JURNAL INFOTEL
VOL. 17, NO. 4, NOVEMBER 2025, PP. 730–749. DOI:10.20895/INFOTEL.V17I4.1318

RESEARCH ARTICLE

Design and Evaluation of a CMS-Integrated Academic Chatbot Using Gemini AI

Bunga Laelatul Muna[1], Muhammad Lulu Latif Usman[2,*], Sudianto[3], and Bachtiar Herdianto[4]

[1,2,3]Department of Informatics, Telkom University, Purwokerto, Indonesia Department, Institution, 53144, Indonesia
[4]Lab STICC IMT atlantique, Brest, France

*Corresponding email: muhlulu@telkomuniversity.ac.id

Received: January 23, 2025; Revised: July 01, 2025; Accepted: August 20, 2025.

Abstract: Efficient academic services are a crucial component, and their implementation is the responsibility of higher education institutions. Telkom University Purwokerto (TUP) faces challenges in providing responsive academic services, especially in conventional online services. This research proposes the development and integration of an artificial intelligence-based academic chatbot, 'Akif', utilizing the Gemini 1.5 Flash model, which is linked to the institution's Content Management System (CMS). This integration enables the retrieval of real-time information and automatic updates to the model. The tuning and evaluation process was conducted using the BLEU metric, with a value of 0.88 being reached, indicating a fairly good level of agreement between the generated answers and the reference. Although the results are promising, the system still faces limitations, particularly the risk of hallucination, which is a common challenge with generative models. Additionally, the use of BLEU as an initial evaluation metric overlooks aspects of semantic depth and user satisfaction. This research contributes a modular integration framework between generative AI and institutional systems, and highlights its potential and limitations in academic service automation.

Keywords: academic services, chatbot, content management system (CMS), gemini, system integration

 ISSN: 2085-3688; e-ISSN: 2460-0997

1 Introduction

Every higher education institution is obligated to ensure the quality of academic services to facilitate the smooth operation of the educational and administrative processes for students [1]. At Telkom University Purwokerto (TUP), academic services are provided through two main modes: onsite and online. Online services currently only rely on WhatsApp communication, which is managed directly by academic staff. Based on interviews with relevant staff, it is known that this method is considered not optimal, due to the high workload associated with face-to-face services. As a result, online queries are often delayed or overlooked.

This condition indicates the limitations of conventional academic service systems, especially in terms of information management efficiency. One potential solution to address this challenge is the use of academic chatbots. A chatbot is an autonomous machine agent that can communicate with users in natural language, either through text, speech, or based on query commands [2]. The research of Uzoka et al. (2024) [3] and Putra et al. (2022) [4] proves that chatbots can help manage routine questions. Chatbots can be integrated into various platforms, one of which is Telegram, which is flexible in bot customization [5].

Along with the development of Natural Language Processing (NLP), Chatbots have evolved from rule-based systems to AI-based ones, specifically Large Language Models (LLMs) [6]. LLM can understand and generate text in natural language through deep learning [7, 8]. Some examples of LLMs that can be used for chatbots are ChatGPT [9, 10], BERT [10–12] and Gemini [13–15]. ChatGPT (Generative Pretrained Transformer) has the ability for interactive conversations, but the use of ChatGPT requires large computing resources and is prone to hallucinations [16, 17]. Then, Google's first model, BERT (Bidirectional Encoder Representations from Transformers), is a bidirectional-based model that can understand context more deeply than GPT. However, BERT is less suitable for interactive conversations [18, 19]. Alternatively, Google developed another model, Gemini, which is easy to use and capable of generating text to handle conversations interactively, like humans [20, 21]. Thus, Gemini will be very suitable for use as an academic chatbot model, but it will still face the challenge of hallucination [21]. Based on a study by the Gemini team (2023), hallucination is a significant challenge in LLM models because most models are trained only once and cannot update information in real-time [22].

To overcome these limitations, the researcher proposes integrating academic chatbots with the Content Management System (CMS) as a medium for chatbot information sources. A Content Management System supports the proper implementation of chatbots for academic services [23, 24]. CMS is a software application designed to efficiently create, manage, and publish digital content [25, 26]. At TUP, CMS serves as a database for storing important information related to academic services, including lecturer data, exam schedules, and other frequently updated details. The integration of a Chatbot with CMS allows the chatbot to access and deliver real-time information to students. This will ensure accuracy and efficiency in the delivery of information, optimizing academic services.

Researchers have conducted several studies to support the arguments presented in this research, as noted by Kumar (2024) [27] and Velasques et al. (2024) [28] who designed an academic chatbot but did not utilize Gemini AI and did not integrate it with a CMS. Research by Dimitrios et al. (2022) [29], examines the integration of chatbots with CMS. However, the study focuses more on sending user data through chatbots to CMS without using AI-based chatbots. This means that the chatbot in that research cannot accept dy-

namic data changes. Based on these studies, a gap is identified that highlights the need for further research on integrating Gemini AI with CMS for academic services, which can expedite information updates and provide more accurate and real-time responses.

Based on this identified gap, this research aims to develop and integrate the AI Gemini chatbot with the Content Management System (CMS) to help academic online services at Telkom University Purwokerto (TUP). With this integration, it is expected that the chatbot will receive data updates automatically through the CMS, eliminating the need for retuning and allowing students to access more accurate and up-to-date information directly.

The main contribution of this research is to provide a more efficient and responsive system for delivering real-time academic information, which not only reduces dependence on human staff but also increases student satisfaction in obtaining the information they need. The steps taken in this research include preprocessing the dataset again to ensure the cleanliness and readability of the information, which involves tokenization, removing stop words, and punctuation; adjusting the parameters of the Gemini model to set the level of variation in the chatbot response so it can provide relevant and natural answers to users; tuning the model using fine-tuning techniques so the chatbot can deliver responses that are more appropriate to the academic context and handle more complex questions; assessing the similarity between the answers generated by the chatbot and the expected answers using the BLEU Score; and finally changing the chatbot model into a Telegram bot and integrating it with the TUP academic CMS.

Through this approach, this research contributes to the development of a more efficient chatbot that can be integrated with the CMS for academic services at TUP, improving the accuracy of information provided to students, and reducing the workload of administrative staff.

2 Research Method

The steps for modeling and integrating a chatbot with an academic CMS for services are illustrated in Figure 1.

2.1 Data Collection

The data source for this research was obtained from interviews with academic services related to questions frequently asked by students, FAQs from the Merdeka Campus system, and Telkom University Purwokerto lecturer data. The total dataset used for model tuning consists of 250 patterns, which include both user and model tags. This research did not require a large amount of training data, as it could be added through the inferred model in the content management system being built. The sample data used in this research can be seen in Table 1.

2.2 Preprocessing

Preprocessing is the alteration of the data itself to produce more optimized performance [30]. In general, when using LLM models such as Gemini AI, preprocessing is already done automatically by the model [31]. However, to obtain more optimal model results, researchers need to preprocess the data again. Here were some preprocessing samples that

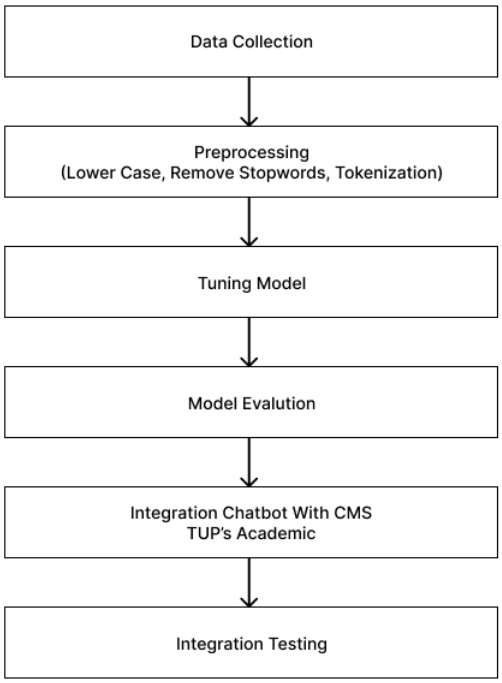

Figure 1: Research diagram.

had been done: Lower Casing, where all characters in the text were converted to lowercase to avoid the difference between uppercase and lowercase letters. Tokenization, where the text was broken into tokens or individual words using the Natural Word Tokenizer to help separate each word from the complete sentence. Then Remove Punctuation and Stopwords, where punctuation marks such as periods, commas, exclamation marks, and others were removed from each token to ensure only words remain, while removing stopwords involved eliminating common words that did not provide much information, such as "and," "or," "is," etc.

2.3 Gemini AI

Gemini AI is a collection of large language models (LLMs) hosted on Google Cloud's Vertex AI platform, featuring multimodal capabilities that enable the processing of text, images, audio, and code [22, 32]. As a generative AI model, Gemini can generate new data based on existing training data. With the encoder-decoder architecture, Gemini analyzes inputs through encoders to generate numerical representations that are then translated by decoders into outputs [33]. This architecture utilizes self-attention and multi-head attention mechanisms to understand the relationship between words in a sentence [34]. The architecture diagram can be seen in Figure 2.

The Gemini AI response flow is illustrated in Figure 3. The Gemini AI workflow process starts with the user asking a question, such as, "Where is the Language Center Room?". The

Table 1: Sample of student FAQs

Questions	Answer
How do I apply for a supervisor?	To apply for a supervisor, you can fill out the test form and submit it to the academic room at DC-205.
How to request a TAK/SKPI acc?	To get ACC TAK you can first choose the best 5 TAK points, then submit it to the guardian lecturer. If the guardian lecturer has not responded or has encountered issues, you can go directly to the academic room at Rek 401.
Where is the Language Center Room?	The Language Center Room is in Rek-401.
Who is dpp msib batch 6 at Telkom University Purwokerto?	The MSIB DPP for this batch is Mr. Dadiek, and his room is in Rec 405.
What is the structure of writing a proposal/manuscript/research thesis for the Final Project/TA?	The structure of writing the Final Project proposal can be accessed at the link: https://v.ff.it.telkom-pwt.ac.id/wp-content/uploads/2023/10/Panduan-Tugas-Akhir-revisi-26-April-2023-Edisi-Revisi-1.pdf

Table 2: MBKM question sample

No	Questions	Answer
1	What is MBKM?	MBKM is a program from the government for students to study and work during the lecture period.
2	What are the requirements for participating in MBKM?	One of the requirements for participating in the Independent Campus program is that student data must be registered with PDDikti and recorded at a tertiary institution under Kemendikbudristek.
3	What are the Kampus Merdeka programs?	There is an Independent Campus program from the Ministry of Education and Culture, the Ministry of Social Affairs, and BUMN.

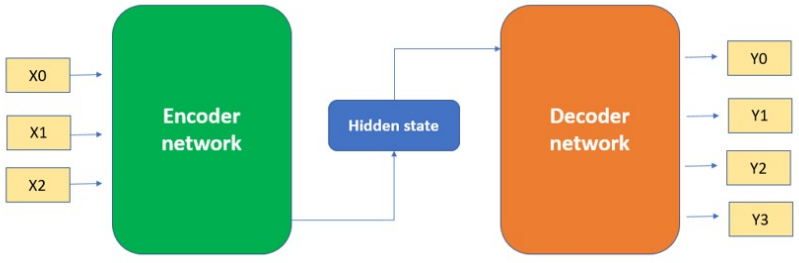

Figure 2: Encoder-Decoder architecture [34].

question then undergoes preprocessing, specifically tokenization, which breaks the text into word units, for example, [“where”, “language”, “cente”, “room”]. Next, the model searches for relevant answers based data train, such as “The Languange Center Room is in Rek 401” or “Languange Center room in Rek 405.” In the output sampling stage, the model evaluates the existing answer options, selects the most appropriate one, and finally produces the final answer, e.g., “The Languange Center Room is in Rek 401,” which is presented to the user.

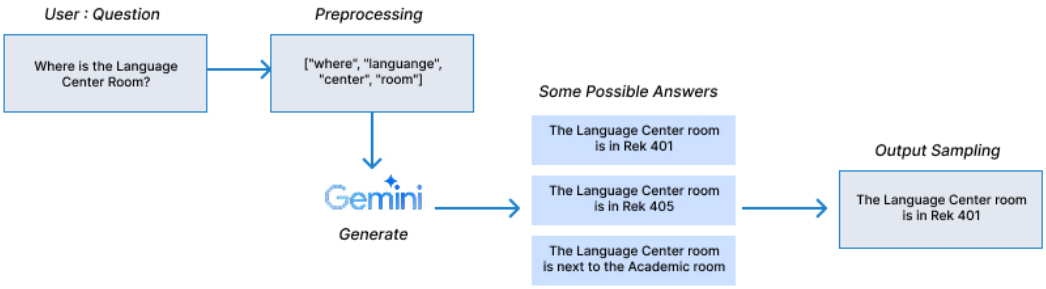

Figure 3: Illustration of the Gemini AIs response generation in this research.

2.4 Tuning Model

The tuning parameters for the Gemini 1.5 Flash model focus on four main variables: temperature, top-K, top-P, and Blockness [22]. Temperature controls the degree of novelty and variation in the output. Top-K limits token selection to a set of K tokens with the highest probability, thereby maintaining the focus of the response. Top-P (nucleus sampling) determines the cumulative probability threshold of selected tokens, thus regulating the balance between response variety and relevance. Blockness serves as a filter mechanism to prevent inappropriate output, but was disabled in this research to allow the model to respond to sensitive questions such as room and lecturer names. The selection of this configuration is based on research by Zhang et al. (2021) which showed that Temperature and Top-P can affect the level of hallucinations, and a low Temperature will reduce the hallucinations of the model [35]. An exploration of tuning parameters can be seen inTable 3

Table 3: Tuning parameter configuration

Scenario	**Temperature**	***Top-K***	***Top-P***	**Bleu**
Default	1.0	20	1.0	0.55
Low Temperature	0.6	20	0.6	0.71
Controlled Diversity	0.7	40	0.7	0.68
High Filtering	0.4	10	0.4	0.62

Initial testing results using BLEU scores showed that the Default scenario produced highly varied but less relevant responses, with the lowest BLEU score of 0.55, indicating

a tendency towards hallucination. In the Low Temperature scenario, the model responses became more focused and contextually appropriate, as evidenced by the increase in BLEU score to 0.71. The Controlled Diversity and High Filtering scenarios yield better results than the Default scenario, but are still inferior to the Low Temperature scenario. Considering the balance between relevance and response variety, as well as the reduction of hallucination, the Low Temperature scenario (Temperature 0.6, Top-K 20, Top-P 0.6, no Blockness) will be chosen as the optimal configuration for the academic chatbot.

2.5 Evaluation

The chatbot performance evaluation in this study utilized the BLEU score (Bilingual Evaluation Understudy) as the primary metric. The choice of BLEU was based on its suitability for evaluating short text outputs, such as responses from Q&A-based chatbots. BLEU was designed for automated testing, making it fast, cheap, and language-independent [36]. Although other metrics, such as ROUGE or METEOR, offer alternatives that consider recall or semantic proximity, BLEU was deemed adequate for an initial evaluation based on structure and content accuracy. BLEU score testing compared the expected answer in the dataset with the predicted answer from the bot. In this research, testing would be carried out using 20 test data sets and 20 validation data sets, divided into 250 data sets with a ratio of 3:1:1. The BLEU score evaluation results would be used to measure the model's accuracy in providing answers as expected. The BLEU Score is represented in Equations (1), (2), and (3). Table 4 display the test data or questions used in this research.

$$\text{Precision} = \frac{\text{Count of matching } n\text{-grams}}{\text{Count of candidate } n\text{-grams}}, \tag{1}$$

$$\text{BP} = \begin{cases} 1, & \text{if length}_{\text{candidate}} > \text{length}_{\text{reference}}, \\ e^{\left(1-\frac{\text{length}_{\text{candidate}}}{\text{length}_{\text{reference}}}\right)}, & \text{otherwise.} \end{cases} \tag{2}$$

$$\text{BLEU} = \text{BP} \times \exp\left(\sum(\text{precisions} \times \text{Weight})\right), \tag{3}$$

where BP represents *Brevity Penalty* which the step function computes based on the length of the candidate.

2.6 System Integration and Testing

The system integration stage in this research aimed to implement the developed model as a Telegram-based chatbot named Akif and integrate it with TUP's academic content management system (CMS). The purpose of this integration was to make the model usable in the form of a Telegram chatbot, enabling it to respond to changes in information on the CMS. Additionally, testing would be used to ensure that the system functions properly within the academic ecosystem. The integration process consisted of two main components: the CMS server and the Telegram chatbot server, which were subsequently combined into a single pipeline to enable real-time operation. Testing would use a Top-Down incremental testing approach, which involves starting from the highest-level component (CMS) and then proceeding to lower-level components (Telegram Chatbot) [36].

Table 4: Question for testing.

Label	Question
0	How do I register for the judiciary?
1	How do I check my eprt score?
2	What are the requirements to get a skip?
3	How do I apply for trial registration?
4	What are the types of SMEs?
5	What are the requirements for the IA1 Session?
6	How do I participate in social activities?
7	Is there a community on campus?
8	How do I join a community?
9	What if the value of D repeats?
10	How do I access the sports facilities?
11	Are there counseling services on campus?
12	How do I get counseling services?
13	What is mbkm?
14	How to access healthcare?
15	Where is the canteen located on campus?
16	What are the requirements to pass the TOEFL?
17	Where is the student space?
18	Where is the library?
19	Is there parking on campus?
20	Where can I get signature of dean?

This research would be divided into three levels. Level 1 ensured that data from the CMS was received and processed correctly by the server, with a focus on the accuracy and integrity of the data transfer. Level 2 tested the ability of the 'Akif' chatbot to receive a response from the server based on the latest data added via the CMS. Level 3 was a thorough test to ensure that the entire system, from the server and CMS to the chatbot, was functioning properly and that the latest data could be processed correctly by all components. Through this approach, each element was tested and validated in stages before the system was fully integrated, thereby supporting a more structured development process for the system.

3 Results

3.1 Result Of Model

After loading the training data and setting the parameters, the model base was ready to be tuned using a batch size of 4 and 30 epochs, as it only had a relatively small amount of data. The tuning model in this research was named 'Model-akif,' and the tuning process produced a graph of decreasing the average value of loss validation for 30 epochs, as shown in Figure 4.

Figure 4 illustrates the decreasing trend of the loss value, indicating the model's capacity to recognize patterns in the training data. A steady decreased in the loss value suggests that the model was continuously reducing its prediction error rate. The small fluctuation

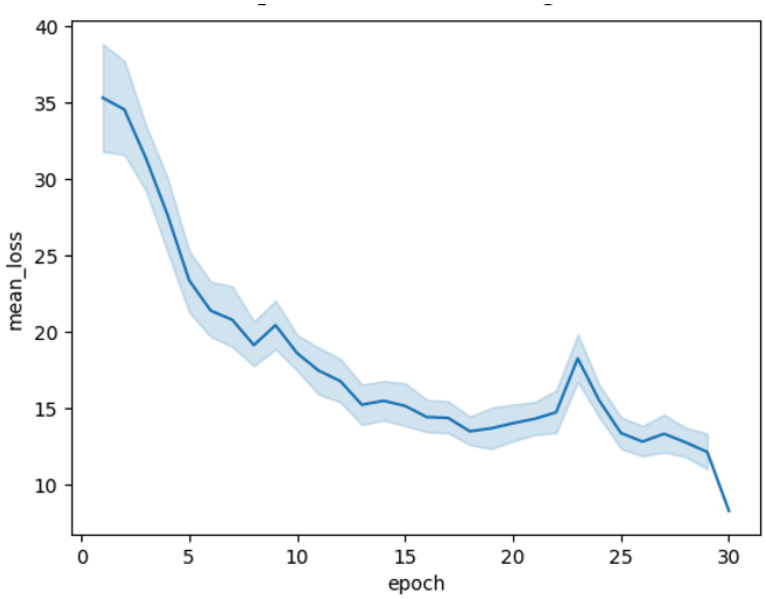

Figure 4: Validation loss over 30 epochs.

at the 25th epoch suggested the presence of possible outliers or noise in the data. This indicated that the labeling or preprocessing process could still be improved. At the end of the training, the graph showed signs of stabilization, indicating that the model had reached a convergence point, where further epochs did not yield significant performance improvements. Thus, it could be concluded that the model could be successfully trained without overfitting or underfitting, which also confirmed that the number of epochs and batch size used was optimal for limited-scale data. However, validation was still needed by testing the model with the BLEU Score. The model that had been built was tested using the BLEU Score with the test data presented in Table 4. Table 5 shows the test calculation.

Table 5: Result of evaluation

Label	Expected Answer	Generated Answers	Score
0	To register for the judiciary you need to prepare a file first. The files that need to be prepared can be accessed on bit.ly/ panduanpendaftaranta2425	To register for the judiciary you need to prepare a file first. The files that need to be prepared can be accessed on //bit.ly/ panduanpendaftaranta2425	1.0
1	You can check the eprt score on the igracias page in the language center menu section.	You can check the eprt value on igracias in the language center menu section	0.92

Label	Expected Answer	Generated Answers	Score
2	For skpi approval, you need to send your best 5 taks and then confirm them to the student affairs room in rek 407.	For skpi approval, you need to send your best 5 taks and then confirm them to the student affairs room in rek 407. Hope it is useful	0.75
3	You need to access the https://bit.ly/registrationguide2425page	For complete information you need to access the https://bit.ly/registrationguide2425page	0.86
4	Ukm at telkom university purwokerto includes ski, basketball, sircle, karate, wapala and mapala ukms.	The ukm at telkom university purwokerto is very diverse, including ski, basketball, sircle, karate, wapala and mapla ukms.	0.81
5	The requirements for fy 1 are at the following link https://bit.ly/registrationguideanta2425	The requirements for fy 1 are at the following link https://bit.ly/registrationguideanta2425	1.0
6	You just need to follow ukm or study program organizations, faculties or institutes. Or you can also take part in community service with lecturers.	You just need to follow ukm or study program organizations, faculties or institutes. Or you can also take part in community service with lecturers.	0.82
7	On-campus communities in the form of learning communities such as sircle and aws.	On-campus communities are learning communities like sircle and aws.	1.0
8	You can directly contact the community coordinator	You can directly contact the community coordinator	0.92
9	Yes, if you get a d grade in one of the courses you take, then you need to repeat it in the next semester.	Yes, if you get a d grade in one of the courses you take, then you need to repeat it in the next semester.	1.0
10	Mbkm is a government program that provides students with the opportunity to study and work while studying in the industrial world	Mbkm is a government program that provides students with the opportunity to study and work while studying in the industrial world	1.0
11	If you want to do counseling, you can come directly to the student room at rek 407	If you want to do counseling, you can come directly to the student room at rek 407	0.92
12	Of course, you can go to the telcomedika room near the telkom prayer room	Of course, you can go to the telcomedika room near the telkom prayer room	0.80
13	You just have to come to telkomedika which is near the telkom prayer room	You just have to come to telkomedika near the telkom prayer room. Hope it helps	0.75

Label	Expected Answer	Generated Answers	Score
14	There are 2 locations in the canting, namely next to the corridor of the dc building and the tt building	There are 2 locations in the canting, next to the corridor of the dc building and the tt building	0.86
15	Pass a minimum of 450 toefl	Pass the toefl with a minimum score of 450	0.86
16	The student room is in rek 407	Student affairs room in rek 407	0.86
17	The library is on the 1st floor of the rector	Library in the rector on the 1st floor	0.86
18	You can park in the area in front of the rectorate, dsp and behind the rectorate	You can park in the area in front of the rectorate, dsp and behind the rectorate. Remember not to forget to orginized.	0.70
19	You can go directly to the faculty dean's room	You can go directly to the dean's room of your faculty	0.86
Average			**0.88**

In addition to using the evaluation test using the Python library, researchers also conducted a manual calculation of the BLEU Score, which can be seen in Table 6.

Table 6: Manual calculation of BLEU score for selected samples

Label	Precision/n-gram	Score
0	(1.0), (1.0), (1.0), (1.0)	1.0
1	(1.0), (0.97), (0.94), (0.92)	0.92
2	(0.83), (0.80), (0.75), (0.66)	0.75
3	(0.93), (0.91), (0.89), (0.86)	0.86
4	(0.92), (0.86), (0.79), (0.71)	0.81
Average		**0.86**

The evaluation results indicated a slight difference between the manual and automatic methods (0.86 vs. 0.88). This difference may be caused by the different tokenization methods or n-gram calculation strategies used in the manual method and the automatic library. However, these differences were still within reasonable limits and generally demonstrate consistency in the accuracy and relevance to context of the answers generated by the model. With a BLEU score of 0.88, the model could create responses that were relevant to the context of the question asked. However, it should be noted that this score may decrease over time if the model was not updated with the latest information, given the changing dynamics of academic data. In addition, this evaluation was conducted numerically through the system, without involving validation by end-users (e.g., academic staff and students), so the aspect of context correctness from a human perspective had not been fully tested.

Another factor that influenced score variation was the nature of generative models, such as Gemini 1.5 Flash's parameter settings and optimal preprocessing, which allowed the model to still generate language variation without losing meaning. However, the BLEU Score did not fully capture meaning congruence, especially if the model experienced hal-

lucinations. This was a limitation of generative models, such as Gemini. Therefore, this model still required further development, such as integration with dynamic data sources (CMS), to consistently retrieve the latest information without the need to retrain from scratch.

3.2 Integration With CMS

In addition, for this model to be utilized directly by students, the next step is to integrate it with a platform-based system, in this case, a Telegram chatbot, and connect it with Telkom University Purwokerto's internal Content Management System (CMS).

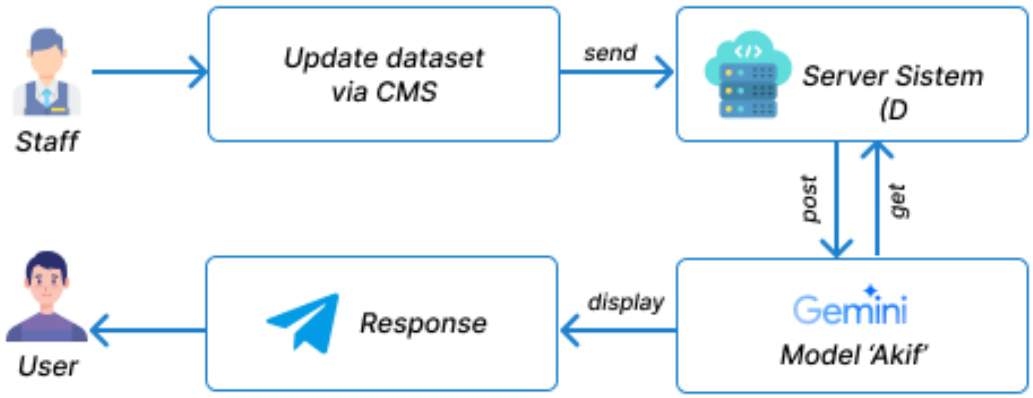

Figure 5: System integration workflow.

This integration ensured that the information provided by the chatbot was always up-to-date and accurate. This process began with staff updates to the CMS. The updated data would be sent to the server, which would then be accessed by the model for inference and retraining, automatically based on the latest data. This integration process was equipped with an Auto-Update dataset feature that eliminated the need for retraining. This feature was important given the changing dynamics of academic information. With this integration, the model can be updated regularly without the need for retraining from scratch, thereby supporting system efficiency. The results of this automated training were then implemented in the Telegram chatbot, ensuring that the chatbot could provide relevant and accurate responses based on the latest information. This process not only accelerated the data update cycle but also enhanced the overall sustainability of the system.

3.3 Integration Testing

Integration testing is carried out in 3 levels: Level 1, Level 2, and Level 3, as shown in Figure 6.

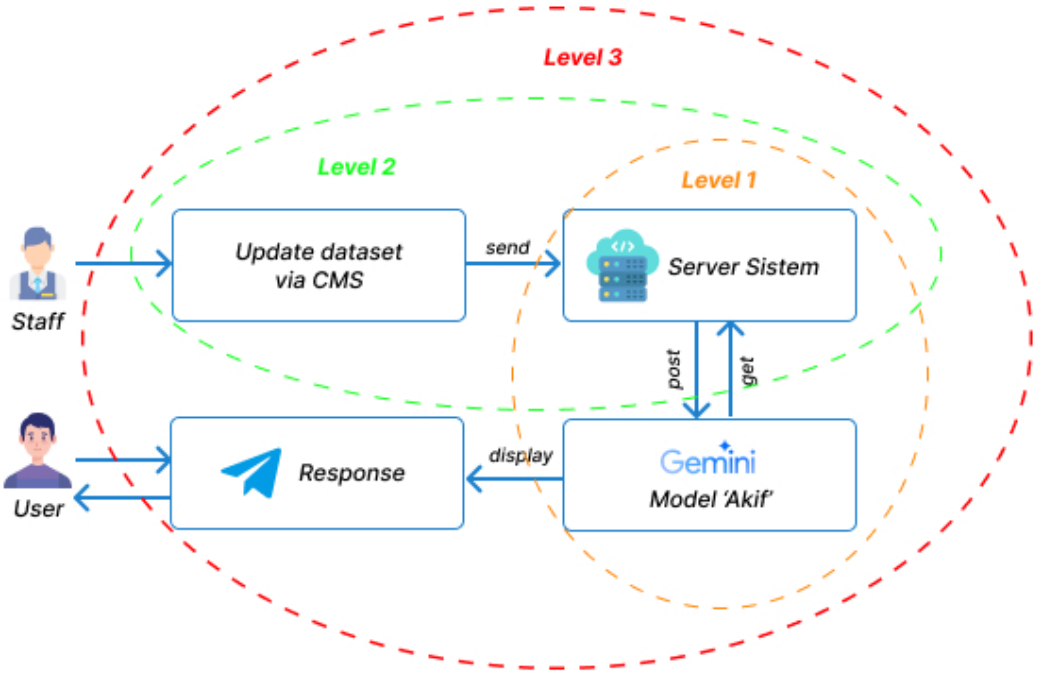

Figure 6: Integration level testing scheme.

3.3.1 Level 1 Testing

This test was conducted to ensure that data from the CMS could be received and processed by servers utilizing the Gemini model. The researchers manually modified the `data_training.json` file and then monitored the process through the server terminal.

In level 1 testing, the server successfully received the new data, and the Gemini model successfully retrieved it from the server. The new data would then be processed by Gemini to be used as new training data.

3.3.2 Level 2 Testing

At this level, testing was done to ensure that the Telegram Chatbot 'Akif' could receive a response from the server based on the data sent by the CMS. This test was similar to Level 1, but the difference was that data input is done through the CMS. The researcher entered new data through the CMS interface and verified the response through the server terminal. Figure 7 shows the process of new data added through CMS. According to Level 2 test, the server was still able to read new data even though it was input through the CMS interface, and the Gemini model was also able to retrieve new data from the server.

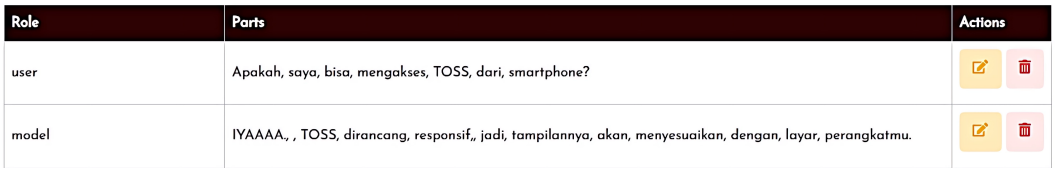

Figure 7: Adding new data via CMS.

3.3.3 Level 3 Testing

This test was an overall test, covering Level 2 with the results displayed on the Telegram chatbot. Testing was carried out with the following scenario in Table 7.

Table 7: Integration results

Function	Scenario	Expected Result	Status
Ask Chatbot	Giving questions outside the dataset to the chatbot	Chatbot responds to user questions dengan asal atau berhalusinasi	Valid
Create Dataset	Add a new dataset through the CMS	New dataset successfully added	Valid
Read Dataset	See if a new dataset exists in a CMS table	The dataset in the CMS is displayed correctly	Valid
Update Dataset	See if the dataset file is synced with the chatbot	The modified dataset is synchronized and validated by the chatbot	Valid
Ask Chatbot	Give questions according to the new dataset to the chatbot	The chatbot is able to respond to user questions with answers according to the new dataset.	Valid

From Table 7, it could be concluded that five positive cases had successfully obtained a valid status, indicating that the chatbot had successfully read and utilized the new dataset added through CMS. The following was a comparison of chatbot responses before and after the addition of the new dataset.

3.3.4 Chatbot response before adding dataset

The chatbot was unable to answer specific questions related to the new dataset and provide contextually relevant answers.

3.3.5 Chatbot response after adding new data

The chatbot could provide relevant answers based on new datasets that had been uploaded. The addition of this dataset showed that the integration between the CMS and chatbot was going well and meets the testing criteria.

Table 8: Chatbot response before the dataset is created

User	Chatbot Response
When is the registration for the 2024/2025 thesis defense period?	The registration for the 2024/2025 thesis defense period will be held from January 4th to January 10th, 2025.

Table 9: Chatbot response after the dataset is created

User	Chatbot Response
When is the registration for the 2024/2025 thesis defense period?	The registration for the 2024/2025 thesis defense period will be held from January 4th to January 10th, 2025.

The results of this test indicated that the CMS-chatbot integration system performed satisfactorily. The chatbot was automatically updated with the latest data. The responses provided by the chatbot correspond to what was entered into the CMS. The chatbot's responses changed slightly yet remain in the same context thanks to the use of a tailored model. This model's parameter adjustment aimed to mitigate the effect of overfitting by utilizing solely the inference or training model based on the new input-output from the CMS.

4 Discussion

This research aimed to develop a Gemini 1.5 Flash-based chatbot integrated with a CMS system to improve academic information services automatically and contextually. The model tuning results showed a consistent trend of decreasing loss over 30 epochs, indicating that the model was well-suited to adapt to the training data. Additionally, the BLEU score evaluation results suggested that the Gemini 1.5-flash model provided relevant answers, achieving a score of 0.88. The score was obtained after tuning and setting the parameters with the Low Temperature scenario (temperature 0.6, Top-K 20, and Top-P 0.6), which was adjusted for academic needs. The results of the exploration of tuning parameters also support research by Zhang et al. (2021) [35], that low temperature could reduce hallucinations, so that an initial score of 0.71 to a final score of 0.88 could be achieved.

This result also surpassed several previous studies, including Ghassemiazghandi (2024), which utilized ChatGPT with a BLEU score of 0.82 [37], and a survey by Chantoui & Ata (2021), which achieved a score of 0.80 [38]. The achievement of a score of 0.88 in this study indicated that the Gemini 1.5 Flash approach, with tuning and system integration, could provide significant performance improvements. However, using BLEU as the only evaluation metric had limitations. BLEU only measured the similarity between model responses and reference answers based on n-grams, thereby not directly capturing human judgment. In some cases, the model provides answered with the right meaning but contained some sentence variations that differ from the reference. For example, in label 18 (BLEU 0.70), additional sentences such as "Remember not to forget to organize" and in label 13, "Hope it helps." These variations were considered minor hallucinations that lower

the BLEU score, even though they were contextually relevant. This suggested the need for additional evaluations that considered human meaning and judgment so that the quality of answers could be measured more thoroughly.

An innovation brought by this research was the integration of a chatbot system with a CMS that enabled auto-training and real-time data updates. This feature allowed the chatbot to respond based on the most recent information entered into the CMS. The advantages of this approach were clearly evident compared to static chatbot systems, as noted in a study by Chaubey (2024) [39] and Sudan Prada (2022) [40], which states that a chatbot trained only once would experience a decrease in response accuracy due to outdated data. While the integration was effective on a pilot scale, the challenge of system scalability remained to be fully addressed. Testing was still limited to a small environment, and it haD not been tested whether the system was capable of serving thousands of users simultaneously. The system required strengthening of backend infrastructure, such as workload management and response time optimization, to operate efficiently at scale.

This research overall demonstrated great potential in the application of generative AI to support academic information services, particularly in reducing staff workload and facilitating faster access to accurate and real-time information for students. However, some aspects still required further development: (1) the limitation of evaluation metrics that only rely on BLEU; (2) the potential for mild hallucination in responses to ambiguous questions; and (3) the lack of testing of the system in real large-scale usage scenarios. Therefore, it was recommended that further research incorporate the use of meaning-based and human evaluation metrics, as well as live testing of the system by students and academic staff, to ensure the accuracy, relevance, and overall quality of the user experience.

5 Conclusion

This research successfully developed a chatbot based on Gemini 1.5 Flash, integrated with a Content Management System (CMS), to support academic information services at Telkom University, Purwokerto. Parameter selection with the 'Low Temperature' scenario successfully improved accuracy, and model tuning showed a consistent decrease in loss value over 30 epochs, indicating that the model adapted well to the training data without exhibiting symptoms of overfitting. Evaluation using the BLEU score yielded a value of 0.88, indicating that the chatbot can provide relevant and contextually appropriate responses despite variations in sentence structure. The main advantage of this system lies in its automatic integration capability between the chatbot and the CMS, which enables real-time data updates without requiring manual tuning from scratch. Thus, the system can adapt to the changing dynamics of academic information, in contrast to static chatbots that are only trained once.

However, this research has several limitations. First, the model's performance evaluation is only conducted automatically using the BLEU score, without involving human validation, so aspects of semantic understanding and interaction quality have not been fully assessed. Second, the system has not been extensively tested on a large scale, so aspects of scalability and resilience to high user loads remain unknown. Third, the chatbot still shows the possibility of hallucination, especially on questions that are ambiguous or not contained in the training data.

Future research is recommended to involve user-based evaluations (e.g., students or academic staff), utilize semantic metrics such as BERTScore, or employ human judgment, and test the system's performance in massive real-world usage scenarios. Development can also focus on reducing hallucination and improving answer relevance through integration with retrieval-based systems. Overall, although not yet perfect, this research makes a promising initial contribution to the development of an adaptive, informative, and integrated academic chatbot.

References

[1] D. Mustafa, S. C. Ahsan, M. Aris, R. Niswaty, and T. Prasodjo, "Service quality and performance of academic administration employees on student satisfaction," *Jurnal Ilmu-ilmu Sosial dan Humaniora Vol. 24, No. 3, November 2022: 335-342*, 2022.

[2] D. Ivanova and V. Atanasov, "Use of a chatbot in automated information system for process management in education: Use of a chatbot in automated information system for process management in education," *Journal scientific and applied research*, vol. 27, no. 1, pp. 105–112, 2024.

[3] U. ABEL, C. EMMANUEL, and U. O. PASCAL, "Leveraging ai-powered chatbots to enhance customer service efficiency and future opportunities in automated support," *Computer Science*, vol. 5, no. 10, pp. 2485–2510, 2024.

[4] H. S. Putra, H. Santoso, and C. Cifran, "Implementation of chatbot customer service features on pt dian prima jayaraya using dialogflow," *Infotech: Journal of Technology Information*, vol. 8, no. 2, pp. 143–148, 2022.

[5] G. F. Avisyah, I. J. Putra, and S. S. Hidayat, "Open artificial intelligence analysis using chatgpt integrated with telegram bot," *Jurnal ELTIKOM*, vol. 7, no. 1, pp. 60–66, 2023.

[6] B. K. Mishra and R. Kumar, "Natural language processing in artificial intelligence," 2014. Technical report / publication year unspecified.

[7] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, "A comprehensive overview of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 16, no. 5, pp. 1–72, 2025.

[8] G. K. Vamsi, A. Rasool, and G. Hajela, "Chatbot: A deep neural network based human to machine conversation model," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–7, IEEE, 2020.

[9] J. Liu, D. Shen, Y. Zhang, W. B. Dolan, L. Carin, and W. Chen, "What makes good in-context examples for gpt-3?," in *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd workshop on knowledge extraction and integration for deep learning architectures*, pp. 100–114, 2022.

[10] S. D. Pawar, "Chatgpt: Revolutionizing education administration—a case study analysis," *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, vol. 11, pp. 4447–4454, Apr. 2023.

[11] Z. Liu, W. Lin, Y. Shi, and J. Zhao, *A Robustly Optimized BERT Pre-training Approach with Post-training*. Springer Nature Switzerland AG, 2021.

[12] D. Anastasiou, "Enrich4all: A first luxembourgish bert model for a multilingual chatbot," in *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pp. 207–212, 2022.

[13] K. Peyton and S. Unnikrishnan, "A comparison of chatbot platforms with the state-of-the-art sentence bert for answering online student faqs," *Results in Engineering*, vol. 17, p. 100856, 2023.

[14] R. Islam and I. Ahmed, "Gemini-the most powerful llm: Myth or truth," in *2024 5th Information Communication Technologies Conference (ICTC)*, pp. 303–308, IEEE, 2024.

[15] A. Nazarius, F. Saputra, V. H. Pranatawijaya, *et al.*, "Penerapan gemini ai dalam pembuatan deskripsi produk e-commerce," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 3, pp. 3721–3725, 2024.

[16] M. Idowu, J. Cena, and G. Olaoye, "Exploring the limitations and challenges of gpt models." https://www.researchgate.net/publication/385771524, 2024. Online; accessed 2024.

[17] H. J. Christanto, C. Dewi, S. A. Sutresno, and A. D. K. Silalahi, "Analyzing the use of chat generative pre-trained transformer and artificial intelligence.," *Revue d'Intelligence Artificielle*, vol. 38, no. 4, 2024.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

[19] M. V. Koroteev, "Bert: a review of applications in natural language processing and understanding," *arXiv preprint arXiv:2103.11943*, 2021.

[20] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[21] H. R. Saeidnia, "Welcome to the gemini era: Google deepmind and the information industry," *Library Hi Tech News*, no. ahead-of-print, 2023.

[22] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[23] S. Neupane, E. Hossain, J. Keith, H. Tripathi, F. Ghiasi, N. A. Golilarz, A. Amirlatifi, S. Mittal, and S. Rahimi, "From questions to insightful answers: Building an informed chatbot for university resources," in *2024 IEEE Frontiers in Education Conference (FIE)*, pp. 1–9, IEEE, 2024.

[24] L. Mehnen and B. Pohn, "Supporting academic teaching with integrating ai in learning management systems: Introducing a toolchain for students and lecturers," in *2024 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, pp. 1–6, IEEE, 2024.

[25] Y. Mardon, "Modernization of higher education system management system: Innovations, challenges, and opportunities," *International Journal of Advance Scientific Research*, vol. 4, pp. 60–65, Apr. 2024.

[26] A. Hayrapetyan, R. Erbacher, C. A. Carrillo Montoya, D. M. Newbold, W. Carvalho, N. Karunarathna, M. Górski, M. Sommerhalder, N. Parmar, B. Ujvari, *et al.*, "Springer: The cms statistical analysis and combination tool: Combine," *Comput. Softw. Big Sci.*, vol. 8, no. CMS-CAT-23-001, p. 19, 2024.

[27] S. Kumar, D. Paikar, K. S. Vutukuri, H. Ali, S. R. Ainala, A. M. Krishnan, and Y. Zhang, "Katzbot: Revolutionizing academic chatbot for enhanced communication," *arXiv preprint arXiv:2410.16385*, 2024.

[28] E. D. Velásquez, M. E. Coaguila, H. Apaza, V. Yana, and C. A. Silva, "Implementation and evaluation of recurrent neural network models for an intelligent university admission chatbot," in *2024 43rd International Conference of the Chilean Computer Science Society (SCCC)*, pp. 1–6, IEEE, 2024.

[29] D. Chaskopoulos, J. E. Hægdahl, P. Sagvold, C. Trinquet, and M. Edalati, "Implementing a chatbot solution for learning management system," *arXiv preprint arXiv:2206.13187*, 2022.

[30] S. Programming, "Retracted: Lstm-based attentional embedding for english machine translation," 2023.

[31] V. B. Parthasarathy, A. Zafar, A. Khan, and A. Shahid, "The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities," *arXiv preprint arXiv:2408.13296*, 2024.

[32] M. Imran and N. Almusharraf, "Google gemini as a next generation ai educational tool: a review of emerging educational technology," *Smart Learning Environments*, vol. 11, no. 1, p. 22, 2024.

[33] M. Alsaadi, M. A. Abdulridha, A. H. J. Mohamed, and L. A. Alabbassi, "Magazine of student research virtual agent (chatbot) using open artificial intelligence." https://www.researchgate.net/publication/379507857, 2024. Online; accessed 2024.

[34] B. Bergner, A. Skliar, A. Royer, T. Blankevoort, Y. Asano, and B. E. Bejnordi, "Think big, generate quick: Llm-to-slm for fast autoregressive decoding," *arXiv preprint arXiv:2402.16844*, 2024.

[35] H. Zhang, D. Duckworth, D. Ippolito, and A. Neelakantan, "Trading off diversity and quality in natural language generation," in *Proceedings of the workshop on Human Evaluation of NLP Systems (HumEval)*, pp. 25–33, 2021.

[36] J. Wang, Y. Huang, C. Chen, Z. Liu, S. Wang, and Q. Wang, "Software testing with large language models: Survey, landscape, and vision," *IEEE Transactions on Software Engineering*, vol. 50, no. 4, pp. 911–936, 2024.

[37] 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, 2024.

[38] H. Chatoui and O. Ata, "Automated evaluation of the virtual assistant in bleu and rouge scores," in *2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pp. 1–6, IEEE, 2021.

[39] H. K. Chaubey, G. Tripathi, R. Ranjan, *et al.*, "Comparative analysis of rag, fine-tuning, and prompt engineering in chatbot development," in *2024 International Conference on Future Technologies for Smart Society (ICFTSS)*, pp. 169–172, IEEE, 2024.

[40] S. P. Uprety and S. R. Jeong, "The impact of semi-supervised learning on the performance of intelligent chatbot system.," *Computers, Materials & Continua*, vol. 71, no. 2, 2022.