Recent Science and Technology

Journal homepage: https:// www.tci-thaijo.org/index.php/rmutsvrj/index

ISSN: 3057-0786 (Online)

Research Article

RMUTTOBot: Transforming University Admission Services with a TAG-based RAG LLM Chatbot

Vipa Thananant [a] and Saowakhon Nookhao [b *]

[a] *Division of Information Technology, Faculty of Business Administration and Information Technology, Rajamangala University of Technology Tawan-ok : Chakrabongse Bhuvanarth Campus, Bangkok 10400, Thailand.*

[b] *Digital Transformation and Technology Management Program, KMITL Business School, Bangkok 10520, Thailand.*

Article history:

Received: 2025-05-13
Revised: 2025-10-15
Accepted: 2025-10-28

Keywords:

Large language models (LLM);
Retrieval-augmented generation (RAG);
Structured data retrieval;
BERTScore;
Domain-specific chatbot

ABSTRACT

Advancements in artificial intelligence, particularly in large language models (LLMs) and retrieval-augmented generation (RAG) techniques, have improved chatbot capabilities for more natural and domain-specific interactions. However, conventional RAG systems, which retrieve information from unstructured text sources like websites and PDFs, exhibit critical failures when applied to the dynamic and precise nature of university information. This research addresses these gaps through the design and development of RMUTTOBot, a domain-specific chatbot providing admissions support for prospective students at Rajamangala University of Technology Tawan-ok (RMUTTO). We propose a novel, lightweight table-augmented generation (TAG) approach that combines a curated, updatable knowledge base for general information with live database queries for real-time, dynamic data. Performance was evaluated using both automated metrics and human assessments across six criteria: semantic similarity, retrieval effectiveness, relevance, fluency, coverage, and consistency. Experimental results show that the TAG-based RAG system significantly outperformed both the baseline LLM-only configuration and PDF-based RAG system, achieving a 12.76% higher BERTF1 score compared to a PDF-based RAG. Human evaluation confirmed the system's high response relevance and linguistic fluency, with strong inter-rater reliability (Krippendorff's $\alpha > 0.835$). These findings demonstrate that combining structured data augmentation with RAG substantially enhances chatbot accuracy, contextual grounding, and completeness, offering a robust framework for intelligent conversational systems in academic domains. The source code and implementation details are publicly available at https://github.com/vipa-thananant/RMUTTOBot.

1. Introduction

This research is situated within the significant evolution of chatbots, which have transformed through distinct developmental stages from simple tools into powerful learning partners (Alkishri *et al*., 2025). Early systems relied on pattern-matching, while subsequent versions integrated AI and natural language processing (NLP) for more context-aware interactions. Today, generative AI and Large Language Models (LLMs) represent the state-of-the-art. Within the LLM paradigm, solutions generally fall into three categories: fine-tuned models, retrieval-augmented generation (RAG) systems, and hybrid approaches (Ren *et al*., 2025; Wan *et al*., 2025). Fine-tuning involves adapting a pre-trained model using domain-specific datasets to improve accuracy and tone alignment (Doumanas *et al*., 2025). RAG-based systems, in contrast, retrieve relevant external content during inference without requiring retraining (Liang *et al*., 2025; Uhm *et al*., 2025). RAG is particularly effective when dealing with unstructured sources such as documents, websites, and PDF files. It functions by encoding such content into embeddings stored in a vector database, enabling efficient semantic search and integration into the generation process (Fan *et al*., 2024;

* **Corresponding author.**

E-mail address: *saowakhon.no@kmitl.ac.th*

Cite this article as:

Thananant, V. and Nookhao, S. 2026. RMUTTOBot: Transforming University Admission Services with a TAG-based RAG LLM Chatbot. **Recent Science and Technology** 18(1): 267581.

https://doi.org/10.65411/rst.2026.267581

Arslan *et al*., 2024). Hybrid models combine both strategies, leveraging fine-tuning for personalization and tone consistency while using retrieval mechanisms to provide fresh, topic-specific information (Budakoglu and Emekci, 2025).

While RAG has advanced the capabilities of university chatbots, its foundational reliance on semantic search over unstructured documents creates critical failures in the precision and reliability required for this high-stakes domain (Barnett *et al*., 2024). This approach struggles significantly with queries that demand keyword-level accuracy, such as requests for a specific course code like CSE101. A search may incorrectly link this query to general "introduction to computer science" documents, misdirecting students and preventing them from finding essential information. Furthermore, the chunking process used to index documents fragments context, leading to incomplete answers. A retrieved passage stating, "A minimum GPA of 3.5 is required," is not just incomplete but actively misleading if the preceding, unretrieved chunk specifies, "For the Faculty of Engineering applicants," causing potential applicants to miss critical requirements. This ambiguity is compounded when students ask for broad information like the "admissions policy," where a standard RAG system often returns a generic policy without distinguishing between the distinct requirements for undergraduate, graduate, or international students.

Most critically, conventional RAG is ill-suited for the dynamic, multi-faceted queries common in a university setting. Its core weakness is its reliance on static documents, such as admissions brochures, which cannot be updated in real time. For example, consider a prospective student asking, "Is the Bachelor of Nursing program still accepting applications for the fall semester?" A RAG system retrieving from a brochure published months prior might incorrectly state that applications are open until the official deadline. However, if that popular program has already reached its capacity and closed early, the chatbot provides false hope and misleads the student into wasting time preparing an invalid application. An accurate response requires a real-time query to the live admissions database to check the program's current status. This limitation makes static document retrieval unreliable for the critical, time-sensitive needs of prospective students.

To address these specific failures, this research introduces RMUTTOBot, a domain-specific chatbot built on our novel table-augmented generation (TAG) approach. Unlike heavyweight systems that generate complex SQL or reason over graphs, our lightweight TAG-based RAG operates on a hybrid model. It retrieves foundational knowledge from a database of QA pairs and uses LLM-native function calling to trigger simple, pre-written queries for specific, real-time data. Our TAG-based RAG approach is an evolution of this RAG framework, specifically adapting it for the structured and dynamic data environment of a university, thereby addressing a key gap in current methodologies.

1.1 Related work

This section reviews the evolution of university chatbots to contextualize the contribution of our TAG-based RAG approach. The narrative traces the progression from traditional NLP systems to the current state-of-the-art in LLM-powered RAG, highlighting the persistent challenges that motivate our research.

Early university chatbots relied on traditional NLP techniques like intent recognition and entity extraction to handle user queries. Systems developed for Petrozavodsk State University (Shchegoleva *et al*., 2021) and Universitas Stikubank, which used the RASA framework (Hadiono *et al*., 2024), employed rule-based dialogue management to address frequently asked questions. While functional, these systems were limited in flexibility; as Pothuri (2024) notes, rule-based intent recognition struggles with linguistically diverse queries and often fails to maintain coherent multi-turn dialogue. The advent of LLMs marked a significant paradigm shift, offering more robust natural language understanding and adaptability that overcame these constraints (Karanikolas *et al*., 2025).

The predominant architecture in this new era is LLM-powered RAG, which enhances LLMs with external, domain-specific context. A common strategy in educational settings, as explored by Alsafari *et al*. (2024), is to build a knowledge base from unstructured documents like course materials, websites, and student handbooks. This approach is exemplified by systems like JayBot at Johns Hopkins University (Odede and Frommholz, 2024) and a similar chatbot at the University of Mosul (Sharief and Ersayyem, 2024). While these systems demonstrate increased flexibility, studies consistently highlight a critical shared vulnerability: their accuracy is entirely dependent on the currency of the static documents in their knowledge base, and they lack automated pipelines for ingesting new data. To address precision issues, some researchers have developed hybrid retrieval methods. URAG, for instance, combines semantic vector search with keyword-based search to retrieve precise terms (Nguyen and Quan, 2025), while a system at XJTLU integrated TF-IDF for the same purpose (Xu and Liu, 2024). However, these innovations still operate on text and inherit its fundamental limitations, particularly an inability to address personalized or confidential queries requiring database lookups.

Recognizing the limitations of text-only retrieval for factual precision, recent research has focused on adapting RAG for structured and tabular data. For instance, frameworks like Binder demonstrate how LLMs can be bound to symbolic languages, enabling them to execute SQL queries directly against relational databases for high-fidelity data retrieval (Cheng *et al*., 2023). Other approaches, such as StructGPT, construct knowledge graphs from structured data to perform complex, multi-hop reasoning across interconnected information (Jiang *et al*., 2023).

While these structured-data-aware architectures significantly improve factual accuracy, they introduce trade-offs in complexity, computational cost, and real-time adaptability that limit their practicality for chatbots. The complex pipelines required for on-the-fly text-to-SQL generation (Cheng *et al.*, 2023), dynamic graph construction (Jiang *et al.*, 2023), or iterative multi-stage retrieval are often computationally intensive and require significant engineering effort. They can also rely on static data representations, making them ill-suited for the dynamic university environment where information like course availability and admission statuses change frequently.

This leaves a critical gap for a system that can achieve the factual reliability of structured-data reasoning without the high overhead of these complex frameworks. Our research addresses this gap by proposing RMUTTOBot, a lightweight TAG-based RAG system that leverages LLM-native function calling. Instead of tasking the model with generating complex SQL or reasoning over a graph, our approach uses the LLM as an intelligent router to trigger pre-defined, highly-optimized database queries. This method provides a practical, scalable, and accurate solution tailored to the specific needs of university information systems by achieving factual reliability without the high engineering and computational overhead of more complex reasoning frameworks.

The structure of this paper is organized as follows. Section 2 describes the methodology of the study, including the design and components of the RMUTTOBot framework, as well as the evaluation strategies adopted to assess its effectiveness. Section 3 presents the results and discussion, incorporating both quantitative evaluations using automated metrics and qualitative assessments based on human judgment. Section 4 concludes the paper by summarizing key findings and suggesting directions for future research.

2. Materials and Methods

2.1 RMUTTOBot framework

The RMUTTOBot system utilizes a lightweight TAG-based RAG approach built on a hybrid, dual-source retrieval strategy. This architecture is designed to ensure both comprehensive context and real-time factual accuracy by distinguishing between two distinct but interconnected information sources: a knowledge base and a direct interface to live institutional databases.

The foundation of the system is a knowledge base of over 600 pre-verified QA pairs, which contains general, foundational information about the university. Critically, this is not a static repository. The QA pairs are stored in a database and are managed through an application that allows administrators to easily add, update, or remove questions and answers. This ensures the chatbot's foundational knowledge remains current without needing to retrain or re-index complex documents. For highly dynamic or volatile information, the system employs a dynamic data interface using LLM-native function calling. This allows the chatbot to query live institutional databases for real-time data such as course availability, admission statuses, or specific tuition fees.

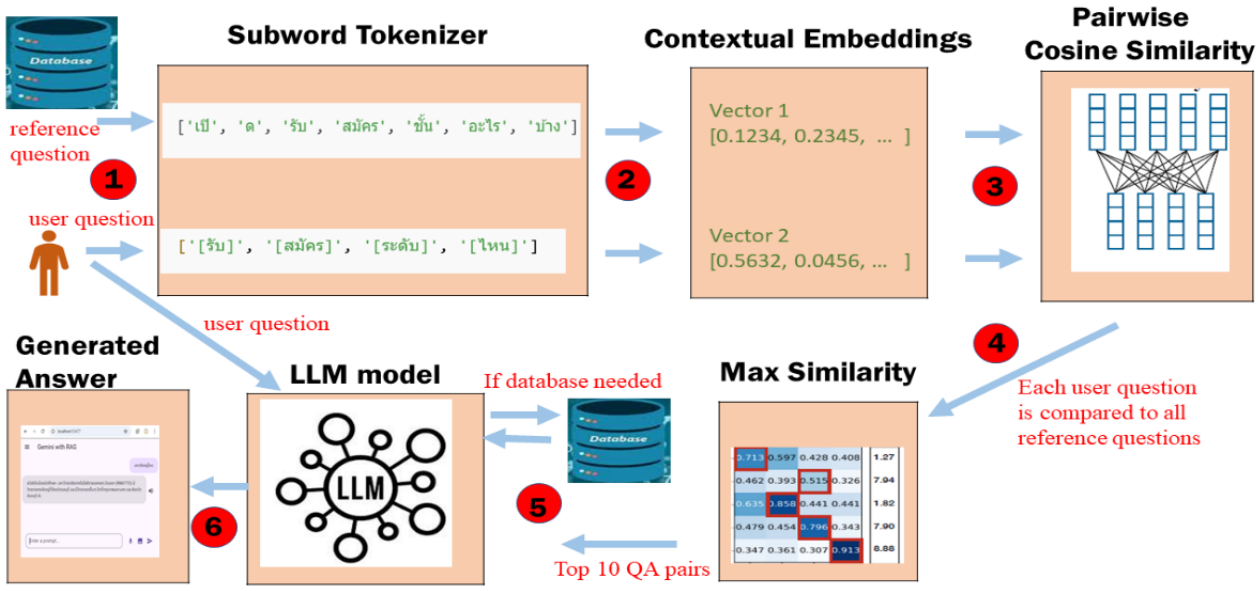


Figure 1 Over all Framework of the Proposed Methodology

The proposed methodology is comprised of four primary components: data preparation and tool definition, hybrid retrieval and query dispatch, context consolidation and prompt engineering, and response generation.

Step 1 Data Preparation and Tool Definition: The development process begins with the creation of the foundational knowledge base. Using the GPT-4 Turbo model, approximately 600 QA pairs were automatically generated. This state-of-the-art model was selected to ensure both scale and quality, enabling the rapid production of a comprehensive set of questions covering a wide range of university-specific topics and diverse linguistic phrasings for common inquiries. This automated generation provided a high-quality baseline for the initial knowledge base.

To enable the LLM to trigger database-related functions, we deliberately modified approximately 20% of the initially generated QA pairs. With a modified query such as "What is the tuition fee for computer engineering?", the LLM is designed to call the get_tuition_fee (major_name) function, which then retrieves tuition fee of computer engineering from the institutional database. This deliberate inclusion of function-triggering queries ensures that the system can effectively handle both knowledge-based and dynamic data-driven interactions, supporting real-time information retrieval through predefined database functions.

To ensure reliability and factual consistency, each generated QA pair underwent manual review by a team of three domain experts. These experts meticulously cross-checked every answer against verified institutional sources, including the university website, official databases, student handbooks, and admissions brochures. This rigorous verification process was designed to identify and correct hallucinated or inaccurate information, thereby guaranteeing alignment with official institutional data.

Following validation, the all QA pairs were stored in a Firebase Firestore database, forming the system's primary knowledge repository. In parallel, a set of function tools was defined for the LLM to support real-time data retrieval. These tools correspond to high-value database lookups, which allow the system to access dynamic institutional database as needed.

Step 2 Hybrid Retrieval and Query Dispatch: As shown in Figure 1, the retrieval process is initiated when a user question is submitted through the chatbot interface (① Figure 1). The system concurrently retrieves reference questions from the QA database and processes both inputs using a Subword Tokenizer (②), which decomposes text into subword units to accommodate linguistic variability across languages.

Next, the tokenized sequences are encoded into contextual embeddings (③) using the bert-base-multilingual-cased model. This model was selected over alternative embedding methods such as Sentence-BERT, word2vec, or fastText due to its strong multilingual capability including full support for the Thai language contextual sensitivity, and robustness across semantically diverse text corpora. Unlike static word embeddings, bert-base-multilingual-cased generates context-aware vector representations that maintain semantic meaning even in paraphrased or linguistically complex queries. Its multilingual pretraining ensures effective cross-lingual generalization, allowing the chatbot to process queries in both English and Thai, which are commonly used in the university context.

These embeddings capture the semantic meaning of each question, enabling robust cross-lingual matching. The system then calculates pairwise cosine similarity between the user question vectors and the stored reference vectors (④), facilitating semantic retrieval of the most relevant QA pairs.

The top 10 QA pairs with the highest similarity scores are selected (⑤). This threshold was empirically validated in a pilot evaluation of 50 paraphrased queries, where the correct reference consistently appeared within the top ten retrieved results. The chosen limit optimizes both accuracy and computational efficiency, ensuring a high signal-to-noise ratio, minimizing token length for prompt construction, and improving response latency.

Concurrently, the Gemini 1.5 Flash model serves as a reasoning engine that inspects the user query to determine whether real-time data retrieval is required. This model was selected over alternative LLMs due to its exceptional balance between reasoning performance, inference speed, and cost efficiency, making it well suited for real-time chatbot applications. Compared with larger or slower models such as Gemini Pro or GPT-4 Turbo, Gemini 1.5 Flash demonstrates significantly lower latency while maintaining competitive accuracy in structured reasoning and function-calling tasks. Its optimized architecture allows the system to process queries rapidly without compromising contextual understanding an essential factor in delivering responsive and reliable user interactions within the university chatbot environment.

If the query matches a predefined function tool (e.g., tuition fees or application deadlines), the system triggers the corresponding database function (⑤). Otherwise, the retrieved QA pairs proceed directly to the next stage (Step 3).

Step 3 Context Consolidation and Prompt Engineering: As shown in Figure 1, The system aggregates all relevant contextual information, consisting of the top ten retrieved QA pairs and, when applicable, real-time database outputs, into a structured input for the LLM. (⑤). This process ensures factual accuracy and prevents the generation of unsupported or outdated information.

The retrieved content and contextual information are consolidated into a structured prompt. This prompt design provides the model with rich contextual grounding while maintaining efficiency in token usage. In addition, system-level instructions and fallback directives are embedded within the prompt to control the model's reasoning behavior, ensuring consistent tone, factuality, and adherence to institutional guidelines. The

completed prompt is then forwarded to the response generation module for synthesis (⑥).

Step 4 Response Generation and Delivery: As depicted in Figure 1, the final stage involves submitting the structured prompt to the Gemini 1.5 Flash model for natural language synthesis (⑥). The model generates a coherent, contextually grounded, and factually reliable response that draws from both the QA knowledge base and any dynamic data obtained through database queries. The output is then delivered through the Flutter-based chatbot interface, providing the user with a seamless and interactive experience.

The flowchart illustrating from query input to response delivery is presented in Figure 2.

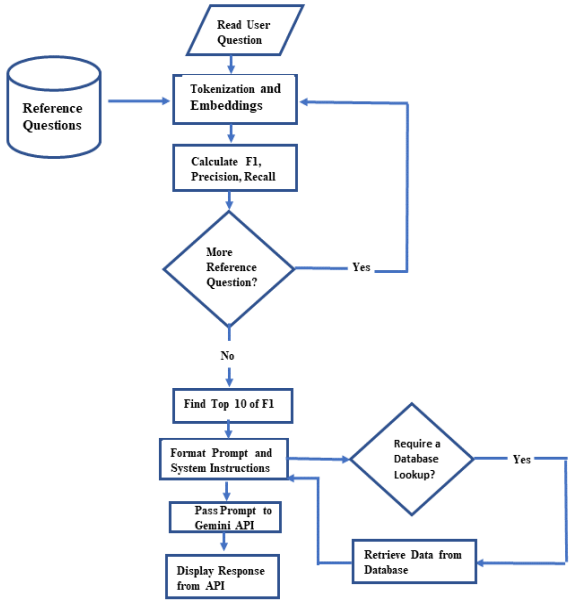


Figure 2 RMUTTOBot flow chart

2.2 Evaluation strategies

To ensure a fair and reproducible assessment of the proposed chatbot system (RMUTTOBot), a structured evaluation procedure was implemented. The evaluation was conducted in two phases automated testing and human evaluation under identical operating conditions. The chatbot was deployed on a controlled environment using the Gemini 1.5 Flash API integrated with a Firebase Firestore knowledge base. All evaluations were performed through the same user interface to eliminate bias introduced by differing platforms or communication modes.

Evaluating LLMs–based chatbots within a RAG framework requires a multidimensional approach to ensure both semantic fidelity and practical usability. Six key criteria provide a comprehensive foundation for assessing such systems: semantic similarity, retrieval effectiveness, response relevance and accuracy, linguistic fluency and coherence, content coverage and completeness, and robustness and consistency (Abeysinghe and Circi, 2024; Joshi, 2025; Oro *et al*., 2024).

Semantic similarity measures how closely the chatbot's responses preserve the intended meaning of the reference answers in the knowledge base. Retrieval effectiveness assesses the informativeness and relevance of the top-N retrieved texts that support answer generation an essential factor influencing factual precision. Response relevance and accuracy determine whether the generated output is factually correct and contextually aligned with the retrieved materials. Linguistic fluency and coherence capture grammaticality, readability, and logical flow, while content coverage and completeness gauge whether the system's output thoroughly addresses user questions using available information. Finally, robustness and consistency evaluate the chatbot's ability to generate stable, logically consistent answers across similar or ambiguous queries.

The following two assessment strategies, automated metrics and human evaluation, were used to address these criteria:

1. Automated metrics evaluation: The criteria of semantic similarity, retrieval effectiveness, and robustness and consistency were evaluated using automated metrics.

For the evaluation of semantic similarity and robustness and consistency, a representative of 200 reference questions was randomly selected from a pool of 600 QA entries in RMUTTOBot's structured QA database. Furthermore, to facilitate a thorough assessment of dynamic information retrieval, a 20% subset of these questions (40 in total) was specifically selected to activate the LLM's pre-defined database queries. To simulate realistic variations in user input, each selected question was paraphrased using GPT-4 Turbo, introducing lexical and syntactic diversity while preserving semantic intent. Each paraphrased query was then processed under three system configurations for comparison: (1) baseline LLM-only (Without RAG): The chatbot relied solely on its base language model without external data augmentation. This configuration, implemented using the Flutter framework, served as the baseline. (2) TAG-based RAG: This configuration employs a hybrid dual-source retrieval strategy, leveraging structured knowledge base and LLM-native function calling. It was implemented using the Flutter framework. (3) PDF-based RAG: In this configuration, retrieval relied exclusively on external PDF documents. The ChatDOC platform was used for evaluation due to its integration of retrieval-augmented generation (RAG) with high-quality semantic embeddings, citation mapping, and support for multi-document queries. ChatDOC's research-oriented design enables more reliable academic information retrieval compared to conventional PDF-QA tools.

Performance was measured using BERTScore, which evaluates semantic equivalence by comparing contextual embeddings (Zhang *et al*., 2019). This metric was chosen over traditional lexical overlap metrics like BLEU, METEOR, and ROUGE, which primarily evaluate text quality by counting the number of shared words and phrases (n-grams) between a

candidate and a reference sentence. A key limitation of these lexical methods is their inability to recognize synonyms or paraphrasing. BERTScore produces three key measures: BERTPrecision, BERTRecall, and BERTF1, computed as shown in Equations 1–4.

$$BERTPrecision = \frac{1}{m}\sum_{i=1}^{m} \max_{j} sim(c_i, r_j) \quad (1)$$

$$BERTRecall = \frac{1}{n}\sum_{j=1}^{n} \max_{i} sim(c_i, r_j) \quad (2)$$

$$BERTF1 = \frac{2PR}{P+R} \quad (3)$$

$$sim(c_i, r_j) = \cos\left(\phi(c_i), \phi(r_j)\right) \quad (4)$$

where c = [c1,c2,…,cm] = tokens of the candidate sentence
r = [r1,r2,…,rn] = tokens of the reference sentence
ϕ (c_i) and ϕ (r_j) = BERT embeddings of tokens c_i and r_j, respectively
$sim(c_i, r_j)$ = cosine similarity between two vectors ϕ (c_i) and ϕ (r_j)

For the evaluation of retrieval effectiveness, a test set of 50 paraphrased user queries was used to validate the retrieval stage of the TAG-based RAG framework. The system retrieved the top k most semantically similar reference questions for each query. Performance was measured using Top-k Accuracy and Mean Reciprocal Rank (MRR).

Top-k Accuracy measures the proportion of queries for which the correct reference question appears within the top-k retrieved candidates, defined in Equations 5:

$$Top - k\ Accuracy = \frac{\#\ correct\ results\ within\ top\ k}{N} \quad (5)$$

where N = the total number of test queries
k = the number of results being considered.

MRR evaluates how highly the correct reference question is ranked, computed in Equations 6:

$$MRR = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{rank_i} \quad (6)$$

where $rank_i$ denotes the position of the correct reference question for the i-th query.

2. Human evaluation: The criteria of response relevance and accuracy, linguistic fluency and coherence, and content coverage and completeness were assessed through human evaluation. For this task, 25 high-school students from Surasakmontree school were recruited as raters. Each evaluator assessed 25 paraphrased queries, comparing the responses generated by the TAG-based RAG (RMUTTOBot Flutter Framework) and PDF-based RAG (ChatDOC) systems against reference answers.

Responses were rated using a 5-point Likert scale, with detailed scoring rubrics provided in Tables 1–3

Table 1 Detailed scoring procedures for response relevance and accuracy criterion.

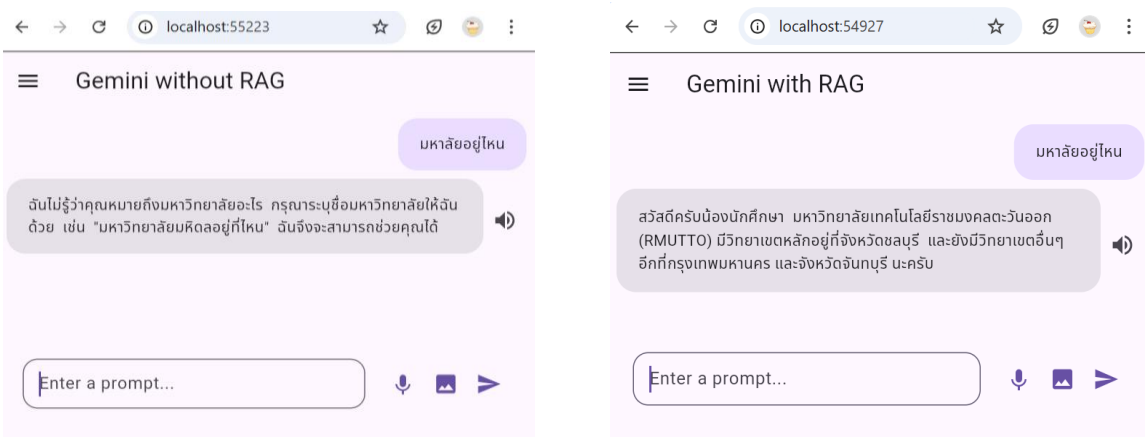
Score	Description
1	The answer is completely irrelevant to the question and factually incorrect
2	The answer is mostly off-topic or contains significant factual inaccuracies
3	The answer is somewhat related to the question and includes some correct information, but it also contains notable errors or omissions
4	The answer is generally relevant and factually accurate, with only minor issues or inaccuracies
5	The answer is highly relevant and entirely accurate. It fully addresses the user's question using correct and well-grounded information

Table 2 Detailed scoring procedures for linguistic fluency and coherence criterion.

Score	Description
1	Poor grammar; hard to understand
2	Some grammar issues; awkward phrasing affects clarity
3	Understandable but contains some unnatural or disjointed language
4	Mostly fluent with minor readability issue
5	Clear, coherent, grammatically correct, and easy to read

Table 3 Detailed scoring procedures for content coverage and completeness criterion.

Score	Description
1	Completely incomplete; no meaningful content
2	Covers only a small part; key details missing
3	Partially complete; addresses the main point but lacks supporting information
4	Mostly complete; minor omissions
5	Fully complete; all parts of the question are answered thoroughly



(a) without RAG (baseline LLM-only) (b) TAG-based RAG

Figure 3 Comparision Chatbot Responses: Without RAG vs. TAG-based RAG Approach via a Flutter Application

To ensure credibility and minimize subjective bias, the central tendency measures median and mode were computed for each criterion to capture consensus tendencies, while Krippendorff's Alpha (α) was calculated to assess inter-rater reliability, ensuring consistency and reliability across evaluators. Reliability interpretations followed established conventions: $\alpha \geq 0.80$ = strong agreement, $0.67 \leq \alpha < 0.80$ = acceptable agreement, and $\alpha < 0.67$ = low agreement (Krippendorff, 2011).

By integrating both automated and human evaluations, this mixed-method approach ensures breadth, depth, and reproducibility.

3. Results and Discussion

3.1 Evaluation of semantic similarity, robustness and consistency

To evaluate across two core criteria: semantic similarity, and robustness and consistency, the three system configurations Without RAG, PDF-based RAG, and TAG-based RAG were tested against a set of 200 paraphrased user queries. Qualitative comparisons revealed significant performance differences between these configurations. As illustrated in Figures 3 and 4, the without RAG baseline configuration (Figure 3a) failed to provide a substantive answer to the query "Where is the university located?". In contrast, both TAG-based RAG (Figure 3b) and PDF-based RAG (Figure 4) configurations successfully retrieved accurate information, demonstrating the core benefit of RAG. Further, Figure 5 highlights a key distinction in retrieval precision: when asked "How much is the tuition for Agricultural Engineering?", the PDF-based RAG (Figure 5a) produced an incorrect value of 12,700, despite the correct figure 13,500 being present in the PDF file. Conversely, the TAG-based RAG (Figure 5b) returned the correct answer, indicating superior retrieval precision.

The chatbot's responses across all three configurations were benchmarked against the reference answers using BERTScore. As shown in Table 4 and visualized in Figure 6, the integration of RAG markedly improved performance across all metrics. The baseline (without RAG) yielded the lowest scores. While the PDF-based RAG approach showed clear improvement, the TAG-based RAG approach demonstrated the most significant performance gains, achieving a BERTF1 score of 0.7361. The relative improvements of 13.44% in BERTPrecision, 11.92% in BERTRecall, and 12.76% in BERTF1 over the PDF-based approach underscore the advantage of structured data in facilitating more accurate and contextually relevant responses. These findings highlight that TAG-based RAG substantially enhances both semantic similarity and response robustness.

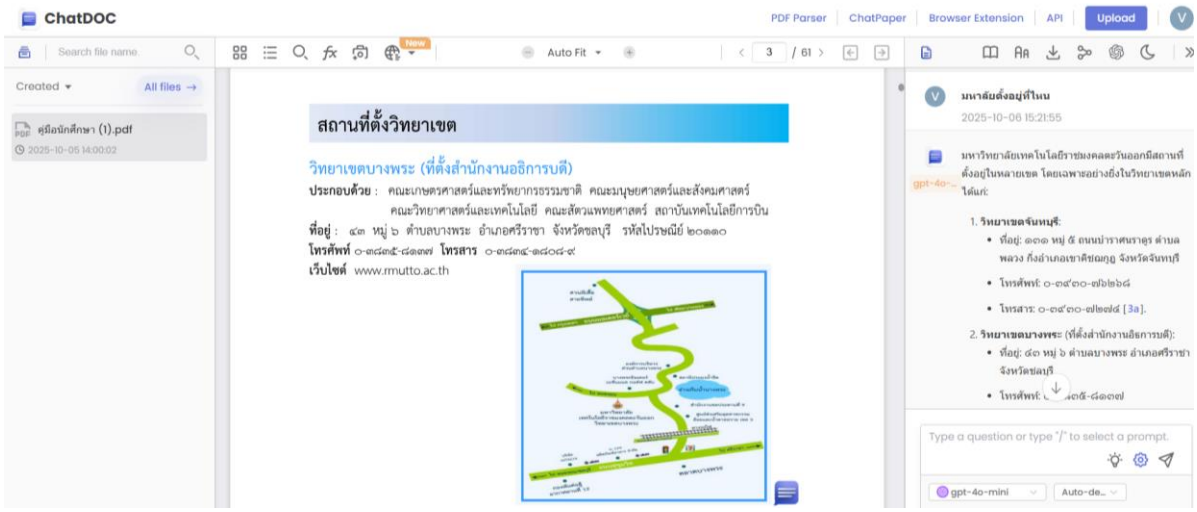
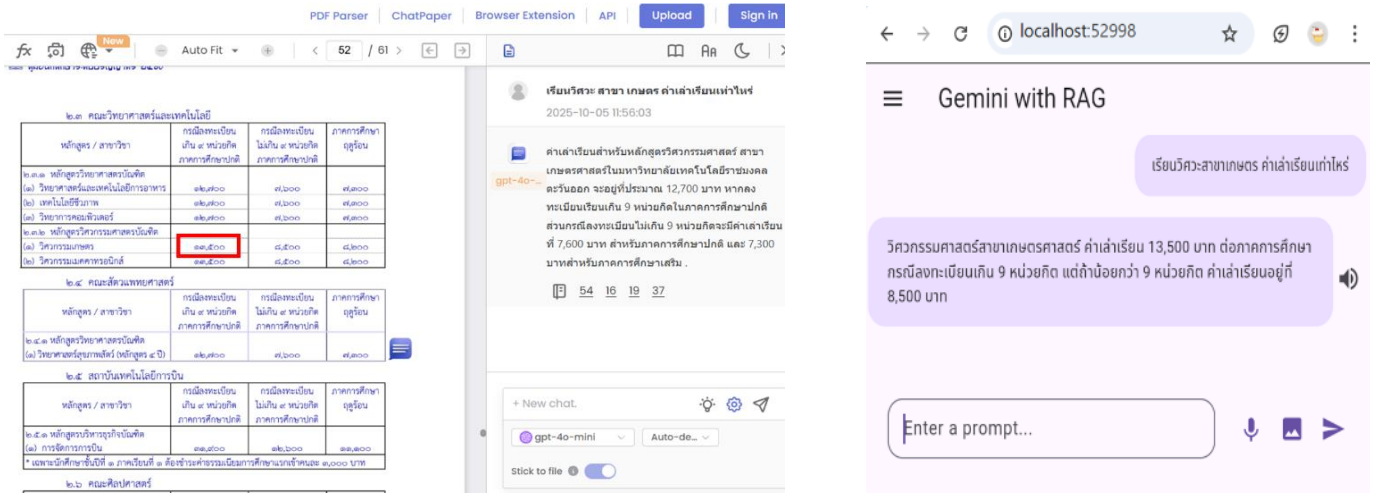


Figure 4 PDF-Based RAG Chatbot response via ChatDOC platform



(a) PDF-based RAG (b) TAG-based RAG

Figure 5 Comparision Chatbot Responses: PDF-based RAG vs. TAG-based RAG Approach

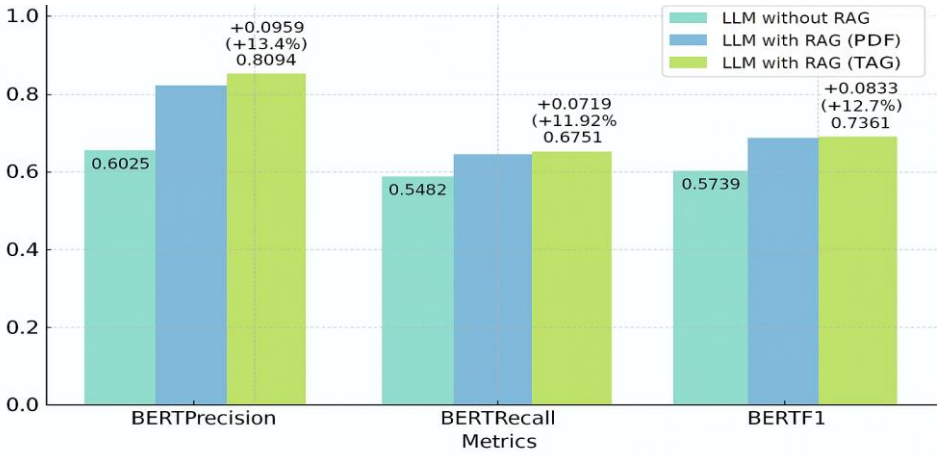


Figure 6 BERTScore Comparison: without RAG vs. PDF-based RAG vs. TAG-based RAG.

Table 4 BERTScore Comparison: without RAG vs. PDF-based RAG vs. TAG-based RAG.

Metric Score	without RAG	PDF-based RAG	TAG-based RAG	Difference TAG and PDF based	% Improvement TAG over PDF based
BERTPrecision	0.6025	0.7135	0.8094	+0.0959	+13.44%
BERTRecall	0.5482	0.6032	0.6751	+0.0719	+11.92%
BERTF1	0.5739	0.6528	0.7361	+0.0833	+12.76%

Table 5 Summary of Rank Distribution Across 50 Queries

Rank of Correct Reference Question	Number of Queries	Reciprocal Rank (1/rank)	Weighted Contribution
1	35	1.0000	35.0000
2	7	0.5000	3.5000
3	3	0.3333	1.0000
4	2	0.2500	0.5000
5	1	0.2000	0.2000
6	1	0.1667	0.1667
7-10	1	0.1250	0.1250
Total	**50**	**-**	**40.4917**

3.2 Evaluation of retrieval effectiveness

To evaluate TAG-based RAG retrieval performance, a test set of 50 paraphrased user queries was selected. The semantic similarity between each paraphrased query and all reference questions in the QA database was calculated using BERTScore. For each query, the system retrieved the top 10 most semantically similar reference questions, ranked by descending similarity scores. To provide a more comprehensive evaluation, two additional retrieval effectiveness metrics were applied: Top-10 Accuracy and MRR.

As presented in Table 5, the retrieval system achieved a Top-10 accuracy of 100%, indicating that the correct reference question was consistently identified within the top ten retrieved results for all 50 paraphrased queries. This result demonstrates the robustness of the semantic retrieval component in maintaining contextual equivalence despite variations in query phrasing.

Examination of the rank distribution in Table 5 shows that the correct reference question appeared at Rank 1 in 70% of cases and within the top four ranks in 94% of cases, reflecting strong ranking precision and stability. Only three queries (6%) were ranked between fifth and tenth positions, suggesting minimal retrieval deviation. The mean reciprocal rank (MRR) was 0.8098 (40.4917/50 Table 5), indicating that, on average, the correct match appeared within the top two retrieved results. Overall, these findings confirm the high reliability and semantic discrimination capability of the retrieval mechanism, supporting its effectiveness as a foundation for accurate and contextually grounded response generation within the TAG-based RAG framework.

Two representative examples are presented to illustrate retrieval behavior. Table 6 displays the results for the query "What are the available admission channels?". The corresponding reference question, "Which channels can be used to apply for admission?", was retrieved at rank 6 with a BERTF1 score of 0.8099. Although several higher-ranked items were not exact matches, they remained topically related (e.g., required documents, admission rounds), indicating that the retrieval model effectively grouped semantically coherent content.

Similarly, Table 7 presents the retrieval results for the query "Which educational programs are open for admission?". The correct reference question, "What levels of study are open for admission?", was ranked first, achieving a BERTF1 score of 0.7835, reflecting optimal semantic alignment.

To compare the retrieval effectiveness of the PDF-based and TAG-based RAG systems, both configurations were evaluated using the same set of 50 paraphrased user queries and BERTScore metrics, as shown in Table 8. Across all measures BERTPrecision, BERTRecall, and BERTF1 the TAG-based configuration outperformed the PDF-based one.

BERTPrecision improved from 0.7835 to 0.8272 (+5.58%), indicating that TAG retrieves more semantically aligned context. BERTRecall showed a modest increase of 0.57%, while BERTF1 rose from 0.7326 to 0.7534 (+2.85%), reflecting balanced gains in both precision and recall.

These results collectively demonstrate that integrating structured data within the TAG-based RAG framework leads to more precise and semantically consistent retrieval compared to traditional unstructured PDF-based methods. The findings reinforce that TAG-based RAG enhances contextual relevance and retrieval robustness key factors that contribute to downstream improvements in the accuracy and reliability of the chatbot's generated responses.

Table 6 Top 10 Retrieval Results for Query “What are the available admission channels?”

Rank	Reference Question	BERTF1
1	การสมัครเรียนต้องใช้เอกสารอะไรบ้าง	0.8675
2	เอกสารที่ต้องใช้ในการสมัครมีอะไรบ้าง	0.8372
3	ทุนเรียนดีมีอะไรบ้าง	0.8323
4	เปิดรับสมัครชันอะไรบ้าง	0.8243
5	รอบการรับสมัครมีช่วงไหนบ้าง	0.8157
6	**สามารถสมัครเรียนผ่านช่องทางไหนได้บ้าง**	**0.8099**
7	ทุนการศึกษาต้องใช้เอกสารอะไรบ้าง	0.8088
8	ต้องใช้คะแนนสอบอะไรบ้างในการสมัคร	0.7969
9	มหาวิทยาลัยมีกิจกรรมทางวิชาการอะไรบ้าง	0.7943
10	ใช้คะแนนสอบอะไรบ้าง	0.7926

Table 7 Top 10 Retrieval Results for Query “Which Educational Programs Are Open for Admission?”

Rank	Reference Question	BERTF1
1	**เปิดรับสมัครชันอะไรบ้าง**	**0.7835**
2	รอบการรับสมัครมีช่วงไหนบ้าง	0.7824
3	มีกำหนดการรับสมัครช่วงไหนบ้าง	0.7780
4	สามารถสมัครเรียนผ่านช่องทางไหนได้บ้าง	0.7645
5	นักศึกษาสามารถฝึกงานได้ที่ไหนบ้าง	0.7615
6	การสมัครเรียนต้องใช้เอกสารอะไรบ้าง	0.7549
7	มีทุนการศึกษาสำหรับนักศึกษาใหม่หรือไม่	0.7527
8	นักศึกษาสามารถเข้าร่วมโครงการสังคมหรือจิตอาสาได้ไหม	0.7483
9	นักศึกษาสามารถฝึกงานได้ที่ไหนบ้าง	0.7459
10	ทุนการศึกษาต้องใช้เอกสารอะไรบ้าง	0.7441

Table 8 BERTScore Retrieval Effectiveness Comparison: PDF-based RAG vs. TAG-based RAG.

Metric Score	PDF-based RAG	TAG-based RAG	Difference TAG and PDF based	% Improvement TAG over PDF based
BERTPrecision	0.7835	0.8272	+0.0437	+5.58%
BERTRecall	0. 6883	0.6922	+0.0039	+0.57%
BERTF1	0.7326	0.7534	+0.0209	+2.85%

3.3 Evaluation of response relevance and accuracy, linguistic fluency and coherence, content coverage and completeness

A human evaluation was conducted to compare the qualitative performance of the two RAG-based systems PDF-based RAG and TAG-based RAG across three core criteria: response relevance and accuracy, linguistic fluency and coherence, and content coverage and completeness.

A total of 25 high school students participated as independent evaluators. Each rater assessed 25 representative queries for both systems, resulting in 625 ratings per evaluation. For each query, paraphrased inputs were submitted to both the RMUTTOBot (TAG-based RAG) and ChatDOC (PDF-based RAG) systems. The generated responses were compared against the reference answers using a 5-point Likert scale, where higher scores indicated stronger performance.

Table 9 summarizes the quantitative results and corresponding qualitative interpretations.

The evaluation results demonstrate that both systems achieved high overall performance, with almost perfect inter-rater reliability (Krippendorff's $\alpha \geq 0.90$) for most criteria, underscoring the robustness and objectivity of the findings.

For response relevance and accuracy, both RAG configurations were rated at the highest level, achieving median and mode scores of 5, corresponding to “highly relevant and entirely accurate” answers. This indicates that both systems were equally effective in retrieving and presenting correct, on-topic responses aligned with the reference answers.

In terms of linguistic fluency and coherence, both systems maintained strong performance. The median score of 4 (“mostly fluent with minor readability issues”) alongside a mode of 5 (“clear and coherent”) suggests that while the generated responses were generally well-structured and readable, occasional phrasing or grammatical inconsistencies were noted.

Table 9 Summary of Human Evaluation Results across Three Criteria

Criterion	System	Median	Mode	Krippendorff's Alpha	Key Finding
response relevance and accuracy	PDF-based RAG	5	5	0.912 (Almost Perfect)	Excellent performance with very high rater agreement.
	TAG-based RAG	5	5	0.904 (Almost Perfect)	Excellent performance, comparable to the PDF system.
linguistic fluency and coherence	PDF-based RAG	4	5	0.908 (Almost Perfect)	High fluency, though not consistently rated as perfect.
	TAG-based RAG	4	5	0.912 (Almost Perfect	High fluency, with performance similar to the PDF system.
content coverage and completeness	PDF-based RAG	4	4	0.904 (Almost Perfect)	Good performance, but responses were typically rated 'mostly' complete.
	TAG-based RAG	5	5	0.835 (Substantial)	Excellent performance, outperforming the PDF system on this criterion.

A more distinct difference emerged in content coverage and completeness. The TAG-based RAG system exhibited superior performance, achieving median and mode scores of 5, indicating that its responses were consistently judged as “fully complete.” By contrast, the PDF-based RAG configuration achieved median and mode scores of 4, suggesting that its responses were “mostly complete” but occasionally omitted supplementary details.

Overall, these findings affirm that while both systems perform strongly across relevance, fluency, and completeness, the TAG-based RAG approach offers a notable advantage in delivering comprehensive, contextually enriched, and factually complete answers demonstrating the value of structured data integration in enhancing RAG system.

3.4 Overall Discussion

Collectively, the evaluation results demonstrate the clear superiority of the TAG-based RAG framework. The strong performance in retrieval effectiveness (Section 3.2), evidenced by a high MRR of 0.8098 and 100% Top-10 accuracy, serves as the foundation for the system's overall success. This highly precise retrieval of structured information directly contributed to the significant improvements in semantic similarity and robustness, as measured by BERTScore (Section 3.1), where the TAG-based RAG approach outperformed both the baseline and the PDF-based RAG. This quantitative superiority was corroborated by the human evaluation (Section 3.3). While both RAG systems were perceived as accurate and fluent, the ability of the TAG-based RAG to consistently provide "fully complete" answers highlights its key advantage. The structured nature of the TAG-based RAG ensures that all relevant data points are retrieved and synthesized, preventing the information omissions sometimes observed in the PDF-based RAG. This synthesis of automated and human evaluations confirms that integrating a robust, TAG-based RAG mechanism with a powerful LLM creates a system that is not only semantically aligned but also factually precise, comprehensive, and reliable

4. Conclusion

This research presented the design, development, and evaluation of RMUTTOBot, integrating a TAG-based RAG framework that employs dynamic data interfaces and LLM-native function calling to enable real-time retrieval from institutional databases. The performance of RMUTTOBot was rigorously assessed using a combination of automated metrics and human evaluations across six criteria: semantic similarity, robustness and consistency, retrieval effectiveness, response relevance and accuracy, linguistic fluency and coherence, and content coverage and completeness.

The experimental findings demonstrate that RAG substantially improves the quality and reliability of chatbot responses compared with a baseline language model operating without retrieval support. Quantitative results based on BERTScore metrics indicated that the TAG-based RAG configuration achieved the highest overall performance, with increases of 13.44%, 11.92%, and 12.76% in BERTPrecision, BERTRecall, and BERTF1, respectively, compared to the PDF-based RAG approach. The semantic retrieval evaluation further confirmed that the TAG-based RAG framework consistently identified semantically equivalent questions, successfully locating the correct reference question within the top 10 retrieved results for all test cases.

Human evaluation results corroborated these findings. Both RAG configurations demonstrated strong performance with high inter-rater reliability ($\alpha > 0.835$). While the PDF-based and TAG-based RAG systems performed comparably in response relevance and linguistic fluency, the TAG-based model outperformed in content coverage and completeness, reflecting its ability to dynamically access structured data for more

comprehensive answers. These results collectively validate the effectiveness of the TAG-based RAG framework in improving retrieval precision, contextual grounding, and overall response quality in chatbot systems.

Although the findings confirm the efficacy of the proposed framework, several directions remain for future research. First, scaling and generalization should be explored by extending the TAG-based RAG framework to other institutional domains and integrating multilingual support to enhance accessibility. Second, improvements in retrieval ranking algorithms and context filtering could further optimize performance, reducing redundancy and ensuring that only the most relevant information is provided to the LLM. Third, incorporating user interaction analytics and reinforcement learning from human feedback (RLHF) may enable adaptive refinement of chatbot responses based on real usage patterns. Finally, future iterations of RMUTTOBot could leverage knowledge graphs or structured ontologies to enable more interpretable, explainable, and semantically grounded responses.

Overall, the study contributes a practical, empirically validated framework for developing domain-specific chatbots. The proposed TAG-based RAG approach demonstrates that combining structured data augmentation with dynamic retrieval significantly enhances both the accuracy and completeness of LLM-generated responses an important step toward more intelligent, reliable, and context-aware conversational systems for academic institutions.

5. Acknowledgements

The authors would like to express sincere gratitude to Rajamangala University of Technology Tawan-ok (RMUTTO) for providing valuable data and institutional support used in the development and evaluation of the chatbot system for university admission services. This research would not have been possible without their contributions.

6. References

Abeysinghe, B. and Circi, R. 2024. **The challenges of evaluating LLM applications: An analy sis of automated, human, and LLM-based approaches.** Computation and Language. Available Source: https://doi.org/10.48550/arXiv.2406.03339, October 1, 2025.

Alkishri, W., Yousif, J.H., Al Husaini, Y.N. and Al-Bahri, M. 2025. Conversational AI in Education: A General Review of Chatbot Technologies and Challenges. **Journal of Logistics, Informatics and Service Science** 12(3): 264-282.

Alsafari, B., Atwell, E., Walker, A. and Callaghan, M. 2024. Towards effective teaching assistants: From intent-based chatbots to LLM-powered teaching assistants. **Natural Language Processing Journal** 8: 100-101.

Arslan, M., Ghanem, H., Munawar, S. and Cruz, C. 2024. A survey on RAG with LLMs. **Procedia Computer Science** 246: 3781-3790.

Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z. and Abdelrazek, M. 2024. Seven failure points when engineering a retrieval augmented generation system, pp. 194-199. *In* **Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering – Software Engineering for AI (CAIN '24).** Association for Computing Machinery.

Budakoglu, G. and Emekci, H. 2025. Unveiling the power of large language models: A comparative study of retrieval-augmented generation, fine-tuning, and their synergistic fusion for enhanced performance. **IEEE Access** 13: 30936-30951.

Cheng, Z., Xie, T., Shi, P., Li, C., Nadkarni, R., Hu, Y., Xiong, C., Radev, D., Ostendorf, M., Zettlemoyer, L., Smith, N. and Yu, T. 2023. **Binding language models in symbolic languages.** Computation and Language. Available Source: https://doi.org/10.48550/arXiv.2210.02875, October 1, 2025.

Doumanas, D., Soularidis, A., Spiliotopoulos, D., Vassilakis, C. and Kotis, K. 2025. Fine-Tuning Large Language Models for Ontology Engineering: A Comparative Analysis of GPT-4 and Mistral. **Applied Sciences** 15(4): 2146.

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.S. and Li, Q. 2024. A survey on RAG meeting LLMs: Towards retrieval-augmented large language models, pp. 6491-6501. *In* **The 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24).** Association for Computing Machinery, United States.

Hadiono, K., Andreas, F., Supriyanto, A. and Irawan, S. 2024. Chatbots implementation for students admission. **Journal of Software Engineering and Simulation** 10: 44-52.

Jiang, J., Zhou, K., Dong, Z., Ye, K., Zhao, X. and Wen, J.R. 2023. StructGPT: A general framework for large language model to reason over structured data, pp. 9237- 9251. *In* **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing.** Association for Computational Linguistics.

Joshi, S. 2025. **Evaluation of large language models: Review of metrics, applications, and methodologies.** Artificial Intelligence and Machine Learning. Available Source: https://www.preprints.org/manuscript/202504.0369/v1, October 1, 2025.

Karanikolas, N.N., Manga, E., Samaridi, N., Stergiopoulos, V., Tousidou, E. and Vassilakopoulos, M. 2025. Strengths and Weaknesses of LLM-Based and Rule-Based NLP Technologies and Their Potential Synergies. **Electronics** 14(15): 3064.

Krippendorff, K. 2011. **Computing Krippendorff's alpha-reliability.** Computer Science. Available Source: https://www.researchgate.net/publication/260282682_Computing_Krippendorff's_Alpha-Reliability, October 1, 2025.

Liang, H., Zhou, Y. and Gurbani, V.K. 2025. Efficient and verifiable responses using retrieval augmented generation (RAG), pp. 1-6. *In* **The 4th International Conference on AI-ML Systems (AIMLSystems '24).** Association for Computing Machinery (ACM).

Nguyen, L.S.T. and Quan, T.T. 2025. URAG: Implementing a Unified Hybrid RAG for Precise Answers in University Admission Chatbots – A Case Study at HCMUT, pp. 82-93. *In* **Proceedings of the 2024 International Symposium Information and Communication Technology.** Springer.

Odede, J. and Frommholz, I. 2024. JayBot Aiding university students and admission with an LLM-based chatbot, pp. 391-395. *In* **Proceedings of the 2024 Conference on Human Information Interaction and Retrieval (CHIIR '24).** Association for Computing Machinery.

Oro, E., Granata, F.M., Lanza, A., Bachir, A., Grandis, L.D. and Ruffolo, M. 2024. **Evaluating Retrieval-Augmented Generation for Question Answering with Large Language Models.** Available Source: https://ceur-ws.org/Vol-3762/495.pdf, January 27, 2025.

Pothuri, V. 2024. Natural language processing and conversational AI. **International Research Journal of Modernization in Engineering Technology and Science** 6: 436-440.

Ren, R., Ma, J. and Zheng, Z. 2025. Large language model for interpreting research policy using adaptive two-stage retrieval augmented fine-tuning method. **Expert Systems with Applications** 278: 127330.

Sharief, B. and Ersayyem, Y. 2024. LLM and RAG powered chatbot for the College of Computer Science and Mathematics at the University of Mosul. **International Research Journal of Innovations in Engineering and Technology** 8: 59-61.

Shchegoleva, L., Burdin, G. and Attia, A. 2021. Chatbot for Applicants on University Admission Issues, pp. 491-494. *In* **Proceedings of the 29th Conference of Open Innovations Association.** FRUCT.

Uhm, M., Kim, J., Ahn, S., Jeong, H. and Kim, H. 2025. Effectiveness of retrieval augmented generation-based large language models for generating construction safety information. **Automation in Construction** 170: 105926.

Wan, Y., Chen, Z., Liu, Y., Chen, C. and Packianather, M. 2025. Empowering LLMs by hybrid retrieval-augmented generation for domain-centric Q&A in smart manufacturing. **Advanced Engineering Informatics** 65(B): 103212.

Xu, L. and Liu, J. 2024. A chat bot for enrollment of Xi'an Jiaotong-Liverpool University based on RAG, pp. 125-129. *In* **Proceedings of the 2024 International Workshop on Control Engineering and Advanced Algorithms (IWCEAA).** IEEE.

Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q. and Artzi, Y. 2019. BERTScore: Evaluating text generation with BERT. *In* **Proceedings of the International Conference on Learning Representations.** ICLR.