Efficient Multi-Agent Collaboration for Medical Question Answering with Quantized LLMs under Resource Constraints

Cuong Do[2] *, Phu Quy Quach[1], Dang Quach[1],
Tin Huynh Ngoc [1], Khuong Vu[1], Lan Vu[3],
Tuan Do[1] , Nam-Tran Nguyen[1], Thanh Ha [1],
Chi Pham [4], Duc-Nam Hoang [1],
Quang-Anh Nhu [1] , An-Vu Hoang [1] , Thanh Duong [1] , Huy Do [1],
Duc-Viet Hoang [1], Hana Tran [1], Thieu Anh Tran [1], Quynh Y Vo [1],
Tuan Le [1], Nhan-Khanh Tran [1], Dung Le [1], Kien Le [1]
[1]George Washington Institute (ISODS), Lexington, MA, USA
[2]University of Connecticut, Storrs, CT, USA
[3]Hanoi University of Public Health, Hanoi, Vietnam
[4]University of British Columbia, Canada

November 17, 2025

Abstract

Addressing the critical need for accurate AI-driven medical information in languages like Vietnamese requires architectures that balance performance with computational feasibility. This paper introduces DoctorAI, a novel multi-agent framework designed to provide reliable answers to Vietnamese medical questions by orchestrating specialized agents. The main contribution is our comprehensive performance analysis of several leading Vietnamese-supporting Large Language Models. Our evaluation on single-turn (Med-Single-QA) benchmark reveals a crucial trade-off: The compact VyLinh-3B model excels in efficiency, demonstrating a 3.2x reduction in latency and superior context retrieval precision. Conversely, the larger Qwen2.5-7B-Instruct model achieves higher generative quality, with superior factual correctness in complex disease detection. These findings validate the efficacy of our multi-agent approach in real-world healthcare in resource-constrained environments.

Keywords: RAG ,LLM ,Quantized ,Medical

1 Introduction

The advent of Large Language Models (LLMs) has introduced transformative potential for medical question answering, promising to democratize access to healthcare information. However, using these models in real-world clinical settings faces a trifecta of significant challenges. Firstly, the inherent risk of factual hallucination in LLMs necessitates architectures such as Retrieval-Augmented Generation

*Professor Cuong Do, Professor Tin Huynh, Professor Khuong Vu, Professor Lan Vu, Quy Quach and Dang Quach contributed equally to this work and share first authorship.

(RAG) to ground responses in verifiable medical knowledge[1]. Secondly, the substantial computational demands of state-of-the-art LLMs conflict with the typical resource-constrained environments in healthcare facilities, making quantized models an essential alternative for practical implementation [2]. Finally, the complexity of clinical diagnostics, which often involves multi-step reasoning, dynamic information gathering, and nuanced patient interaction, exceeds the capabilities of single agent systems. This limitation motivates the shift towards multi-agent collaboration[3].

These challenges are further compounded in non-English contexts, such as Vietnamese, which suffer from a scarcity of specialized medical LLMs and curated corpora. This creates the needs for reliable, efficient, and linguistically-aware AI healthcare assistants. To address this multifaceted problem, we introduce DoctorAI, a novel framework for medical question answering that uniquely integrates efficient multi-agent collaboration with quantized LLMs tailored for the Vietnamese language. Our system orchestrates multiple specialized agents, including a ReAct-based Diagnostic Agent for interactive reasoning [4] and a Doctor Agent for evidence-based synthesis, all within a resource-aware architecture. By deliberately decoupling the reasoning and retrieval phases and leveraging the efficiency of the GGUF model format, DoctorAI is designed for reliable deployment under significant computational constraints. This paper presents the architecture, implementation, and rigorous evaluation of DoctorAI. The key contributions of this work are threefold: (1)A novel multi-agent collaborative framework that balances sophisticated reasoning with the practical limitations of quantized models in a high-stakes domain; (2)A comprehensive performance and efficiency analysis of leading quantized Vietnamese LLMs (Qwen2.5- 7B-Instruct and VyLinh-3B), revealing a critical trade-off between reasoning depth and computational feasibility; and (3)The development of a complete data processing pipeline for building a Vietnamese medical knowledge base, utilizing specialized tools to handle the unique challenges of local medical documents.

2 Data Description & Preprocess Data

2.1 Knowledge Base Construction

The foundation of our RAG system is a high-quality knowledge base, the construction of which is particularly challenging in the Vietnamese medical domain due to the scarcity and fragmentation of digitized, authoritative sources compared to English. To overcome this, we undertook a meticulous manual curation process to build a specialized corpus that is both comprehensive and reliable. Our primary goal was to gather documents reflecting modern medical practices that could serve as a trustworthy source for clinical question-answering. We established strict selection criteria, prioritizing recent publications to ensure our system's knowledge reflects the latest clinical guidelines and therapeutic advancements. This rigorous process yielded a final corpus of 60 medical books and documents, totaling nearly 12 thousand pages. The content spans seven key medical topics as detailed in Table 1.

Class
Diabetes
Hypertension
Heart failure
Respiratory diseases
COVID-19
Specialized surgical
Cancer treatments

Table 1: List of classes in the dataset

2.2 Pre-processing and Structuring Medical Documents

Effective pre-processing and structuring of medical documents are important steps in AI systems, particularly those utilizing RAG. The medical domain presents unique challenges due to the complexity and size of documents such as clinical notes, research papers, and treatment guidelines [5]. These documents are often long, detailed, and contain critical information that needs to be properly segmented to improve the relevance and accuracy of LLM's responses. The initial task is to divide these documents into smaller, more manageable sections, enabling the retrieval system to identify the most relevant chunks of information for effective synthesis.

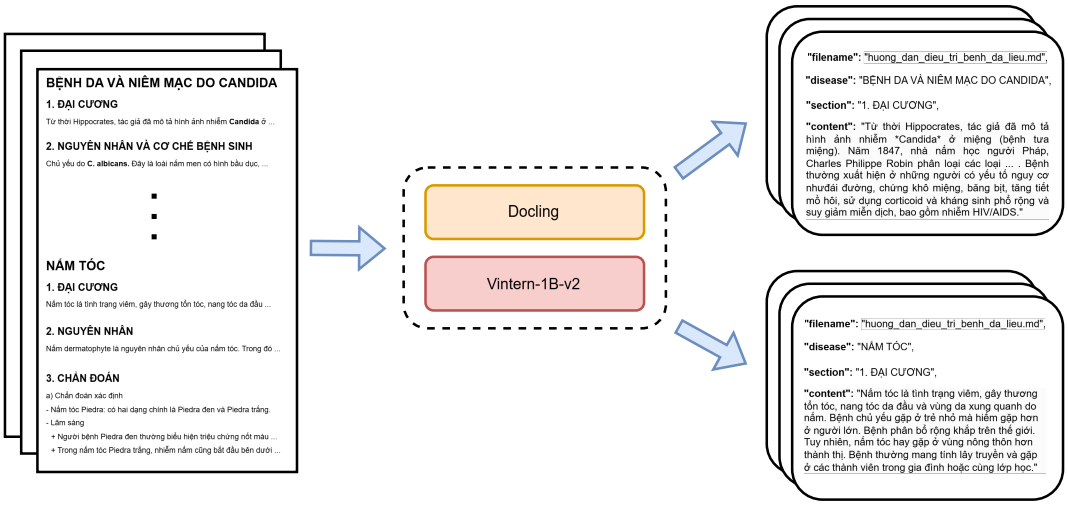

Figure 1: The workflow for processing medical documents. Medical PDF files (in Vietnamese) are extracted and converted into Markdown format. The system identifies and segments the content based on headings, creating distinct sections (for example "DAI CUONG", meaning "Overview"). These sections are structured into standardized JSON objects for easier retrieval and analysis.

To achieve segmentation and structural analysis of medical documents, structural elements such as layout, reading order, tables, and figures are extracted using Docling [6], For image-based documents, a vision-language model (VLM) as an OCR system is used to extract text from non-digital content. Markdown is chosen as the target format due to its simplicity, human-readability, and high compatibility with downstream AI systems. Markdown preserves the semantic structure of headings and sections while remaining lightweight and easily convertible into formats such as JSON or HTML, making it ideal for organizing documents in ways that improve information retrieval.

In our pipeline shown in Figure 1, medical PDF files are processed, in which layouts are analyzed, tables are detected and reconstructed, and figures are filtered, and metadata is extracted (e.g., title, authors, language). The content is then converted into structured Markdown, which can be easily manipulated, searched, and annotated. However, for image-heavy PDF files, such as scanned handwritten notes or low-quality printouts, we found that Docling's built-in OCR system, powered by a VLM called SmolDocling [7], often struggled with Vietnamese content. The OCR system frequently produced inaccurate text and failed to maintain structural consistency. To address this, we integrated Vintern-1B-v2, a Vietnamese VLM, as an alternative solution. Vintern-1B-v2 excels at extracting both textual and visual features from such challenging Vietnamese documents. It performs OCR, layout recognition, and visual understanding, producing well-structured JSON representations that more accurately preserve the document meaning than Docling's default OCR. Additionally, Vintern-1B-v2 is efficient enough to run on accessible hardware (e.g. T4 GPUs), making it a practical solution for resource-constrained environments.

After documents are divided into sections, any section that exceeds the context length limitation of the embedding model is further partitioned into smaller chunks. The optimal chunk size is determined based on the maximum input token capacity of the dedicated embedding model, ensuring efficient vector representation without information loss. These organized chunks are then saved in a

Figure 2: Performance of Vietnamese Embedding Models on STS benchmarks (Spearman Correlation on STSb-vn).

database. ChromaDB [8] was chosen for its advantages in AI-driven applications, particularly when compared to alternatives such as FAISS [9]. While FAISS primarily focuses on search functionality, ChromaDB offers a comprehensive solution with native metadata storage, support for dynamic updates, and multi-modal compatibility, which are all critical features for our implementation in the medical domain. In addition, ChromaDB significantly reduced implementation complexity through its straightforward APIs connection, while its seamless integration with frameworks such as LangChain [10] and LangGraph [11] minimized boilerplate code. Its flexible metadata attachment functionality also enables contextual filtering, enhancing retrieval precision for domain-specific queries.

3 Method

3.1 Embedding Models

Effective information retrieval in the medical domain requires moving beyond traditional keyword-based methods such as TF-IDF [12], which struggle with the semantic complexity and terminological variance of medical language, or non-contextual word embeddings such as Word2Vec [13] To address this, we use a Sentence Transformer (SBERT) model [14] [15]. which fine-tunes BERT-like models using a Siamese network structure, for ingestion and retrieval, allowing for the efficient generation of fixed-size sentence embeddings. Additionally both questions and medical documents are encoded into dense vector representations in a shared semantic space. This approach enables the system to capture deeper contextual meanings and enhances the relevance of retrieved answers.

3.2 Quantized Large Language Models (LLMs)

The generative and reasoning capabilities of our DoctorAI system are powered by modern Large Language Models (LLMs), which have become the cornerstone of recent natural language processing, with the introduction of the transformer model [16] with self-attention mechanism and the powerful decoder-only architectures, such as those in the GPT series [17].

As shown in Figure 3, quantization refers to the process of reducing the numerical precision of model parameters, such as converting 32-bit floating-point weights to lower-bit representations like 8-bit or 4-bit integers [2] [18] [19]. This reduction dramatically decreases the model's memory footprint and computational demands, enabling faster inference and lower energy consumption without significantly compromising accuracy [21-23] [20] [21] [22].

Open-source LLMs provide transparency and control over model internals and data handling, which are essential for compliance with strict patient privacy regulations such as Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR). Unlike proprietary models, open-source alternatives allow healthcare institutions to audit, customize, and securely deploy models on-premises or in controlled cloud environments, minimizing risks associated with data leakage or unauthorized access. This combination of quantization and open-source availability thus facilitates the practical adoption of advanced AI tools in sensitive medical contexts.

We use quantized models with the GGUF [23] format, a compact and self-contained binary for-

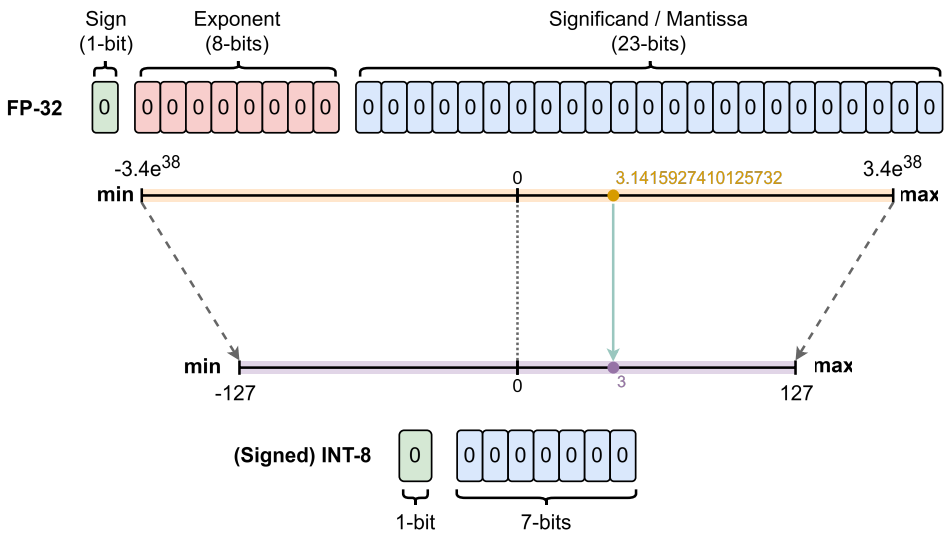

Figure 3: Basic quantization flow for compressing LLMs from 32-bit Floating Point (FP32) to 8-bit integer (INT8), demonstrating reduction in bit precision to decrease memory and computational load.

mat [24] for LLMs to support efficient inference across a diverse range of hardware platforms, with encapsulating both model weights and metadata within a single file.

Recent empirical research supports the effectiveness of low-bit quantized models in maintaining high accuracy. For example, Kurtic et al. [25] demonstrated that 8-bit or even 4-bit quantized versions of LLaMA 3.1 [26] models achieve performance remarkably close to their full-precision FP16 counterparts across multiple benchmarks, with minimal accuracy loss. This negligible trade-off is outweighed by the substantial improvements in inference speed, reduced memory usage, and lower power consumption-factors critical for deploying AI in resource-limited medical environments.

To ensure that our selected models are capable of processing Vietnamese medical content, we focus our evaluation on the VMLU Benchmark (Vietnamese Multitask Language Understanding) [27], a comprehensive framework designed to assess the proficiency of large language models in Vietnamese.

VMLU consists of 10,880 multiple-choice questions across 58 topics, organized into four domains: STEM, Social Sciences, Humanities, and interdisciplinary topics. It includes multiple difficulty levels, ranging from primary school to postgraduate. The benchmark supports both zero-shot and few-shot prompting, with performance evaluated through accuracy-based metrics on a public leaderboard. This makes VMLU a practical and reliable tool for assessing localized language understanding in real-world contexts. Furthermore, given the hardware constraints of our deployment environment, most evaluated models were quantized to 4-bit precision, with an exception being Arcee-VyLnh-3B [28], which was quantized to 8-bit. Despite the higher bit-width, Despite the higher bit-width, it ranks second in VMLU performance, closely following the larger Qwen2.5-7B-Instruct model, and remains efficient on edge devices and low-resource CPUs.

For the system's primary language processing component, we selected fine-tuned Vietnamese models, including Vistral-7B-Chat [29], Arcee-VyLinh-3B [28], and PhoGPT-4B-Chat [30]. We also tested Qwen2.5-7B-Instruct [31] and Qwen2.5-3B-Instruct [31] to assess the performance gap between Arcee-VyLinh-3B (a variant of Qwen2.5-3B-Instruct), and the larger Qwen2.5-7B-Instruct model. As shown in Figure 4, Arcee-VyLinh-3B was optimized specifically for Vietnamese, offering strong reasoning capabilities while remaining deployable within our computational constraints. The training process incorporates challenging questions and iterative Direct Preference Optimization (DPO)[32], enabling the model to perform exceptionally well despite its compact 3B parameter size. The inclusion of Vistral-7B-Chat and PhoGPT-4B-Chat models provided additional insights into the performance differences across different model sizes. Arcee-VyLinh-3B's proficiency with Vietnamese medical terminology and conceptual relationships suggests excellent transfer learning potential for our specialized domain. Additionally, its large token context window (maximum context window up to 32K tokens) supports

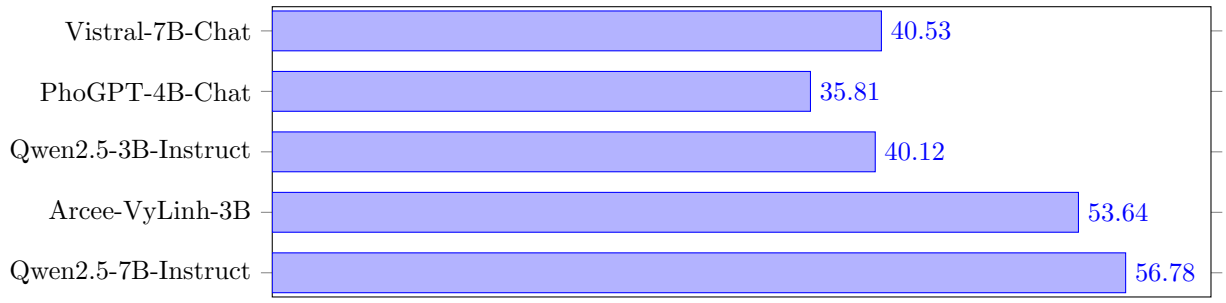

Figure 4: The performance of quantized LLMs on the VMLU benchmark. All models were quantized to 4-bit and 8-bit where applicable. Each result is averaged over three runs to mitigate output fluctuations caused by hallucinations.

comprehensive analysis of patient histories and medical documentation, making it highly suitable for real-world healthcare applications.

3.3 Multi-agent Architecture for Agent Collaboration

3.3.1 Multi-agent Collaboration Concepts

Recently, the integration of multi-agent systems with LLMs has garnered significant attention within the artificial intelligence research community. Talebirad et al. [3] introduced a collaborative architecture in which multiple LLM agents, each with specialized roles, work together to solve complex tasks that exceed the capabilities of a single agent. This design mirrors human team-based problem solving, with agents assigned roles such as planning, execution, verification, and response generation. By incorporating API calls and inter-agent communication, such systems extend the utility of LLMs beyond simple question answering to more sophisticated tasks requiring deep reasoning and multi-step coordination. This approach represents a substantial advancement toward building autonomous, collaborative AI systems with potential applications in artificial general intelligence (AGI) and process automation. In the context of medical question answering, a multi-agent LLM system is suitable for managing diverse and intricate user queries.

3.3.2 System Flow and Integrated Agents

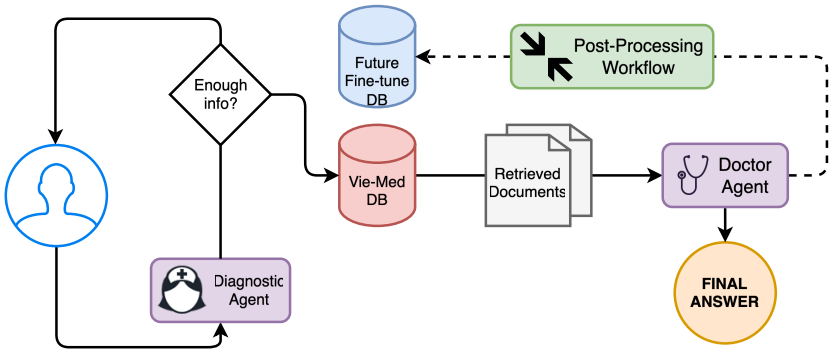

Figure 5: Overall architecture of multi-agent in Doctor AI.

The Doctor AI system shown in Figure 5 uses a multi-agent architecture designed to provide medical assistance. Patient interaction initiates the process through a query submission interface. Upon receiving a query, the Diagnostic Agent serves as the primary interface between the patient and the system. This agent implements the Reasoning and Acting (ReAct) prompting [4], as illustrated in Figure 6, which enables structured reasoning interleaved with action steps. Prior to ReAct, Chain of

Thought (CoT) prompting [33] was widely used to enhance reasoning in language models by encouraging them to decompose complex problems into sequential thought steps. While effective in many domains, CoT is limited to internal reasoning, which only allows models to articulate their logic, but not to interact with the environment such as gather new information or verify assumptions.

In contrast, ReAct builds on CoT by combining thought steps with executable actions, enabling the Diagnostic Agent to ask clarification questions, retrieve relevant medical documents, and refine its reasoning dynamically based on newly acquired information [4]. This interactive loop of thinking and acting supports more adaptive and context-aware diagnostic reasoning. It also closely mimics the iterative nature of real-world medical interviews, where clinicians probe and adapt based on patient responses. Empirical evaluations show that ReAct prompting significantly improves diagnostic accuracy compared to non-interactive reasoning approaches [34, 35].

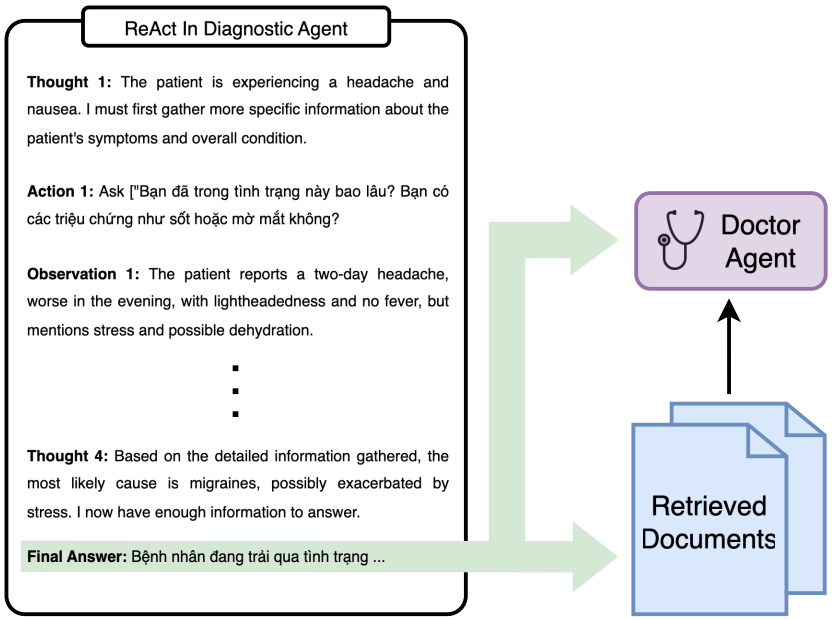

Figure 6: The reasoning and acting process using ReAct prompting continues until a final diagnosis of the patient's symptoms is reached. Once the diagnosis is determined, it will be passed to the RAG system to retrieve the most relevant documents. Finally, both the final diagnosis and the retrieved documents are then fed into the Doctor Agent to generate the best possible response.

In the Diagnostic Agent, document retrieval is not used during the ReAct reasoning phase as shown in [36] that integrating retrieved documents during reasoning can mislead models down incorrect paths, particularly when the documents are irrelevant or subtly misleading. These findings highlight the dangers of entangling reasoning with retrieval, especially in domains where the cost of error is substantial. Specifically the Diagnostic Agent first completes its ReAct reasoning process based solely on patient interaction and its internal knowledge to reach a preliminary diagnosis. Only after this reasoning phase is complete does the system activate the RAG mechanism to retrieve relevant documents that specifically pertain to the identified condition. This sequential, rather than interleaved, approach minimizes the likelihood of retrieval-induced reasoning errors while still leveraging the Vietnamese medical knowledge base for diagnostic verification and treatment planning. This design choice represents a key innovation in our system, prioritizing diagnostic reliability by deferring external information until after the core reasoning has concluded.

Once the relevant documents have been retrieved, the Doctor Agent assumes responsibility for analyzing the curated information and generating refined diagnosis along with recommended treatment protocols. This specialized agent leverages LLM capabilities enhanced through carefully engineered prompting strategies [33, 37]. The Doctor Agent takes in three key inputs: (1) the patient's symptom descriptions captured during the ReAct phase, (2) the retrieved medical documents from the Vietnamese medical knowledge base, and (3) a set of structured reasoning guidelines that enforce medical reasoning principles. This methodology ensures that the generated diagnosis remains grounded in

authoritative Vietnamese medical literature while maintaining relevance to the specific patient presentation.

Once the relevant documents have been retrieved, the Doctor Agent assumes responsibility for analyzing the curated information and generating refined diagnosis along with recommended treatment protocols. This specialized agent leverages LLM capabilities enhanced through carefully engineered prompting strategies [33] [37]. The Doctor Agent takes in three key inputs: (1) the patient's symptom descriptions captured during the ReAct phase, (2) the retrieved medical documents from the Vietnamese medical knowledge base, and (3) a set of structured reasoning guidelines that enforce medical reasoning principles. This methodology ensures that the generated diagnosis remains grounded in authoritative Vietnamese medical literature while maintaining relevance to the specific patient presentation.

To ensure consistent and clinically appropriate responses, the Doctor Agent employs a series of instruction-based guardrails within its prompt template [38, 39]. These include explicit directives to maintain professional medical tone, prioritize evidence-based reasoning, acknowledge diagnostic uncertainty when appropriate, and adhere to Vietnamese medical terminology and protocols. The prompt structure also incorporates format specifications that encourage the systematic presentation of differential diagnoses with supporting evidence, followed by treatment recommendations in accordance with Vietnamese clinical guidelines. The effectiveness of this prompt-based approach is enhanced through regular refinement based on expert feedback from Vietnamese healthcare professionals [40, 41]. By implementing these specialized prompting techniques rather than custom algorithmic implementations, the system achieves flexibility and maintainability while producing diagnostically sound outputs that align with Vietnamese medical standards and practices.

Following the diagnostic decision and treatment recommendation, the Post-processing Workflow processes the complete interaction sequence. It applies privacy-preserving techniques to remove personal information while maintaining clinical relevance [42]. This is achieved through a series of steps: Using a conversation summarization tool to extract key details such as symptoms, diagnoses, and treatment advice, applying a disease-tagging tool, and filtering out personally identifiable information (PII), as illustrated in Figure 7. This workflow is triggered only after a period of user inactivity, ensuring that the conversation is complete before being finalized for archival and learning purposes, the agent then formats the information into a structured JSON output. This structured summary is stored in a dedicated database, facilitating both future diagnostic retrieval and the fine-tuning of the core language model. By organizing data in this format, the system ensures efficient knowledge management while supporting scalable improvements in diagnostic performance [43, 44, 45].

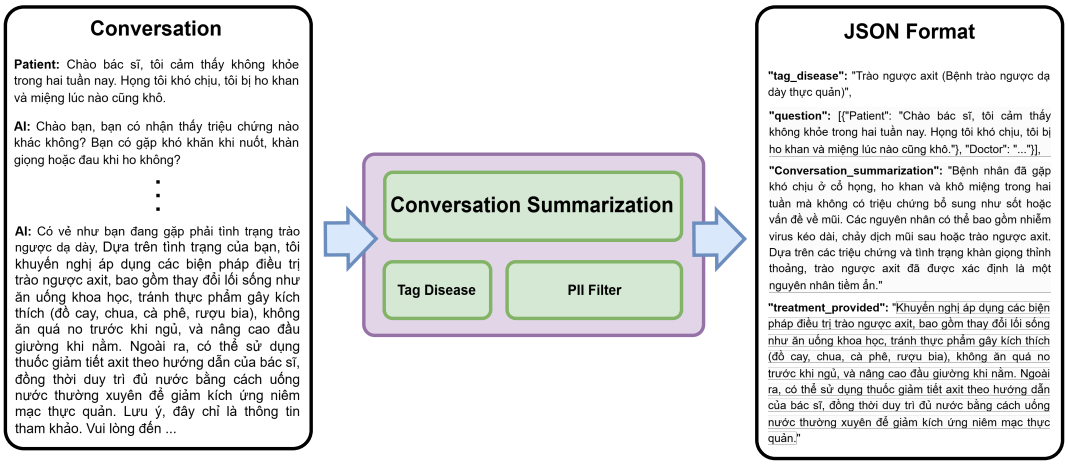

Figure 7: Post-Processing Workflow - Converting Doctor-Patient Conversations into Structured JSON Format. After a period of user inactivity, the Post-Processing Workflow summarizes the conversation, filters PII, tags the relevant disease, and transforms the interaction into a standardized JSON structure for secure storage and future model fine-tuning.

3.3.3 System Evaluation in the Medical Field

To evaluate our system, we constructed a specialized evaluation dataset: Med-Single-QA. For the initial data generation, we leveraged Gemini-2.0-Flash to create a baseline set of questions and answers. The Med-Single-QA dataset consists of 300 medical questions with their corresponding answers, designed to evaluate the system's ability to provide accurate responses to isolated queries.

For our evaluation methodology, we adopted the RAGAS [46] framework, which provides standardized and reproducible metrics. We structured our evaluation into three main categories:

- **Retrieval Evaluation:** We used Context Recall to measure the extent to which the retrieved context aligns with the ground-truth answer, and Context Precision to measure the ratio of relevant to irrelevant items, also known as the signal-to-noise ratio, in retrieved contexts against reference answers, which helped optimize the effectiveness of our top-K retrieval parameters.
- **Generation Evaluation:** Faithfulness is used to measure how factually consistent a generated answer is with the retrieved context. Factual Correctness is used to quantify alignment between our generated answers and established medical reference material, using claim-level analysis to detect even subtle factual discrepancies. Finally, Answer Relevancy is used to assess whether responses directly address the original queries, providing a surface-level quality check for our prompt engineering and instruction-following capabilities.
- **Computational Efficiency:** In addition to quality evaluation, computational efficiency is evaluated using two key metrics: Giga Floating Point Operations (GLOPs) to quantify the total computation cost per response, and end-to-end latency to measure the time required to process a query from input to final response. These metrics helped us better understand the trade-off between performance and responsiveness, particularly important for real-time applications in medical settings.

This comprehensive approach allowed us to identify specific areas for improvement within both our retrieval and generation components while maintaining a reproducible evaluation methodology for our medical AI assistant.

Metrics		**Med-Single-QA**	
		Qwen2.5-7B-Instruct	**VyLinh-3B**
Retrieval-focused metrics	Context Precision	0.883	0.933
	Context Recall	0.850	0.833
Generation-focused metrics	Factual Correctness	0.797	0.734
	Answer Relevancy	0.741	0.724
	Faithfulness	0.794	0.695
Computational Efficiency	GLOPs	553.38	249.84
	Latency	44.00s	13.69s

Table 2: Performance Metrics for Qwen2.5-7B-Instruct and VyLinh-3B on Med-Single-QA

From Table 2, we observe that our system's performance in the Med-Single-QA evaluation highlights both strengths and areas for improvement. The retrieval component performs strongly, with VyLinh-3B showing higher context precision than Qwen2.5-7B-Instruct, while both models maintain solid context recall. This suggests the system effectively retrieves relevant documents while covering most

of the necessary information. However, generation-focused metrics are more mixed. Qwen2.5-7B-Instruct performs better than VyLinh-3B in factual correctness, answer relevancy, and faithfulness, although there is still room for improvement across both models. It can be seen that VyLinh-3B shows lower faithfulness, indicating a higher likelihood of generating content not grounded in the retrieved context. On the other hand, VyLinh-3B is significantly more efficient, requiring fewer computational resources and responding faster, making it a strong candidate for real-time applications.

Nevertheless, we have observed that VyLinh-3B often struggles with long contextual inputs, occasionally generating hallucinated content or mixing Vietnamese with Chinese or English, which introduces ambiguity and reduces accuracy in both diagnosis and recommendation tasks. These limitations underscore the importance of model capacity, language consistency, and contextual reasoning in high-stakes domains such as healthcare.

4 Discussion

Beyond the disparities identified in retrieval versus generation performance, the comparative analysis of Med-Single-QA highlights the profound impact of context complexity on system behavior and output quality. Even Qwen2.5-7B-Instruct, despite its larger size, consistently outperforms VyLinh-3B. However, this improvement comes at the cost of significantly higher computational demands and longer inference times. Its strong performance in factual correctness and recommendation detection underscores the advantages of larger models when dealing with medical ambiguity and generating patient-specific recommendations. Yet, the increased GLOPs and latency of Qwen2.5-7B-Instruct highlight the trade-off between accuracy and efficiency, which is particularly critical in real-time or edge deployments where computational resources are constrained.

On the other hand, VyLinh-3B, while more efficient in terms of resource usage, faces notable challenges in maintaining faithfulness and language consistency during extended conversations. This underscores the need for focused fine-tuning strategies, aiming to not only improve the model's grounding in retrieved context but also to strengthen its cross-lingual capabilities and mitigate hybrid-language artifacts. Such improvements are essential to ensure clinical reliability and enhance performance in multilingual environments.

Across the QA settings, a recurring pattern emerges: the tendency of models to prioritize answer fluency and topical alignment rather than focusing on evidential grounding. The faithfulness metric, particularly low in VyLinh-3B, highlights an implicit bias toward generating overconfident and unsupported content. This behavior reinforces previous concerns about hallucinations and is especially problematic in clinical contexts, where the spread of misinformation could result in adverse outcomes.

In summary, our findings indicate a central dilemma in applied medical AI: The trade-off between the diagnostic accuracy of larger models and the deployment feasibility of smaller, more efficient ones. This study demonstrates that while techniques like RAG and multi-agent architectures provide a robust framework, fundamental challenges in faithfulness and reasoning consistency under resource constraints persist. These specific limitations, identified through our empirical analysis, are not insurmountable barriers but rather define the critical path forward. They form the primary motivation for the future research directions proposed in the next section, which aim to enhance model reliability and bridge the gap between computational efficiency and clinical safety.

5 Conclusion

In this paper, we addressed the multifaceted challenge of creating a reliable and efficient AI medical assistant for the Vietnamese language, operating under significant resource constraints. We introduced DoctorAI, a novel framework that integrates a multi-agent collaborative architecture with quantized Large Language Models and a Retrieval-Augmented Generation pipeline. Our work involved constructing a specialized Vietnamese medical knowledge base and conducting a comprehensive evaluation of the system's performance.

Our central finding reveals a critical trade-off between computational efficiency and diagnostic accuracy, as demonstrated by the comparative analysis of the compact VyLinh-3B and the larger Qwen2.5-7B-Instruct models. While the smaller model offers a practical solution for real-time applications, the larger model provides superior reasoning and faithfulness. This study not only presents a viable architecture for domain-specific, resource-aware AI assistants but also provides a crucial roadmap and empirical insights for researchers and practitioners aiming to deploy safe and effective LLMs in real-world healthcare settings.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," 2021.

[2] T. Dettmers and L. Zettlemoyer, "The case for 4-bit precision: k-bit inference scaling laws," 2023.

[3] Y. Talebirad and A. Nadiri, "Multi-agent collaboration: Harnessing the power of intelligent llm agents," *arXiv preprint arXiv:2306.03314*, 2023.

[4] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," 2023.

[5] R. Yang, Y. Ning, E. Keppo, M. Liu, C. Hong, D. S. Bitterman, J. C. L. Ong, D. S. W. Ting, and N. Liu, "Retrieval-augmented generation for generative artificial intelligence in health care," *npj Health Systems*, vol. 2, no. 1, p. 2, 2025.

[6] C. Auer, M. Lysak, A. Nassar, M. Dolfi, N. Livathinos, P. Vagenas, C. B. Ramis, M. Omenetti, F. Lindlbauer, K. Dinkla, L. Mishra, Y. Kim, S. Gupta, R. T. de Lima, V. Weber, L. Morin, I. Meijer, V. Kuropiatnyk, and P. W. J. Staar, "Docling technical report," 2024.

[7] A. Nassar, A. Marafioti, M. Omenetti, M. Lysak, N. Livathinos, C. Auer, L. Morin, R. T. de Lima, Y. Kim, A. S. Gurbuz, M. Dolfi, M. Farré, and P. W. J. Staar, "Smoldocling: An ultra-compact vision-language model for end-to-end multi-modal document conversion," 2025.

[8] C. Contributors, "Chroma: Open-source embedding database." https://github.com/chroma-core/chroma, 2023. Accessed: 2025-05-17.

[9] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, "The faiss library," 2025.

[10] L. Contributors, "Langchain github repository." https://github.com/langchain-ai/langchain, 2023. Accessed: 2025-05-17.

[11] L. Contributors, "Langgraph github repository." https://github.com/langchain-ai/langgraph, 2024. Accessed: 2025-05-17.

[12] K. Spärck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[13] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," 2013. https://arxiv.org/abs/1310.4546.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018. https://arxiv.org/abs/1810.04805.

[15] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019. https://arxiv.org/abs/1908.10084.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[17] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[18] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, "Gptq: Accurate post-training quantization for generative pre-trained transformers," 2023.

[19] J. Lin, J. Tang, H. Tang, S. Yang, W.-M. Chen, W.-C. Wang, G. Xiao, X. Dang, C. Gan, and S. Han, "Awq: Activation-aware weight quantization for llm compression and acceleration," 2024.

[20] J. Pan, C. Wang, K. Zheng, Y. Li, Z. Wang, and B. Feng, "Smoothquant+: Accurate and efficient 4-bit post-training weightquantization for llm," 2023.

[21] S. Ashkboos, A. Mohtashami, M. L. Croci, B. Li, P. Cameron, M. Jaggi, D. Alistarh, T. Hoefler, and J. Hensman, "Quarot: Outlier-free 4-bit inference in rotated llms," 2024.

[22] D. Lee, S. Choi, and I. J. Chang, "Qrazor: Reliable and effortless 4-bit llm quantization by significant data razoring," 2025.

[23] G. Gerganov and Contributors, "Gguf: Gpt-generated unified format for llms." https://github.com/ggml-org/ggml/blob/master/docs/gguf.md, 2023. Accessed: 2025-05-17.

[24] G. Gerganov and Contributors, "llama.cpp: Inference of llama models in c/c++." https://github.com/ggml-org/llama.cpp, 2023. Accessed: 2025-05-17.

[25] E. Kurtic, A. Marques, S. Pandit, M. Kurtz, and D. Alistarh, "Give me bf16 or give me death? accuracy-performance trade-offs in llm quantization," 2025.

[26] Meta AI, "The llama 3 herd of models." https://ai.meta.com/research/publications/the-llama-3-herd-of-models/, 2024.

[27] ZaloAI-Jaist, "A vietnamese multitask language understanding benchmark suite for large language models." https://vmlu.ai/report, 2025. Available at https://vmlu.ai/.

[28] Arcee.ai, "Vylinh-3b: Vietnamese large language model." https://huggingface.co/arcee-ai/Arcee-VyLinh, 2024. Accessed: 2025-05-17.

[29] N. Research, "Vistral-7b: A vision-aligned language model based on mistral." https://huggingface.co/Viet-Mistral/Vistral-7B-Chat, 2024. Accessed: 2025-05-17.

[30] D. Q. Nguyen, L. T. Nguyen, C. Tran, D. N. Nguyen, D. Phung, and H. Bui, "Phogpt: Generative pre-training for vietnamese," 2024.

[31] Qwen Team, "Qwen2.5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.

[32] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," 2024.

[33] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023.

[34] X. Shi, Z. Liu, L. Du, Y. Wang, H. Wang, Y. Guo, T. Ruan, J. Xu, and S. Zhang, "Medical dialogue: A survey of categories, methods, evaluation and challenges," 2024.

[35] G. Mialon, R. Dessì, M. Lomeli, C. Nalmpantis, R. Pasunuru, R. Raileanu, B. Rozière, T. Schick, J. Dwivedi-Yu, A. Celikyilmaz, E. Grave, Y. LeCun, and T. Scialom, "Augmented language models: a survey," 2023.

[36] Y. Zhang, T. Wang, S. Chen, K. Wang, X. Zeng, H. Lin, X. Han, L. Sun, and C. Lu, "Arise: Towards knowledge-augmented reasoning via risk-adaptive search," 2025.

[37] D. Guo and D. Terzopoulos, "Prompting medical large vision-language models to diagnose pathologies by visual question answering," *Machine Learning for Biomedical Imaging*, vol. 3, p. 59–71, Mar. 2025.

[38] X. Zhang, C. Tian, X. Yang, L. Chen, Z. Li, and L. R. Petzold, "Alpacare:instruction-tuned large language models for medical application," 2025.

[39] J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, and L. Hou, "Instruction-following evaluation for large language models," 2023.

[40] D. Oniani, X. Wu, S. Visweswaran, S. Kapoor, S. Kooragayalu, K. Polanska, and Y. Wang, "Enhancing large language models for clinical decision support by incorporating clinical practice guidelines," 2024.

[41] J. Wang, E. Shi, S. Yu, Z. Wu, C. Ma, H. Dai, Q. Yang, Y. Kang, J. Wu, H. Hu, C. Yue, H. Zhang, Y. Liu, Y. Pan, Z. Liu, L. Sun, X. Li, B. Ge, X. Jiang, D. Zhu, Y. Yuan, D. Shen, T. Liu, and S. Zhang, "Prompt engineering for healthcare: Methodologies and applications," 2024.

[42] I. C. Wiest, D. Ferber, J. Zhu, M. van Treeck, S. K. Meyer, R. Juglan, Z. I. Carrero, D. Paech, J. Kleesiek, M. P. Ebert, D. Truhn, and J. N. Kather, "Privacy-preserving large language models for structured medical information retrieval," *npj Digital Medicine*, vol. 7, no. 1, p. 257, 2024.

[43] V. Schlegel, H. Li, Y. Wu, A. Subramanian, T.-T. Nguyen, A. R. Kashyap, D. Beck, X. Zeng, R. T. Batista-Navarro, S. Winkler, and G. Nenadic, "Pulsar at mediqa-sum 2023: Large language models augmented by synthetic dialogue convert patient dialogues to medical records," 2023.

[44] T. Das, D. Albassam, and J. Sun, "Synthetic patient-physician dialogue generation from clinical notes using llm," 2024.

[45] V. Nair, E. Schumacher, and A. Kannan, "Generating medically-accurate summaries of patient-provider dialogue: A multi-stage approach using large language models," 2023.

[46] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "Ragas: Automated evaluation of retrieval augmented generation," 2023.