MJI
Media Jurnal Informatika
Vol. 18, No. 1, June 2026
p-issn : 2088-2114, e-issn : 2477-2542

AI Persona-Based Student Counseling Chatbot Using Large Language Model, RAG, and Prompt Engineering

Vina Zahrotun Nazah [a,1,*], Rully Pramudita [a,2]

[a] Informatics Engineering Study Program, Faculty of Informatics and Design, Bina Insani University, Jl. Raya Siliwangi No.6, RT.001/RW.004, Sepanjang Jaya, Kec. Rawalumbu, Kota Bks, Jawa Barat 17114
[1] vinazahrotunnazah@gmail.com*; [2] rullypramudita@binainsani.ac.id ;
coresponding author

ARTICLE INFO

Article history
Received 2026-06-22
Revised 2026-06-29
Accepted 2026-06-29

Keywords
AI Persona
Student Counseling Chatbot
Large Language Model
Retrieval-Augmented Generation
Prompt engineering
LLM-as-a-Judge

ABSTRACT

Chatbots are increasingly used in student counseling services because they offer easy access, fast responses, and flexible availability. However, conventional chatbots often produce generic responses, have limited contextual understanding, and provide insufficient emotional support. This study aims to develop an AI Persona-based student counseling chatbot using a Large Language Model (LLM), Retrieval-Augmented Generation (RAG), and prompt engineering to generate relevant, contextual, and empathetic responses. The study uses a Research and Development (R&D) approach with the CRISP-DM framework. The system uses Gemini 2.5 Flash as the generative model, multilingual-e5-small as the embedding model, and FAISS as the vector index. Four institutional documents and campus service data are processed through chunking, embedding, and semantic retrieval. Evaluation was conducted using LLM-as-a-Judge on 45 scenarios, expert validation by a counselor, and User Acceptance Testing (UAT) with 20 students. The LLM-as-a-Judge evaluation produced an average score of 4.47 out of 5, with the highest score in Context Relevance at 4.70. Expert validation showed that all 45 chatbot responses were categorized as Suitable, with an overall suitability percentage of 100%, although the counselor noted that several responses still tended to be repetitive and showed similar response patterns. UAT achieved 91% user acceptance in the very good category, with naturalness and empathy as the highest indicator at 95%. These results indicate that integrating LLM, RAG, and prompt engineering can improve chatbot response quality without fine-tuning, although further development is needed in ablation testing, response variation improvement, multimodal document support, local model deployment, and retrieval mechanism refinement.

1. Introduction

University students are in a transitional stage toward adulthood and often face various academic and non-academic challenges during their studies. In addition to being required to achieve academic performance, students also encounter problems such as adaptation difficulties in the campus environment, time management, academic pressure, career planning, social relationships, and psychological issues that may affect their well-being and academic success [1], [2]. Therefore, guidance and counseling services are important facilities provided by higher education institutions to help students obtain support, direction, and assistance in dealing with these problems [3].

Along with the development of digital technology, various educational institutions have begun to use chatbots as supporting media for counseling services. Chatbots offer advantages such as 24-hour availability, ease of access, and the ability to provide fast responses to users [4]. In addition, several studies show that some

individuals feel more comfortable expressing complaints or personal problems to artificial intelligence-based systems due to anonymity, accessibility, and reduced concerns about social judgment [5], [6]. These conditions indicate that chatbots have the potential to serve as initial companion media in student counseling services.

Universitas Bina Insani has implemented a counseling chatbot service through the Suara BiU feature available on the student affairs portal. The service is intended to support student counseling activities by providing an initial access point for users to express their concerns or obtain information about available counseling services. The existing chatbot adopts a rule-based Natural Language Processing (NLP) approach, with intent classification as its primary mechanism. In this mechanism, user queries are classified into predefined categories to generate corresponding responses. However, based on system observation and exploration, the existing chatbot still has several limitations, particularly in providing specific responses, preserving conversational context, and demonstrating empathy toward users' conditions. In certain conversational scenarios, the generated responses remain relatively generic and do not fully accommodate users' individual needs. These limitations indicate the need for the development of a more contextual and adaptive chatbot system that can generate responses more appropriately aligned with the nature of student counseling services.

The development of Large Language Models (LLMs) opens opportunities to improve the quality of conversational systems through their ability to understand natural language, maintain conversational context, and generate more natural responses compared with conventional chatbot approaches [7]. Various studies have shown the potential use of LLMs in mental health and counseling domains. Chen et al. developed a Structured Dialogue System (SuDoSys) based on Qwen2-7B that improves counseling conversation coherence through a stage-aware approach [8]. Zhang and Luo developed SOULSPEAK by integrating Retrieval-Augmented Generation (RAG) and a dialogue memory mechanism to improve the quality of psychotherapeutic responses [9]. Meanwhile, Guo et al. showed that a combination of prompt engineering and Retrieval-Augmented Generation can produce more relevant and contextual responses without requiring model fine-tuning [10].

Although these studies show promising results, most of them still focus on general mental health or psychotherapy contexts. In addition, many studies rely on fine-tuning processes or complex architectures that require relatively large computational resources [8], [9]. On the other hand, research that specifically develops LLM-based chatbots for student counseling services in higher education with a lighter and easier-to-implement approach remains limited. This research gap indicates the need to develop a system capable of generating more specific, contextual, and empathetic responses according to the characteristics of students' problems without requiring model retraining.

Based on these issues, this study develops an AI Persona-based student counseling chatbot using a Large Language Model with a Retrieval-Augmented Generation (RAG) approach and prompt engineering. Unlike common implementations of RAG-based LLMs, which primarily focus on improving information accuracy, the AI Persona in this study is represented through a communication pattern designed using prompt engineering. This design enables the chatbot to maintain supportive, empathetic, non-judgmental, and contextually appropriate responses for students' initial counseling support needs. The system utilizes institutional documents and campus service information as additional knowledge sources, supported by the Gemini 2.5 Flash generative model integrated with a retrieval mechanism based on multilingual-e5-small and FAISS. The chatbot is developed as an initial support medium to provide assistance and information to students, rather than as a substitute for professional counselors in counseling processes or psychological decision-making.

The contribution of this study lies in the development of an AI Persona-based student counseling chatbot designed to generate more natural, contextual, and empathetic responses through the use of a Large Language Model. In addition, this study applies a combination of Retrieval-Augmented Generation (RAG) and prompt engineering without fine-tuning, making it more efficient to implement. The quality of the generated responses is evaluated using the LLM-as-a-Judge approach and expert validation by a counselor, while User Acceptance Testing (UAT) is conducted to assess user acceptance of the developed system.

2. Method

This study is a Research and Development (R&D) study that aims to develop a student counseling chatbot based on a Large Language Model (LLM) using Retrieval-Augmented Generation (RAG) and prompt engineering. The CRISP-DM framework is adopted because the chatbot development process does not only involve software implementation, but also includes data- and knowledge-oriented stages, such as document collection, preprocessing, chunking, embedding, semantic retrieval, response modeling, and LLM performance evaluation. The AI Persona in this study is not implemented as a visual avatar, but rather as a chatbot communication pattern designed through prompt engineering. This design enables the system to provide responses that are more empathetic, specific, contextual, and appropriate to the needs of initial student counseling support.

2.1 Type and Approach of Research

This study uses a Research and Development (R&D) approach. The development process follows the CRISP-DM framework, which consists of Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

2.2 Object and Scope of Research

The research object is a student counseling companion chatbot developed to support student affairs services at Bina Insani University. The system focuses on Indonesian text-based interactions and is designed to provide initial support through contextual, empathetic, and relevant responses without performing psychological diagnosis, mental health screening, or medical decision-making.

2.3 Data Collection Techniques

The research data were obtained through observations of the chatbot currently used in the Suara BiU service, interviews with the student affairs unit to identify system requirements, and a literature review related to Large Language Models, Retrieval-Augmented Generation (RAG), prompt engineering, and counseling chatbots. In addition, this study utilized four institutional documents, Biro Scholarship Announcement, course registration announcements, payment procedure guidelines, and chatbot information documents, as well as four categories of campus service data covering student affairs, academic services, financial services, and library services as knowledge sources in the RAG mechanism. Both the institutional documents and service data were stored in a database and managed through create, read, update, and delete (CRUD) features, allowing the information to be updated according to institutional needs.

2.4 Tools and Materials Used

The main generative model used is Gemini 2.5 Flash, accessed through an API because it supports large-scale processing with low latency [11]. Semantic representation of documents and user questions uses multilingual-e5-small as a multilingual embedding model [12]. Similarity-based document search is performed using FAISS as the vector index [13]. The backend is developed using FastAPI, the testing interface uses Streamlit, data storage uses PostgreSQL, and the actual implementation is integrated into a React-based interface in the student affairs system.

Table 1 presents the main configuration used in the implementation of Retrieval-Augmented Generation (RAG).

Table 1. RAG Implementation Parameters

Parameter	Value
Generative Model	Gemini 2.5 Flash
Embedding Model	multilingual-e5-small
Vector Database	FAISS
Chunk Size	800 characters
Chunk Overlap	150 characters
Retrieval Method	Semantic Retrieval
Top-K Retrieval	3
Similarity Threshold	0,80
Similarity Metric	Cosine Similarity

These parameters are used to optimize the context retrieval process so that the information provided to the generative model remains relevant to the user's question.

2.5 Research Procedures or Stages

The research stages follow CRISP-DM. In the Business Understanding stage, problems and system requirements are identified. The Data Understanding and Data Preparation stages include collecting, cleaning, chunking, embedding, and storing documents in FAISS. The Modeling stage integrates RAG, prompt engineering, and Gemini 2.5 Flash. The Evaluation stage uses LLM-as-a-Judge, expert validation by a counselor, and UAT, while Deployment is carried out by integrating the system into the student affairs service through an API.

Figure 1 shows the system architecture developed in this study.

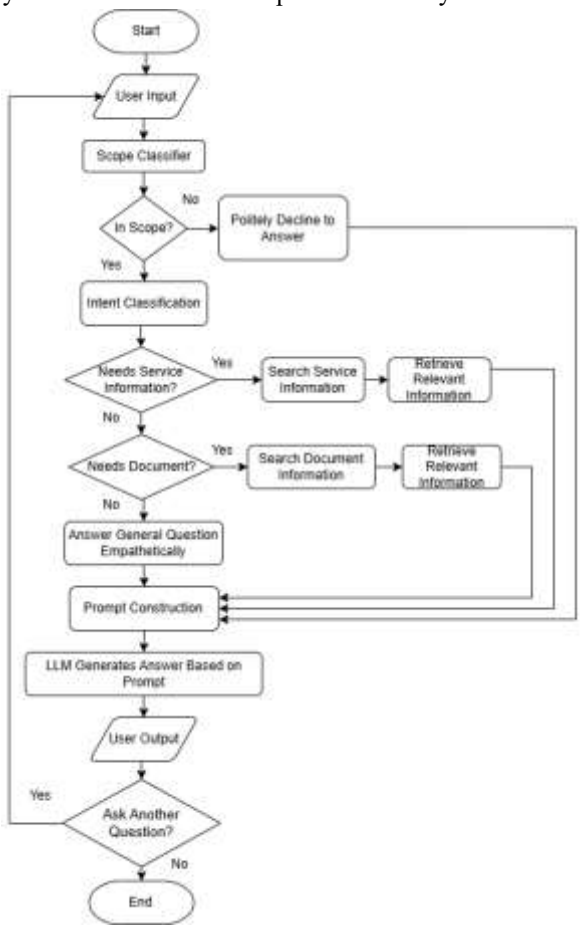

Figure 1. System Architecture

Figure 1 shows the system flow, starting from user input, scope checking, intent classification, document and campus service retrieval, threshold checking, prompt construction based on persona and conversation history, and response generation by Gemini 2.5 Flash.

2.6 Data Analysis Techniques

Response quality analysis was conducted using the LLM-as-a-Judge approach with Llama 3.3 70B Versatile as the evaluator. This approach was used because it enables structured assessment of open-ended conversational responses and has been reported to show good alignment with human judgment [14]. The evaluation was conducted on 45 test scenarios covering general conversation, Retrieval-Augmented Generation (RAG), and campus service inquiries. The metrics used included Empathy, Context Relevance, Specificity, Answer Relevance, and Faithfulness. The evaluator assessed each response using a predefined rubric on a 1–5 scale, based on the user question, chatbot response, and retrieved context when available. To complement the automated evaluation, expert validation was conducted by a counselor as a form of human assessment. The counselor reviewed each chatbot response holistically and provided categorical validation in the form of "Suitable" or "Not Suitable", along with notes when necessary. This validation aimed to ensure that the chatbot

responses were appropriate for the context of initial student counseling support. In addition, user acceptance was analyzed through User Acceptance Testing (UAT) involving 20 students using a 1–5 Likert scale and open-ended questions. However, due to the limited number of respondents, the UAT results were interpreted as an initial indication of user acceptance rather than as a basis for broad generalization.

3. Results and Discussion

This section presents the implementation results of the AI Persona-based student counseling chatbot and the evaluation results of response quality, expert validation, and user acceptance. The discussion focuses on system implementation, LLM-as-a-Judge evaluation results, counselor expert validation, User Acceptance Testing (UAT) results, and analysis of the findings.

3.1 Presentation of Research Results

The AI Persona-based student counseling chatbot was successfully developed by integrating a Large Language Model (LLM), Retrieval-Augmented Generation (RAG), and prompt engineering. The system receives user input in text form, processes the question, retrieves relevant context from institutional documents or campus service information when needed, and then generates responses using Gemini 2.5 Flash. Prompt engineering in the system is applied through three prompt templates according to the routing results, namely a campus service data prompt, a RAG document prompt, and a general prompt. The campus service data prompt is used when the user's question is related to campus service information, such as contact details, service hours, or addresses, so the response is limited to the available service data. The RAG document prompt is used when the question requires information from institutional documents, with instructions for the model to answer only based on the retrieved context and not add information beyond the source. Meanwhile, the general prompt is used for general conversations or student venting, with a casual, empathetic, supportive, and non-judgmental response style while still considering conversation history.

The RAG implementation is used to help the chatbot answer questions that require specific information from knowledge sources. Documents are processed through chunking, embedding using the multilingual-e5-small model, and storage in FAISS. When a user asks a question, the system retrieves the top three contexts based on cosine similarity and applies a similarity threshold of 0.80 to determine context relevance before including it in the prompt. This approach is used because RAG enables generative models to utilize external knowledge sources so that the generated responses can be more relevant and context-based [15].

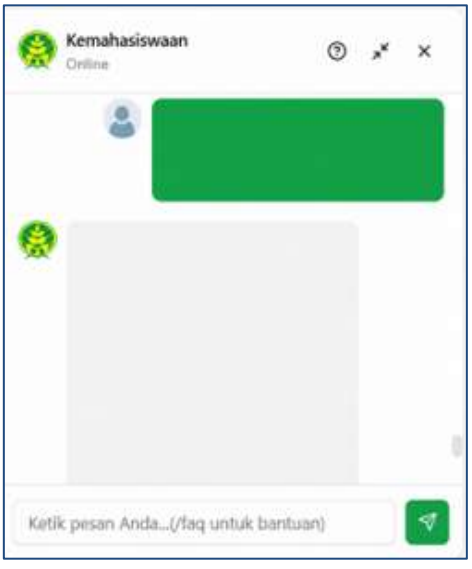

Figure 2. Student Counseling Chatbot Interface

Response quality evaluation was conducted using the LLM-as-a-Judge approach with Llama 3.3 70B Versatile as the evaluator. The test involved 45 scenarios consisting of 15 general conversation scenarios, 20 Retrieval-Augmented Generation (RAG) scenarios, and 10 campus service scenarios. The distribution of test scenarios is shown in Table 2.

Table 2. Distribution of Test Scenarios

Evaluation Category	Number of Scenarios
General Conversation	15

Evaluation Category	Number of Scenarios
Retrieval-Augmented Generation (RAG)	20
Campus Services	10
Total	45

The LLM-as-a-Judge evaluation results for Gemini 2.5 Flash responses are shown in Table 3.

Table 3. LLM-as-a-Judge Evaluation Results for Gemini 2.5 Flash

Evaluation Category	Metric	Average Score
General Conversation	Empathy	4.53
General Conversation	Context Relevance	4.70
General Conversation	Specificity	4.10
Retrieval-Augmented Generation (RAG)	Answer Relevance	4.40
Retrieval-Augmented Generation (RAG)	Faithfulness	4.62
Campus Services	Answer Relevance	4.50
Average		**4.47**

Based on Table 3, the chatbot obtained an overall average score of 4.47 on a 1–5 scale. The highest score was obtained in the Context Relevance metric at 4.70, followed by Faithfulness at 4.62 and Empathy at 4.53. These results indicate that the system can generate responses that are relevant to the conversational context, sufficiently consistent with the information sources used, and still demonstrate empathy in interactions with users. Meanwhile, the Specificity score of 4.10 was the lowest, although it remains in the good category. This indicates that the chatbot responses are sufficiently specific but can still be improved, especially in conversation scenarios that require more detailed guidance.

In addition to the LLM-as-a-Judge evaluation, expert validation was conducted by a counselor to provide human assessment of the chatbot responses. This validation was used to ensure that the responses generated by the chatbot were appropriate for the context of initial student counseling support. Unlike the LLM-as-a-Judge evaluation, which used numerical scores for each metric, the counselor validation was conducted holistically using categorical assessment, namely “Suitable” and “Not Suitable”.

The counselor evaluated the responses based on the characteristics of each test category. For general conversation scenarios, the responses were considered suitable when they demonstrated empathy, contextual relevance, non-judgmental language, and sufficiently specific guidance. For RAG scenarios, the responses were considered suitable when they were relevant to the user’s question and faithful to the retrieved context. Meanwhile, for campus service scenarios, the responses were considered suitable when they provided clear and relevant information according to the available service data. The results of the counselor validation are shown in Table 4.

Table 4. Expert Validation Results by Counselor

Evaluation Category	Number of Scenarios	Suitable	Not Suitable	Percentage
General Conversation	15	15	0	100%
Retrieval-Augmented Generation (RAG)	20	20	0	100%
Campus Services	10	10	0	100%
Total	**45**	**45**	**0**	**100%**

Based on Table 4, all 45 chatbot responses were categorized as Suitable by the counselor, resulting in an overall suitability percentage of 100%. This result indicates that the chatbot responses were considered appropriate from the perspective of a counseling expert for use as initial student counseling support. However, the counselor also noted that some chatbot responses tended to be repetitive and showed similar response patterns across several conversations. This note suggests that although the responses were generally suitable, further refinement is still needed to improve response variation and reduce repetitive wording in future development.

In addition to automatic evaluation, testing was also conducted using User Acceptance Testing (UAT) involving 20 students from Bina Insani University. The UAT instrument consisted of three closed-ended questions using a 1–5 Likert scale and three open-ended questions. In this evaluation, Chatbot A refers to the existing Suara BiU chatbot, which uses a rule-based NLP approach with intent classification to generate responses based on predefined categories. Meanwhile, Chatbot B refers to the proposed AI Persona-based chatbot developed in this study using LLM, RAG, and prompt engineering. The use of a Likert-scale survey is relevant for measuring user perceptions and satisfaction with AI-based chatbots because it allows respondents to provide graded assessments of interaction experience aspects [16]. The UAT results are shown in Table 5.

Table 5. User Acceptance Testing (UAT) Results

Indicator	Total Score	Percentage	Category
Chatbot B provides more specific answers than Chatbot A	87	87%	Very Good
Chatbot B better understands the question context than Chatbot A	91	91%	Very Good
Chatbot B responses feel more natural and empathetic than Chatbot A	95	95%	Very Good
Average Score		**91%**	**Very Good**

The UAT results show that the developed chatbot obtained an average user acceptance percentage of 91%, which falls into the very good category. The highest indicator was response naturalness and empathy at 95%, while the answer specificity indicator obtained 87%. These results show that users perceived the developed chatbot as more natural, empathetic, and better able to understand context than the previous chatbot.

3.2 Analysis of Findings

The LLM-as-a-Judge evaluation results show that the developed chatbot system obtained an average score of 4.47 on a 1–5 scale. This value indicates that the integration of a Large Language Model, Retrieval-Augmented Generation, and prompt engineering can produce good-quality responses in the context of student counseling services. The highest score was obtained in the Context Relevance metric at 4.70. This indicates that the system can generate responses that are aligned with the user's question context. This capability is influenced by the use of conversation history in prompt construction and the retrieval mechanism that provides additional context from institutional documents and campus service information.

In the RAG category, the Faithfulness metric obtained a score of 4.62. This value indicates that the generated responses are sufficiently consistent with the information contained in the knowledge sources used. This finding shows that the RAG mechanism helps the model reduce full dependence on the LLM's internal knowledge by providing relevant external information as additional context. This is in line with the concept of Retrieval-Augmented Generation, which utilizes external knowledge bases to improve LLM response quality, particularly in reducing the limitations of the model's internal knowledge, improving answer relevance, and supporting domain-specific knowledge-based tasks [15], [17].

The Empathy metric obtained a score of 4.53, indicating that the system can provide responses that are sufficiently empathetic toward the user's condition. This result shows that prompt engineering applied in the system plays a role in shaping the chatbot's communication style to be more supportive, non-judgmental, and aligned with the character of a student counseling companion. In this study, prompts are not only used to guide model answers, but also to shape the chatbot persona so that the generated responses are more natural and attentive to users' emotional conditions. This approach is consistent with research on prompt engineering, which states that prompts can be used to guide model behavior without changing the model's core parameters [18].

Although all metrics obtained good scores, the Specificity score of 4.10 was the lowest in the LLM-as-a-Judge evaluation. This indicates that in several scenarios, chatbot responses can still be improved to be more specific to the user's condition or needs. This condition may be influenced by variations in question forms, the limited coverage of documents in the knowledge base, and the nature of counseling conversations, which are often open-ended and require deeper understanding of the user's situation.

The User Acceptance Testing (UAT) results support the automatic evaluation results. The average user acceptance of 91% shows that the developed chatbot was rated very good by respondents. The naturalness and empathy indicator obtained the highest score at 95%, indicating that users experienced an improvement in interaction quality with the developed chatbot. This finding is consistent with Borsci et al., who emphasized that chatbot evaluation needs to consider interaction experience quality, including aspects that are not always covered in usability measurements for non-conversational systems [19].

The context understanding indicator obtained 91%, while the answer specificity indicator obtained 87%. This pattern is consistent with the LLM-as-a-Judge results, where Context Relevance was the highest score, while Specificity was the lowest. Thus, both automatic evaluation and user evaluation show the same tendency: the system can understand conversational context well, but still needs improvement in providing more detailed and specific answers.

Based on respondents' descriptive feedback, most users stated that the developed chatbot provided responses that were easier to understand, more natural, and more aligned with the question context than the previous chatbot. Some respondents also considered the chatbot to feel more like a conversation with a human because of its less rigid and more empathetic response style. This is consistent with the finding that human-like cues, such as conversational tone, empathy, and adaptability, can increase social presence, trust, and user experience in AI-based chatbots [20].

Several respondents also provided development suggestions, such as adding a voice feature, increasing response variety, improving the interface design, and expanding the knowledge base. These suggestions indicate that the system has been well accepted but still has room for development to provide a more comprehensive interaction experience.

3.3 Implications of the Results

The results of this study show that the combination of a Large Language Model, Retrieval-Augmented Generation, and prompt engineering can be used to improve the quality of student counseling chatbots without fine-tuning. This approach provides a lighter alternative for system development because response quality is improved through prompt construction, the use of conversational context, and external knowledge sources. This is consistent with the concept of prompt engineering, which enables control over model instructions and output style without changing the model's core parameters [18], as well as the RAG approach, which utilizes external knowledge to improve response relevance in knowledge-based tasks [15], [17].

Practically, the developed chatbot can be used as an initial companion medium for students to express concerns, obtain information, and receive guidance related to campus services. The system is not intended to replace the role of professional counselors, but it can help provide faster and more accessible initial assistance. The UAT result of 91% shows that users gave positive evaluations of the system, particularly its ability to provide responses relevant to the problems presented, demonstrate empathy toward the user's condition, and maintain conversational context consistently. This finding indicates that students need not only informative answers, but also interactions that make them feel understood and emotionally supported. Therefore, user experience quality is an important factor in the application of AI-based chatbots in student assistance services [19], [16].

Academically, this study shows that an LLM-based AI Persona can be applied in the context of student counseling services in higher education. The persona formed through prompt engineering helps the model generate responses that are more natural, supportive, and empathetic according to user needs in consultation situations. These results indicate that persona configuration not only maintains communication style consistency but also helps the model provide more contextual responses to various student problems, ranging from academic pressure to the need for campus service information. This finding also supports the view that human-like cues, such as conversational tone, empathy, and adaptability, can increase social presence, trust, and user experience in AI-based chatbots [20]. Thus, the results of this study can serve as a basis for developing LLM-based chatbots for education, counseling, and student support services in higher education environments.

3.4 Limitations of the Study

This study has several limitations that need to be considered. First, the developed system still depends on an external API service, namely Gemini 2.5 Flash as the main generative model. This dependence means that system performance may be affected by service availability, quota limits, rate limits, latency, and usage policies from the service provider. In addition, API use also needs to consider cost and access stability when the system is used on a larger scale.

Second, the knowledge sources in the Retrieval-Augmented Generation (RAG) mechanism are still limited to text-based documents. The system has not been designed to process information in the form of images, scanned documents, complex tables, or other visual formats. This limitation may affect the completeness of the knowledge base, especially when important information in institutional documents is presented in non-text forms. Therefore, future development may consider support for multimodal document processing or Optical Character Recognition (OCR) integration to expand the system's knowledge coverage.

Third, the number of LLM-as-a-Judge test scenarios is still limited to 45 scenarios, consisting of general conversation, Retrieval-Augmented Generation (RAG), and campus services. This number does not fully represent all possible variations of student conversations in real conditions. In addition, the number of general conversation scenarios is smaller than other categories because responses in this category tend to be longer and must adjust to token limitations in the evaluation process.

Furthermore, the developed system still focuses on Indonesian text-based interactions and does not yet support voice input or other media. The system is also not designed to perform psychological diagnosis, mental health screening, or replace the role of professional counselors. Future development can focus on improving

the model's ability to provide more empathetic and more specific responses according to user context and needs, optimizing answer quality by refining the retrieval mechanism, and deploying the model locally (on-premise) to improve user data security and privacy while reducing dependence on external API services.

4. Conclusion

This study successfully developed a student counseling chatbot based on a Large Language Model (LLM) using Retrieval-Augmented Generation (RAG) and prompt engineering. The developed system is designed as an initial companion for student counseling services to generate more relevant, contextual, and empathetic responses. By utilizing Gemini 2.5 Flash as the generative model, multilingual-e5-small as the embedding model, and FAISS as the vector index, the chatbot can use institutional documents and campus service information as additional knowledge sources in the response construction process.

The evaluation using the LLM-as-a-Judge approach shows that the chatbot obtained an average score of 4.47 on a 1–5 scale. The highest score was obtained in the Context Relevance metric at 4.70, indicating that the system can generate responses that align with the conversational context. In addition, expert validation by a counselor showed that all 45 chatbot responses were categorized as Suitable, with an overall suitability percentage of 100%. This result indicates that the responses were considered appropriate from the perspective of a counseling expert for use as initial student counseling support. However, the counselor also noted that several responses still tended to be repetitive and showed similar response patterns across several conversations. Furthermore, the User Acceptance Testing (UAT) results involving 20 students showed a user acceptance level of 91%, which falls into the very good category. However, the UAT results should be interpreted as an initial indication of user acceptance due to the limited number of respondents.

Based on these results, the implementation of RAG and prompt engineering in an LLM-based chatbot can be considered a promising alternative for improving the quality of initial student assistance services without requiring fine-tuning. Nevertheless, this study still has several limitations. First, this study has not conducted an ablation study to separately measure the contribution of each component, such as LLM without RAG, LLM with RAG without persona, LLM with persona without RAG, and LLM with RAG and persona. Therefore, the results cannot yet determine the individual contribution of each component to the overall response quality. Second, the system still depends on external API services and text-based knowledge sources. Third, the evaluation scenarios and UAT respondents were still limited. Future research can focus on ablation testing, local model deployment, knowledge base expansion, support for multimodal documents, and refinement of prompt engineering strategies and retrieval mechanisms to generate responses that are more specific, varied, relevant, and aligned with the context of students' problems.

Declarations

Author contribution. Vina Zahrotun Nazah contributed to research conceptualization, system development, data collection, evaluation, analysis, and manuscript preparation. Rully Pramudita contributed to supervision, methodology review, validation, and manuscript revision.
Funding statement. This research received no external funding.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this article.

Data and Software Availability Statements

The data used in this study consist of four knowledge-source documents for the Retrieval-Augmented Generation (RAG) mechanism, namely the Biro scholarship announcement, KRS filling announcement, payment procedure document, and chatbot-related information document. In addition, this study uses campus service data covering four categories: student affairs services, academic services, financial services, and library services. The evaluation data consist of 45 LLM-as-a-Judge test scenarios and User Acceptance Testing (UAT) responses from 20 students of Bina Insani University.

The data used are not publicly available because they are related to internal institutional documents and user response data. However, the data may be provided by the corresponding author upon reasonable request and with institutional approval. The chatbot prototype and related implementation files are not publicly archived and may be accessed upon request with institutional approval.

References

[1] A. N. Rahman, "PENERAPAN BK DI PERGURUAN TINGGI DALAM MENGATASI KEJENUHAN BELAJAR MAHASISWA MENGGUNAKAN TEKNIK SELF HEALING," *Jurnal Pelayanan Bimbingan dan Konseling*, vol. 7, no. 1, Feb. 2024, Accessed: Mar. 07, 2026. [Online]. Available: https://ppjp.ulm.ac.id/journals/index.php/jpbk/article/view/14805

[2] R. N. Gultom, E. Yakub, K. Khadijah, P. Studi, B. Konseling, and U. Riau, "Penyesuaian Diri Mahasiswa Baru Tahun Pertama Fkip UNRI (Jurusan Ilmu Pendidikan) dalam Menghadapi Kehidupan

Perkuliahan," *Jurnal Pendidikan Tambusai*, vol. 7, no. 2, pp. 15242–15249, Aug. 2023, doi: 10.31004/JPTAM.V7I2.8798.

[3] Sukarman and Aminullah, "Problematika Bimbingan dan Konseling pada Perguruan Tinggi," *JISHUM (Jurnal Ilmu Sosial dan Humaniora)*, vol. 3, no. 4, pp. 671–680, 2025, doi: https://doi.org/10.57248/jishum.v3i4.644.

[4] M. Fahmi Ajiz, M. Faza, S. Ramadan, H. Dzalfa Mutia, and P. D. Yanuari, "Pengembangan Aplikasi Chatbot Informasi Akademik Berbasis Web Menggunakan Metode Artificial Intelligence Markup Language (AIML)," *Media Jurnal Informatika*, vol. 15, no. 2, pp. 143–148, Dec. 2023, doi: 10.35194/MJI.V15I2.3316.

[5] A. Nurkhairani, O. D. Arwansyah, and R. Ginting, "Tabularasa : Jurnal Ilmiah Magister Psikologi Menemukan Kenyamanan dalam Algoritma : Fenomena Curhat ke AI dalam Era Digital Finding Convenience in Algorithms : The Phenomenon of Venting to AI in the Digital Age," vol. 7, no. 2, pp. 80–89, 2025, doi: 10.31289/tabularasa.v7i2.5801.

[6] G. Yoseppin, P. A. M. Nagita Dewi, and Y. K. Purba, "Fenomena Chatbot AI Sebagai Teman Curhat: Implikasi Pada Hubungan Antarpribadi di Era Digital," *Calathu: Jurnal Ilmu Komunikasi*, vol. 7, no. 1, pp. 45–53, May 2025, doi: 10.37715/CALATHU.V7I1.5376.

[7] S. Minaee *et al.*, "Large Language Models: A Survey," Mar. 2025, Accessed: Mar. 12, 2026. [Online]. Available: http://arxiv.org/abs/2402.06196

[8] Y. Chen *et al.*, "Structured Dialogue System for Mental Health: An LLM Chatbot Leveraging the PM+ Guidelines," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 15170 LNAI, pp. 262–271, 2024, doi: 10.1007/978-981-96-1151-5_27.

[9] X. Zhang and Z. Luo, "Advancing Conversational Psychotherapy: Integrating Privacy, Dual-Memory, and Domain Expertise with Large Language Models," 2024, [Online]. Available: http://arxiv.org/abs/2412.02987

[10] Q. Guo, J. Tang, W. Sun, H. Tang, Y. Shang, and W. Wang, "SouLLMate: An Application Enhancing Diverse Mental Health Support with Adaptive LLMs, Prompt Engineering, and RAG Techniques," Oct. 2024, Accessed: Jun. 22, 2026. [Online]. Available: https://arxiv.org/pdf/2410.16322

[11] "Gemini 2.5 Flash | Gemini API | Google AI for Developers." Accessed: Jun. 22, 2026. [Online]. Available: https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash?hl=id

[12] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual E5 Text Embeddings: A Technical Report," Feb. 2024, Accessed: Jun. 22, 2026. [Online]. Available: https://arxiv.org/pdf/2402.05672

[13] M. Douze *et al.*, "the Faiss Library," *IEEE Trans. Big Data*, Jan. 2025, doi: 10.1109/TBDATA.2025.3618474.

[14] L. Zheng *et al.*, "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," *Adv. Neural Inf. Process. Syst.*, vol. 36, Jun. 2023, Accessed: May 25, 2026. [Online]. Available: https://arxiv.org/pdf/2306.05685

[15] Y. Gao *et al.*, "Retrieval-Augmented Generation for Large Language Models: A Survey," *Proceedings - 2024 Conference on AI, Science, Engineering, and Technology, AIxSET 2024*, pp. 166–169, Dec. 2023, doi: 10.1109/AIxSET62544.2024.00030.

[16] C. G. Møller, K. E. Ang, M. de Lourdes Bongiovanni, M. S. Khalid, and J. Wu, "Metrics of Success: Evaluating User Satisfaction in AI Chatbots," *ICAAI 2024 - Conference Proceedings of the 2024 8th International Conference on Advances in Artificial Intelligence*, pp. 168–173, Mar. 2025, doi: 10.1145/3704137.3704182.

[17] S. Wu *et al.*, "Retrieval-Augmented Generation for Natural Language Processing: A Survey," Jul. 2024, Accessed: Jun. 22, 2026. [Online]. Available: https://arxiv.org/pdf/2407.13193

[18] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, "A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications," Feb. 2024, Accessed: Jun. 22, 2026. [Online]. Available: https://arxiv.org/pdf/2402.07927

[19] S. Borsci *et al.*, "The Chatbot Usability Scale: the Design and Pilot of a Usability Scale for Interaction with AI-Based Conversational Agents," *Personal and Ubiquitous Computing 2021 26:1*, vol. 26, no. 1, pp. 95–119, Jul. 2022, doi: 10.1007/S00779-021-01582-9.

[20] T. Liu *et al.*, "The Illusion of Empathy: How AI Chatbots Shape Conversation Perception," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 13, pp. 14327–14335, Nov. 2024, doi: 10.1609/aaai.v39i13.33569.