URAG: Implementing a Unified Hybrid RAG for Precise Answers in University Admission Chatbots – A Case Study at HCMUT

Long S. T. Nguyen[1,2] and Tho T. Quan[1,2](✉)

[1] URA Research Group, Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), Ho Chi Minh City, Vietnam
qttho@hcmut.edu.vn
[2] Vietnam National University Ho Chi Minh City, Ho Chi Minh City, Vietnam

Abstract. With the rapid advancement of Artificial Intelligence, particularly in Natural Language Processing, Large Language Models (LLMs) have become pivotal in educational question-answering (QA) systems, especially university admission chatbots. Concepts such as Retrieval-Augmented Generation (RAG) and other advanced techniques have been developed to enhance these systems by integrating specific university data, enabling LLMs to provide informed responses on admissions and academic counseling. However, these enhanced RAG techniques often involve high operational costs and require the training of complex, specialized modules, which poses challenges for practical deployment. Additionally, in the educational context, it is crucial to provide accurate answers to prevent misinformation, a task that LLM-based systems find challenging without appropriate strategies and methods. In this paper, we introduce the Unified RAG (URAG) Framework, a hybrid approach that significantly improves the accuracy of responses, particularly for critical queries. Experimental results demonstrate that URAG enhances our in-house, lightweight model to perform comparably to state-of-the-art commercial models. Moreover, to validate its practical applicability, we conducted a case study at our educational institution, which received positive feedback and acclaim. This study not only proves the effectiveness of URAG but also highlights its feasibility for real-world implementation in educational settings.

Keywords: Question-Answering Systems · Retrieval-Augmented Generation · University Admission Chatbots

1 Introduction

Artificial Intelligence (AI) has become a fundamental component of modern technological advancements, transforming industries across the board, including education. Among its various applications, *Natural Language Processing* has

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025
W. Buntine et al. (Eds.): SOICT 2024, CCIS 2351, pp. 82–93, 2025.
https://doi.org/10.1007/978-981-96-4285-4_7

proven particularly valuable in the development of chatbots designed to assist with university admissions and provide comprehensive institutional information [11]. These chatbots play an essential role in bridging the communication gap between universities and prospective students, ensuring that inquiries are met with timely and accurate responses. Recent breakthroughs in AI, such as the introduction of the *Attention Mechanism* [21] and the rise of *Large Language Models* (LLMs), have revolutionized educational chatbots by enabling them to generate human-like text and perform complex tasks [7]. However, despite these advancements, LLM-based chatbots are still prone to generating inaccurate or misleading responses, commonly referred to as *hallucinations* [9]. This limitation is particularly critical in high-stakes contexts such as university admissions, where precise and reliable information about application deadlines and program details is essential. To mitigate these risks, the *Retrieval-Augmented Generation* (RAG) approach has emerged as a potential solution [12]. *Naive RAG* combines retrieval-based mechanisms with generative models, enabling chatbots to consult external sources of information before generating responses. While this approach helps reduce hallucinations and improves accuracy, early implementations of RAG have faced limitations, such as noise in retrieval results, a disconnect between retrieval and generation processes, and difficulties in managing longer contexts [25]. These limitations can still result in inaccurate responses and hallucinations. Furthermore, more advanced RAG systems [2,8,22,24], though promising, often introduce greater complexity and operational costs, making their deployment in real-world educational settings less feasible.

In response to these challenges, we propose the ***Unified RAG*** (URAG) framework, specifically designed to improve lightweight LLMs for use in university admission chatbots. URAG integrates the reliability of rule-based systems with the adaptability of RAG, creating a two-tiered approach. The first tier leverages a comprehensive *Frequently Asked Questions* (FAQ) system to provide accurate responses to common queries, especially those involving sensitive or critical information. If no match is found in the FAQ, the second tier retrieves relevant documents from an augmented database and generates a response through an LLM. To enhance this process, we propose two key mechanisms, URAG-D for augmenting the original document database and URAG-F for generating an enhanced FAQ. These mechanisms not only enrich the database but also improve the retrieval process in both tiers, as shown in Fig. 1.

Our experiments highlight the effectiveness of URAG when paired with an in-house developed Vietnamese lightweight LLM. We benchmarked URAG's performance against *state-of-the-art* (SOTA) commercial chatbots, including GPT-4o[1], Gemini 1.5 Pro[2], and Claude 3.5 Sonnet[3]. To further validate URAG's practical application, we integrated it into HCMUT Chatbot[4], where it has

[1] https://openai.com/chatgpt/.
[2] https://gemini.google.com/.
[3] https://claude.ai/.
[4] https://www.ura.hcmut.edu.vn/bk-tvts/.

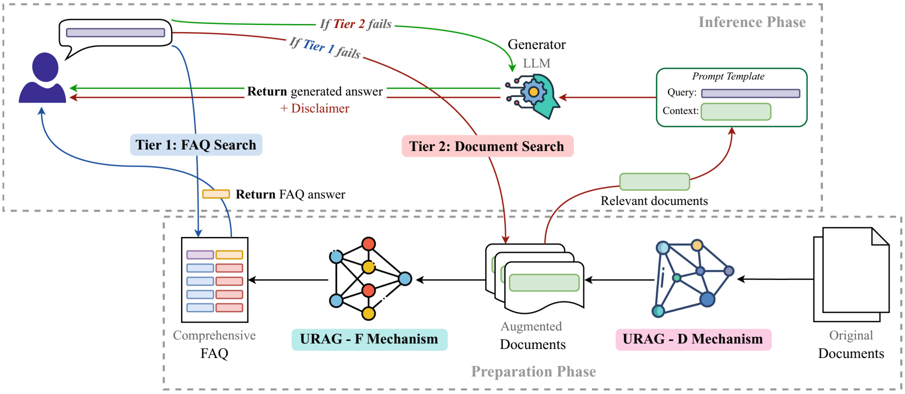


Fig. 1. The architecture of URAG framework illustrating the two-tiered approach for improving LLM performance in university admission chatbots.

since gained recognition for its positive impact on university admissions at *Ho Chi Minh City University of Technology* (HCMUT).

In summary, our contributions are as follows.

- We introduce URAG, a hybrid system that integrates rule-based and RAG approaches to enhance the performance of lightweight LLMs tailored for educational chatbots.
- We collected a real-world dataset of university admission questions from high school students and conducted a comprehensive evaluation of URAG against leading commercial chatbots, demonstrating its competitive performance.
- We successfully implemented URAG in a practical deployment at HCMUT, showcasing its effectiveness through a functional product that continues to address the university's needs.

2 Related Work

2.1 Strategies for Enhancing a RAG Pipeline

Naive RAG implementations, while effective, face significant limitations in complex question-answering tasks such as university admission chatbots [12]. Enhancements to RAG pipelines are typically categorized into input refinement, retriever improvements, generator optimization, result validation, and overall pipeline integration [25]. Advanced methods, including *Self-Reflective RAG* (Self-RAG) [2], *Corrective RAG* (CRAG) [24], and *Adaptive RAG* [8], enrich inputs through web searches and detect hallucinations for response correction. Another notable approach, *Speculative RAG* [22], generates multiple response drafts by processing query-document clusters in parallel and selecting the best answer. While these methods improve accuracy, they also introduce additional complexity. Iterative feedback loops for response refinement increase processing times and computational demands. Although speculative execution reduces latency, the need for separate module training renders these advanced RAG models resource-intensive and challenging for practical educational applications.

2.2 Practical Approaches to University Admission Chatbots

University admission chatbots are crucial in providing immediate support to prospective students, with *Natural Language Understanding* enabling intent classification and entity recognition, and *Natural Language Generation* creating responses [1]. Platforms like *Rasa* [3] facilitate robust NLU training and dialogue management, handling FAQs effectively but struggling with complex or out-of-scope queries [13,15]. To address these limitations, advanced models integrate Sequence-to-Sequence architectures with Attention Mechanisms [6] or simplified RAG structures fine-tuned on domain-specific data [5]. Some chatbots also utilize APIs like GPT-3.5 Turbo [14,16], enhancing capabilities but introducing challenges such as external data handling risks and high operational costs. Despite these advancements, reliance on LLMs alone often leads to inaccuracies in addressing sensitive, context-specific queries. This highlights the need for a lightweight, efficient solution that combines LLM capabilities with mechanisms ensuring accurate responses tailored to the demands of university admissions.

3 URAG: A Unified RAG for Precise University Admission Chatbots

3.1 Overview of URAG Architecture

The operational flow of the URAG framework is systematically outlined in Algorithm 1, with key terms defined in Table 1.

3.2 The Truth Behind "Unified" in URAG Architecture

The URAG architecture integrates two key mechanisms during its preparatory phase to optimize deployment performance. One critical prompt-based learning technique in this process is *Chain-of-Thought* (CoT) Prompting [23], which enhances the reasoning capabilities of LLM-based components. By leveraging few-shot prompting, CoT employs triples <*input, chain of thought, output*> to generate reliable and contextually appropriate responses.

URAG-D Mechanism: Document Database Augmentation. URAG-D enhances document retrieval by segmenting documents into coherent chunks and rewriting them to ensure consistency and contextual relevance, addressing the limitations of processing entire documents as in Naive RAG, as outlined in Algorithm 2. A core component of URAG-D is Semantic Chunking, a novel technique proposed by [10], which adaptively determines chunk boundaries based on embedding similarity. By analyzing sentence embeddings $\mathcal{E}(s_r)$

Table 1. Notation

Symbol	Description
$\mathbb{T}$	The textual domain
$\mathcal{E} : \mathbb{T} \rightarrow \mathbb{R}^m$	A mapping function that transforms a text $t \in \mathbb{T}$ into an m-dimensional embedding vector $E(t) \in \mathbb{R}^m$, facilitated by an Embedding Model
$\mathcal{L} : \mathbb{T} \rightarrow \mathbb{T}$	A generation function that maps a prompt $p \in \mathbb{T}$ to a response $\mathcal{L}(p) \in \mathbb{T}$, generated by a Large Language Model
$f_i = \{a_i, b_i\}$	A *question-answer* (Q-A) pair from the enriched FAQ list F, where $a_i \in \mathbb{T}$ is a question and $b_i \in \mathbb{T}$ is its corresponding answer
d_i	A document segment within the augmented corpus D
$t_{\text{FAQ}}, t_{\text{Doc}}$	**Thresholds** defining the acceptance criteria for FAQ and Document search tiers, respectively
w_i	A warning or disclaimer from a predefined set $W \subseteq \mathbb{T}$

Algorithm 1. URAG Operational Framework

```
Input: User query $q \in \mathbb{T}$
Output: Final answer $a_G \in \mathbb{T}$
Query Embedding: Compute $q' = \mathcal{E}(q) \in \mathbb{R}^m$
Tier 1: FAQ Search
Identify the top $k$ FAQ pairs $F_k = \{f_i\}_{i=1}^{k}$ such that each $f_i = \{a_i, b_i\}$ satisfies $\cos(\theta(q', \mathcal{E}(a_i))) \geqslant t_{\text{FAQ}}$, where $a_i \in f_i \in F$ and $i \in \{1, 2, .., k\}$
if $|F_k| \geqslant 1$ then
    Select answer $b_j$ from the pair $f_j = \{a_j, b_j\}$ with the highest score in $F_k$
    Return $a_G = b_j$
else
    Proceed to Tier 2
end if
Tier 2: Document Search
Retrieve the top $K$ document segments $D_K = \{d_i\}_{i=1}^{K}$ that meet the threshold $t_{\text{Doc}}$, meaning $\cos(\theta(q', \mathcal{E}(d_i))) \geqslant t_{\text{Doc}}$, where $d_i \in D$ and $i \in \{1, 2, .., K\}$
if $|D_K| \geqslant 1$ then
    Construct prompt $p = \text{concat}(q, D_K)$, where $p \in \mathbb{T}$
    Generate $a_G = \mathcal{L}(p)$
    Return $a_G := \text{concat}(\mathcal{L}(p), w_s)$
else
    Fallback: Generate $a_G = \mathcal{L}(q)$
    Return $a_G := \text{concat}(\mathcal{L}(q), w_r)$
end if
```

within a document o_i, semantically similar sentences are grouped into cohesive chunks $c_{i_j} = \{s_r\}_r$. It is important to note that each chunk d_{i_j} is treated as a distinct *document* within the system, collectively forming the *augmented document corpus*, denoted as D.

Algorithm 2. URAG-D Workflow

```
Input: Set of original documents $O = \{o_i\}_{i=1}^{K}$, where $O \subseteq \mathbb{T}$
Output: Augmented context corpus $D$, used in Tier 2 of Algorithm 1
Initialize $D = \emptyset$
for $i = 1$ to $K$ do
    Semantic Chunking: Apply Semantic Chunking on $o_i$ to produce coherent chunks $\{c_{i_j}\}_{j=1}^{L}$, where $L = f(o_i)$ is adapted based on the content of $o_i$
    Context Extraction: Extract the general context $g_i$ from $o_i$ using $\mathcal{L}(o_i)$ to capture overarching themes
    Chunk Reconstruction:
    for $j = 1$ to $L$ do
        Rewrite the chunk $c_{i_j}$ using the context $g_i$ to form $t_{i_j} = \mathcal{L}(c_{i_j} \mid g_i)$
        Condense $t_{i_j}$ into a succinct representation $h_{i_j} = \mathcal{L}(t_{i_j})$
        Combine to produce the final chunk $d_{i_j} = \text{concat}(h_{i_j}, t_{i_j})$, where $d_{i_j} \in \mathbb{T}$
    end for
    Append to the augmented corpus: $D := D \cup \{d_{i_j}\}_j$
end for
```

URAG-F Mechanism: FAQ Database Enrichment. URAG-F, as described in Algorithm 3, enriches the FAQ database by leveraging the augmented document corpus D from URAG-D and the initial FAQ set. It generates new Q-A pairs and paraphrases them into multiple variations to enhance linguistic diversity, particularly for languages like Vietnamese, where questions can be expressed in many ways. This approach expands the FAQ collection, improving coverage and response quality.

Algorithm 3. URAG-F Workflow

```
Input: Context corpus $D$ and initial FAQ set $F_0$
Output: Enriched FAQ collection $F$, used in Tier 1 of Algorithm 1
Initial FAQ Expansion:
Utilize $F_0$ to generate an expanded FAQ set: $F = F_0 \cup \mathcal{L}(F_0)$
FAQ Generation from Documents:
for $i = 1$ to $|D|$ do
    Extract Q-A pairs $F_{D_i} = \mathcal{L}(d_i)$ from documents $d_i$
    Update the FAQ collection: $F := F \cup F_{D_i}$
end for
FAQ Enrichment:
for $j = 1$ to $|F|$ do
    Generate paraphrased variants $\{f_{j_k}\}_k$ of each Q-A pair $f_j$
    Enrich the FAQ set with the paraphrased versions: $F := F \cup \{f_{j_k}\}_k$
end for
```

4 Experimentation

4.1 Experiment Setup

We conducted an experiment to evaluate the accuracy of the URAG system in answering questions related to university admissions.

Dataset Preparation. The dataset comprises 500 curated questions collected from high school students during three HCMUT admission events. It includes 85.4% Factual and 14.6% Reasoning types, reflecting typical inquiries in the admissions process. Questions with readily available answers were sourced from the official HCMUT website[5] and other reliable online platforms.

URAG Baseline. For the Embedding Model, we used the SOTA Vietnamese model dangvantuan/vietnamese-embedding, which leverages *Sentence-BERT* (SBERT) [18] for sentence-level semantic understanding. Semantic Chunking was performed using LangChain [17], a robust framework for LLM applications. The Large Language Model generator is powered by the lightweight Vietnamese model ura-hcmut/ura-llama-7b [20], pretrained on LLaMA [19] with an extensive Vietnamese dataset. Additionally, we prepared 300 documents on general information, events, personnel, admissions, and other HCMUT-related topics. Together, these components form HCMUT Chatbot, a fully integrated system for addressing university admission inquiries.

Comparative Systems. To benchmark URAG's performance, we compared it with leading commercial chatbots, including GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro, using the prepared dataset. Evaluations were conducted on their official platforms, enabling full use of advanced features like Internet access and reasoning engines. These systems were configured with prompts to perform online searches, mirroring CRAG systems that use external searches to enhance response accuracy.

Evaluation Metric. Each question from the evaluation dataset was systematically posed to each model, and the responses were assessed by experts for correctness. *Accuracy*, calculated as shown in Eq. 1, was chosen as the primary evaluation metric due to its relevance in measuring the correctness of chatbot answers.

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n}\text{Correct}_i, \tag{1}$$

where Correct_i equals 1 if the answer to the i-th question is correct and 0 otherwise, and n represents the total number of questions in the dataset. This metric is particularly suitable for university admission chatbots, where the priority is on the correctness of the answers rather than their phrasing.

[5] https://hcmut.edu.vn/.

4.2 Results and Analysis

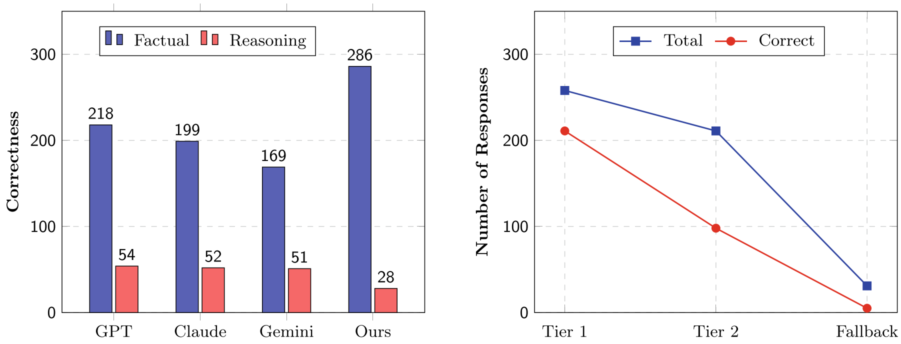
Table 2 demonstrates that HCMUT Chatbot, powered by the URAG system, achieved the highest accuracy among all evaluated systems. This result underscores the effectiveness of URAG's two-tier architecture in enhancing lightweight model performance through efficient information retrieval.

Performance by question types, depicted in Fig. 2a, shows that HCMUT Chatbot excelled in Factual questions but exhibited limited gains in Reasoning tasks. This is attributed to its reliance on a lightweight Vietnamese language model, compared to commercial chatbots equipped with advanced reasoning engines and larger parameter scales.

The impact of URAG's tiered design is illustrated in Fig. 2b. URAG-F and URAG-D successfully resolved most questions within the first two tiers while minimizing fallback cases. Responses were evenly distributed between Tier 1 and Tier 2, with slight accuracy declines in lower tiers, reflecting the structured and efficient hierarchy of the retrieval process.

Table 2. Performance Metrics of Evaluated Chatbot Systems

System	Correctness	Accuracy
GPT-4o	272	0.544
Claude 3.5 Sonnet	251	0.502
Gemini 1.5 Pro	220	0.440
HCMUT Chatbot	**314**	**0.628**



(a) Performance Evaluation of Chatbot Models by Response Type

(b) Tier-Based Distribution and Accuracy of Responses in URAG

Fig. 2. Comprehensive Analysis of Chatbot Model Performance and URAG's Tiers

4.3 Case Study: Deployment of the Admission Chatbot at HCMUT

The HCMUT Chatbot, powered by the URAG system, was deployed on the university's official website at ura.hcmut.edu.vn/bk-tvts. Over four months of deployment, the chatbot recorded high levels of interaction, particularly among high school students, as shown in Fig. 3. Interaction peaks were observed in early June, coinciding with the end of the high school year, and late August, just before the university term began. These trends align with the expected demand for admission information during critical periods. The chatbot consistently delivered fast responses under optimal conditions, with only minor delays for more complex queries. These results demonstrate the system's stability and practical applicability, particularly in resource-constrained educational environments.

A key strength of the URAG system lies in its ability to effectively handle concise or incomplete queries, a common challenge when interacting with prospective students. The examples below highlight how the chatbot provides accurate answers, even with minimal input, by leveraging its enriched FAQ database.

- ***Query:*** Address?
  ↪ ***Response:*** Ho Chi Minh City University of Technology is located at 268 Ly Thuong Kiet Street, Ward 14, District 10, Ho Chi Minh City, Vietnam.
- ***Query:*** Founding year?
  ↪ ***Response:*** Ho Chi Minh City University of Technology was established in 1957 as the National Centre for Technical Education.

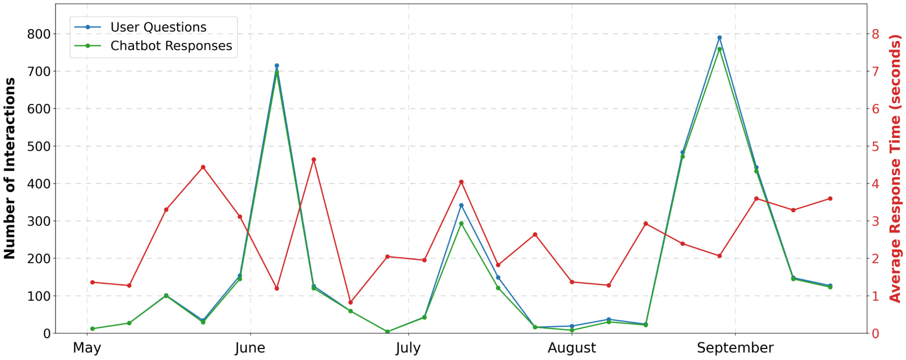


Fig. 3. Interaction and Response Time Statistics of HCMUT Chatbot

5 Discussion, Limitations, and Future Work

This paper introduced the URAG system as an efficient and lightweight solution tailored for university admission chatbots, successfully deployed at HCMUT with a steadily growing user base. The system's two-tier architecture demonstrates significant improvements in response accuracy, particularly for high-priority queries, by effectively mitigating hallucination issues commonly observed in LLM-based systems. Furthermore, the traceability features of URAG-D and URAG-F enhance the management of enriched FAQ databases by enabling seamless linkage between questions, variations, and their original sources. This design simplifies the updating process, ensuring the chatbot remains accurate and aligned with evolving information needs.

Despite these advancements, URAG's reliance on a lightweight generative model, while reducing deployment costs and enhancing security, introduces certain performance limitations compared to larger models or third-party APIs. These limitations are particularly evident in handling general queries outside the admissions domain [20]. Future work could explore integrating RAG with cross-data *Knowledge Graphs*, which are well-suited for managing complex and interconnected educational datasets [4]. This approach could significantly expand the chatbot's ability to address a broader range of queries while maintaining domain specificity and relevance. Nonetheless, it is essential to acknowledge that certain inquiries will continue to require the nuanced judgment and contextual understanding of human advisors.

References

1. Aloqayli, A., Abdelhafez, H.A.: Intelligent chatbot for admission in higher education. Int. J. Inf. Educ. Technol. (2023)
2. Asai, A., et al.: Self-RAG: learning to retrieve, generate, and critique through self-reflection (2023)
3. Bocklisch, T., et al.: Rasa: Open Source Language Understanding and Dialogue Management (2017)
4. Bui, T., et al.: Cross-data knowledge graph construction for LLMenabled educational question-answering system: a case study at HCMUT. In: Proceedings of the 1st ACM Workshop on AI-Powered Q&A Systems for Multimedia. AIQAM 2024, pp. 36–43. Association for Computing Machinery, Phuket, Thailand (2024). https://doi.org/10.1145/3643479.3662055
5. Cabezas, D., et al.: Integrating a LLaMa-based chatbot with augmented retrieval generation as a complementary educational tool for high school and college students. In: Proceedings of the 19th International Conference on Software Technologies (ICSOFT 2024), pp. 395–402, January 2024. https://doi.org/10.5220/0012763000003753
6. Chandra, Y.W., Suyanto, S.: Indonesian chatbot of university admission using a question answering system based on sequenceto- sequence model. Procedia Comput. Sci. **157** (2019). The 4th International Conference on Computer Science and Computational Intelligence (ICCSCI 2019): Enabling Collaboration to Escalate Impact of Research Results for Society, pp. 367–374. https://doi.org/10.1016/j.procs.2019.08.179

7. Ganesan, M., et al.: A survey on chatbots using artificial intelligence. In: 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), pp. 1–5 (2020). https://doi.org/10.1109/ICSCAN49426.2020.9262366
8. Jeong, S., et al.: Adaptive-RAG: learning to adapt retrieval-augmented large language models through question complexity. In: Duh, K., Gomez, H., Bethard, S. (eds.) Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Association for Computational Linguistics, Mexico City, Mexico, pp. 7036–7050, June 2024. https://doi.org/10.18653/v1/2024.naacl-long.389
9. Ji, Z., et al.: Survey of hallucination in natural language generation. In: ACM Comput. Surv. **55**(12) (2023). https://doi.org/10.1145/3571730
10. Kamradt, G.: 5 levels of text splitting (2024). https://github.com/FullStackRetrievalcom/RetrievalTutorials
11. Lende, S.P., Raghuwanshi, M.M.: Question answering system on education acts using NLP techniques. In: 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave), pp. 1–6 (2016). https://doi.org/10.1109/STARTUP.2016.7583963
12. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS 2020. Curran Associates Inc., Vancouver, BC, Canada (2020)
13. Nguyen, M.-T., et al.: Building a chatbot for supporting the admission of universities. In: 2021 13th International Conference on Knowledge and Systems Engineering (KSE), pp. 1–6 (2021). https://doi.org/10.1109/KSE53942.2021.9648677
14. Nguyen, T.-H., et al.: AI-powered university: design and deployment of robot assistant for smart universities. J. Adv. Inf. Technol. (2022)
15. Nguyen, T.T., et al.: NEU-chatbot: chatbot for admission of National Economics University. Comput. Educ. Artif. Intell. **2**, 100036 (2021). https://doi.org/10.1016/j.caeai.2021.100036
16. Odede, J., Frommholz, I.: JayBot – aiding university students and admission with an LLM-based chatbot. In: Proceedings of the 2024 Conference on Human Information Interaction and Retrieval, CHIIR 2024, pp. 391–395. Association for Computing Machinery, Sheffield, United Kingdom (2024). https://doi.org/10.1145/3627508.3638293
17. Pandya, K., Holia, M.S.: Automating customer service using LangChain: building custom open-source GPT chatbot for organizations (2023)
18. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: Inui, K., et al. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLPIJCNLP), pp. 3982–3992. Association for Computational Linguistics, Hong Kong, China, November 2019. https://doi.org/10.18653/v1/D19-1410
19. Touvron, H., et al.: LLaMA: open and efficient foundation language models (2023)
20. Truong, S., et al.: Crossing linguistic horizons: finetuning and comprehensive evaluation of Vietnamese large language models. In: Duh, K., Gomez, H., Bethard, S. (eds.) Findings of the Association for Computational Linguistics: NAACL 2024, pp. 2849–2900. Association for Computational Linguistics, Mexico City, Mexico, June 2024. https://doi.org/10.18653/v1/2024.findings-naacl.182
21. Vaswani, A., et al.: Attention is all you need. In: Guyon, I. et al. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017)

22. Wang, Z., et al.: Speculative RAG: Enhancing Retrieval Augmented Generation through Drafting (2024)
23. Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS 2022, Curran Associates Inc., New Orleans, LA, USA (2024)
24. Yan, S.-Q., et al.: Corrective Retrieval Augmented Generation (2024)
25. Zhao, P., et al.: Retrieval-augmented generation for AI-generated content: a survey (2024)