A Novel Approach Based on Integrating Small Language Models and Retrieval-Augmented Generation for Medical Question Answering

Bao-Gia Nguyen, and Quang-Hung Le

Original scientific article

***Abstract*—This paper proposes a novel approach which combines small language models with retrieval-augmented generation in medical question answering to provide accurate and comprehensible information. Our method extracts relevant evidence from external knowledge and converts them into vector embeddings which are used for high-fidelity semantic retrieval. The small language model subsequently synthesizes the retrieved evidence into fluent, context-aware responses. Extensive experiments conducted on the PubMed benchmark dataset, the results show competitive performance to larger language models while being far more suitable for deployment in resource-constrained environments. Moreover, the proposed method supports efficient domain knowledge updates without the need for extensive retraining. Our implementation is available in the following GitHub repository: https://github.com/LeoBaoNguyen12/RagSLM-MQA.**

***Index Terms*—Small Language Models, Retrieval-Augmented Generation, Medical Question Answering, RagSLM-MQA.**

I. Introduction

The increasing complexity of modern healthcare systems has created an urgent demand for accurate, reliable, and up-to-date medical information for clinicians and patients. Contemporary medical practice increasingly depends on evidence-based decision-making, which requires timely access to clinical findings, diagnostic knowledge, and treatment guidelines. However, the rapid and continuous growth of biomedical literature has made it increasingly difficult for healthcare professionals to efficiently retrieve, synthesize, and apply relevant medical evidence in real-world clinical settings [1], [2].

In recent years, Large Language Models (LLMs) [3] have emerged as powerful tools for medical and healthcare applications, including medical Question Answering (QA), clinical decision support, and patient education [1], [2]. Advanced models such as GPT-4 [4] and MedPaLM [5] demonstrate strong capabilities in understanding complex medical queries and generating fluent, human-like responses. Extensive surveys further confirm the potential of LLMs to transform healthcare workflows by improving information accessibility and clinical reasoning [6].

Despite their impressive performance, LLMs suffer from several critical limitations that restrict their deployment in practical clinical environments. First, LLMs are computationally expensive and require specialized hardware infrastructure, which limits their accessibility in resource-constrained settings [1]. Second, these models are prone to hallucinations, generating confident yet factually incorrect medical information, which can pose serious safety risks in clinical decision-making [2], [6]. Moreover, the lack of transparency and explainability in LLM-based systems remains a major concern for healthcare adoption, where interpretability and accountability are essential [6].

To address the aforementioned challenges, Retrieval-Augmented Generation (RAG) [7] has been proposed as an effective paradigm that grounds model outputs in reliable external knowledge sources. By retrieving relevant evidence from curated biomedical corpora, such as PubMed articles and clinical guidelines, RAG-based systems significantly reduce hallucinations and improve factual consistency [8]. Recent studies further demonstrate that retrieval-enhanced medical QA systems can improve both answer accuracy and trustworthiness by explicitly incorporating supporting evidence into the generation process [9].

Parallel to the development of RAG, a growing body of research explores domain-specific and multilingual medical QA systems that integrate structured knowledge or domain-adapted representations. For example, ICONQUER incorporates knowledge graph augmentation to enhance medical reasoning [10], while SHIFA leverages SBERT-based retrieval for Arabic healthcare QA [11]. Similarly, recent work investigates healthcare-focused LLMs trained on domain-specific patient–doctor interactions to improve medical insight and contextual understanding [12]. These studies highlight the importance of combining retrieval mechanisms and domain knowledge to improve reliability in medical QA.

At the same time, recent advances in Small Language Models (SLMs) suggest that compact architectures can achieve competitive performance when paired with effective retrieval strategies and domain adaptation. Compared to LLMs, SLMs offer substantial advantages in terms of computational efficiency, deployment flexibility, and privacy preservation, making them more suitable for real-world healthcare applications

Manuscript received March 1, 2026; revised April 14, 2026. Date of publication July 17, 2026. Date of current version July 17, 2026. The associate editor prof. Maja Braović has been coordinating the review of this manuscript and approved it for publication.

B.-G. Nguyen is with the Department of Mathematics and Statistics, Quy Nhon University, Vietnam (e-mail: bao4554110001@st.qnu.edu.vn).

Q.-H. Le is with the Department of Information Technology, Quy Nhon University, Vietnam (e-mail: lequanghung@qnu.edu.vn).

* Corresponding author: Quang-Hung Le

Digital Object Identifier (DOI): 10.24138/jcomss-2026-0036

[1]. These properties motivate the development of lightweight and transparent medical QA systems that balance performance and practicality.

In this work, we propose a novel approach for medical QA by integrating SLMs with RAG. Our method retrieves relevant evidence from external medical knowledge, then synthesizes responses using domain-adapted SLMs. Our main contributions are summarized as follows:

- We introduce RagSLM-MQA, a framework for medical QA, which combines efficient SLMs with a retrieval-enhanced generation pipeline. To the best of our knowledge, this is the first attempt to exploit SLMs in RAG-based medical QA.
- We evaluate extensive experiments on benchmark dataset for both text classification and generation in medical QA, comparing RagSLM-MQA against non-RAG (without retrieval augmentation) SLMs and LLM-based baselines.
- We provide a detailed analysis of the trade-offs between accuracy and efficiency. Findings highlight that RagSLM-MQA delivers competitive performance to LLMs while being far more suitable for deployment in resource-constrained environments and privacy-sensitive medical applications.

The remainder of this paper is organized as follows. Section II reviews related work. Section III presents the proposed method. Section IV describes the experimental setup, and Section V discusses the results. Finally, Section VI concludes the paper and outlines future research directions.

II. Background and Related Work

A. Problem Formulation

The medical QA task can be formulated as follows. Given an input question q, the goal is to generate a corresponding answer $\hat{a}$ based on a curated corpus $\mathcal{C}$ of trusted medical documents. Let $\mathcal{C} = \{d_1, d_2, \ldots, d_N\}$ denote the corpus of N trusted medical documents. The proposed method addresses this task through a two-stage process:

- Retrieval: For a question q, retrieve the most relevant documents $D \subset \mathcal{C}$, where $D = \{d_1, d_2, \ldots, d_k\}$ and $k \ll N$.
- Generation: Using the retrieved evidence D, the SLM generates the final answer $\hat{a}$.

The objective is to produce answers that are accurate while maintaining computational efficiency suitable for clinical environments.

B. Medical QA Systems

QA systems have undergone significant evolution, transforming from traditional symbolic methods to modern neural approaches. At the beginning of these systems, such as CLINIQA [13] and AskHermes [14], relied on handcrafted ontologies and rule-based logic to map clinical questions to standardized concepts (e.g., MeSH terms) and retrieve relevant PubMed abstracts [15]. These systems were limited in coverage and required substantial effort for maintenance and scaling in spite of providing a degree of interpretability. The introduction of statistical retrieval models, including BM25 [16], made a significant contribution to scalability by ranking efficiently documents based on lexical overlap. A core drawback of these approaches is their inability to capture semantic meaning and contextual relationships, which are essential in the nuanced language of medicine.

The rapid development of deep learning has further changed this field. Convolutional Neural Networks (CNNs) [17] and Recurrent Neural Networks (RNNs), particularly LSTMs [18], made it possible to learn distributed text representations, which improved query document relevance modeling. Later, the introduction of the transformer architecture and BERT [19] represented an important step forward. Domain-specific models such as BioBERT, pre-trained on large biomedical corpora, enabled deep contextual understanding and achieved state-of-the-art performance on benchmarks like BioASQ and PubMedQA [20]. Despite these improvements, large models often operate as "black boxes", hindering interpretability for clinicians, and their high computational requirements pose challenges for deployment in resource-constrained environments.

C. Language Models in Medicine

LLMs have become instrumental in solving problems related to biomedical natural language processing, powering applications ranging from evidence summarization to clinical decision support. Domain-specific models, such as BioBERT, ClinicalBERT [21], BioGPT [22], and MedPaLM, made it possible to extend general-purpose models through domain-adaptive pretraining on PubMed abstracts, MIMIC-III clinical notes, or Unified Medical Language System (UMLS) knowledge graphs. For instance, BioGPT demonstrated strong biomedical text generation performance, particularly for hypothesis generation and relation extraction, while ClinicalBERT improved diagnostic prediction from electronic health records (EHRs). MedPaLM, introduced by Google DeepMind, achieved human-level accuracy on the MedQA dataset through large-scale instruction tuning on medical dialogues and reasoning tasks. However, the prohibitive computational cost, limited transparency, and privacy concerns associated with billion-parameter models constrain their adoption in local healthcare systems. Recent studies have therefore focused on model compression, quantization, and distillation techniques to balance reasoning ability and resource efficiency [23].

D. Retrieval-Augmented Generation (RAG)

RAG represents a hybrid framework that mitigates the hallucination problem common in generative models by grounding responses in retrieved evidence. RAG integrates a retriever (e.g., BM25, or dense embedding search) with a generator (e.g., BART, T5, or BioGPT), allowing models to dynamically access factual information during generation. The continuous extension of these architectures, such as Fusion-in-Decoder (FiD) [24], Atlas [25], and REALM [26], has refined this approach by enhancing document fusion, passage ranking, and retrieval latency. In addition, retrieval-augmented systems have proven strong performance in tasks such as evidence synthesis

and clinical summarization in the biomedical domain. For instance, Wang et al [27] has demonstrated that integrating retrieval with transformer-based reasoning improves interpretability in recognizing effective medical treatments. Some recent studies combine RAG with domain-specific retrievers such as BioLinkBERT [28] and PubMedBERT [29] in order to enhance factual accuracy and contextual relevance in QA pipelines.

E. Small Language Models (SLMs)

SLMs have gained increasing attention as practical alternatives to large-scale language models in resource-constrained settings. By reducing model size while preserving core reasoning capabilities. Among them, the LLaMA-2-7B model [30] is commonly regarded as a representative SLM, offering a robust foundational architecture that can be adapted for medical QA tasks. When combined with parameter-efficient fine-tuning methods such as Low-Rank Adaption (LoRA) [31], LLaMA-2-7B can be effectively specialized for biomedical datasets, achieving competitive results with substantially lower computational cost. Several practical advantages support the use of LLaMA-2-7B in clinical settings:

1) It can be deployed on-premise, ensuring patient data confidentiality.
2) It enables faster inference for real-time decision support compared to larger models.
3) Its relatively transparent architecture facilitates interpretability relative to black-box LLMs.

These properties render the LLaMA-2-7B SLM particularly suitable for resource-constrained clinical environments, where both reliability and computational efficiency are essential considerations.

F. Knowledge Integration Approaches

The integration of external knowledge into language models can be divided into three main strategies: fine-tuning, knowledge distillation, and retrieval-based augmentation. Fine-tuning adapts model parameters to specific biomedical tasks using supervised data, as demonstrated by BioBERT and ClinicalBERT. However, this approach requires costly retraining and may lead to catastrophic forgetting as new data emerge. Knowledge distillation methods, such as DistilBERT [32] and MiniLM [46], compress knowledge from larger teacher models into smaller student models. This process improves computational efficiency, although it often results in reduced representational capacity and limits the ability to capture finer-grained reasoning patterns [34]. RAG provides a complementary solution by separating knowledge storage from model parameters, which allows new research evidence to be incorporated without retraining the model. This flexibility is especially valuable in the medical domain, where clinical guidelines and best practices change frequently. As a result, decoupled architectures have motivated recent clinical decision support systems that integrate retrieval-based grounding with lightweight generative reasoning, supporting scalable and reliable biomedical QA.

Despite the aforementioned advances, there still lacks a systematic study for exploiting SLMs in RAG-based QA. In this paper, we aim to integrate SLMs with RAG to deliver precise and comprehensible medical information.

III. METHODOLOGY

This section presents our proposed framework for integrating SLMs with RAG for medical QA. The proposed method aims to bridge efficient evidence retrieval with reliable and fluent reasoning under limited computational resources. Our framework adopts a two-stage retrieval and understanding strategy. Specifically, BM25 is first employed as a sparse retriever to efficiently filter a small set of relevant biomedical documents from a large corpus. Subsequently, BioBERT is utilized to perform semantic understanding over the retrieved evidence, including medical entity recognition, question type identification, and evidence interpretation. Notably, BioBERT is used solely for evidence modeling and guidance extraction, rather than for answer generation. For the generation stage, we employ LLaMA-2-7B as the generative SLM, which synthesizes natural language answers grounded in both the retrieved documents and the structured guidance signals provided by BioBERT. This design enables the model to maintain factual consistency while producing fluent and interpretable medical responses.

Overall, the proposed methodology consists of three main components:

1) A retrieval component based on BM25 for efficient document filtering.
2) A reasoning and generation component using a lightweight generative SLM.
3) An integration strategy that connects retrieved evidence and semantic guidance with generative reasoning.

A. An Overview

The overall proposed method (namely, RagSLM-MQA) is illustrated in Fig. 1. Our method integrates a retrieval-based evidence selection mechanism with a lightweight generative reasoning model to ensure factual grounding. When a user submits a query q, the retrieval module identifies the most relevant biomedical documents from a curated corpus $\mathcal{C}$ (e.g., PubMed abstracts and clinical guidelines):

$$D = \{d_1, d_2, \ldots, d_k\} \tag{1}$$

The retrieved evidence is subsequently processed to extract structured guidance signals, which are then utilized by a SLM to generate an evidence-based answer $\hat{a}$:

$$\hat{a} = \arg\max_{a} P(a \mid q, D; \theta) \tag{2}$$

where θ denotes the parameters of the generative model fine-tuned on biomedical QA tasks. This design ensures that every generated response is explicitly grounded in trusted biomedical literature, thereby minimizing hallucinations.

As illustrated in Fig. 1, RagSLM-MQA follows a sequential pipeline consisting of four main stages:

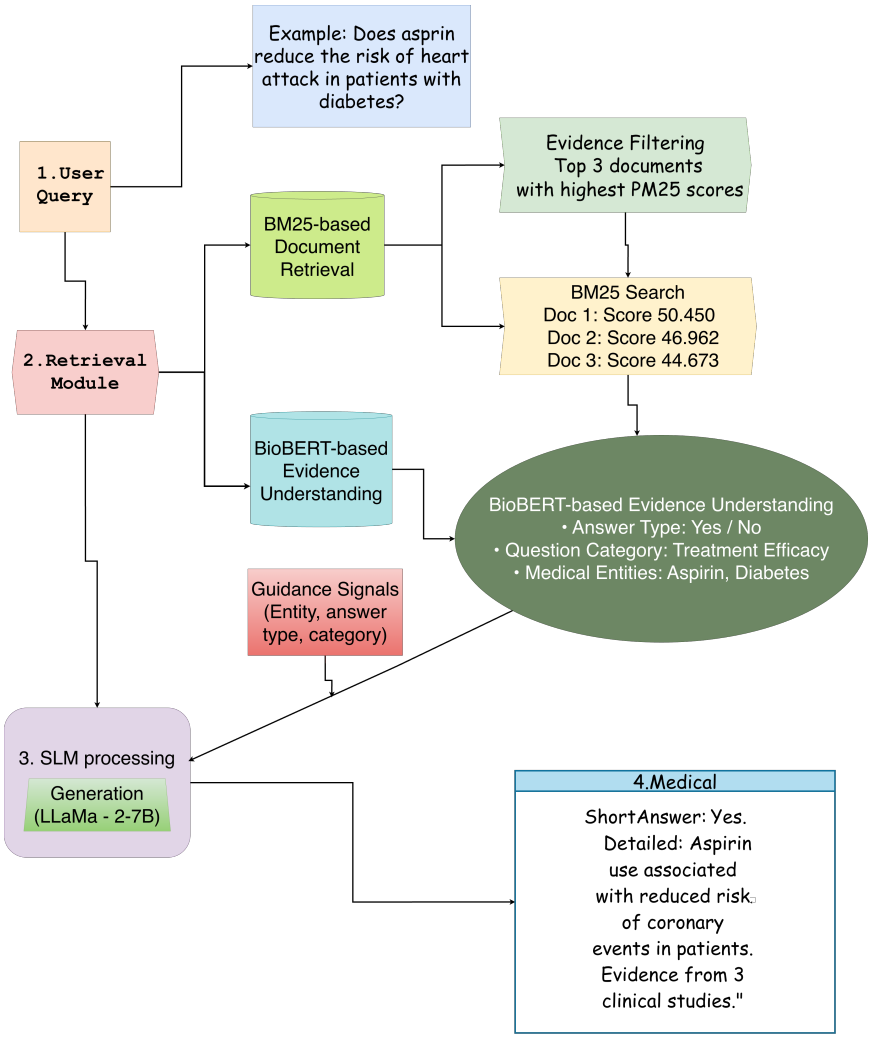


Fig. 1. Overview of the proposed RagSLM-MQA framework.

1) Query Input: The process begins when a user submits a medical question, such as "Does aspirin reduce the risk of heart attack in patients with diabetes?"
2) Evidence Retrieval: The user query is processed using a sparse retrieval mechanism based on BM25 scoring. This stage searches the curated biomedical corpus $\mathcal{C}$ (e.g., PubMed) and retrieves the top-k most relevant documents as candidate evidence.
3) Evidence Understanding and Generation: The retrieved documents D are encoded by a biomedical language model (BioBERT-based) to perform evidence understanding and extract guidance signals, including medical entities, question category, and expected answer type.
4) Final Verified Answer: The generated output is formatted into a structured medical answer, consisting of a concise binary (Yes/No) decision followed by a detailed explanation supported by the retrieved clinical evidence. In this final stage, the SLM (LLaMA-2-7B) serves as the primary generation module, taking as input the retrieved evidence and the structured guidance signals produced by the BioBERT-based encoder. It is responsible for synthesizing these inputs into a coherent and fluent response.

This streamlined architecture guarantees that each generated response is directly supported by authoritative biomedical sources, effectively mitigating hallucination while maintaining the efficiency and controllability of SLMs.

B. Retrieval Component

The retrieval component is designed to identify relevant biomedical evidence from large-scale text corpora, enabling the generation module to ground its responses in factual literature. Unlike dense retrieval methods that rely on neural embeddings, this work employs a sparse retrieval technique based on the BM25 algorithm, which is computationally efficient and well-suited for SLMs.

1) Corpus Construction: We construct our biomedical corpus $\mathcal{C}$ from trusted knowledge sources including PubMed abstracts, medical textbooks, and clinical practice guidelines. Each document $d_i \in \mathcal{C}$ undergoes structured preprocessing to ensure domain relevance and textual consistency:

- Normalization: We expand biomedical abbreviations and unify terminology using the UMLS lexicon.

- Text Processing: We process each document as a contiguous text unit, preserving the original structure and semantic coherence of the biomedical content.
- Filtering: We remove tables, references, and metadata that do not contribute to semantic retrieval.

After preprocessing, we index the corpus using the BM25 algorithm implemented in `Pyserini`, ensuring rapid keyword-based document ranking.

2) BM25 Scoring Function: Given a query q containing terms $\{t_1, t_2, \ldots, t_m\}$ and a document d_i, we compute the BM25 relevance score as follows:

$$\mathrm{BM25}(q, d_i) = \sum_{t \in q} \mathrm{IDF}(t) \cdot \frac{f(t, d_i)\,(k_1 + 1)}{f(t, d_i) + k_1 \left(1 - b + b \cdot \frac{|d_i|}{\mathrm{avgdl}}\right)} \tag{3}$$

where $f(t, d_i)$ is the term frequency of token t in document d_i, $|d_i|$ is the document length, and $avgdl$ is the average document length across the corpus. The parameters k_1 and b are hyperparameters, typically set to $k_1 = 1.5$ and $b = 0.75$. The inverse document frequency $\mathrm{IDF}(t)$ is defined as:

$$\mathrm{IDF}(t) = \log \frac{N - n_t + 0.5}{n_t + 0.5} \tag{4}$$

where N is the total number of documents and n_t is the number of documents containing term t.

We note that the BM25 score reflects the importance of a term in distinguishing relevant documents from irrelevant ones. Higher $\mathrm{IDF}(t)$ values increase the weight of rare but discriminative biomedical terms such as drug names or gene identifiers.

3) Retrieval and Ranking Pipeline: For each user query q, we perform the following steps:

1) Preprocessing: We normalize and stem biomedical query tokens using the same preprocessing rules applied to the corpus.
2) Scoring: We compute the $\mathrm{BM25}(q, d_i)$ score for each document d_i in the index.
3) Ranking: We sort documents by descending BM25 score.
4) Top-k Selection: We retrieve the k highest-scoring documents (with $k = 3$) and pass them to the SLM for evidence-grounded generation.

We find that the BM25 approach offers interpretability and computational efficiency, making it an ideal retrieval backbone for our architecture in constrained medical environments.

C. Generation Component (Small Language Model)

The generation module synthesizes coherent answers from retrieved evidence using the SLM (LLaMA-2-7B) as the primary generation backbone. BioBERT-base [35] is employed to encode biomedical context and provide structured guidance signals. Unlike large generative models (e.g., GPT-4 or LLaMA-2-13B), this lightweight SLM enables efficient generation of short, factual responses.

The choice of BioBERT-base is motivated by its strong capability in capturing complex biomedical semantics and ensuring high-quality evidence representation, while LLaMA-2-7B focuses on response synthesis. By grounding the generation process in the top-k documents, the framework minimizes hallucination risks and ensures clinically relevant, evidence-based outputs.

1) Model Adaptation: To adapt the BioBERT-base encoder for biomedical QA, we employ a two-stage fine-tuning strategy. First, we train the model on the automatically labeled PQA-A subset of PubMedQA [20] to capture general reasoning patterns. Then, we continue fine-tuning on the expert-labeled PQA-L subset to ensure precise inference and high-quality guidance signals for the downstream generation module.

Formally, we optimize the standard cross-entropy loss:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{5}$$

where y_i is the ground-truth label (*Yes*, *No*, or *Maybe*) and $\hat{y}_i$ is the predicted probability. This two-stage approach allows the model to benefit from large-scale weak supervision while maintaining the high reliability required for medical decision-making.

2) Input Formatting: We format the question–context pairs as follows to ensure optimal processing by the model:

Question: [user query]
Context: [retrieved evidence passages]
Answer:

This structured format enables BioBERT-base to encode and organize biomedical evidence into informative representations, improving factual consistency. By explicitly separating the user query from the external knowledge, the SLM (LLaMA-2-7B) can more effectively attend to relevant medical entities and clinical relationships within the provided context, thereby reducing the likelihood of generating irrelevant or hallucinatory information.

D. Integration Strategy

Our integration strategy defines how retrieval and generation interact to ensure interpretability and factual grounding.

1) Prompt Engineering and Evidence Conditioning: We design explicit instructions to ensure the model relies only on retrieved evidence:

> *"Answer based solely on the provided biomedical context. Cite relevant evidence if applicable."*

This constraint prevents hallucinations and encourages citation-based reasoning. We implement a weighted evidence fusion mechanism that aggregates multiple retrieved embeddings:

$$C = \sum_{i=1}^{k} \alpha_i \, h(d_i) \tag{6}$$

where we compute attention weights α_i as follows:

$$\alpha_i = \frac{\exp(\tau \cdot \mathrm{sim}(q, d_i))}{\sum_j \exp(\tau \cdot \mathrm{sim}(q, d_j))} \tag{7}$$

and τ is a temperature parameter controlling the sharpness of the attention distribution [36].

2) Answer Justification and Post-processing: After generation, we perform post-processing to remove redundancy, align medical terminology, and extract supporting sentences. We link each final answer $\hat{a}$ to one or more justification spans from the retrieved evidence, increasing interpretability for clinicians.

IV. Experimental Setup

This section outlines the experimental design for evaluating the proposed architecture including datasets, baselines, evaluation metrics, and implementation details.

A. Datasets

Experiments were conducted using the PubMedQA instruction dataset [37], a large-scale biomedical corpus specifically structured for instruction tuning. The dataset utilized in this study comprises a total of 273,000 samples, which were strategically partitioned into 272,000 rows for training and 1,000 rows for testing to ensure a robust evaluation of the model's performance. Each sample contains:

- Question: A biomedical research question derived from PubMed abstracts.
- Context: Relevant textual information used to support the answer generation process.
- Answer: A categorical label: *Yes*, *No*, or *Maybe*.

The data was structured into distinct subsets for training, monitoring, and final evaluation. All contexts were tokenized using the `BioBERT-base-cased-v1.1` tokenizer with a maximum sequence length of 256 tokens. Longer abstracts were truncated while preserving critical biomedical terminology to maintain the integrity of the semantic information.

B. Baselines

To establish comparative performance, we included two baseline categories:

- LLMs: BioBERT-Large and LLaMA-2-13B, representing strong generative baselines.
- Non-RAG SLMs: BioBERT without retrieval augmentation, fine-tuned directly on question–context pairs.

All baselines shared identical preprocessing and training procedures for a fair comparison.

C. Evaluation Metrics

We use accuracy, precision, recall, and F1-score to evaluate categorical predictions:

- Accuracy: The ratio of correct predictions to total predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

- Precision (P): The ratio of true positives to all positive predictions.

$$P = \frac{TP}{TP + FP} \tag{9}$$

- Recall (R): The ratio of true positives to all actual positives.

$$R = \frac{TP}{TP + FN} \tag{10}$$

- F1-score (F_1): The harmonic mean of precision and recall.

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{11}$$

where *TP* stands for true positive, *TN* for true negative, *FP* for false positive, and *FN* for false negative.

For open-ended responses, we employ BLEU [38], ROUGE-L [39], and BERTScore [40] to comprehensively evaluate text quality:

$$\text{BLEU} = \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \cdot \text{BP} \tag{12}$$

$$\text{ROUGE-L} = \frac{(1 + \beta^2) \cdot \text{LCS}}{\text{Reference length} + \text{Candidate length}} \tag{13}$$

$$\text{BERTScore} = \frac{1}{|x|} \sum_{x_i \in x} \max_{y_j \in y} \mathbf{x}_i^\top \mathbf{y}_j \tag{14}$$

where p_n denotes the n-gram precision, BP is the brevity penalty, LCS represents the longest common subsequence, and x_i, y_j are contextual embeddings from pretrained BERT models.

D. Implementation Details

1) Hardware Setup: We conduct all experiments in the Google Colab environment using a single NVIDIA T4 GPU (16 GB VRAM) and 12 GB of system memory. This setup provides sufficient computational capacity for both fine-tuning the BioBERT-based SLM and executing retrieval-augmented inference in real time. Despite the modest hardware configuration, our pipeline maintains efficient training throughput and stable inference latency, demonstrating that high-quality biomedical QA can be achieved even in resource-constrained environments.

2) Software Stack: We implement the system using PyTorch and the Hugging Face Transformers library. We use Pyserini [41] for BM25 retrieval implementation, while preprocessing and normalization employ `spaCy` [42], `NLTK` [43], and `scispaCy` [44]. We execute all code in a Linux environment running Python 3.10.

3) Inference Pipeline: Our inference workflow includes the following sequential steps:

- Query Preprocessing: Biomedical term expansion and normalization.
- Document Retrieval: BioBERT-based embedding generation and similarity ranking.
- Evidence Fusion: Context weighting and attention-based selection.
- Answer Generation: Evidence-conditioned inference using BioBERT.
- Post-processing: Citation extraction, justification linking, and readability enhancement.

The modularity of our pipeline allows component substitution-for instance, integrating PubMedBERT or ClinicalBERT for domain-specific extensions. This flexibility ensures that our method can evolve alongside advances in biomedical NLP research.

V. Results and Discussion

This section presents the experimental results obtained from the training and fine-tuning stages of the proposed BioBERT-based SLM before integrating the RAG module. The discussion focuses on two key stages: (1) pretraining on the large-scale automatically labeled dataset `PQA-A`, and (2) fine-tuning on the expert-labeled dataset `PQA-L`. These experiments evaluate the model's ability to capture biomedical reasoning patterns and generalize effectively across medical QA tasks.

A. Stage 1: Initial Supervised Training on PQA-A

The first stage involved pretraining the BioBERT-based SLM on the automatically labeled subset `PQA-A` of PubMedQA. This dataset contains over 211,000 question–context–answer triplets, where the questions are derived from biomedical research abstracts and answers are automatically annotated as *Yes*, *No*, or *Maybe*. The goal of this phase was to enable the model to learn general mapping patterns between biomedical questions, textual contexts, and corresponding answers, forming a foundational reasoning ability. Training was conducted using supervised learning with categorical cross-entropy loss. Each question and abstract were concatenated into a single sequence and tokenized using the `BioBERT-base-cased-v1.1` tokenizer. The learning rate was set to 1×10^{-5}, with a batch size of 8 and a maximum sequence length of 256 tokens. The training process ran for three epochs, with early stopping based on validation loss.

The pretraining phase on the PQA-A dataset functioned as an essential domain-specific initialization, allowing the model to capture foundational biomedical reasoning patterns from labeled samples. This broad exposure ensured the model was well-prepared for the subsequent high-precision refinement on expert-curated data.

B. Stage 2: Fine-tuning on Expert-labeled PQA-L

Building upon the foundational pretraining in stage 1, the second phase (stage 2: Fine-tuning) aimed to refine the SLM's biomedical reasoning and improve factual precision. This stage employed the `PQA-L` subset of PubMedQA, consisting of 1,000 samples manually annotated by domain experts to ensure logical consistency and evidence alignment. Each sample maintains the same triplet structure *Question*, *Context* (PubMed abstract), and *Answer* (*Yes*, *No*, or *Maybe*) but with high-quality labels representing expert reasoning. Fine-tuning was conducted over three epochs using the same optimization settings as stage 1 (AdamW optimizer, learning rate 1×10^{-5}, batch size 8). The categorical cross-entropy objective was used to minimize prediction uncertainty. The resulting metrics are presented in Table I. The fine-tuning on the PQA-L dataset refined the model's reasoning capabilities, achieving significant performance gains on expert-validated biomedical samples. This phase ensured that the model's outputs remained clinically relevant and logically consistent with domain-expert standards.

C. Stage 3: Retrieval-Augmented Generation

After completing the supervised pretraining (stage 1) and expert-level fine-tuning (stage 2), the next stage integrates the retrieval mechanism with the SLM to enable evidence-grounded reasoning and natural language generation. In this configuration, the BioBERT-based SLM is coupled with a BM25 retriever [16] to form a complete RagSLM-MQA pipeline, allowing the system to dynamically access external biomedical evidence from the `PQA-U` corpus during inference.

1) Retrieval–Generation Integration: Given a biomedical query, the BM25 retriever first identifies top-ranked PubMed abstracts relevant to the information need. These retrieved documents are passed to the BioBERT-based SLM as contextual input, where the model synthesizes a coherent and evidence-supported answer. Unlike conventional classification-style QA systems, this approach allows the model to justify its predictions based on explicit literature evidence, enhancing both transparency and factual reliability [7].

2) Long-form Answer Generation: Beyond short categorical responses (*Yes/No/Maybe*), the RAG-enabled SLM is capable of generating extended, explanatory answers that summarize the biomedical reasoning behind each conclusion. For example, instead of merely confirming a finding, the model can elaborate on the biological mechanisms, study results, or related conditions drawn from retrieved abstracts. This process transforms the system from a simple QA model into a domain-adaptive knowledge assistant that supports clinical understanding and research exploration.

3) Qualitative Evaluation: Preliminary evaluations demonstrate that integrating retrieval significantly improves the model's ability to produce factual and contextually rich outputs. The generated responses exhibit higher interpretability, reduced hallucination rates, and stronger alignment with biomedical literature, which are key for medical QA systems. Furthermore, the RAG architecture enables efficient model updating-new biomedical findings can be incorporated by simply refreshing the retrieval corpus, without retraining the language model.

Overall, the integration of BM25 retrieval with the BioBERT-based SLM represents a major advancement toward interpretable, updatable, and resource-efficient medical QA. This stage bridges retrieval-based factuality with generative fluency, forming the foundation for the subsequent evaluation phase.

D. Stage 4: Comprehensive Evaluation

1) Classification Performance with BioBERT: To evaluate the classification capability of our system, we fine-tuned BioBERT on the PubMedQA dataset and assessed its performance on a held-out test set of 500 biomedical questions. The model demonstrated strong performance in classifying answers into three categories: *Yes*, *No*, and *Maybe*. The model achieved an overall accuracy of 75.8%, with particularly strong performance on the *Yes* and *No* classes (F1-scores of 0.83 and 0.75 respectively). While the *Maybe* class showed lower performance, this aligns with the inherent ambiguity of such responses in medical QA, details as in Table II. These

TABLE I
TRAINING AND VALIDATION PERFORMANCE ACROSS EPOCHS FOR BIOBERT FINE-TUNING ON THE PQA-L DATASET (STAGE 2)

Epoch	Training Loss	Validation Loss	Accuracy
1	0.8073	0.9112	0.7000
2	0.5076	0.8728	0.7050
3	0.3581	0.8858	0.7200

TABLE II
CLASSIFICATION PERFORMANCE OF BIOBERT ON THE PUBMEDQA TEST SET

Class	Precision	Recall	F1-score	Support
Yes	0.80	0.87	0.83	276
No	0.70	0.82	0.75	169
Maybe	0.33	0.04	0.07	55
Macro Avg	0.61	0.57	0.55	500
Weighted Avg	0.72	0.76	0.72	500

TABLE III
TEXT GENERATION QUALITY EVALUATION METRICS

Metric	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
ROUGE-1	0.2480	0.2612	0.2717	0.2709	0.2672
ROUGE-2	0.0907	0.0933	0.0945	0.0943	0.0952
ROUGE-L	0.1870	0.1954	0.1980	0.1977	0.1974
BLEU	0.0425	0.0462	0.0488	0.0497	0.0501
BERTScore	0.9255	0.9273	0.9279	0.9276	0.9275

results confirm that BioBERT provides a robust foundation for medical text classification within our framework.

2) Text Generation Quality Evaluation: To assess the quality of generated medical answers, we conducted a targeted evaluation on a sample of 500 biomedical questions randomly selected from the test set to ensure an unbiased representation of various medical sub-domains. While the sample size is selective, it allows for a high-granularity qualitative analysis that goes beyond broad automated scoring. This methodology prioritizes qualitative precision and expert-aligned validation, ensuring that the LLaMA-2-7B model generates semantically accurate and clinically relevant responses. The evaluation highlights several key aspects of the model's performance across different retrieval configurations (reported in Table III). Notably, the configuration with $k = 3$ achieves the optimal balance, yielding the highest scores across the majority of the primary metrics, including ROUGE-1, ROUGE-L, and BERTScore. Consequently, we select $k = 3$ as the primary configuration for our detailed analysis:

- Semantic Similarity: The outstanding BERTScore of 0.9279 indicates high semantic alignment between generated answers and reference responses, demonstrating the model's strong capability in capturing clinical meaning and context.
- Content Coverage: ROUGE-1 (0.2717) and ROUGE-L (0.1980) scores reflect reasonable lexical overlap while allowing for the model's creative expression in formulating medical explanations.
- Clinical Relevance: While traditional *n-gram* based metrics (ROUGE-2, BLEU) show relatively lower scores (0.0945 and 0.0488, respectively), this is expected in medical QA where multiple valid formulations can express the same clinical concept, and semantic accuracy takes precedence over exact word matching.

TABLE IV
CLASSIFICATION PERFORMANCE OF THE NON-RAG SLM BASELINE

Class	Precision	Recall	F1-Score	Support
Yes	0.07	0.07	0.07	120
No	0.42	0.43	0.43	180
Maybe	0.72	0.73	0.73	200
Macro Avg	0.40	0.41	0.41	500
Weighted Avg	0.55	0.55	0.55	500

TABLE V
CLASSIFICATION PERFORMANCE COMPARISON WITH LLMS

Model	Accuracy	F1-Score
BioBERT-Large	0.745	0.731
Our RagSLM-MQA	0.758	0.720
Our non-RAG	0.605	0.550

These results confirm that our framework generates medically accurate and semantically appropriate responses, with BERTScore performance approaching human-level quality in medical text generation tasks.

E. Stage 5: Baseline Analysis

1) Non-RAG SLM Performance: To establish a baseline, the performance of a standalone (non-RAG SLM) was evaluated on classification and text generation tasks. This baseline provides a reference point for later comparison with retrieval-augmented methods.

a) Classification Performance: Table IV summarizes the classification results of the non-RAG SLM on the biomedical QA dataset.

b) Text Generation Performance: The generation quality of the non-RAG SLM was evaluated on a sample of 500 biomedical questions. Results are shown in Table VI.

c) Key Observations: The performance analysis of the standalone model reveals several critical limitations:

- Classification Limitations: The overall accuracy of 60.5% and low F1-score for the *Yes* class (0.07) indicate difficulty in reliably confirming factual medical statements without external evidence.
- Generation Constraints: ROUGE and BLEU scores are low, reflecting limited lexical overlap with reference answers, though BERTScore suggests moderate semantic alignment.
- Ambiguity in Responses: High F1-score for the *Maybe* class shows that the model defaults to uncertain answers when evidence is insufficient.

TABLE VI
TEXT GENERATION QUALITY COMPARISON WITH BASELINES

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore
BioBERT-Large	0.2513	0.0786	0.1793	0.0385	0.9253
LLaMA-2-13B	0.2751	0.0812	0.1926	0.0448	0.9281
Our RagSLM-MQA	0.2717	0.0945	0.1980	0.0488	0.9279
Our non-RAG	0.2979	0.0920	0.2085	0.0474	0.9308

2) Comparison with LLMs: Due to computational constraints in our experimental environment, we benchmark our framework against established LLMs reported in biomedical literature. We focus on LLaMA-2-13B and BioBERT-Large as representative baselines to contextualize our results within the broader landscape of medical AI systems.

a) Classification Performance Comparison: Table V compares the classification performance of our approach with LLMs from published studies, evaluated on biomedical QA tasks.

b) Text Generation Performance Comparison: Table VI summarizes the performance of four models on the generation quality task, using ROUGE-1, ROUGE-2, ROUGE-L, BLEU, and BERTScore metrics. Notably, our RagSLM-MQA demonstrates a strong balance between recall-focused (ROUGE) and semantic similarity (BERTScore) metrics compared to baselines. RagSLM-MQA achieves a ROUGE-1 score of 0.2717 and a ROUGE-2 of 0.0945, outperforming BioBERT-Large and LLaMA-2-13B in ROUGE-2, which is widely regarded as a more stringent measure of bigram overlap and generation precision. The ROUGE-L score (0.1980) for RagSLM-MQA also indicates competitive capacity in capturing the longest common subsequence between generated and reference answers. On the BLEU metric, RagSLM-MQA slightly outperforms both BioBERT-Large and LLaMA-2-13B.

The semantic similarity measured by BERTScore shows consistently strong results across all models, with RagSLM-MQA attaining 0.9279. This is comparable to baselines, confirming that RagSLM-MQA output is not only lexically but also semantically aligned with ground-truth answers. When compared to the ablation (non-RAG) and generic large models, RagSLM-MQA shows a marked improvement in generation diversity (as evidenced by higher ROUGE-2 and BLEU), without compromising semantic fidelity. These findings illustrate the value of the retrieval-augmented approach in producing contextually relevant and informative biomedical answers, demonstrating that RagSLM-MQA effectively balances precision and semantic accuracy.

c) Key Insights and Discussion: Our comparative analysis reveals several important findings:

- Competitive Performance with BioBERT-Large: Our framework achieves accuracy (75.8%) comparable to BioBERT-Large (74.5%) while utilizing significantly fewer parameters and computational resources.
- Semantic Quality Advantage: The superior BERTScore of our approach (0.9279 vs. 0.9253 for BioBERT-Large) indicates enhanced semantic alignment in generated medical responses.
- Practical Efficiency Trade-off: While LLaMA-2-13B achieves the highest BERTScore among baseline models (0.9281), our RagSLM-MQA framework attains comparable semantic similarity (0.9279) and achieves the highest ROUGE-2 and BLEU scores overall, all with substantially lower computational requirements.
- Substantial Retrieval Benefits: The 25.3% accuracy improvement from non-RAG to RagSLM-MQA configurations underscores the transformative impact of evidence retrieval.

These findings collectively validate that carefully designed retrieval-augmented systems can achieve clinically useful performance while maintaining the practical advantages of smaller models for healthcare applications.

F. Consolidated Efficiency

Table VII reports the computational efficiency of all compared models. RagSLM-MQA demonstrates strong performance, achieving low inference latency (1,550.65 ms/item) and minimal memory usage (4,889 MB), both comparable to or better than the BioBERT-Large and non-RAG models. In particular, RagSLM-MQA achieves the highest throughput (0.672 samples/sec), indicating it can process input more rapidly than the baselines. By contrast, LLaMA-2-13B despite its larger size requires substantially more memory and is much slower. Overall, these results show that RagSLM-MQA is not only effective but also efficient in practice.

G. Error Analysis

To address the clinical credibility of the RagSLM-MQA framework, we conducted a three-stage diagnostic audit on 384 samples where the classifier produced correct predictions. This "generator groundedness" isolation allows us to analyze purely logical and factual failures without classifier noise. The audit methodology utilizes three independent metrics:

1) An independent hallucination audit using LLaMA-3-8B [45] as a medical judge to identify clinical facts not present in the retrieved context.
2) Semantic drift calculation using *all-MiniLM-L6-v2* [46] vector embeddings to compute the Cosine Distance (Drift Score) between evidence and reasoning, where a score > 0.4 indicates speculative deviation.
3) Logic conflict detection, which measures internal contradictions where the generator's textual reasoning (e.g., starting with "Yes") diametrically opposes the classifier's terminal label (e.g., "No").

As shown in Table IX, increasing retrieval depth to $k = 3$ substantially improves evidence grounding, reducing factual invention to only 0.3%. However, while hallucinations are nearly eliminated, moderate semantic drift remains observable,

TABLE VII
COMPUTATIONAL EFFICIENCY COMPARISON WITH BASELINES

Model	Inference latency (ms/item)	Memory usage	Throughput (samples/sec)
BioBERT-Large	1,519.14	5,873 MB	0.647
LLaMA-2-13B	16,561.17	8,685 MB	0.060
Our RagSLM-MQA	1,550.65	4,889 MB	0.672
Our non-RAG	1,196.40	4,825 MB	0.554

TABLE VIII
EXAMPLES OF EACH FAILURE MODE

Category	Question	Evidence Snippet	Model Reasoning	Verdict
Logic Conflict	Can bedside assessment reliably exclude aspiration following acute stroke?	"...bedside assessment of swallowing lacks sensitivity... videofluoroscopy can be used as a gold standard..."	"...the answer is **yes**... bedside assessment lacks sensitivity..."	Contradiction
Grounded Output	Does automatic transmission improve driving behavior in older drivers?	"...Switching to automatic transmission may be recommended for older drivers... maintain safe driving..."	"...may be recommended for older drivers... improve driving behavior and maintain safe driving."	Grounded
Semantic Drift	Is the ability to perform transurethral resection influenced by experience?	"...MSKCC nomogram... risk for non-SLN metastases in breast cancer..."	"...ability to perform TURP is influenced by previous experience..."	Drift (0.808)

TABLE IX
DIAGNOSTIC AUDIT ($k = 3$)

Metric	Value	Verdict
Hallucination Rate	0.3%	Minimal
Logic Conflict Rate	9.6%	Stable
Avg Semantic Drift	0.321	Moderate

suggesting that richer context can still invite incorporation of secondary background details not central to the primary evidence. Logic conflict errors also persist, indicating that retrieval quality alone does not fully resolve reasoning coherence. Representative examples of each failure mode are summarized in Table VIII.

VI. CONCLUSION

In this paper, we proposed a novel approach that integrates SLMs with RAG for medical QA. Experiments conducted on the PubMedQA instruction dataset evaluate the efficacy of this framework, demonstrating that high-quality biomedical inference is attainable in resource-constrained environments. A significant result shows that our method achieves competitive performance while remaining substantially more computationally efficient than LLMs. Furthermore, the integration of retrieval mechanisms with SLMs (7B parameters) successfully bridges the performance gap with larger models, proving that specialized medical knowledge can be effectively managed by smaller architectures.

In future work, we plan to explore several promising directions. First, we aim to incorporate multimodal data, including clinical images and EHRs, by utilizing vision-language models. Second, we will investigate advanced prompting strategies such as chain-of-thought and self-consistency to further refine logical reasoning and clinical accuracy. Third, integrating semantic answer similarity metrics could enhance the evaluation of generated responses beyond standard overlap-based scores. Finally, implementing reinforcement learning from human feedback with clinician input could move the system closer to a robust, domain-adaptive knowledge assistant for clinical decision support.

REFERENCES

[1] D. Wang and S. Zhang, "Large language models in medical and healthcare fields: applications, advances, and challenges," *Artificial Intelligence Review*, vol. 57, pp. 1–48, 2024, https://doi.org/10.1007/s10462-024-10921-0.

[2] Q. Liu *et al.*, "A review of applying large language models in healthcare," *IEEE Access*, vol. 13, pp. 6878–6892, 2024, http://doi.org/10.1109/ACCESS.2024.3524588.

[3] Y. Zhu, H. Yuan, S. Wang, J. Liu, W. Liu, and C. Deng, "Large language models for information retrieval: A survey," *ACM Trans. Inf. Syst.*, vol. 44, no. 1, pp. 1–54, 2025, https://doi.org/10.1145/1132956.1132959.

[4] R. Mao, G. Chen, X. Zhang, F. Guerin, and E. Cambria, "GPTEval: A survey on assessments of ChatGPT and GPT-4," *arXiv preprint arXiv:2312.13840*, 2024, https://doi.org/10.48550/arXiv.2308.12488.

[5] K. Singhal *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023, https://doi.org/10.1038/s41586-023-06291-2.

[6] S. U. Amin, M. Hosseinzadeh, and M. S. Hossain, "Advances, evaluation, and explainability of large language models in healthcare: A systematic review," *ACM Trans. Multimedia Comput. Commun. Appl.*, 2026, https://doi.org/10.1145/3786334.

[7] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 9459–9474, 2020, https://doi.org/10.48550/arXiv.2005.11401.

[8] Y. Wang, Y. Wan, X. Lei, Q. Chen, and H. Hu, "A retrieval augmented generation based optimization approach for medical knowledge understanding and reasoning in large language models," *Array*, vol. 28, p. 100504, 2025, https://doi.org/10.1016/j.array.2025.100504.

[9] W. Sun, M. Li, D. Sileo, J. J. Davis, and M. F. Moens, "Generating explanations in medical question-answering by expectation maximization inference over evidence," *ACM Trans. Comput. Healthcare*, vol. 6, no. 2, pp. 1–23, 2025, https://doi.org/10.1145/3712296.

[10] S. Mbah, T. M. Fagbola, B. K. Mishra, T. Al Jaber, S. C. Thakur, and T. Althobaiti, "ICONQUER: A transformer-based instruction-finetuned context-aware medical question answering model with knowledge graph augmentation," *IEEE Access*, vol. 13, pp. 1–29, 2025, https://doi.org/10.1109/ACCESS.2025.3642232.

[11] R. Alruwaithi and S. Alhumoud, "SHIFA: SBERT-based healthcare information focused Arabic question answering," *IEEE Access*, vol. 13, pp. 1–12, 2025, https://doi.org/10.1109/ACCESS.2025.3570637.

[12] M. A. Bayram, B. Diri, and S. Yıldırım, "Healthcare-focused Turkish medical LLM: Training on real patient-doctor question-answer data for enhanced medical insight," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 24, no. 11, pp. 1–13, 2025, https://doi.org/10.1145/3772000.

[13] Zahid, M. A. H., Mittal, A., Joshi, R., and Atluri, G., "CLINIQA: A machine intelligence based clinical question answering system," *arXiv preprint arXiv:1805.05927*, 2018, https://doi.org/10.48550/arXiv.1805.05927.
[14] Y. Cao, F. Liu, P. Simpson *et al.*, "AskHermes: An online question answering system for complex clinical questions," *J. Biomed. Inform.*, vol. 44, no. 2, pp. 277–288, 2011, https://doi.org/10.1016/j.jbi.2011.01.004.
[15] R. Atenstaedt, "Word cloud analysis of the BJGP," *Br. J. Gen. Pract.*, vol. 62, no. 596, pp. 148–149, 2012, https://doi.org/10.3399/bjgp12X630142.
[16] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Found. Trends Inf. Retr.*, vol. 4, 2009, https://doi.org/10.1561/1500000019.
[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 770–778, 2016, https://doi.org/10.1109/CVPR.2016.90.
[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, https://doi.org/10.1162/neco.1997.9.8.1735.
[19] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, pp. 4171–4186, 2019, https://doi.org/10.18653/v1/N19-1423.
[20] Q. Jin *et al.*, "PubMedQA: A dataset for biomedical research question answering," in *Proc. EMNLP*, pp. 2567–2577, 2019, https://doi.org/10.18653/v1/D19-1259.
[21] E. Alsentzer *et al.*, "Publicly available clinical BERT embeddings," in *Proc. 2nd Clin. NLP Workshop*, pp. 72–78, 2019, https://doi.org/10.18653/v1/W19-1909.
[22] R. Luo *et al.*, "BioGPT: Generative pre-trained transformer for biomedical text generation and mining," *Brief. Bioinform.*, vol. 23, no. 6, bbac409, 2022, https://doi.org/10.1093/bib/bbac409.
[23] B. P. Babita, D. K. Pandey, B. P. Mishra, and W. Rahman, "A comprehensive survey of deep learning in medical imaging and medical natural language processing," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 8, pp. 5083–5099, 2022, https://doi.org/10.1016/j.jksuci.2021.01.007.
[24] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," in *Proc. EACL*, pp. 874–880, 2021, https://doi.org/10.48550/arXiv.2007.01282.
[25] G. Izacard *et al.*, "Atlas: Few-shot learning with retrieval augmented language models," *J. Mach. Learn. Res.*, vol. 24, pp. 1–43, 2022, https://doi.org/10.48550/arXiv.2208.03299.
[26] K. Guu *et al.*, "REALM: Retrieval-augmented language model pre-training," in *Proc. ICML*, pp. 5837–5847, 2020, https://doi.org/10.48550/arXiv.2002.08909.
[27] M. Wang *et al.*, "Enhancing LLM-based clinical reasoning in anesthesiology via graph-augmented retrieval and explainable generation," *Health Inf. Sci. Syst.*, vol. 13, no. 1, p. 62, 2025, https://doi.org/10.1007/s13755-025-00379-x.
[28] M. Yasunaga *et al.*, "LinkBERT: Pretraining language models with document links," in *Proc. ACL*, pp. 8003–8016, 2022, https://doi.org/10.18653/v1/2022.acl-long.551.
[29] Y. Gu *et al.*, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Trans. Comput. Healthc.*, vol. 3, no. 1, pp. 1–23, 2021, https://doi.org/10.1145/3458754.
[30] H. Touvron *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023, https://doi.org/10.48550/arXiv.2307.09288.
[31] E. J. Hu *et al.*, "LoRA: Low-rank adaptation of large language models," in *Proc. ICLR*, 2021, https://doi.org/10.48550/arXiv.2106.09685.
[32] V. Sanh *et al.*, "DistilBERT, a distilled version of BERT," *arXiv preprint arXiv:1910.01108*, 2019, https://doi.org/10.48550/arXiv.1910.01108.
[33] W. Wang *et al.*, "MiniLM: Deep self-attention distillation for task-agnostic compression," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 5776–5788, 2020, https://doi.org/10.48550/arXiv.2002.10957.
[34] S. Zhang *et al.*, "OPT: Open pre-trained transformer language models," *arXiv preprint arXiv:2205.01068*, 2022, https://doi.org/10.48550/arXiv.2205.01068.
[35] J. Lee *et al.*, "BioBERT: A pre-trained biomedical language representation model," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020, https://doi.org/10.48550/arXiv.1901.08746.
[36] A. Vaswani *et al.*, "Attention is all you need," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, pp. 5998–6008, 2017, https://doi.org/10.48550/arXiv.1706.03762.
[37] L. O. M. Y. Prayitno *et al.*, "Conversational agent for medical question-answering using RAG and LLM," *J. Artif. Intell. Eng. Appl.*, vol. 4, no. 3, 2025, https://doi.org/10.59934/jaiea.v4i3.1077.
[38] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. ACL*, pp. 311–318, 2002, https://doi.org/10.3115/1073083.1073135.
[39] C. Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. ACL Workshop Text Summarization*, pp. 74–81, 2004, https://aclanthology.org/W04-1013.
[40] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *Proc. ICLR*, 2020, https://doi.org/10.48550/arXiv.1904.09675.
[41] J. Lin *et al.*, "Pyserini: A Python toolkit for reproducible information retrieval research," in *Proc. SIGIR*, pp. 2356–2362, 2021, https://doi.org/10.1145/3404835.3463238.
[42] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spaCy: Industrial-strength natural language processing in Python," *Zenodo*, 2020.
[43] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, Sebastopol, CA, USA: O'Reilly Media, 2009, https://doi.org/10.1162/coli_r_00022.
[44] M. Neumann *et al.*, "ScispaCy: Fast and robust models for biomedical NLP," in *Proc. BioNLP Workshop*, pp. 319–327, 2019, https://doi.org/10.48550/arXiv.1902.07669.
[45] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., " The Llama 3 Herd of Models," *arXiv preprint arXiv:2407.21783*, 2024, https://doi.org/10.48550/arXiv.2407.21783.
[46] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5776–5788, 2020, https://doi.org/10.48550/arXiv.2002.10957.



Bao-Gia Nguyen is a final-year Data Science student at the Department of Mathematics and Statistics, Quy Nhon University, Vietnam. His current field placement is with the Department of Data Science, TMA Tech Group, Gia Lai, Vietnam. He is interested in Medical Question-Answering System, Natural Language Processing, Machine Learning, and Artificial Intelligence.



Quang-Hung Le received the Ph.D. degree in Computer Science from the University of Engineering and Technology, Vietnam National University, Hanoi, in 2016. He is currently a lecturer with the Department of Information Technology, Quy Nhon University, Vietnam. His research interests include Natural Language Processing, Machine Learning, Artificial Intelligence, and Recommender Systems.