wang *et al*., "Survey on factuality in large language models: Knowledge, retrieval and domain-specificity," arXiv preprint, arXiv:2310.07521, 2023.
[14] C. Niu *et al*., "Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models," in *Proc. the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024, pp. 10862–10878.
[15] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, "Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1–17, Jan. 2023.
[16] L. Tamine and L. Goeuriot, "Semantic information retrieval on medical texts: Research challenges, survey, and open issues," *ACM Computing Surveys*, vol. 54, no. 7, pp. 1–38, Sep. 2021.
[17] J. D. Brown, A. Raudaschl, D. J. R. Johnson, and I. Operations, "Anchoring global security: Autonomous shipping with mind reading AI, GPT-core and MAMBA-core agents, RAG-fusion, AI communities, hive-AI, and the human psyche," *Research and Operations, Indaiatuba*, vol. 10, pp. 1–12, Dec. 2023.
[18] B. Chen and A. L. Bertozzi, "AutoKG: Efficient automated knowledge graph generation for language models," in *Proc. 2023 IEEE International Conf. on Big Data (BigData)*, 2023, pp. 3117–3126.
[19] Z. Zhang, A. Zhang, M. Li, and A. Smola, "Automatic chain of thought prompting in large language models," in *Proc. ICLR 2023*, 2023, pp. 1–32.
[20] S. Yao *et al*., "React: Synergizing reasoning and acting in language models," arXiv preprint, arXiv:2210.03629, 2023.
[21] Y. Jiang *et al*., "Adaptive domain interest network for multi-domain recommendation," in *Proc. the 31st ACM International Conf. on Information & Knowledge Management*, 2022, pp. 3212–3221.
[22] J. Pei, Z. Jiang, A. Men, L. Chen, Y. Liu, and Q. Chen, "Uncertainty-induced transferability representation for source-free unsupervised domain adaptation," *IEEE Transactions on Image Processing*, pp. 2033–2048, 2023.
[23] X. Liu *et al*., "WebGLM: Towards an efficient web-enhanced question answering system with human preferences," in *Proc. the 29th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining*, 2023, pp. 4549–4560.
[24] J. Pavlopoulos, V. Kougia, and I. Androutsopoulos, "A survey on biomedical image captioning," in *Proc. the Second Workshop on Shortcomings in Vision and Language*, 2019, pp. 26–36.
[25] Y. Luo *et al*., "Fairclip: Harnessing fairness in vision-language learning," in *Proc. the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2024, pp. 12289–12301.
[26] A. E. Johnson *et al*., "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Sci. Data*, vol. 6, no. 1, 317, Dec. 2019.
[27] C. Y. K. Williams, B. Y. Miao, and A. J. Butte, "Evaluating the use of GPT-3.5-turbo to provide clinical recommendations in the emergency department," *medRxiv*, 2023. doi: 10.1101/2023.10.19.23297276
[28] Z. Li *et al*., "Label supervised llama finetuning," arXiv preprint, arXiv:2310.01208, 2023. doi: 10.48550/arXiv.2310.01208
[29] K. I. Roumeliotis, N. D. Tselikas, and D. K. Nasiopoulos, "Llama 2: Early adopters' utilization of meta's new open-source pretrained model," *Preprints*, 2023072142, 2023. doi: 10.20944/preprints202307.2142.v2
[30] Y.-F. Shea, C. M. Y. Lee, W. C. T. IP, D. W. A. Luk, and S. S. W. Wong, "Use of GPT-4 to analyze medical records of patients with extensive investigations and delayed diagnosis," *JAMA Netw Open*, vol. 6, no. 8, pp. e2325000–e2325000, Aug. 2023.
[31] A. Q. Jiang *et al*., "Mistral 7B," arXiv Print, arXiv:2310.06825, 2023. doi: 10.48550/arXiv.2310.06825
[32] A. Eliseev and D. Mazur, "Fast inference of mixture-of-experts language models with offloading," arXiv preprint, arXiv:2312.17238, 2023.
[33] Z. Luo, H. Yan, and X. Pan, "Optimizing transformer models for resource-constrained environments: A study on model compression techniques," *Journal of Computational Methods in Engineering Applications*, vol. 3, no. 1, pp. 1–12, Nov. 2023. doi: 10.62836/jcmea.v3i1.030107
[34] H. Yan and D. Shao, "Enhancing transformer training efficiency with dynamic dropout," arXiv preprint, arXiv:2411.03236, 2024.
[35] Y. Liu and J. Wang, "AI-driven health advice: Evaluating the potential of large language models as health assistants," *Journal of Computational Methods in Engineering Applications*, vol. 3, no. 1, pp. 1–7, Nov. 2023. doi: 10.62836/jcmea.v3i1.030106