Optimizing Retrieval-Augmented Generation through Agentic RAG Ecosystem Based on Fine-Tuned BERT Cross Encoder and GPT-4 Model

Arya Jayavardhana[1], Faustine Ilone Hadinata[2], and Samuel Ady Sanjaya[3]

[1,2,3]Universitas Multimedia Nusantara, Tangerang Indonesia
samuel.ady@umn.ac.id

Abstract. Education plays a fundamental role in personal and professional growth, yet many students struggle with selecting the right major due to insufficient guidance, leading to dissatisfaction in their academic and career paths.

To address this, we propose an Agentic Retrieval-Augmented Generation (RAG) system that enhances chatbot-based academic advising by integrating a BERT-based agent to filter and validate retrieved information, ensuring contextually relevant and factually accurate responses. Additionally, GPT-4 is employed as the Natural Language Generation (NLG) component to produce fluent, structured answers.

Experimental results show that incorporating the agent significantly enhances response accuracy and relevance, where from 11 majors the METEOR Score resulted at 74.33%, Jaccard similarity at 58.77%, and Cosine similarity at 94.13%, improving by 6.54%, 5.57%, and 3.57%, respectively. The BERT Relevancy score remains consistently high at 96.91%. Deployment using Django is also implemented to allow real-use scenarios. Although promising, it is suggested that the next research involved a larger dataset consisting of tens of thousands of rows if possible to reduce bias and enable a more fine-tuned agent.

Keywords: Agent AI, BERT, Chatbot, GPT-4, Retrieval-Augmented Generation

1 Introduction

Education plays a vital role in shaping individuals, supporting social development, and driving economic progress, with higher education serving as a key stage in this journey. Universities not only provide access to knowledge and research opportunities but also help students build the skills and mindset needed for professional success [1][2]. A crucial part of this process is selecting an appropriate major, a decision that significantly affects both academic engagement and long-term career outcomes. Making an informed choice requires students to understand their own interests and strengths, as well as the structure and content of each study

program. While academic advisors are generally available to offer guidance [3], a considerable number of students and graduates continue to report dissatisfaction with their chosen fields [4]. According to a survey by Forbes, only 31% of respondents expressed satisfaction and enthusiasm in their current jobs, with many associating their dissatisfaction to having selected an unsuitable major during university [5]. This issue is often linked to limited access to relevant information or ineffective academic advising systems that fail to provide personalized and accurate support [6].

Advancements in artificial intelligence and machine learning have enabled the development of chatbots capable of assisting students by offering academic program guidance. These systems are intended to address limitations in traditional one-on-one counseling, which, while personal, is often constrained by limited availability, inconsistent quality, and a lack of standardized information delivery across large student populations [7][8][9][10][11]. Recent studies further highlight the transformative potential of AI chatbots in education, particularly in encouraging student participation and supporting more informed academic decision-making [12][13][14][15]. However, many existing solutions remain rule-based, such as those built with platforms like Dialogflow or Expert Systems, following predefined decision trees that restrict their ability to handle diverse or nuanced queries [16][17]. In contrast, more modern chatbots powered by machine learning employ Natural Language Understanding (NLU) to interpret user intent, but they remain limited in several ways. These models often rely heavily on publicly available datasets, which may not align with the specific needs of students at a given institution, making it difficult to provide tailored responses. Moreover, they struggle to adapt dynamically to evolving user inputs without frequent retraining, which can be both time-consuming and resource-intensive [18][19]. A more advanced approach, Retrieval-Augmented Generation (RAG), improves response accuracy by supplementing large language models (LLMs) with external documents, allowing them to retrieve supporting content instead of depending solely on internal knowledge [20]. However, even baseline RAG-based systems are prone to hallucinations where the model generates responses that appear correct but are factually inaccurate, irrelevant, or based on unreliable sources [21][22][23]. These issues arise particularly when retrieved content is not properly validated before being passed to the language model.

Therefore, this study proposes the development of an Agentic RAG architecture that integrates Retrieval-Augmented Generation with a BERT-based agent acting as a Dialogue Manager (DM). Unlike conventional RAG pipelines that pass retrieved content directly to a language model, the proposed approach introduces an intermediary filtering mechanism that evaluates the relevance and quality of the retrieved information before it is used in response generation. The BERT model is used for its ability to perform deep contextual understanding and classification, ensuring that only contextually appropriate passages are retained [24]. Supporting this, GPT-4 is used as the Natural Language Generation (NLG) component, tasked with producing fluent and informative responses that align with the reviewed content [25]. This hybrid design combining BERT's filtering capabilities with GPT-4's generative strengths—differentiates the proposed model from both standalone LLMs and traditional RAG systems, particularly by minimizing hallucinations and enhancing factual accuracy. By providing more reliable, context-aware responses, the system aims to support students in making

better-informed academic decisions. The remainder of this paper is structured as follows: Section II reviews related work and relevant models; Section III describes the methodology and system design; and Section IV presents experimental results and evaluation with the conclusion.

2 Preliminary Works

The use of chatbots in education is still limited and requires further exploration. Conventional chatbots still struggle in handling complex contexts and specific information requests that are often necessary in the academic world [16]. The traditional RAG system is a more promising approach which can access and leverage internal data, but still often faces challenges in terms of low retrieval accuracy [21]. For instance, a study from Cornell University reported a METEOR score of 22.2% and a cosine similarity of only 54.5% [26]. To address these limitations, this study considers an agent-based approach, leveraging BERT-based models, which offer advantages in understanding bidirectional context and handling complex queries. Among these, RoBERTa is one of the outstanding models, which managed to achieve 98.96% accuracy in chatbot text classification, outperforming models like BERT, DistilBERT, and XLM [27]. This stems from its dynamic masking strategy, larger training data, and removal of Next Sentence Prediction (NSP), which collectively improve its ability to understand and classify user intent. However, one identified challenge with RoBERTa is its computational inefficiency as a large model with high costs, making it impractical for real-time filtering of vast amounts of retrieved text.

Meanwhile, MiniLM is a model with lightweight architecture that excels in semantic similarity tasks, achieving an accuracy of 0.82 and a recall of 0.95 [28]. Unlike RoBERTa, MiniLM is designed for efficient sentence embedding and retrieval, utilizing only 6 transformer layers instead of 12 in BERT-base. Therefore, combining the 2 models can help reduce excessive cost by letting MiniLM ranks and scores documents based on semantic similarity, ensuring only the relevant passages are passed to RoBERTa instead of letting the latter model process all retrieved documents which is expensive and slow. MiniLM also has the ability to compute dense vector representations with minimal cost, which is compatible with RAG systems that rely on vector database outputs.

Given these strengths, we propose a hybrid agentic approach that combines MiniLM for retrieval and RoBERTa for classification and ranking. While the combination of these two models as agents in such a framework has not been widely explored, our study aims to demonstrate that leveraging their complementary strengths not only reduces computational overhead but also improves the quality of retrieved information, making the RAG pipeline more effective and scalable. Furthermore, with the rapid development of LLMs, this study proposes utilizing GPT-4 as part of the chatbot system. GPT-4 is considered superior in providing more in-depth and relevant answers by leveraging advanced natural language processing and computational power. Therefore, the combination of Agentic RAG and GPT-4 in this study is expected to offer a more optimal solution for guiding prospective students, addressing the weaknesses of previous models, and making a significant contribution to AI-driven education.

3 Methodology

Therefore, the experiment that this study proposes is also aimed to help reduce hallucinations in RAG systems and provide more factual based content instead of using general knowledge of the LLMs. It is shown below the detailed thought process of the model used, from dataset retrieval all the way to overall pipeline which is shown in Figure 1.

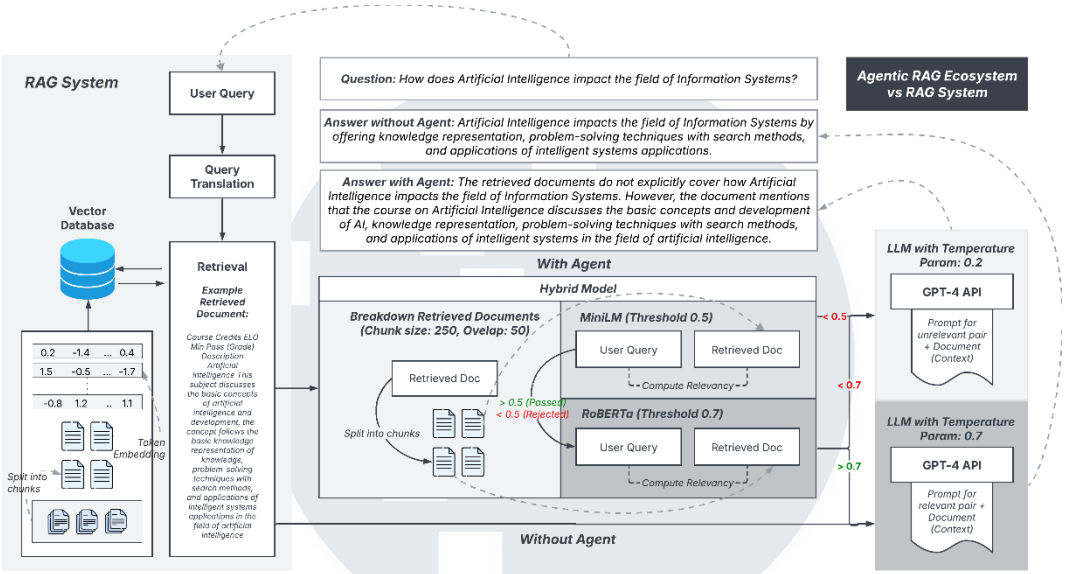

Fig. 1. *Agentic RAG Ecosystem Pipeline*

3.1 Context/Dataset Retrieval

Several academic handbooks related to different majors were collected and downloaded, resulting in a total of 19 documents. To facilitate processing, these documents were segmented into over 800 chunks. From these chunks, a representative sample was selected to generate 366 questions, allowing for a diverse range of queries that reflect real-world student inquiries. This approach ensures that the chatbot is trained on relevant and comprehensive academic content.

3.2 Bidirectional Encoder Representations from Transformers (BERT)

BERT (Bidirectional Encoder Representations from Transformers) in this Agentic RAG system acts as an agent that analyzes the retrieved information and filtering out irrelevant data [29]. BERT can function as a cross-encoder, where it processes both the queries and candidate responses simultaneously, rather than encoding them separately. These models work by merging a pair of sentences into a single input sequence and processing them together using a pre-trained model. Analyzing both sentences at the same time allows the model to recognize intricate relationships and dependencies between them, allowing for highly accurate predictions [30]. This ability makes cross-encoder well-suited for chatbots, as it can assess user queries in conjunction with potential responses. Unlike bi-encoders, which may lose fine-grained word-level interactions, cross-encoders allow for deeper contextual comprehension, making them ideal for improving chatbot

dialogue coherence and response ranking. Figure 2 illustrates how differently bi-encoders and cross-encoder process user queries.

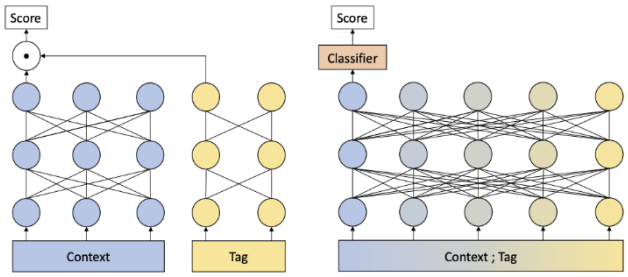

Fig. 2. *Bi-Encoder vs Cross-Encoder Illustration [32]*

The two versions of BERT utilized in this study are MiniLM & RoBERTa. MiniLM is a lightweight language model designed for efficiency while maintaining strong NLP performance. It reduces computational requirements while preserving response quality, making it suitable for applications with limited resources. Meanwhile, the RoBERTa is an extension of the BERT architecture, trained on a 160GB dataset consisting of English-language corpora. This allows RoBERTa to match or surpass the performance of post-BERT models. RoBERTa also implements Byte-Pair Encoding (BPE), a subword tokenization method that helps process 'rare' words by breaking them down into smaller, more frequent subword units. BPE improves language model generalization and reduces vocabulary size, enhancing performance across various NLP tasks [31].

To ensure the Agentic RAG system performs efficiently without sacrificing accuracy, a carefully structured configuration is implemented. First, the retrieved academic content is segmented into chunks, each limited to 256 tokens. This length offers a balance between contextual richness and computational feasibility. Additionally, a 50-token overlap is applied between segments to maintain coherence and avoid cutting off important contextual links. For relevance filtering, MiniLM acts as the first-pass filter, generating dense vector embeddings for both the user query and the retrieved chunks. Using cosine similarity, it scores the semantic alignment between them. MiniLM is assigned a threshold of 0.5, allowing it to perform a quick scan and filter out irrelevant responses without being over restrictive. This number is referred from a previous research of multilingual news article similarity, where setting the threshold at 0.5 is reasonable as it successfully achieves the highest accuracy of approximately 82% [33]. In contrast, chunks that pass this initial screening are evaluated by RoBERTa, which functions as a cross-encoder.

Unlike bi-encoders, RoBERTa jointly encodes the query and the candidate text, allowing for deeper contextual interaction. It then assigns a confidence score that reflects how well the two texts align semantically, a stricter threshold of 0.7 is applied here. This aligns with previous research using MaxProb as a confidence-based thresholding method, with HellaSwag and SocialIQA as benchmark datasets. Through extensive evaluations, it was found that HellaSwag had a mean confidence score of 78.0%, while SocialIQA had a lower mean confidence of 69.0% [34]. With this, a 0.7 threshold was chosen as a reasonable balance point. Setting the confidence threshold at 0.7 or higher led to an optimal trade-off between accuracy

and abstention, ensuring that the model only made predictions when it was sufficiently confident. When the confidence in a given answer fell below 0.7, the model was more likely to produce incorrect predictions.

3.3 Pipeline

Figure 1 compares the Agentic RAG Ecosystem with a standard RAG system, highlighting how the agent improves response accuracy by filtering retrieved documents. In the standard RAG system, retrieved documents are directly passed to GPT-4, which may generate misleading or hallucinated responses. The Agentic RAG System enhances this by introducing a hybrid model that evaluates document relevance using MiniLM (threshold: 0.5) and RoBERTa (threshold: 0.7), discarding low-relevance chunks before sending them to GPT-4. By adjusting the temperature parameter (0.2 for unfiltered, 0.7 for relevant content), the system balances factual accuracy with fluency, ensuring responses remain grounded in reliable information.

3.4 User Interface

The chatbot implementation is crucial for it to be more accessible and impactful. Numerous institutions have also started adopting advanced technologies, such as the Ruby on Rails (RoR) framework for developing data-driven web applications in the education sector [35]. RoR is known for its fast development and flexibility, but often seen as less structured, making it difficult for developers to maintain code consistency and project scalability. As a result, Django, equally popular, emerges as a very promising alternative, as it simplifies database management, offers robust security, and has a well-structured architecture, allowing developers to focus more on business logic without worrying about potential system vulnerabilities [36]. Django also excels in integrating with AI technologies like NLP. With this, the integration pipeline can be implemented modularly.

4 Experimental Results

To help evaluate the results of the hybrid model, qualitative and quantitative methods have been adopted to ensure that the hybrid model is fairly evaluated. The explanation of each can be found below.

4.1 Quantitative (METEOR, BERT Relevancy, Cosine Similarity, and Jaccard Similarity)

The METEOR Score, Bert Relevancy, Cosine Similarity, and Jaccard Similarity are used to quantitatively evaluate the performance of the hybrid model. Below are a brief explanation about what each metrics are used for:

1) **METEOR Score**
   The formula below calculates a penalty-adjusted score for METEOR. The term *(1 − Pen)* applies the penalty to the *Fmean* score, lowering it when the candidate text is highly fragmented. If there is no fragmentation *(Pen = 0)*, the score remains equal to *Fmean*, but as fragmentation increases,

the penalty reduces the score. Equation (1) shows the final formula to compute METEOR score [37].

$$score = (1 - Pen) \cdot F_{mean} \quad (1)$$

2) BERT Relevancy

BERT relevancy is a metric designed to represent the average relevancy score assigned by the hybrid agent model, which combines MiniLM and RoBERTa. It is calculated as the mean of the relevancy scores produced by both models, effectively capturing their joint assessment of how well a generated response aligns with the ground truth. Rather than serving as a standalone evaluation measure, BERT relevancy is presented primarily to provide insight into the agent's decision-making process, showcasing the overall relevance score determined by the hybrid approach.

3) Cosine Similarity

The given formula in Equation (2) represents Cosine similarity, a metric used to measure how close two vectors are aligned in a multi-dimensional space [38]. The numerator calculates the dot product of two vectors by summing the product of their corresponding elements. The denominator consists of the magnitudes (or norms) of both vectors, computed by summing the squares of their elements and taking the square root. The Cosine similarity score is obtained from dividing the dot product by the product of these magnitudes, ranging from -1 (completely opposite) to 1 (identical), with 0 indicating no similarity [39].

$$cos\theta = \frac{\sum_1^n a_i \; b_i}{\sqrt{\sum_1^n a_i^2}\sqrt{\sum_1^n b_i^2}} \quad (2)$$

4) Jaccard Similarity

Jaccard similarity is a metric for evaluating the overlap between a generated response and the ground truth by comparing their word sets [40]. It is defined as the ratio of the intersection to the union of words in both texts. Equation (3) shows the formula which measures the proportion of common elements between two sets relative to their total unique elements. In this formula, s and t represent two sets, such as tokenized words from two text samples. The numerator | s ∩ t | represents the number of shared elements between the two sets, while the denominator | s ∪ t | represents the total number of unique elements across both sets. The resulting value ranges from 0 to 1, where 0 indicates no overlap and 1 represents identical sets [39].

$$Jac(s,t)_{token} = \frac{|s \cap t|}{|s \cup t|} \quad (3)$$

Table I shows an example case of hallucination in the retrieval process, where the system without agent fabricates an answer based on partial knowledge rather than adhering to the ground truth. Although coherent, the result can include irrelevant or even misleading information that appears correct but lacks factual grounding. This issue is reflected in the low METEOR score (32.6%), Jaccard similarity (25.0%), and Cosine similarity (71.2%), indicating poor alignment with the actual answer. In contrast, the with agent response acknowledges the absence of direct information while highlighting relevant context from the document. This results in a significantly higher METEOR score (72.3%), Jaccard similarity (40.2%), and Cosine similarity (93.9%), demonstrating better factual alignment. The BERT relevancy score in the agent-assisted response also manages to reach a high number of 97.4%.

Table 1. *Response Comparison*

Question	Answer	Evaluation Metrics	Value
What are the biggest challenges in the Communication Science industry today?	**Without Agent**		
	The biggest challenges in the Communication Science industry today, according to the document, involve anticipating and adapting to digital disruption, as well as ensuring a balance between scientific and practical perspectives, especially with regard to digital technology development and practical applications.	METEOR Score	32.6%
		BERT Relevancy	84.3%
		Jaccard Similarity	25.0%
		Cosine Similarity	71.2%
	With Agent		
	The provided document does not explicitly mention the biggest challenges in the Communication Science industry today. However, it does highlight a shift in professional roles that demand mastery of advanced strategic knowledge rather than just operational skills, and the need to anticipate digital disruption. It also suggests that Communication Science education in Indonesia, particularly at the master's level, may not be adequately addressing aspects of digital technology development and practical perspectives.	METEOR Score	72.3%
		BERT Relevancy	97.4%
		Jaccard Similarity	40.2%
		Cosine Similarity	93.9%

The results from the perspective of different majors can be seen from Table II, where it compares retrieval performance with and without an agent across 11 majors. The METEOR score improves from 67.79% to 74.33% with the agent, marking an 6.54% increase. METEOR is a metric designed to evaluate the fluency and alignment of generated responses with the ground truth by considering synonym matches, stemming, and word order. The higher score with the agent suggests that the responses are not only more lexically accurate but also better structured and contextually appropriate, reducing disjointed or incomplete answers.

Table 2. *Evaluation Result Based on Major*

Major	Evaluation Metrics							
	Without Agent				With Agent			
	METEOR Score	BERT Relevancy	Jaccard Similarity	Cosine Similarity	METEOR Score	BERT Relevancy	Jaccard Similarity	Cosine Similarity
Journalism	54.6%	94.0%	41.1%	91.2%	63.8%	97.4%	48.8%	94.5%
Communication Science	56.4%	93.9%	40.7%	88.0%	65.9%	95.4%	47.6%	92.5%
Film	64.6%	93.5%	49.2%	86.5%	74.1%	95.5%	58.4%	92.7%
Information System	75.3%	97.2%	63.8%	90.4%	82.0%	97.6%	68.7%	94.7%
Visual Communication Design	64.7%	94.5%	46.5%	85.6%	74.2%	95.6%	54.3%	91.9%
Accounting	74.2%	96.8%	56.9%	94.9%	80.4%	97.5%	64.7%	96.2%
Architecture	70.2%	96.4%	53.0%	95.7%	71.4%	96.9%	54.1%	95.9%
Electrical Engineering	64.1%	96.3%	53.8%	92.7%	71.1%	97.8%	58.0%	95.1%
Informatics	71.3%	96.2%	61.1%	88.5%	80.3%	97.6%	67.8%	95.3%
Management	62.3%	94.3%	45.7%	89.3%	70.6%	97.5%	56.3%	93.5%
D3 Hospitality	88.0%	96.7%	73.4%	93.3%	83.8%	97.2%	67.8%	93.1%
Average	**67.79%**	**95.44%**	**53.20%**	**90.55%**	**74.33%**	**96.91%**	**58.77%**	**94.13%**
Difference					**+6.54%**	**+1.47%**	**+5.57%**	**+3.57%**

Further reinforcing the agent's impact, Jaccard similarity improves from 53.20% to 58.77%, while Cosine similarity increases from 90.55% to 94.13%. The Jaccard similarity boost indicates that the agent helps retain more key terms from the ground truth, improving lexical accuracy with a slight 5.57% increase. Meanwhile, the slight rise in Cosine similarity suggests that the agent enhances the semantic alignment of responses, ensuring that generated answers remain contextually close to the intended meaning. Together, these results demonstrate that the agent improves deeper semantic understanding with the generated responses more precise, relevant, and reliable.

Another significant observation is the BERT Relevancy score, reaching an average of 96.91%. This metric reflects the relevancy assessment made by the hybrid agent model (MiniLM + RoBERTa), indicating how the generated responses align with the expected answers. A high BERT Relevancy score reinforces the effectiveness of the agent in filtering irrelevant or ambiguous responses, ensuring that the output remains informative and contextually appropriate.

In addition to the observed improvements across quantitative metrics, it is important to acknowledge certain edge cases where the system underperforms, particularly in handling user queries that lack sufficient context. One notable example that was observed involves the question: “What are the compulsory courses in Semester 5?”. Without additional information specifying the intended major, the system proceeded to retrieve and generate a response based on the most semantically relevant document. The answer produced was: “The compulsory courses in Semester 5 are Media and Politics (JR 349), History of Journalism (JR 214), and English for Journalism (JR 411).” While this output is factually correct based on the Journalism program, it illustrates a broader limitation namely, the system’s inability to recognize ambiguity and request further clarification. This behavior may lead to unintended or misaligned responses, especially when the user’s actual intent pertains to a different academic program. Such cases specify the need for an intent-awareness mechanism capable of detecting under-specified queries and initiating follow-up interactions to gather missing context before retrieving and generating an answer. Addressing this issue in future iterations would

enhance the system's robustness and overall user trust, particularly in real-world educational settings where accuracy and relevance are critical.

4.2 Qualitative (Human Evaluation)

For the qualitative evaluation, human assessment was conducted to determine the clarity and completeness of responses generated by the system. Evaluators reviewed a set of 196 user queries, identifying cases where the responses were ambiguous, incomplete, or unanswerable. The results showed that the agent-assisted system produced 25 such instances, whereas the baseline system without the agent had 33. This suggests that incorporating an agent leads to a reduction in unclear or insufficient answers, highlighting its role in improving response reliability.

A human-in-the-loop evaluation was also conducted using a Google Form completed by 46 users who tested the chatbot. Participants rated four aspects, namely clarity, accuracy, fluency, and response time, on a 5-point Likert scale. Most users found the responses fluent and easy to understand, with fluency and clarity receiving the highest ratings, mainly 5 and 4, respectively. Accuracy received more moderate scores, with a majority rating it 3, indicating room for improvement in response relevance. Response time was generally viewed as satisfactory, with most users giving it a score of 4.

4.3 Deployment

The Agentic RAG system is deployed into a web-based application using Django, enabling users to interact with it through an interactive chat interface powered by the RAG pipeline. The frontend is built with HTML, CSS, and JavaScript, supporting both guest and logged-in users. Logged-in users can access saved chat histories organized by session, while guest users can chat without storing their messages. Additional features include the option to print chat conversations as PDF documents. The prototype can be found in Figure 3.

During the deployment process, there were a few technical adjustments needed, such as aligning the Python and MySQL versions used across different environments like Anaconda and Visual Studio Code. These issues were resolved by managing dependencies and setups carefully. Another challenge is handling cases where users were not logged in. The system is adjusted to skip saving chat histories when no user is associated, avoiding issues with missing user data in the database. In the end, the deployment has successfully made the RAG system usable in a real-world setting.

5 Conclusion

In conclusion, the integration of the agent into the system successfully enhances response quality across multiple evaluation metrics. The METEOR score rose from 0.668 to 0.738, while the Jaccard similarity score increased from 0.523 to 0.584,

and the Cosine similarity score went up from 0.903 to 0.941. Additionally, the high BERT Relevancy score of 0.970 reinforces the agent's ability to filter out irrelevant or ambiguous outputs, making responses more reliable. These improvements highlight the effectiveness of the Agentic RAG approach in refining chatbot performance. Furthermore, the successful deployment of the prototype demonstrates its practical applicability for real-world implementation.

Acknowledgements

This research received support from the Institution of Research and Community Services at Universitas Multimedia Nusantara. We extend our appreciation to our colleagues at the Big Data Laboratory and Student Development within the Information Systems Department at Universitas Multimedia Nusantara, whose valuable input and expertise greatly enriched the research.

References

[1] B. Bing, "The impact of higher education on high quality economic development in China: A digital perspective," *PLoS ONE*, vol. 18, no. 8, p. e0289817, Aug. 2023, doi: 10.1371/journal.pone.0289817.

[2] A. Novakovic, E. N. Patrikakou, and M. S. Ockerman, "School Counselor Perceptions of preparation and importance of college and career readiness counseling," *Professional School Counseling*, vol. 25, no. 1, Mar. 2021, doi: 10.1177/2156759x21998391.

[3] I. Junita, F. Kristine, S. Limijaya, and T. E. Widodo, "A Study of Undergraduate Students' Perception about Academic Advising in an Indonesian University," *Humaniora*, vol. 11, no. 2, pp. 129–135, Jul. 2020, doi: 10.21512/humaniora.v11i2.6490.

[4] M. C. Sáiz-Manzanares, R. Marticorena-Sánchez, L. J. Martín-Antón, I. González Díez, and L. Almeida, "Perceived satisfaction of university students with the use of Chatbots as a tool for self-regulated learning," *Heliyon*, vol. 9, no. 1, Jan. 2023. doi: 10.1016/j.heliyon.2023.e12843

[5] R. Ellison, "How choosing the right major can lead to more success and less college debt," *Forbes*, Jan. 08, 2021. [Online]. Available: https://www.forbes.com/sites/forbescoachescouncil/2019/12/26/how-choosing-the-right-major-can-lead-to-more-success-and-less-college-debt/?sh=13eb42afa83e [Accessed Feb. 1, 2025].

[6] "Factors influencing students' choice of programme of study at the College of Distance Education, University of Cape Coast: Curriculum implication," *International Journal of Social Sciences and Educational Studies*, vol. 5, no. 2, Jan. 2018, doi: 10.23918/ijsses.v5i2p205.

[7] D. Akiba and M. C. Fraboni, "Ai-supported academic advising: Exploring CHATGPT's current state and future potential toward student empowerment," *Education Sciences*, vol. 13, no. 9, p. 885, Aug. 2023. doi: 10.3390/educsci13090885

[8] M. Dawood, "Assessing the effectiveness of Chatbots in providing personalized academic advising and support to higher education students: A narrative literature review," *Studies in Technology Enhanced Learning*, vol. 4, no. 1, Oct. 2024. doi: 10.21428/8c225f6e.7140f8f4

[9] M. M. Thottoli, B. H. Alruqaishi, and A. Soosaimanickam, "Robo academic advisor: Can chatbots and Artificial Intelligence replace human interaction?," *Contemporary Educational Technology*, vol. 16, no. 1, Jan. 2024. doi: 10.30935/cedtech/13948

[10] O. Iatrellis, N. Samaras, K. Kokkinos, and T. Panagiotakopoulos, "Leveraging generative AI for sustainable academic advising: Enhancing educational practices through AI-driven recommendations," *Sustainability*, vol. 16, no. 17, p. 7829, Sep. 2024. doi: 10.3390/su16177829

[11] G. Bilquise, S. Ibrahim, and S. M. Salhieh, "Investigating student acceptance of an academic advising chatbot in higher education institutions," *Education and Information Technologies*, vol. 29, no. 5, pp. 6357–6382, Aug. 2023. doi: 10.1007/s10639-023-12076-x

[12] X. Chen, D. Zou, H. Xie, and F. L. Wang, "Educational chatbot research: Text mining and Bibliometrics," *Interactive Learning Environments*, pp. 1–19, Nov. 2024. doi: 10.1080/10494820.2024.2430632

[13] N. F. Davar, M. A. Dewan, and X. Zhang, "Ai Chatbots in education: Challenges and opportunities," *Information*, vol. 16, no. 3, p. 235, Mar. 2025. doi: 10.3390/info16030235

[14] O. Yetişensoy and H. Karaduman, "The effect of AI-powered Chatbots in Social Studies Education," *Education and Information Technologies*, vol. 29, no. 13, pp. 17035–17069, Feb. 2024. doi: 10.1007/s10639-024-12485-6

[15] H. Jo, "From concerns to benefits: A comprehensive study of CHATGPT usage in education," *International Journal of Educational Technology in Higher Education*, vol. 21, no. 1, Jun. 2024. doi: 10.1186/s41239-024-00471-4

[16] A. Alkhoori, M. A. Kuhail, and A. Alkhoori, "UniBud: A virtual academic adviser," *2020 12th Annual Undergraduate Research Conference on Applied Computing (URC)*, 2020. doi: 10.1109/urc49805.2020.9099191

[17] O. Dogan and O. F. Gurcan, "Enhancing e-business communication with a hybrid rule-based and extractive-based chatbot," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 19, no. 3, pp. 1984–1999, Aug. 2024. doi: 10.3390/jtaer19030097

[18] A. Jayavardhana and S. A. Sanjaya, "A Systematic Literature review: A comparison of available approaches in Chatbot and Dialogue Manager development," *International Journal of Science Technology & Management*, vol. 4, no. 6, pp. 1441–1450, Nov. 2023, doi: 10.46729/ijstm.v4i6.983.

[19] S. K. Assayed, M. Alkhatib, and K. Shaalan, "A systematic review of Conversational AI chatbots in academic advising," *Lecture Notes in Civil Engineering*, vol. 473, pp. 346–359, 2024. doi: 10.1007/978-3-031-56121-4_33

[20] S. Wollny *et al.*, "Are we there yet? - A systematic literature review on Chatbots in Education," *Frontiers in Artificial Intelligence*, vol. 4, 2021. doi: 10.3389/frai.2021.654924

[21] J. Song *et al.*, "RAG-HAT: A Hallucination-Aware Tuning Pipeline for LLM in Retrieval-Augmented Generation," *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1548–1558, Jan. 2024, doi: 10.18653/v1/2024.emnlp-industry.113.

[22] U. H. Khan, M. H. Khan, and R. Ali, "Large language model based educational virtual assistant using RAG framework," *Procedia Computer Science*, vol. 252, pp. 905–911, 2025. doi: 10.1016/j.procs.2025.01.051

[23] Z. Chen, D. Zou, H. Xie, H. Lou and Z. Pang, "Facilitating university admission using a chatbot based on large language models with retrieval-augmented generation," Educational Technology & Society, vol. 27, no. 4, pp. 454–470, Oct, 2024. doi: 10.30191/ETS/202410_27(4).TP02

[24] Gon, Anudeepa & Mukherjee, Gunjan & Chanda, Kaushik & Nandi, Subhadip & Ganguly, Aryabhatta.. BERT Model: A Text Classification Technique Applications of Computational intelligence in law and Criminology pp. 313-321, Dec. 2024

[25] R. Islam and O. M. Moushi, "GPT-4o: The Cutting-Edge Advancement in Multimodal LLM," *Preprint*, Jul. 2024, doi: 10.36227/techrxiv.171986596.65533294/v1.

[26] R. Lakatos, P. Pollner, A. Hajdu, and T. Joo, "Investigating the performance of Retrieval-Augmented Generation and fine-tuning for the development of AI-driven knowledge-based systems," *arXiv (Cornell University)*, Mar. 2024, doi: 10.48550/arXiv.2403.09727

[27] J. J. Bird, A. Ekárt, and D. R. Faria, "Chatbot Interaction with Artificial Intelligence: human data augmentation with T5 and language transformer ensemble for text classification," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 4, pp. 3129–3144, Aug. 2021, doi: 10.1007/s12652-021-03439-8.

[28] C. Yin and Z. Zhang, "A study of sentence similarity based on the All-MiniLM-L6-V2 model with 'Same semantics, different structure' after fine tuning," in *Advances in computer science research*, 2024, pp. 677–684. doi: 10.2991/978-94-6463-540-9_69.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv (Cornell University)*, Jan. 2018, doi: 10.48550/arxiv.1810.04805.

[30] H. S. Lee *et al.*, "Cross Encoding as augmentation: towards Effective Educational text Classification," *arXiv (Cornell University)*, Jan. 2023, doi: 10.48550/arxiv.2305.18977.

[31] B. Richardson and A. Wicaksana, "Comparison of indobert-lite and roberta in text mining for Indonesian language question answering application," *International Journal of Innovative Computing, Information and Control*, vol. 18, no. 6, pp. 1719–1734, Dec. 2022, doi: 10.24507/ijicic.18.06.1719.

[32] H. Déjean, S. Clinchant, and T. Formal, "A thorough comparison of Cross-Encoders and LLMs for reranking SPLADE," *arXiv (Cornell University)*, Mar. 2024, doi: 10.48550/arxiv.2403.10407.

[33] N. Goel and R. R. Bommidi, "SEMEVAL-2022 Task 8: Multi-lingual News article similarity," *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pp. 1129–1135, Jan. 2022, doi: 10.18653/v1/2022.semeval-1.159.

[34] K. Shen and M. Kejriwal, "Quantifying confidence shifts in a BERT-based question answering system evaluated on perturbed instances," *PLoS ONE*, vol. 18, no. 12, p. e0295925, Dec. 2023, doi: 10.1371/journal.pone.0295925.

[35] D. V. Waghmare and P. Adkar, "Agile development using Ruby on Rails Framework - IRE Journals," *IRE Journals*, vol. 2, no. 9, pp. 62–67, Mar. 2019, [Online]. Available: https://www.irejournals.com/formatedpaper/1701034.pdf

[36] N. M. Kumar and N. D. R. Nandal, "Python's Role in Accelerating Web Application Development with Django," *Deleted Journal*, vol. 2, no. 06, pp. 2092–2105, Jun. 2024, doi: 10.47392/irjaem.2024.0307.

[37] R. Musaev, "Contextual Clarity: Generating Sentences with Transformer Models using Context-Reverso Data," *arXiv (Cornell University)*, Mar. 2024, doi: 10.48550/arxiv.2403.08103.

[38] H. Steck, C. Ekanadham, and N. Kallus, "Is Cosine-Similarity of Embeddings Really About Similarity?," *ACM Web Conference 2024*, doi: 10.48550/arXiv.2403.05440

[39] T. P. Rinjeni, A. Indriawan, and N. A. Rakhmawati, "Matching Scientific Article Titles using Cosine Similarity and Jaccard Similarity Algorithm," *Procedia Computer Science*, vol. 234, pp. 553–560, Jan. 2024, pp. 1–9, Mar 2024, doi: 10.1016/j.procs.2024.03.0

[40] M. Azam et al., "A comprehensive evaluation of large language models in mining gene relations and pathway knowledge," *Quantitative Biology*, vol. 12, no. 4, pp. 360–374, Jun. 2024, doi : 10.1002/qub2.57