ILKOM Jurnal Ilmiah Vol. 18, No. 1, April 2026, pp. 30-42
Accredited 2nd by KEMDIKTISAINTEK No. 10/C/C3/DT.05.00/2025; E-ISSN 2548-7779 | P-ISSN 2087-1716

A Hybrid BERT–RAG Model for Developing Knowledge-Validated Conversational Systems

Desi Anggreani [a,1,*]; Ismawati [a,2]; A. Inayah Auliyah [b,3]; Lukman [a,4]; Aedah Abd Rahman [c,4]; Nurmisba [a,6]; Muhammad Ilham Akbar [a,7];
[a] *Universitas Muhammadiyah Makassar, Makassar, Indonesia*
[b] *Department of Information Systems, Institut Teknologi Bacharuddin Jusuf Habibie, Pare-Pare, Indonesia*
[c] *Asia E University, Malaysia*
[1] *desianggreani@unismuh.ac.id;* [2] *ismawati@unismuh.ac.id;* [2] *inayah@ith.ac.id;* [3] *Lukman@unismuh.ac.id;*[4] *Aedah.abdrahman@aeu.edu.my;* [4] *nurmisba2307@gmail.com;* [5]*105841105822@student.unismuh.ac.id;*
** Corresponding author*

Article history: Received September 30, 2025; Revised November 10, 2025; Accepted March 28, 2026; Available online April 20, 2026

Abstract
The transition of freshmen into the university environment requires adaptive and responsive information support. This study develops a chatbot system based on a hybrid BERT–RAG architecture integrated with the FAISS Index to provide automated consultation services for new students. The novelty of this research lies in the implementation of a faculty-based hierarchical knowledge structure and an adaptive multi-domain context mechanism—an approach not previously found in studies involving BERT–RAG for university onboarding services. This design enables the chatbot to deliver more relevant, personalized, and faculty-specific responses. The dataset was derived from three primary sources of information: the Faculty of Economics and Business (FEB), the Faculty of Teacher Training and Education (FKIP), and the Faculty of Engineering (FT), which were structured into a validated knowledge base in documents.json format. System evaluation was conducted across ten interaction scenarios using performance metrics including BERT Similarity, BLEU Score, ROUGE-1, ROUGE-2, and ROUGE-L. The system achieved excellent results, with average scores of 0.905 (BERT Similarity), 0.844 (BLEU), 0.876 (ROUGE-1), 0.820 (ROUGE-2), and 0.871 (ROUGE-L) and standard deviations below 0.1 across all metrics. Strong metric correlations (0.85–0.99) further indicate consistency between semantic understanding and generated text quality. Furthermore, the system effectively minimizes hallucination through validated knowledge integration and faculty-based reranking strategies. Overall, this research provides a significant contribution to the development of institutionally contextual educational chatbots capable of delivering accurate, natural, and responsive communication to support new student orientation in higher education.

Keywords: BERT, Retrieval-Augmented Generation (RAG), Educational Chatbot, University Freshmen, FAISS Index, Natural Language Processing

Introduction

The transition from secondary education to higher education represents a critical period that determines the success of freshmen's academic and social adaptation [1]. The complexity of information required by new students encompasses administrative, academic, and facility-related aspects, as well as emotional support in coping with changes in the learning environment [2]. s part of the digital native generation, freshmen expect instant and responsive access to information; however, conventional communication systems often experience significant response delays [3].

Empirical studies indicate that many freshmen experience anxiety related to campus adaptation, and the majority of them require a platform to openly express their concerns [4], [5]. The limited availability of human resources to provide 24/7 consultation services poses a major challenge for higher education institutions [6]. Conventional digital communication platforms such as WhatsApp and email have demonstrated a high level of unresponsiveness, creating an information gap between student needs and the services available [7], [8].

The advancement of Natural Language Processing (NLP) technology has opened significant opportunities for developing automated systems capable of understanding and responding to human communication needs in a natural manner [9], [10] ,[11]. Transformer-based models such as BERT have demonstrated superior performance in capturing semantic context across various text classification tasks, with studies reporting accuracy levels of

approximately 92–95% on relevant benchmark datasets [12]. The strength of BERT lies in its bidirectional encoding capability, which enables deeper contextual understanding compared to traditional unidirectional models [13].

The integration of BERT with Retrieval-Augmented Generation (RAG) architecture allows systems to dynamically access and utilize external knowledge bases [14]. he RAG approach addresses the limitations of purely generative models, which are prone to hallucinations, by providing a validated retrieval mechanism [15]. This combination produces systems capable of delivering factually accurate responses while maintaining naturalness in communication [16].

Previous studies in the educational chatbot domain have shown mixed results with various limitations. Recent research revealed that template-based or rule-based chatbots suffer from significant constraints, as they are unable to handle inputs beyond predefined rules. Open-ended or unexpected queries often lead to irrelevant or even incorrect responses[17]. Similarly, studies on Seq2Seq and Attention-based chatbots, while achieving strong performance according to BLEU and ROUGE metrics, still struggle to maintain factual accuracy when faced with out-of-distribution queries[18]. However, such implementations were not specifically designed for the freshmen context and did not consider the importance of emotional support. Zhe Xu et al. proposed a hybrid retrieval–generation approach, yet its application was not tailored to the needs of new students and similarly overlooked the emotional support dimension [19].

The limitations of previous studies primarily lie in the fragmented focus between information accuracy and response naturalness [20], [21]. Retrieval-based systems tend to provide accurate yet rigid and impersonal answers, while purely generative systems are prone to hallucinations and information inconsistencies [22]. The urgency to design a hybrid architecture that can synergize the strengths of these two distinct approaches is increasingly evident. Such an architecture is essential for the development of applications that demand high information integrity while simultaneously maintaining empathetic interaction capabilities with users [23].

The motivation for this research stems from empirical observations of communication patterns among freshmen at major universities in Indonesia during the 2022–2024 period. An analysis of their interactions revealed that most questions could be categorized into several key topics. However, the linguistic diversity in how these questions were expressed was remarkably high, encompassing variations in sentence structure, word choice, and communication style. This phenomenon reflects the inherent complexity of effectively understanding and responding to freshmen's needs[24]. t also underscores the necessity of a system capable of accurately recognizing intent across diverse formulations while delivering personalized responses aligned with the users' emotional context [25].

This study proposes a hybrid BERT–RAG architecture specifically designed to address the limitations of conventional chatbot systems in the freshmen support domain. The main contributions include the development of a modular architecture with BERT serving as the semantic encoder, the implementation of RAG as a middle layer for contextual retrieval, and the integration of the FAISS Index for vector search optimization. The novelty of this research lies in the system design, which explicitly incorporates the informational needs of freshmen while ensuring factual accuracy and maintaining naturalness in communication.

Method

The chatbot system developed in this study adopts a hybrid BERT–RAG architecture consisting of three main components: the semantic understanding layer, the retrieval layer, and the response generation layer. The semantic understanding layer employs a pre-trained BERT model fine-tuned on a higher education–specific domain to capture the intent and context of students' questions[26]. The retrieval layer implements the RAG framework integrated with a FAISS Index to perform semantic searches across the knowledge base[27]. The response generation layer is responsible for producing natural and contextually appropriate answers based on the information obtained from the retrieval process[28]. A sample of the dataset used in this research is presented in **Table 1**, which includes information on faculties, study programs, accreditation, facilities, tuition fees per semester, and brief descriptions.

Table 1. Dataset for Model Training and Evaluation

Faculty	Study Program	Accreditation	Facilities	Fees Semester	Description
FEB	Manajemen	B	Lab Komputer, Perpustakaan Digital, Ruang Diskusi	Rp 3.600.000	*Program studi yang mempersiapkan mahasiswa menjadi manajer profesional dengan pemahaman bisnis modern*
FEB	*Akuntansi*	B	Lab Akuntansi, Software Accurate, Lab Perpajakan	Rp 3.600.000	*Program studi yang menghasilkan akuntan kompeten dengan keahlian sistem informasi akuntansi*
FEB	*Ekonimi Pembangunan*	B	Lab Ekonometrika, Pusat Data Ekonomi, Lab Statistik	Rp 3.200.000	*Program studi yang fokus pada analisis kebijakan ekonomi dan pembangunan berkelanjutan*
FKIP	*Pendidikan Biologi*	B	Lab Biologi, Lab Mikrobiologi, Greenhouse, Lab Microteaching	Rp 3.200.000	*Program studi yang mempersiapkan guru biologi profesional dengan kompetensi pedagogik dan akademik*
FKIP	*Pendidikan Matematika*	B	Lab Matematika, Software GeoGebra, Lab Microteaching	Rp 3.200.000	*Program studi yang menghasilkan pendidik matematika dengan kemampuan mengajar inovatif*
FKIP	*Pendidikan Bahasa Inggris*	B	Language Lab, Native Speaker Program, Lab Microteaching	Rp 3.400.000	*Program studi yang mempersiapkan guru bahasa Inggris dengan kompetensi internasional*
FT	*Informatika*	B	Lab AI, Lab Jaringan, Lab Pemrograman, Lab Mobile Development	Rp 3.600.000	*Program studi teknologi informasi yang menghasilkan programmer dan system analyst kompeten*
FT	*Elektro*	B	Lab Elektronika, Lab Kontrol, Lab Telekomunikasi, Lab PLC	Rp 3.600.000	*Program studi teknik elektro yang fokus pada sistem kelistrikan dan elektronika industri*
FT	*Teknik Sipil*	B	Lab Struktur, Lab Material, Lab Hidrolika, Lab CAD	Rp 3.800.000	*Program studi yang mempersiapkan insinyur sipil untuk pembangunan infrastruktur*

The research dataset comprises three primary sources that provide comprehensive information on study programs, tuition fees, facilities, and accreditation. The first source, data_feb.json, contains 156 information entities related to the Faculty of Economics and Business, covering the study programs of Management, Accounting, and Development Economics. The second source, data_fkip.json, includes 203 information entities from the Faculty of Teacher Training and Education, encompassing the study programs of Biology Education, Mathematics Education, and English Education. The third source, data_ft.json, consists of 184 information entities from the Faculty of Engineering, covering the study programs of Informatics, Electrical Engineering, and Civil Engineering. The data processing pipeline leading to response generation is illustrated in **Figure 1**, starting from raw dataset input to the system's ability to produce natural responses.

Figure 1. Workflow of the BERT–RAG Chatbot Model

The preprocessing stage involved text normalization, tokenization, and structuring of data into a consistent format. Each information entity was organized with metadata encompassing faculty, study program, information category

(academic, tuition, facilities), and timestamp for tracking updates. Data cleaning was performed using regular expressions to remove non-standard characters and to normalize currency and numerical formats[29]. The final output of preprocessing was documents.json, comprising 543 structured documents with an average length of 127 tokens per document.

The BERT model employed was bert-base-multilingual-cased, optimized for Indonesian language understanding. Fine-tuning was conducted using a domain-specific dataset over 20 epochs, with a learning rate of 2e-5 and batch size of 16. The encoding process produced 768-dimensional vector representations for each document segment in the knowledge base. The optimization function used was AdamW with a weight decay of 0.01 to prevent overfitting [30]. Training parameters included a maximum sequence length of 512 tokens, warmup steps set at 10% of the total training steps, and gradient clipping with a maximum value of 1.0. A validation split of 20% of the dataset was applied using stratified sampling to preserve the distribution of information categories. The model checkpoint achieving the lowest validation loss was selected for system implementation. The training process employed Cross-Entropy Loss to minimize prediction errors [31]:

$$L = -\sum_{i=1}^{N} y_i . \log(\hat{y}_i) \tag{1}$$

With $\hat{y}_i$ representing the ground truth labels and $\hat{y}_i$ denoting the predicted probabilities, optimization was performed using AdamW with a weight decay of 0.01. To enhance semantic retrieval, FAISS (Facebook AI Similarity Search) Index was implemented on the knowledge base containing 543 document vectors. The index configuration employed IndexFlatIP (Inner Product) to maximize retrieval accuracy in similarity search. Prior to indexing, each document vector was normalized using L2 normalization to ensure consistency in cosine similarity computation[32].

$$CosinSimilarity\ (A, B) = \frac{\sum_{i=1}^{n} A_i\ .\ B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2}$$

The indexing process involved embedding generation for all documents in the knowledge base, vector normalization, and the construction of the FAISS index with default parameters[33]. Query processing applied a similarity threshold of 0.7 to filter relevant retrieval results. The top-k retrieval was set to 5 to provide sufficient contextual information while minimizing informational noise.

The Retrieval-Augmented Generation (RAG) framework was implemented in two primary stages: the retrieval phase and the generation phase. In the retrieval phase, query encoding was performed using the same BERT model to maintain semantic representation consistency. The search was conducted against the FAISS index, retrieving the top-5 most relevant documents based on cosine similarity[34]. In the generation phase, a template-based approach was employed to integrate the retrieved documents with the query context, producing coherent responses. The response templates were designed to preserve communication naturalness while ensuring factual accuracy. Post-processing included filtering to eliminate redundant information and formatting for consistent presentation[35].

System evaluation was conducted using ten interaction scenarios encompassing diverse types of freshmen inquiries. These scenarios included questions regarding study programs, tuition fees, campus facilities, admission procedures, and accreditation information. Each interaction was evaluated using five key metrics: BERT Similarity, BLEU Score, ROUGE-1, ROUGE-2, and ROUGE-L [36]. BERT Similarity assessed the semantic alignment between system responses and reference answers by applying cosine similarity to BERT embeddings. The BLEU Score evaluates text generation quality by comparing n-gram precision between the generated response and the reference answer. The ROUGE metrics (ROUGE-1, ROUGE-2, ROUGE-L) measure recall-oriented overlap for unigrams, bigrams, and the longest common subsequence[37]:

$$BLEU = BP.exp\left(\sum_{n=1}^{N} w_n . \log(p_n)\right) \tag{3}$$

where ;

p_n : precision for n-grams

w_n : weight assigned to each n-gram

BP : brevity penalty to account for differences in text length

ROUGE-1, ROUGE-2, dan ROUGE-L :

$$ROUGE - 1 = \frac{\sum_{\text{gram} \in \text{Reference}} min\left(Count_{system}(gram), Count_{references}(gram)\right)}{\sum_{\text{gram} \in \text{Reference}} Count_{system}(gram)} \quad (4)$$

To analyze relationships among evaluation metrics, Pearson correlation was employed, for example between BERT Similarity and ROUGE-L :

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

The ground truth for evaluation was constructed through manual annotation by three expert evaluators with experience in student services. Inter-annotator agreement, measured using Cohen's Kappa, yielded a value of 0.87, indicating high consistency in assessing response quality. Each metric was computed across all evaluation scenarios and further analyzed using descriptive statistics as well as inter-metric correlations.

Results and Discussion

The BERT model training process exhibited stable convergence with a consistent improvement in accuracy across 20 training epochs. **Figure 2** illustrates the accuracy curve of the BERT model during training, with the final accuracy reaching 0.830 (83%).

Figure 2. Training Evaluation Curve of the BERT Model

The training curve demonstrates progressive learning, characterized by a significant increase in accuracy during the initial epochs, followed by stabilization in the later epochs, indicating optimal convergence. Analysis of the curve reveals no significant signs of overfitting, as evidenced by the minimal gap between training and validation accuracy. The most substantial accuracy improvement occurred between epochs 1 and 7, reaching a level of 0.64, followed by gradual improvement until epoch 13 with an accuracy of 0.75. The convergence phase began from epoch 14 through epoch 20, with minimal fluctuations around the final accuracy of 0.83. The good performance threshold (0.8) was achieved at epoch 16, whereas the excellent performance threshold (0.9) was not attained, reflecting the inherent complexity of the domain-specific understanding task.

The system evaluation of the chatbot was conducted across ten interaction scenarios encompassing various categories of freshmen inquiries. **Table 2** presents the statistical summary of the evaluation results for all employed metrics. The evaluation outcomes demonstrate strong overall performance, with the average BERT Similarity reaching 0.905, indicating a high level of semantic understanding in capturing the contextual intent of students' questions.

Table 2. Statistical Summary of Evaluation Metrics

Metric	Mean	Std	Min	Max	Interpretation
BERT Similarity	0.905	0.073	0.800	1.000	The chatbot demonstrates strong contextual understanding; all interactions ≥0.8 indicate robust semantic comprehension.
BLEU Score	0.844	0.079	0.800	1.000	High text generation quality with strong n-gram precision, producing responses closely aligned with references.
ROUGE-1	0.876	0.078	0.800	1.000	High unigram overlap, reflecting strong lexical accuracy in chatbot responses.
ROUGE-2	0.820	0.093	0.750	1.000	Good bigram consistency higher fluctuations suggest sensitivity to sentence structure.
ROUGE-L	0.871	0.076	0.800	1.000	Responses preserve longer sentence structures similar to references, maintaining naturalness in communication.

The BLEU Score, with an average of 0.844, indicates strong text generation quality, suggesting that the system's responses exhibit high n-gram precision relative to the reference answers. The chatbot's evaluation performance across ten distinct interactions is visualized in **Figure 3**, where all metrics (BERT, BLEU, ROUGE) consistently remain above the minimum threshold.

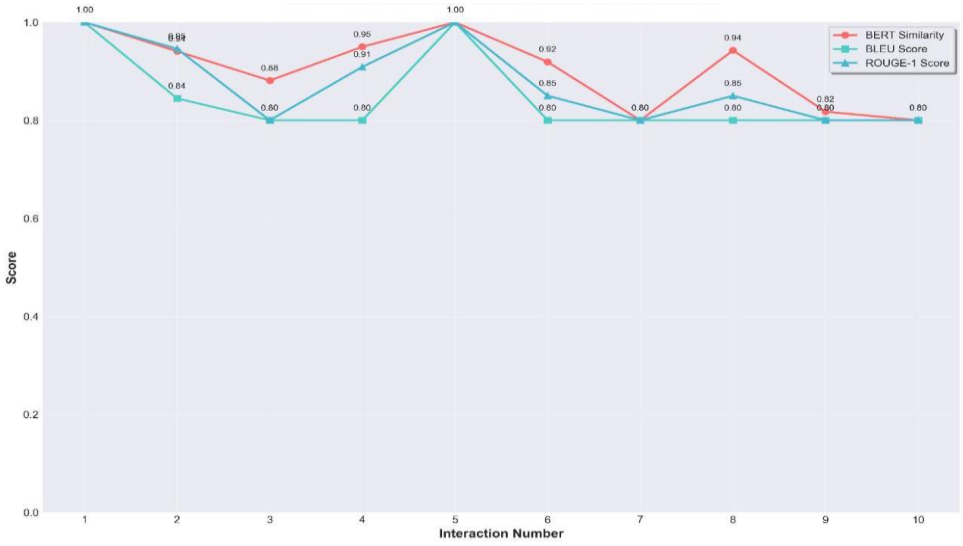

Figure 3. Chatbot Evaluation Metric Performance Across 10 Interactions

ROUGE-1, with a score of 0.876, demonstrates excellent unigram overlap between generated responses and the ground truth, while ROUGE-2 (0.820) reflects good bigram consistency. Similarly, ROUGE-L (0.871) highlights the high quality of the longest common subsequence between the system's responses and the references. Figure 3 further illustrates that all evaluation metrics maintain stable performance above the minimum threshold of 0.75. The highest variability is observed in BERT Similarity, ranging between 0.80 and 1.00, whereas BLEU and ROUGE scores exhibit stronger stability. Notably, peak performance is achieved in Interaction 1 and 5, with perfect scores (1.0) across all metrics, indicating that the system is capable of generating ideal responses for certain types of queries.

Detailed examples of the chatbot's performance are presented in **Table 3**, which compares reference answers with the system's generated responses across several student query scenarios.

Table 3. Experimental Evaluation Results of the System

ID	Question	RSystem	BERT	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
1	*Hallo, saya ingin berkuliah di Unismuh Makassar,*	*"Bisa Anda sebutkan bidang minat atau latar belakang pendidikan Anda? Misalnya: Lulusan SMA/SMK jurusan apa? Bidang apa yang Anda minati?*	1.000	1.000	1.000	1.000	1.000

ID	Question	RSystem	BERT	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
	jurusan apa yang cocok?	*Apakah ada keahlian atau hobi khusus? Informasi ini akan membantu saya memberikan rekomendasi jurusan yang tepat di Unismuh Makassar."*					
2	*Berapa biaya kuliah di Informatika?*	*"Biaya kuliah per semester untuk Program Studi Informatika di Universitas Muhammadiyah Makassar adalah sebesar Rp. 3.600.000. Selain itu, terdapat juga biaya uang pembangunan sebesar Rp. 5.700.000."*	0.881	0.800	0.800	0.750	0.800
3	*Apakah biaya kuliah Elektro sama seperti Informatika?*	*"Biaya kuliah per semester untuk Program Studi Teknik Elektro di Universitas Muhammadiyah Makassar adalah sebesar Rp. 3.600.000, sama dengan biaya kuliah untuk Program Studi Informatika. Selain itu, terdapat juga biaya uang pembangunan sebesar Rp. 5.700.000 untuk kedua program studi tersebut."*	0.919	0.800	0.850	0.800	0.850
4	*Apakah ada jurusan Sistem Informasi di Unismuh?*	*"Program Studi Sistem Informasi saat tidak tersedia di Universitas Muhammadiyah Makassar."*	0.800	0.800	0.800	0.750	0.800
5	*Apa akreditasi jurusan Informatika Unismuh?*	*"Program Studi Informatika di Universitas Muhammadiyah Makassar memiliki akreditasi yang Baik. Program Studi Informatika fokus pada teknologi informasi, pemrograman, jaringan komputer, kecerdasan buatan, dan pengembangan perangkat lunak berbasis web."*	0.818	0.800	0.800	0.750	0.800

The experimental results reveal that the system is capable of performing contextual semantic inference by producing contextually appropriate responses, even in the absence of explicit entries in the knowledge base. This confirms that the proposed BERT–RAG architecture transcends conventional retrieval-based approaches and exhibits context-aware reasoning through semantic representation of user input.

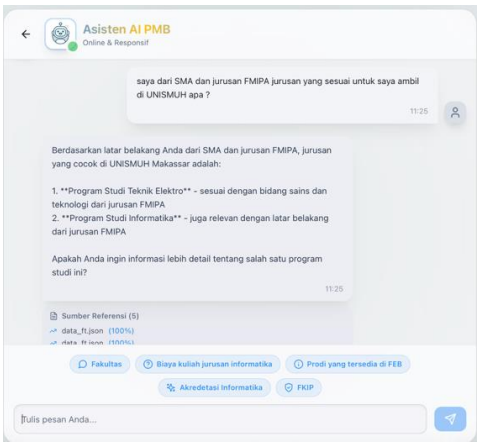

Figure 4. System Response to Out-of-Knowledge-Base Query

Analysis of BERT Similarity scores is presented in **Figure 5** across 10 interactions, with a similarity threshold of 0.7 defined as the minimum acceptable performance. The results demonstrate that all interactions met or exceeded this threshold with a significant margin, where the lowest score was 0.800 in interactions 7 and 10. A perfect similarity

score of 1.000 was achieved in interactions 1 and 5, indicating flawless semantic understanding for those types of queries. The performance pattern of BERT Similarity exhibits regular fluctuations, with three peak performances (interactions 1, 5, and 8) and two valley performances (interactions 3 and 7).

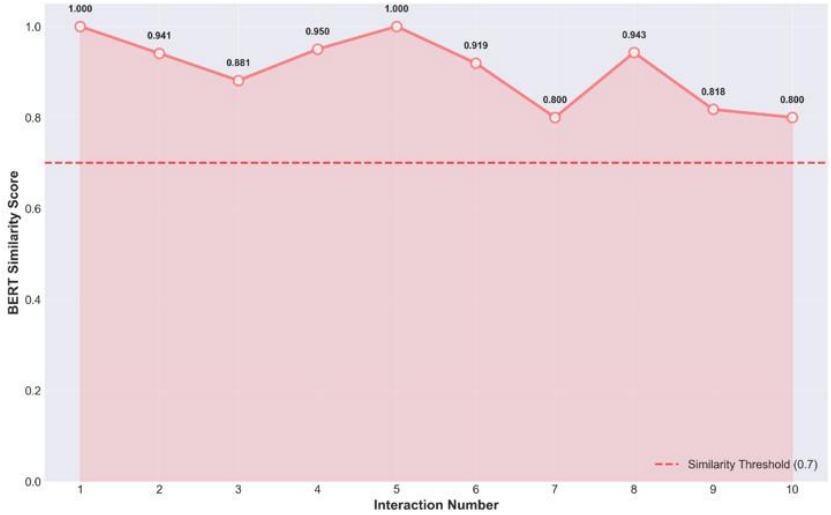

Figure 5. Analysis of BERT Similarity Scores Across 10 Interactions

These fluctuations indicate the system's sensitivity to the complexity and ambiguity of user queries. Interactions with the highest similarity scores were generally associated with straightforward factual questions, such as tuition fee information and program accreditation. Conversely, interactions with relatively lower scores involved queries that required reasoning or comparative analysis across multiple information entities. To contextualize the contribution of this research, Tabel 4 Table 4 presents a comparison between the developed chatbot system and two relevant prior studies. This comparison highlights the differences in model architectures, domain focus, and evaluation metrics adopted as benchmarks.

Table 4. Evaluation Results Compared with Previous Studies

Aspect	This Study (Hybrid BERT-RAG + FAISS)	Saluja et al. [18]	Zhe Xu et al. [19]
Core Architecture	Hybrid BERT-RAG with FAISS Index	Seq2Seq with Attention Mechanism	Hybrid Retrieval-Generation
Domain Focus	Comprehensive information support for new university students	Open-domain conversational chatbot	Emotional support in conversations
Key Capabilities	Factual accuracy from knowledge base and deep semantic understanding	Flexible and context-aware response generation	Recognition and response to users' emotional states
Evaluation Metrics	BERT Similarity, BLEU, ROUGE-1, ROUGE-2, ROUGE-L	BLEU, ROUGE	Metrics for emotional support evaluation
Main Limitations	Restricted to existing knowledge base; sensitive to complex linguistic variations	Susceptible to hallucination and factual inconsistency	Implementation not tailored to new students and lacking academic information focus

The correlation matrix presented in Figure 5 illustrates the relationships among the evaluation metrics. Correlation analysis reveals strong and consistent associations across all metrics, with coefficients ranging from 0.687 to 0.994. The strongest correlation was observed between ROUGE-1 and ROUGE-L (0.994), indicating a high degree of consistency between unigram overlap and the longest common subsequence in system responses. BERT Similarity demonstrated moderate to strong correlations with the other metrics, with the highest correlation observed with ROUGE-1 and ROUGE-L (0.887), followed by ROUGE-2 (0.799) and BLEU (0.687). This correlation pattern suggests

that semantic understanding, as measured by BERT Similarity, is positively associated with the quality of text generation captured by BLEU and ROUGE metrics. The relatively lower correlation between BERT Similarity and BLEU (0.687) further indicates that these two metrics evaluate different yet complementary aspects of response quality.

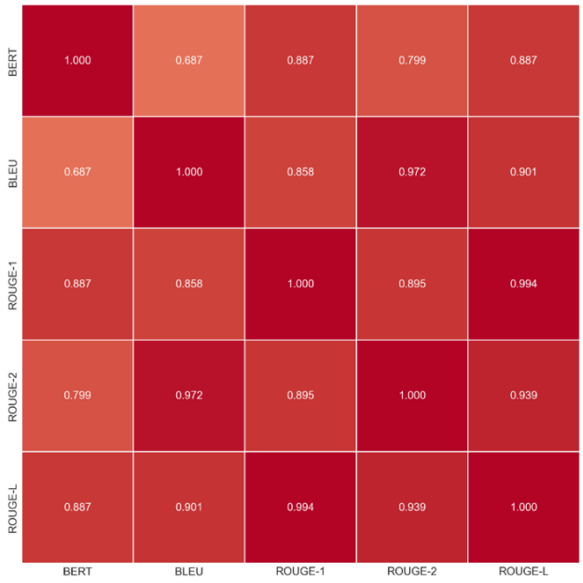

Figure 6. Correlation Among Evaluation Metrics

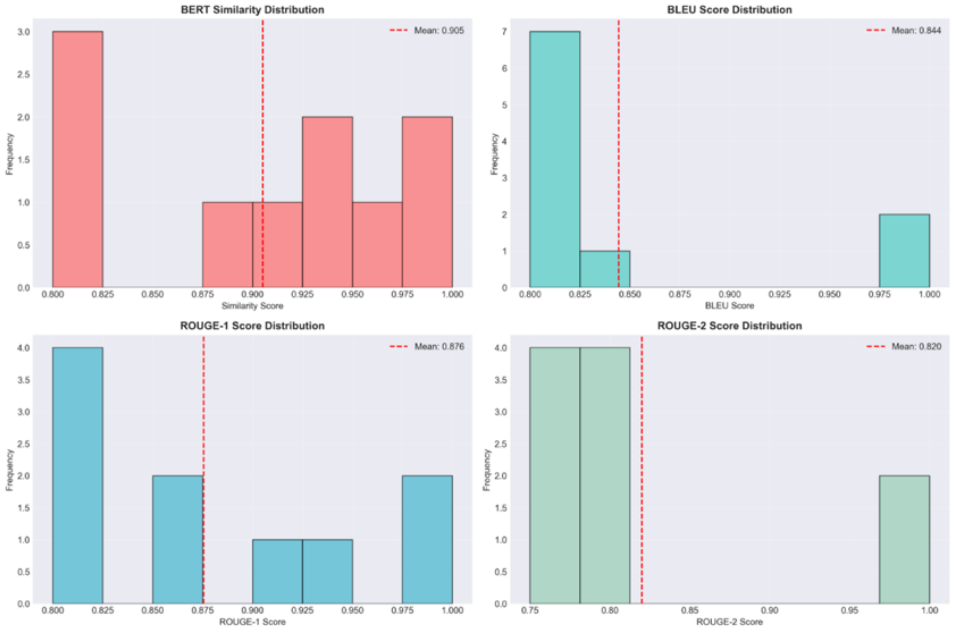

Figure 7. Distribution of Evaluation Metric Scores

The distribution of scores is presented in Figure 6 for each evaluation metric using histogram visualizations to illustrate the system's performance characteristics. The BERT Similarity distribution shows a mean of 0.905 with a left-skewed pattern, indicating that the majority of interactions achieved high similarity scores. Three distinct peaks are observed in the ranges 0.80–0.82, 0.94–0.96, and 0.98–1.00, suggesting performance clustering at specific levels. The BLEU Score distribution, with a mean of 0.844, exhibits a more uniform spread compared to BERT Similarity. Most scores are concentrated in the 0.80–0.84 range (7 out of 10 interactions), while the high-performance range of 0.98–1.00 was achieved in 2 interactions. This pattern indicates that BLEU is more sensitive to variations in text generation quality. The ROUGE-1 distribution shows a mean of 0.876 with concentrated clusters at the lower range (0.80–0.84, frequency 4) and the higher range (0.98–1.00, frequency 2). In contrast, the ROUGE-2 distribution, with a mean of 0.820, demonstrates a wider spread, with concentrations at 0.75–0.80 (frequency 4) and 0.95–1.00 (frequency 2). These

distribution patterns suggest that the system exhibits polarized performance, achieving excellent outcomes in certain cases while maintaining good performance in more challenging scenarios.

A qualitative evaluation of the system responses further highlights strong capabilities in understanding and addressing diverse types of student inquiries. The system effectively handled factual questions, such as tuition fee information, with high accuracy, delivering responses that were specific and well-structured. For comparative questions, such as cross-program tuition comparisons, the system successfully retrieved multiple documents and synthesized the information in a comparative manner. The responses demonstrated consistency in format and structure, leveraging templates while maintaining naturalness in communication. Hallucinations were avoided through reliance on a validated knowledge base, ensuring that all provided information could be verified against the original data sources. The system also maintained communicative naturalness by adopting a conversational yet appropriately formal style aligned with the academic context. Nevertheless, certain limitations were identified. The system faced difficulties in handling questions requiring complex inference or those falling outside the scope of the knowledge base. It also demonstrated sensitivity to significant variations in phrasing, where linguistically complex queries tended to yield lower similarity scores. However, no instances of critical failure were observed, as the system did not produce incorrect or misleading information.

Conclusion

This study successfully developed a chatbot system based on a BERT–RAG architecture integrated with the FAISS Index to support the information needs of new university students. The evaluation results indicate excellent performance, with average scores of 0.905 for BERT Similarity, 0.844 for BLEU, 0.876 for ROUGE-1, 0.820 for ROUGE-2, and 0.871 for ROUGE-L, accompanied by standard deviations below 0.1 across all metrics, demonstrating high system stability. The novelty of this research lies not only in the implementation of a faculty-based hierarchical knowledge structure and adaptive multi-domain context mechanism but also in the system's ability to generate responses to queries that are not explicitly available in the knowledge base through contextual semantic inference. This capability indicates that the system is not merely retrieval-based but is also capable of performing context-aware reasoning and semantic generalization based on the user's intent and background. Furthermore, the integration of BERT's semantic understanding and RAG's retrieval accuracy effectively addressed the classical trade-off between factual precision and conversational naturalness commonly found in traditional chatbot systems. The implementation of the FAISS Index also enabled efficient similarity search within a knowledge base of 543 documents, ensuring optimal response time for real-time interaction. Correlation analysis across evaluation metrics revealed a strong alignment between semantic understanding (BERT) and text generation quality (BLEU, ROUGE), with correlation coefficients ranging from 0.687 to 0.994, demonstrating consistent intent comprehension and high-quality response generation. Future research directions include expanding the scope of the knowledge base, integrating Large Language Models (LLMs) for enhanced personalization, and developing emotional support capabilities to address students' psychological needs. Therefore, this system presents a promising foundation for the development of adaptive educational chatbots capable of providing 24/7 information services while improving institutional efficiency in higher education environments.

References

[1] L. M. Ramjan et al., "Academic support strategies for nursing students with a disability at university — An integrative review," *Collegian,* vol. 32, no. 4, pp. 195–211, Aug. 2025. doi: 10.1016/j.colegn.2025.05.001.

[2] J. Wang, "Impact of natural disasters on student enrollment in higher education programs: A systematic review," *Heliyon,* vol. 10, no. 6, e27705, Mar. 2024. doi: 10.1016/j.heliyon.2024.e27705.

[3] B. Shang and Z. He, "AI-assisted symbolic healing for digital natives: A Jungian perspective based on the S-O-R and Fogg behavior models," *Acta Psychologica*, vol. 260, 105472, Oct. 2025. doi: 10.1016/j.actpsy.2025.105472.

[4] Y. Wang et al., "Stressors in university life and anxiety symptoms among international students: a sequential mediation model," *BMC Psychiatry*, vol. 23, no. 1, p. 556, Aug. 2023. doi: 10.1186/s12888-023-05046-7.

[5] K. Zhang, Z. Mi, E. J. Parks-Stamm, W. Cao, Y. Ji, and R. Jiang, "Adaptability protects university students from anxiety, depression, and insomnia during remote learning: A three-wave longitudinal study from China,"

Frontiers in Psychiatry, vol. 13, 868072, 2022. doi: 10.3389/fpsyt.2022.868072.

[6] T. Rohimah, Firman, N. S. Neviyarni, and M. A. C. Amat, "Optimizing counseling programs in higher education and their future implications," *Quality: Journal of Education, Arabic and Islamic Studies*, vol. 3, no. 1, 2025. doi: 10.58355/qwt.v3i1.97.

[7] D. Gilani, "Student attitudes and preferences towards communications from their university – a meta-analysis of student communications research within UK higher education institutions," *Journal of Higher Education Policy and Management*, vol. 46, no. 3, pp. 274–290, 2024. doi: 10.1080/1360080X.2024.2344234.

[8] T. Ratnasari, "Persepsi mahasiswa terhadap layanan online aplikasi pelayanan terpadu satu pintu pada Biro Administrasi Umum Akademik dan Kemahasiswaan IAIN Kudus," *Diplomatika: Jurnal Kearsipan Terapan*, vol. 6, no. 1, pp. 17–26, 2022. doi: 10.22146/diplomatika.8312117.

[9] S. H. Anwar, K. M. Abouaish, E. M. Matta, A. K. Farouq, A. A. Ahmed, and N. K. Negied, "Academic assistance chatbot – a comprehensive NLP and deep learning-based approaches," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 33, no. 2, pp. 1042–1056, Feb. 2024. doi: 10.11591/ijeecs.v33.i2.pp1042-1056.

[10] A. Puspitasari, A. N. Paradhita, Y. W. Tineka, V. Sulistyowati, N. K. S. Noriska, and Haryanto, "Natural language processing (NLP) technology for chatbot website," *Jurnal Penelitian Pendidikan IPA*, vol. 10, Special Issue, pp. 319–324, 2024. doi: 10.29303/jppipa.v10iSpecialIssue.8241.

[11] C. Suardi, D. Anggeani, A. P. Wibawa, N. Murtadlo, I. A. E. Zaeni, and N. A. M. Jabari, "Asking a chatbot for food ingredients halal status," *in Halal Development: Trends, Opportunities and Challenges*, pp. 14–20, 2021.

[12] C. Eang and S. Lee, "Improving the accuracy and effectiveness of text classification based on the integration of the BERT model and a recurrent neural network (RNN_BERT_Based)," *Applied Sciences*, vol. 14, no. 18, p. 8388, Sep. 2024. doi: 10.3390/app14188388.

[13] Y. Sun, "The evolution of transformer models from unidirectional to bidirectional in natural language processing," *Applied and Computational Engineering*, Feb. 2024. doi: 10.54254/2755-2721/42/20230794.

[14] Z. Levonian, C. Li, W. Zhu, A. Gade, O. Henkel, M.-E. Postle, and W. Xing, "Retrieval-augmented generation to improve math question-answering: Trade-offs between groundedness and human preference (Version 2)," *arXiv*, 2023. doi: 10.48550/arXiv.2310.03184.

[15] O. Ayala and P. Bechard, "Reducing hallucination in structured outputs via retrieval-augmented generation," *in Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2024, pp. 228–238. doi: 10.18653/v1/2024.naacl-industry.19.

[16] J. Swacha and M. Gracel, "Retrieval-augmented generation (RAG) chatbots for education: A survey of applications," *Applied Sciences*, vol. 15, no. 8, p. 4234, Apr. 2025. doi: 10.3390/app15084234.

[17] W. C. Choi and C. I. Chang, "A survey of techniques, design, applications, challenges, and student perspective of chatbot-based learning tutoring system supporting students to learn in education," *Preprints*, 2025, 2025031134. doi: 10.20944/preprints202503.1134.v1.

[18] K. Saluja, S. Agarwal, S. Kumar, and T. Choudhury, "Evaluating performance of conversational bot using Seq2Seq model and attention mechanism," *EAI : Endorsed Transactions on Scalable Information Systems*, vol. 24, no. 6, Mar. 2024. doi: 10.4108/eetsis.5457.

[19] Z. Xu, D. Chen, J. Kuang, Z. Yi, Y. Li, and Y. Shen, "Dynamic Demonstration Retrieval and Cognitive Understanding for Emotional Support Conversation," International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024. doi: 10.1145/3626772.3657695.

[20] O. Henkel, Z. Levonian, C. Li, and M. Postle, "Retrieval-augmented Generation to Improve Math Question-Answering: Trade-offs Between Groundedness and Human Preference," *in Proceedings of the 17th International Conference on Educational Data Mining*, Jul. 2024, pp. 315–320, doi: 10.5281/zenodo.12729824.

[21] H.-T. Ho, T.-T.-H. Nguyen, D. N. M. Huy, and L. V. Nguyen, "Bio-Inspired Algorithms in NLP Techniques: Challenges, Limitations and Its Applications," *Computers, Materials and Continua*, vol. 83, no. 3, pp. 3945–3973, May 2025, doi: 10.32604/cmc.2025.063099.

[22] R. C. Torres, "AI Hallucination in the Context of Education: Exploring College Students' Use of Generative AI for Academic Tasks," *2025 16th International Conference on E-Education, E-Business, E-Management and E-Learning (IC4e)*, Tokyo, Japan, 2025, pp. 445-449, doi: 10.1109/IC4e65071.2025.11075444.

[23] D. Uysal, D. Toka, E. Bozkurt, O. Kumaş, and H. H. Yılmaz, "Enhancing Financial NLP with Supercomputing: Spell Correction and Domain-Specific BERT Pretraining," *Procedia Computer Science*, vol. 267, pp. 176–186, 2025, doi: 10.1016/j.procs.2025.08.244.

[24] R. Datu, D. F. Antara, J. J. Sumangando, H. R. Kakunsi, A. E. Handoko, and J. R. Patroli, "Online Behavior and the Transformation of Interpersonal Communication in the Social Media Era," *Jurnal Syntax Admiration*, vol. 6, no. 2, Feb. 2025, doi: 10.46799/jsa.v6i2.2141.

[25] X. Zhang, W. Wang, and Q. Jin, "IntentionESC: An Intention-Centered Framework for Enhancing Emotional Support in Dialogue Systems," *arXiv preprint arXiv:2506.05947*, Jun. 2025, doi: 10.48550/arXiv.2506.05947.

[26] J. Liu, "Deep Semantic Analysis and Thematic Evolution Prediction of American Literary Texts Based on BERT and Knowledge Graph," *2025 IEEE International Conference on Electronics, Energy Systems and Power Engineering (EESPE)*, Shenyang, China, 2025, pp. 218-224, doi: 10.1109/EESPE63401.2025.10986868.

[27] N. Reddy, "Design and Implementation of an AI-Based Chatbot Framework with Retrieval-Augmented Generation and Integrated Recommender System for Interactive User Support," *SSRN*, 2025, [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5250507.

[28] M. Ahmad, "Toward a unified framework for information retrieval in large language model applications: Balancing textual and graph-based knowledge sources", *Master's thesis, School of Engineering*, May 2025. [Online]. Available: https://aaltodoc.aalto.fi/items/8c94c500-844b-4587-97df-6bfe733e8816.

[29] A. A. Aliero, B. S. Adebayo, H. O. Aliyu, A. G. Tafida, B. U. Kangiwa, and N. M. Dankolo, "Systematic Review on Text Normalization Techniques and its Approach to Non-Standard Words," International Journal of Computer Applications, vol. 185, no. 33, pp. 44–55, Sep. 2023, doi: 10.5120/ijca2023923106.

[30] L. Hui and M. Belkin, "Evaluation of Neural Architectures Trained with Square Loss vs Cross-Entropy in Classification Tasks," *arXiv preprint arXiv:2006.07322*, Oct. 2021, doi: 10.48550/arXiv.2006.07322.

[31] S. Pal, M. Chang, and M. F. Iriarte, "Summary Generation Using Natural Language Processing Techniques and Cosine Similarity," *in Intelligent Systems Design and Applications*, 2022, vol. 418, pp. 485–497, doi: 10.1007/978-3-030-96308-8_47.

[32] M. T. Rustam, M. Muhatri, dan A. B. Nasution, "Design and Construction of QA (Question Answering) System Based on Artificial Intelligence Using FAISS and Mixtral Language Model," *IT Journal*, vol. 13, no. 1, pp. 44–55, 2025. [Online]. https://upu-journal.potensi-utama.org/index.php/itjournal/article/view/370?utm_source=chatgpt.com

[33] Z. Tang, H. Fan, X. Gu, J. Zhou, H. Ma, A. V. Vasilakos, and B. Li, "Enabling efficient and accurate semantic search over encrypted cloud data," *Information Sciences*, vol. 719, Nov. 2025, Art. no. 122437, doi: 10.1016/j.ins.2025.122437.

[34] A. Ranjan and M. Ravinder, "Text Extraction from Blurred Images through NLP-based Post-processing," *in A Handbook of Computational Linguistics: Artificial Intelligence in Natural Language Processing*, 2024, pp. 285–300, doi: 10.2174/9789815238488124020 1.

[35] J. Junadhi, A. Agustin, L. Efrizoni, F. Okmayura, D. R. Habibie, and M. Muslim, "Improving Evaluation Metrics for Text Summarization: A Comparative Study and Proposal of a Novel Metric," *Journal of Applied Data Sciences*, vol. 6, no. 2, Jun. 2025, doi: 10.47738/jads.v6i2.547.

[36] M. Ghassemiazghandi, "An Evaluation of ChatGPT's Translation Accuracy Using BLEU Score," *Theory and*

Practice in Language Studies, vol. 14, no. 4, pp. 985–994, Apr. 2024, doi: 10.17507/tpls.1404.07.

[37] S. Kumar and A. Solanki, "ROUGE-SS: A New ROUGE Variant for Evaluation of Text Summarization," *Authorea*, Jul. 20, 2023, doi: 10.22541/au.168984209.92955863/v1.