International Journal of Scientific Engineering and Science
Volume 10, Issue 4, pp. 70-76, 2026. ISSN (Online): 2456-7361

Application of Retrieval-Augmented Generation (RAG) in University Admissions Question-Answering Systems: A Case Study at Thuongmai University

Do Thi Thanh Tam[1*], Nguyen Hung Long[1], Vu Thi Le[2]
[1]Department of Informatics, Faculty of Economic Information System and E-Commerce, Thuongmai University, Vietnam
[2]Thai Binh University of medicine and pharmacy, Vietnam
Email address: tam.dtt@tmu.edu.vn[1*], ntthlong@tmu.edu.vn[1], levt@tbump.edu.vn[2]

Abstract— *Question answering systems based on large language models (LLMs) have demonstrated strong capabilities in natural language understanding and generation. However, when applied to domain-specific contexts, these systems still face significant challenges, particularly due to hallucination and the lack of access to up-to-date, contextually relevant knowledge. In the context of university admissions in Vietnam, existing support systems are primarily based on frequently asked questions (FAQs) and rule-based approaches, which limit their ability to handle diverse user queries and adapt to annually changing admission policies. To address these limitations, this study proposes a Retrieval-Augmented Generation (RAG)-based question answering system for university admission consulting, with a case study at Thuongmai University. The proposed system integrates LLMs with an external knowledge retrieval mechanism, leveraging admission-related data collected from 2021 to 2025. Experimental results demonstrate that the combination of LLMs and RAG significantly outperforms baseline LLM-only approaches, with improvements across all evaluation metrics ranging from 0.02% to 98.7%. The study contributes a practical RAG-based framework for domain-specific applications in higher education and highlights the effectiveness of combining knowledge retrieval with generative models to enhance response reliability and timeliness. Specifically, the system achieves a Context Precision of 0.6750, a Context Recall of 0.6177, and a Faithfulness score of 0.7059.*

Keywords— *Retrieval-Augmented Generation, admission question-answering, Large language model.*

I. INTRODUCTION

In recent years, Artificial Intelligence (AI) has driven significant transformations across various domains, particularly in higher education, where the demand for automating student support services is rapidly increasing. In an highly competitive admissions landscape, admission consulting not only serves as an information provision channel but also plays a critical role in enhancing institutions' ability to attract and engage prospective students. During each admission cycle, universities are required to handle a substantial volume of inquiries related to academic programs, admission methods, tuition fees, scholarships, subject combinations, and entry requirements. However, traditional consulting approaches, such as manual support by admission staff or call centers services, often face limitations in terms of service availability, response consistency, and operational costs.

In Vietnam, although the higher education system comprises more than 250 institutions, the adoption of AI-based solutions for admission question answering remains relatively limited, with only approximately 15 institutions (accounting for 6.1%) implementing such systems. Notably, most existing solutions rely on frequently asked questions (FAQs) and rule-based approaches, which constrain their ability to handle diverse queries and adapt to annually evolving admission policies [1-4]. This situation highlights a significant gap in leveraging AI to enhance the effectiveness of admission communication and improve the experience of prospective students.

The Question Answering (QA) problem is a core task in Natural Language Processing (NLP), aiming to develop systems capable of understanding natural language queries and providing accurate answers based on relevant knowledge sources. Over time, QA approaches have evolved through multiple stages, including rule-based systems, traditional information retrieval methods, supervised machine learning, deep learning, and, more recently, Transformer-based models. In recent years, the emergence of large language models (LLMs), such as GPT-based models, has demonstrated substantial potential in advancing QA systems, owing to their ability to generate coherent natural language, capture contextual semantics, and provide flexible responses across diverse knowledge domains [5]. However, despite their strong performance across a wide range of language tasks, LLMs still exhibit critical limitations when deployed in real-world educational settings. First, the phenomenon of hallucination may cause the models to generate inaccurate or non-existent information. Second, the internal knowledge of LLMs cannot be updated in real time to reflect the continuously evolving admission regulations, which vary across academic years, admission methods, and training programs. These limitations pose significant risks when such systems are employed as official advisory tools in university admissions contexts [6]. To address these limitations, Retrieval-Augmented Generation (RAG) has emerged as a promising approach due to its ability to integrate external knowledge retrieval with natural language

generation. In this architecture, the system first retrieves relevant documents from an external knowledge base, and then employs a LLM to generate responses conditioned on the retrieved context. This approach enhances answer accuracy, mitigates hallucination, and enables the incorporation of up-to-date, domain-specific knowledge without requiring costly model retraining [7].

Motivated by these challenges, this study proposes a RAG-based chatbot system for university admission consulting, aimed at automatically answering admission-related queries in higher education settings. The system is developed using admission data collected from 2021 to 2025, with a case study at Thuongmai University. It integrates semantic retrieval with LLMs to generate contextually appropriate responses.

The main contributions of this study are summarized as follows:

• First, we propose a RAG-based chatbot architecture for admission consulting, designed to improve the accuracy of question answering systems in the higher education domain.

• Second, we construct an admission knowledge base from real-world data spanning 2021–2025, enabling semantic retrieval of relevant documents instead of relying on fixed keyword-based approaches.

• Third, we evaluate the effectiveness of the proposed system on a real-world test set of admission-related questions, demonstrating the feasibility of deploying intelligent systems to support university admission services with RAG.

II. Related works

Large language models (LLMs)

Since the release of the first generation of ChatGPT in 2022, the world has witnessed a rapid growth of LLMs in terms of their quantity, scale, and capabilities. These models range from those with a few billion parameters, such as LLaMA, Falcon, mT5, GPT-NeoX-20B, CodeGen, and Flan-T5, to those with hundreds of billions of parameters, including BLOOM, BLOOMZ, GPT-3, GLM, and Galactica. Such models have demonstrated remarkable performance across a wide range of NLP tasks, including text generation, summarization, machine translation, and particularly QA. Owing to their ability to capture deep contextual semantics and perform linguistic reasoning, LLMs can effectively handle factoid questions, open-ended queries, and complex information-seeking tasks that require synthesizing knowledge from multiple sources. Furthermore, these models are highly effective in supporting multi-turn dialogue systems, enabling them to maintain contextual coherence and provide consistent responses throughout extended interactions [8].

For Vietnamese - a language with relatively limited resources compared to English and Chinese - the ecosystem of LLMs is still in its early stages of development. Early research has primarily focused on medium-scale pretrained models, among which PhoBERT stands out as one of the first models to achieve strong performance across a wide range of Vietnamese natural language processing tasks [9]. In recent years, alongside the global advancement of LLMs, several larger-scale Vietnamese models have been developed following two main approaches: training from scratch and fine-tuning from existing foundation models. Notable open-source models include Vietcuna-7B-v3, Vistral, PhoGPT-7B5, PhoGPT-7B5-Instruct, VinaLLaMA, URA-LLaMA, and ViGPT [10]. In addition, several commercial models with Vietnamese language support, such as GPT-3.5 Turbo, GPT-4, and Gemini 1.0, have been widely applied in real-world tasks due to their strong capabilities in analysis, reasoning, and natural language generation in Vietnamese. These applications include question answering in Vietnamese legal domains, chatbots for economic and service-related information, and performance on tasks such as the Vietnamese National High School Examination [11-13].

Moreover, LLMs are increasingly being extended to multimodal domains, including image and audio processing, code generation, and the development of intelligent agent systems. However, in domain-specific question answering tasks, LLMs still face notable limitations, such as hallucination and the inability to access up-to-date, context-specific knowledge. These challenges have driven the development of hybrid approaches that combine generative models with external knowledge retrieval, among which RAG has emerged as a prominent solution [6, 7, 14].

TABLE 1. Comparison of representative large language models, including Vietnamese models and foundation models

Model Group	Model	Model Type	Access Type	Year	Parameters	Characteristics
Baseline	PhoBERT	Encoder (BERT-based)	Open-source	2020	135M / 370M	Strong Vietnamese model for NLP tasks
Vietnamese LLMs	VinaLLaMA	Decoder (LLaMA-based)	Open-weight	2023	7B	Optimized for Vietnamese based on LLaMA
	PhoGPT-7B	Decoder	Open-source	2023	7.5B	Pretrained for Vietnamese language tasks
	Vietcuna-7B	Instruction-tuned LLM	Open-weight	2023	7B	Fine-tuned for conversational tasks
Foundation Models	GPT-4	Proprietary LLM	Closed-source	2023	Not disclosed	Strong reasoning, QA
	Gemini 1.0 Pro	Multimodal LLM	Closed-source	2023	Not disclosed	Native multimodal

Source: Compiled by the authors ***[9, 10, 15, 16]***

Model *Selection* Rationale

In this study, two LLMs are selected to represent different approaches within the RAG framework, including a Vietnamese-specific model (VinaLLaMA) and a commercial model (GPT-4). According to experimental results reported by Quan Nguyen et al., VinaLLaMA-7B demonstrates superior performance compared to other Vietnamese language models, such as ViGPT™-170K, PhoGPT-7B5-Instruct, Vietcuna-7B-v3, URA-LLaMA-13B, SeaLLM-7B-Chat, and BKAI-LLaMA-2 [10]. VinaLLaMA-7B is a model that has been pre-trained and fine-tuned on Vietnamese data, enabling it to effectively capture linguistic characteristics and local contextual nuances. This capability is particularly important in admission consulting tasks, which involve domain-specific terminology, natural user expressions, and non-standard language variations. In addition, VinaLLaMA-7B offers advantages in on-premise deployment, ensuring data privacy and optimizing operational costs. In contrast, GPT-4 represents

a class of large-scale commercial models with strong reasoning capabilities, deep contextual understanding, and superior performance across a wide range of question answering tasks. When integrated with RAG, GPT-4 can effectively leverage retrieved information to generate accurate, coherent, and well-synthesized responses. However, this model depends on external infrastructure and incurs higher usage costs.

The combination and comparison of these two models within a unified RAG framework enable a comprehensive evaluation of system performance, while also clarifying the roles of language specificity and model scale in improving domain-specific question answering. The experimental results thus provide a foundation for selecting appropriate models in real-world university admission consulting systems.

Retrieval-Augmented Generation (RAG)

RAG introduced by Patrick Lewis et al. in 2020, fundamentally differs from traditional fine-tuning approaches. Instead of requiring model retraining when new data becomes available, RAG incorporates an external retrieval mechanism to enable flexible access to up-to-date knowledge. This approach allows the model to leverage newly acquired or domain-specific information without costly retraining. By combining the generative capabilities of LLMs with an external knowledge base, RAG enhances context-aware reasoning. As a result, it produces more accurate and well-grounded responses while significantly reducing hallucination. Consequently, RAG is particularly well-suited for applications that require dynamic and specialized knowledge [17].

The core components of a RAG system can be summarized as follows:

- *Vector database creation (indexing)*: The entire dataset is first converted into vector representations (embeddings) and stored in a vector database for efficient retrieval.
- *User input*: The user provides a natural language query to seek information or obtain a response.
- *Information retrieval*: The retrieval mechanism scans the vector database to identify knowledge chunks that are semantically similar to the user's query. These retrieved paragraphs are then passed to the LLM to enrich the contextual input for answer generation.
- *Context integration*: The retrieved paragraphs are combined with the original user query to form an augmented prompt.
- *Response generation*: The enriched prompt, containing both the query and retrieved context, is processed by the LLM to generate the final response aligned with the provided information.

Evaluation Metrics [18]

To evaluate the performance of the proposed RAG-based chatbot system, we employ a combination of lexical, semantic, and retrieval-based metrics, implemented using the RAGAS framework.

(1) ROUGE Score

ROUGE measures the lexical overlap between the generated answer and the ground-truth reference. We use ROUGE-N and ROUGE-L, which evaluate n-gram overlap and longest common subsequence (LCS), respectively. Scores range from 0 to 1, with higher values indicating better alignment.

(2) Semantic Similarity

Semantic similarity evaluates the meaning-level alignment between the generated answer and the reference answer using embedding-based cosine similarity. This metric captures semantic equivalence beyond surface-level text matching.

(3) Answer Relevance

Answer Relevance measures how well the generated answer addresses the input query. It is computed as the average cosine similarity between the embedding of the original question and a set of reverse-generated questions derived from the answer:

$$\text{Answer Relevance} = \frac{1}{N}\sum_{i=1}^{N}\cos(q, q_i') \quad (1)$$

where q is the original query embedding, q_i' are generated questions, and N is the number of generated questions.

(4) Faithfulness

Faithfulness evaluates whether the generated answer is grounded in the retrieved context. It is defined as:

$$\text{Faithfulness} = \frac{\text{Number of claims in the response supported by the retrieval context}}{\text{Total number of claims in the response}} \quad (2)$$

A higher score indicates that the answer is more factually consistent with the retrieved documents.

(5) Context Precision

Context Precision measures how effectively a retriever orders its results by prioritizing relevant chunks over irrelevant ones for a given query. In particular, it reflects how well the relevant chunks are positioned toward the top of the retrieved list:

$$\text{Context Precision @k} = \frac{\sum_{k=1}^{K}(Precision@k \times v_k)}{Total\ number\ of\ relevant\ items\ in\ the\ top\ K\ resuls} \quad (3)$$

Context Recall

Context Recall evaluates the ability of the retriever to retrieve all relevant information:

$$\text{Context Recall} = \frac{\text{Number of claims in the refernce supported by the retrieved context}}{\text{Total number of claims in the reference}} \quad (4)$$

III. METHODOLOGY

Dataset

In Vietnam, the university admission process typically takes place from June to September each year. According to regulations issued by the Ministry of Education and Training, higher education institutions are required to develop admission plans, establish their own admission policies, and specify detailed regulations for each training program while ensuring consistency with the national admission framework. In addition, all admission-related information must be publicly disclosed in a transparent manner to enable prospective students to access information early, register for admission, and ensure fairness in the selection process.

In this study, the research team collected admission-related data from Thuongmai University, including official admission proposals published between 2021 and 2025 on the university's website (http://www.tmu.edu.vn), as well as data from the university's official admission fanpage over a two-year period

(2024-2025): https://www.facebook.com/tuyensinhdhtm/?locale=vi_VN. The fanpage data, managed by the university's admission office, consist of real-world question-answer interactions between prospective students and admission advisors during the admission cycles.

Implementation process

In our research, there are two main phases: data preparation and RAG system. These phases are described in Figure 1.

Phase 1: Data preparation

The dataset used in this study is collected from two primary sources: formal documents, notably official admission proposals, and informal data consisting of question–answer pairs extracted from the university's admission fanpage. The dataset is inherently heterogeneous in structure, containing both unstructured and semi-structured elements such as tables, images, and emojis, which pose significant challenges for data processing and information extraction. To address these challenges, a data preprocessing pipeline is implemented to standardize and enhance the quality of the input data. Specifically, elements with limited semantic value, including images, emojis, and noisy text (such as informal abbreviations and non-standard "teencode"), are removed. A key component of this stage involves processing tabular data within admission documents. To preserve both structural integrity and contextual meaning, tables are categorized into five schema types: Flat tables, Hierarchical tables, Continuation tables, Matrix tables, and Multi-entity tables. This classification enables the application of tailored prompt design strategies for LLMs corresponding to each table type, thereby improving the accuracy of data transformation and representation.

Phase 2: RAG system

After completing the preprocessing stage, the data becomes ready for subsequent steps in the RAG pipeline, including information extraction, indexing, and retrieval. This process not only ensures data consistency but also establishes a solid foundation for improving the performance of large language model-based question answering systems.

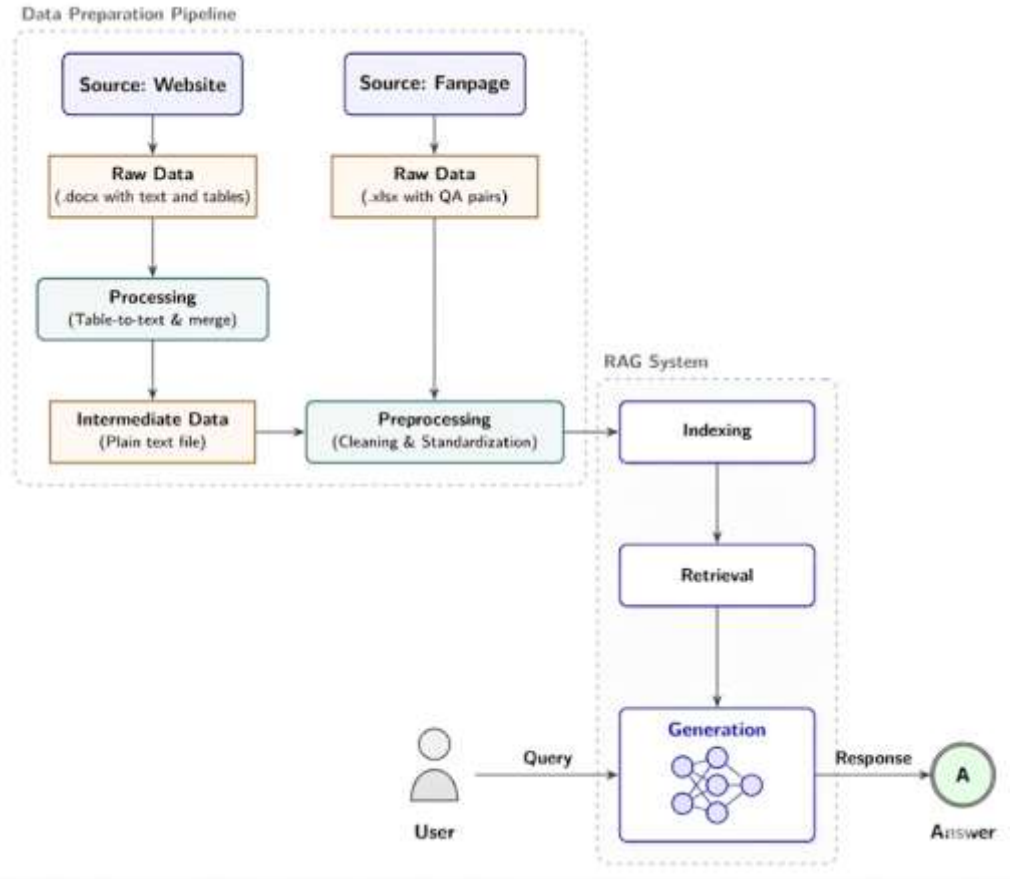

Figure 1. Pipeline admission RAG chatbot

Algorithm: *Parse Table and Generate Text*

The core idea of the algorithm is to iterate over raw document files (e.g., .docx), identify contextual structure based on document headings, and extract tables present within the documents. These tables are then standardized through preprocessing steps, including header normalization, cell merging, and table type identification, before being converted into structured data formats such as .json. Subsequently, the structured data are segmented into smaller chunks and combined with contextual information to construct prompts for large language models (LLMs). The generated outputs are then

post-processed and stored as textual passages, which serve as inputs for downstream retrieval and question answering tasks.

```
Algorithm: Parse Table And Generate Text
INPUT:
    D: Set of .docx files
    L: LLM model instance
    β: Batch size (default: 10)

OUTPUT:
    Γ: Set of generated text elements

BEGIN
    Γ ← ∅
    global_table_id ← 0

    FOR EACH document dᵢ ∈ D DO
       heading_stack ← []

       FOR EACH element ∈ dᵢ.elements DO

         // Update heading hierarchy
         IF element IS Heading THEN
            heading_info ← EXTRACT_HEADING(element)
            UPDATE_HEADING_STACK(heading_stack,
heading_info)

         // Process table
         ELSE IF element IS Table THEN

           // Get context from current heading
           context ← GET_CONTEXT(heading_stack)

           // Extract raw table matrix
           M₀ ← EXTRACT_RAW_MATRIX(element)
           IF |M₀| < 2 THEN CONTINUE

           // Detect table schema
           σ ← DETECT_SCHEMA(M₀)

           // Normalize headers and separate data rows
           H, M' ← NORMALIZE_HEADERS(M₀)

           // Fill merged cells based on schema
           M'' ← FILL_MERGED_CELLS(M', σ)

           // Convert to structured dictionaries
           R ← CONVERT_TO_STRUCTURED(H, M'')

           // Partition rows into batches
           B ← PARTITION(R, β)

           // Generate text for each batch
           FOR EACH batch ∈ B DO
             // Create prompt with context
             π ← CONSTRUCT_PROMPT(batch, σ, context)

             // Generate with LLM
             g ← L.generate(π, temperature=0.3)

             // Clean and store result
             g' ← CLEAN_TEXT(g)
             e ← CREATE_ELEMENT(g', global_table_id,
σ,context)
             Γ ← Γ ∪ {e}

           global_table_id ← global_table_id + 1

       RETURN Γ
    END

    FUNCTION: CONSTRUCT_PROMPT(batch, σ, context)
       template ← SELECT_TEMPLATE(σ)
       data_str ← FORMAT_BATCH(batch)

       π ← "Bạn là chuyên gia viết nội dung tuyển
sinh. Dựa vào thông tin cung cấp, hãy chuyển đổi
nội dung bảng thành văn bản tự nhiên bằng tiếng
Việt. Không thêm thông tin không có trong bảng.\n"
       π ← π + "Ngữ cảnh: " + context.heading + "\n"
       π ← π + "Dữ liệu bảng:\n" + data_str + "\n"
       π ← π + "Viết đoạn văn tự nhiên:"

       RETURN π
    END
```

After the preprocessing stage, the data are segmented into chunks of 1000 tokens with an overlap of 100 tokens. Section-level metadata are inherited and attached to all corresponding chunks. Documents are segmented using a structure-aware chunking strategy, in which headings, subheadings, paragraphs, and tables are preserved as semantic units prior to embedding generation. This approach enhances retrieval quality by maintaining contextual coherence within each chunk. Following chunking, embeddings are generated using models from HuggingFace, and the resulting vector representations are stored in a vector database. The use of FAISS enables efficient storage and retrieval of both vectors and associated metadata, thereby improving semantic search over the internal dataset. During inference, the system retrieves the top-5 most relevant chunks based on the user's query. These retrieved contexts are then combined with the original query to construct an augmented prompt, which is subsequently fed into the large language model to generate the final response.

Prompt sample:
Hệ thống: Bạn là trợ lý tư vấn tuyển sinh đại học. Dựa vào thông tin được cung cấp, hãy trả lời câu hỏi chính xác và đầy đủ bằng tiếng Việt.

Thông tin tham khảo:
{formatted_context}

Câu hỏi: {question}

Trả lời: ...

The configuration for GPT-4 is set with a temperature of 0.7 and a maximum token limit of 512. Similarly, VinaLLaMA-7B is configured with a temperature of 0.7 and a maximum token limit of 512, and is deployed on GPU hardware equivalent to or higher than NVIDIA T4.

IV. Results and Discussion

To facilitate user interaction with the admission system, a web-based interface is developed using Streamlit, allowing prospective students to submit queries and receive responses.

The study allocates 10% of the interaction question dataset as a test set, and the experimental results are presented in Table 2 and Table 3.

International Journal of Scientific Engineering and Science
Volume 10, Issue 4, pp. 70-76, 2026. ISSN (Online): 2456-7361

Figure 2 Admissions QA system for Thuongmai University.

The results in Table 2 indicate that the integration of RAG significantly improves the performance of both models (GPT-4 and VinaLLaMA-7B) across all evaluation metrics. For GPT-4, Rouge-1 increases from 0.3545 to 0.4599 (29.7% improvement), Rouge-2 from 0.1680 to 0.2476 (47.4%), and Rouge-L from 0.2304 to 0.3273 (42.05%). Meanwhile, VinaLLaMA-7B exhibits even more substantial improvements, particularly in Rouge-2, which nearly doubles from 0.0554 to 0.1101 (98.7% increase). In terms of semantic performance, the Semantic Similarity score of GPT-4 shows a marginal increase from 0.4981 to 0.4982 (0.02%), whereas VinaLLaMA-7B improves significantly from 0.4000 to 0.4742 (18.6%). Notably, VinaLLaMA-7B combined with RAG achieves the highest Answer Relevance score (0.7064), outperforming GPT-4 + RAG (0.6694). These findings suggest that RAG is particularly effective in enhancing the performance of smaller models, while also improving the accuracy and relevance of domain-specific question answering systems.

TABLE 2. Comparison of GPT-4 and VinaLLaMA-7B with and without RAG

Metric	GPT-4	GPT-4 + RAG	VinaLLaMA 7B	VinaLLaMA 7B + RAG
Rouge-1	0.3545	0.4705	0.1799	0.2096
Rouge-2	0.1680	0.2476	0.0554	0.1101
Rouge-L	0.2304	0.3273	0.1326	0.1539
Semantic Similarity	0.4981	0.4982	0.4000	0.4742
Answer Relevance	0.6263	0.6694	0.6454	**0.7064**

Table 3 shows that both models, when integrated with RAG, achieve comparable performance in context retrieval, with Context Precision and Context Recall remaining constant at 0.6750 and 0.6177, respectively. However, a notable difference is observed in the Faithfulness metric, where VinaLLaMA-7B + RAG achieves a score of 0.7059, significantly outperforming GPT-4 + RAG (0.6348), corresponding to an improvement of approximately 11.2%. This indicates that although both models exhibit similar retrieval capabilities, VinaLLaMA-7B, when combined with RAG, generates responses that are more faithful and better grounded in the source documents. These findings further reinforce the role of RAG in enhancing the reliability of question answering systems, particularly for smaller-scale models.

TABLE 3. Metrics between GPT + RAG and VinaLLaMa 7B + RAG

Metric	GPT-4 + RAG	VinaLLaMA 7B + RAG
Faithfulness	0.6348	**0.7059**
Context Precision	0.6750	0.6750
Context Recall	0.6177	0.6177

V. Conclusion

This study proposes and implements a RAG-based question answering system for university admissions at Thuongmai University, Vietnam. The research develops a structured data processing pipeline, including table information extraction and hierarchical data organization, to enhance domain-specific knowledge representation. Experimental results demonstrate that the RAG-based approach significantly improves question answering performance across all evaluation metrics in the educational domain. Notably, in terms of Faithfulness, VinaLLaMA-7B outperforms GPT-4 when integrated with RAG, highlighting the potential of Vietnamese language models when combined with external knowledge retrieval. These findings confirm the effectiveness and practical applicability of the proposed system in university admission contexts, contributing to reducing the workload of admission staff and providing timely and accurate support for prospective students.

References

1. Le Thi Hao and Nguyen Thi Thu Trang, *Developing an automated response system software to support admissions counseling at the Vietnam Trade Union University.* Tap chinghien cuu khoa hoc cong doan, 2025. 34(72).
2. Lê, V.S., *Xay dung he tu van tuyen sinh tu dong cho Truong Dai hoc Phan Thiet*. 2023, Truong Dai hoc Ba Ria-Vung Tau.
3. Nguyen, T.T., et al., *NEU-chatbot: Chatbot for admission of National Economics University.* Computers and Education: Artificial Intelligence, 2021. 2: p. 100036.
4. Phan Thị Thanh Nga, et al., *An approach for building a chatbot system for the admission process of Da Lat university.* TNU Journal of science and technology. 227(14): p. 23-32.
5. Ojokoh, B. and E. Adebisi, *A review of question answering systems.* Journal of Web Engineering, 2018. 17(8): p. 717-758.
6. Alansari, A. and H. Luqman, *Large language models hallucination: A comprehensive survey.* arXiv preprint arXiv:2510.06265, 2025.
7. Sobhan, S. and M.A. Haque, *LLM-Assisted Question-Answering on Technical Documents Using Structured Data-Aware Retrieval Augmented Generation.* arXiv preprint arXiv:2506.23136, 2025.
8. Zhao, W.X., et al., *A survey of large language models.* arXiv preprint arXiv:2303.18223, 2023. 1(2): p. 1-124.
9. Nguyen, D.Q. and A.-T. Nguyen. *PhoBERT: Pre-trained language models for Vietnamese*. in *Findings of the association for computational linguistics: EMNLP 2020*. 2020.
10. Nguyen, Q., H. Pham, and D. Dao, *Vinallama: Llama-based vietnamese foundation model.* arXiv preprint arXiv:2312.11011, 2023.
11. Dam, S.K., et al., *A complete survey on llm-based ai chatbots.* arXiv preprint arXiv:2406.16937, 2024.
12. Ngo-Ho, A.-K., K.-D. Vo, and A.-K. Ngo-Ho. *Evaluation of Large Language Models for the Vietnamese Language in Generative Vietnamese Economy Chatbots (GVEC) Services*. in *International Conference on Electronics and Signal Processing*. 2024. Springer.
13. Le, H., et al., *Optimizing answer generator in Vietnamese legal question answering systems using language models.* ACM Transactions on Asian and Low-Resource Language Information Processing, 2025. 24(6): p. 1-17.

14. Nguyen, V.-V., et al., *Are LLMs Good for Low-resource Vietnamese and Other Translations?* 2024.
15. Nguyen, D.Q., et al., *Phogpt: Generative pre-training for vietnamese.* arXiv preprint arXiv:2311.02945, 2023.
16. Achiam, J., et al., *Gpt-4 technical report.* arXiv preprint arXiv:2303.08774, 2023.
17. Lewis, P., et al., *Retrieval-augmented generation for knowledge-intensive nlp tasks.* Advances in neural information processing systems, 2020. **33**: p. 9459-9474.
18. Ragas. *Ragas documentation*. 2025; Available from: https://docs.ragas.io.