A Bilingual Academic Chatbot Based on Semantic Retrieval Using mBERT

Leni Fitriani[1*], Sahrudin Fiqri Muzahidar[2], Ade Sutedi[3], Fitri Nuraeni[4]

[1,2,3,4] *Department of Computer Science, Institut Teknologi Garut, Garut, Indonesia*

*corr-author: leni.fitriani@itg.ac.id

Abstract - This study proposes a bilingual academic chatbot based on a semantic retrieval approach using the Multilingual BERT (mBERT) transformer architecture to support academic information services in higher education. The dataset was constructed from official academic information at Garut Institute of Technology, including new student admissions, academic calendars, institutional profiles, and lecturer and staff data. The data were organized in a bilingual question–and–answer format in Indonesian and English. The mBERT model was fine-tuned using a Sentence-BERT framework to generate sentence embeddings for semantic retrieval tasks, with MultipleNegativesRankingLoss applied during training. Model performance was evaluated using BERTScore to measure semantic similarity between chatbot responses and human reference answers. Experimental results show that the fine-tuned model outperformed the base model, achieving an average F1-score improvement from 0.7638 to 0.8152 for Indonesian and from 0.7556 to 0.8005 for English. The results also demonstrate more stable score distributions, indicating consistent semantic performance. The optimized model was subsequently integrated into a web-based prototype to enable real-time bilingual academic question answering. These findings confirm that combining mBERT with semantic retrieval effectively enhances the relevance and contextual accuracy of chatbot responses, thereby supporting digital transformation and improving the efficiency of academic services in higher education.

Keywords: Academic, bilingual, chatbot, mBERT, semantic retrieval.

I. INTRODUCTION

The use of generative chatbots has expanded across sectors, driven by their ability to provide fast, relevant, and adaptive responses in multiple languages. A survey conducted by Boston Consulting Group involving 21,000 respondents across 21 countries shows that Indonesia ranks sixth globally in ChatGPT usage, leading Southeast Asia with 32% active users [1]. This phenomenon highlights the significant potential of chatbot technology, particularly in higher education, which requires academic and administrative services that are fast, accurate, and easily accessible [2]. Higher education institutions face ongoing challenges in providing academic information, such as course schedules, admission procedures, and other administrative services [3-5]. The implementation of artificial intelligence–based chatbots with bilingual or multilingual capabilities represents an effective solution to improve service efficiency, reduce staff workload, and deliver responsive user interactions. Furthermore, chatbots can support academic activities such as academic advising, consultations, and the dissemination of institutional policies [6-8].

Several previous studies are relevant to this research. The first study developed a transformer-based chatbot to automatically answer questions related to academic and financial information for students. Using the SEMMA (Sample, Explore, Modify, Model, Assess) approach, the system achieved 84% accuracy, 100% precision, and 84% recall. However, the chatbot supported only a single language and did not employ a semantic retrieval approach, limiting its ability to provide relevant answers to complex queries [9]. The second study developed a question-answering system based on *Tafsir Al-Azhar* using a Large Language Model (LLM) integrated with LangChain. The system demonstrated contextual response capabilities with a UAT accuracy of 83.71%, contextual precision of 90%, relevance of 79%, and a hallucination rate of 41%. Despite its strong performance for Islamic literature, the system was not designed for higher education academic services, lacked multilingual support, and did not employ semantic-based evaluation metrics such as BERTScore [10]. The third study developed a multilabel classification model for detecting toxic comments in Indonesian, using IndoBERT as the embedding and mBERT for classification. The dataset consisted of 7,773 comments, and the model achieved an F1-score of 0.9032. This study demonstrated the effectiveness of mBERT in understanding Indonesian; however, its focus was not on chatbot development or interactive academic services [11].

The fourth study applied IndoBERT for Aspect-Based Sentiment Analysis (ABSA) on reviews of the

Baturraden tourist destination. The model classified reviews into four main aspects-attraction, accessibility, facilities, and supporting services—achieving 94.61% accuracy and an F1-score of 88.11%. Although the results were highly accurate, the study was limited to the tourism domain and did not address multilingual interaction in academic chatbots [12]. The fifth study developed a supervisor recommendation system using a BERT-based model to generate semantic representations of thesis titles and measure their similarity to lecturers' publications using cosine similarity. The system processed a dataset of 3,723 titles from 63 lecturers, achieving an accuracy of 90%. While effective as a recommendation system, the study was not designed as an interactive chatbot and did not accommodate multilingual requirements [13].

Based on this review, existing studies reveal limitations in multilingual support, semantic relevance of responses, and applicability to higher education academic services. To address these gaps, this study proposes a bilingual academic chatbot based on the Multilingual BERT (mBERT) transformer architecture using a semantic retrieval approach. The model is trained on a bilingual Indonesian-English question-answer dataset derived from official information from Garut Institute of Technology and evaluated using BERTScore to measure semantic similarity between chatbot responses and human reference answers. This approach aims to produce an adaptive, context-aware, and semantically relevant chatbot that supports digital transformation and improves the efficiency of academic services in higher education. While the foundational components-such as mBERT, Sentence-BERT fine-tuning, and cosine similarity-are well-established in natural language processing, the primary contribution of this research lies in their practical integration and deployment to solve real-world, bilingual academic service challenges.

II. METHOD

This study adopts the Machine Learning Lifecycle (MLLC) framework, which consists of four main stages: problem definition, data, model, and production system. The MLLC framework facilitates a structured process for data preparation, model selection and training, performance evaluation, and integration into real-world systems (Fig. 1). In addition, this approach enables continuous model refinement based on evaluation results, allowing the quality and relevance of the system to be progressively improved [14-16].

The first stage is problem definition, which addresses the need for higher education services to provide fast, accurate, and multilingual information. Conventional services are often constrained by limited operational hours and human resources, reducing accessibility. To address these challenges, this study proposes a bilingual academic chatbot based on the Multilingual BERT (mBERT) architecture with a semantic retrieval approach to enhance answer relevance and support interactions in Indonesian and English. The second stage is data preparation. The dataset was collected from the official website of Garut Institute of Technology, including admissions information, institutional profiles, academic calendars, and lecturer and staff data. The data were structured in a bilingual question-answer format. An initial set of 5,092 Indonesian pairs was translated into English using the *deep_translator* library to form a parallel corpus. After cleaning and duplicate removal, 10,178 valid instances remained, split into 80% for training and 20% for testing, as shown in Table I.

The third stage is model development. The model used in this study is mBERT (*bert-base-multilingual-cased*), a Transformer-based model pre-trained on 104 languages. Fine-tuning was conducted using the Sentence-BERT framework, which adapts mBERT to generate sentence embeddings for semantic retrieval tasks. The loss function used was MultipleNegativesRankingLoss, which compares each positive pair with multiple negative pairs within each training batch. Varying batch sizes and the number of training epochs conducted several experiments. The training configurations are summarized in Table II.

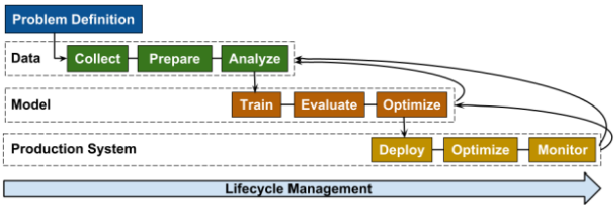

Fig. 1 Machine Learning Lifecycle (MLLC) [14]

TABLE I
DATASET DISTRIBUTION

No	Information Feature	Total_id	Total_en
1	New Student Admissions 2024	848	848
2	ITG Institutional Profile	449	449
3	Academic Calendar	241	241
4	ITG Lecturers and Staff	3,379	3,379
5	Chatbot Information	172	172
	Total	**5,092**	**5,092**

TABLE II
FINE-TUNING PARAMETERS

Experiment	Batch Size	Epoch	Loss Function
exp_1	32	2	MultipleNegativesRankingLoss
exp_2	32	3	MultipleNegativesRankingLoss
exp_3	16	2	MultipleNegativesRankingLoss
exp_4	16	3	MultipleNegativesRankingLoss

The fourth stage is the production system. After model optimization, the best-performing configuration (exp_4) was selected as the final model. This model was then integrated into a web-based prototype application developed using Streamlit. The prototype provides a simple interface that allows users to submit questions in either Indonesian or English and receive responses directly from the model. This stage ensures that the model can be applied in real-world higher education scenarios and continuously improved based on user feedback and additional evaluation results.

III. RESULT AND DISCUSSION

This section presents the experimental results and discusses the performance of the proposed bilingual academic chatbot based on semantic retrieval using Multilingual BERT (mBERT). The discussion focuses on dataset characteristics, fine-tuning configurations, semantic evaluation results, and system implementation, followed by a comparison with related studies.

A. Problem Definition

The problem addressed in this study is the limited effectiveness of conventional academic information services in higher education, which are often constrained by operational hours, limited human resources, and language barriers. The proposed solution-a bilingual academic chatbot-was designed to provide fast, accurate, and semantically relevant responses in both Indonesian and English. The experimental results demonstrate that the chatbot successfully meets these objectives by delivering consistent, context-aware answers across both languages.

B. Data

The data stage is the second phase of this research and comprises three main activities: data collection, data preparation, and data analysis. A detailed explanation of the activities performed at each stage of the data phase is presented in the following subsections.

1) Data Collection: The data collection process was conducted by manually extracting information from the official website of the Garut Institute of Technology. The collected information includes lecturer and staff data, new student admission information, general institutional profile information, the 2023/2024 academic calendar, and additional information related to the chatbot profile. The workflow of the data collection process is illustrated in Fig. 2.

As shown in Fig. 2, the extracted information was converted into question–answer pairs and stored in plain-text (.txt) format. The data collection process was designed to accommodate potential user queries; therefore, a large, diverse dataset was required to address a wide range of possible questions. The collected data were subsequently validated by institutional authorities to ensure their accuracy and legitimacy. Once validated, the data were used in the subsequent stages. At this stage, the dataset consisted of raw data that required further preparation and analysis before being used for model training.

2) Data Preparation: The data preparation stage involves processing the collected data to make it suitable for further analysis. This stage includes data transformation, augmentation, integration, and cleaning. Data transformation was performed by converting the original plain-text format to a pandas DataFrame using the pandas library in Python. This transformation resulted in a structured dataset represented as a dataframe. The transformed dataset is shown in Table III.

Data augmentation was conducted by translating the Indonesian-language dataset into English using the deep_translator library. This process produced two datasets that are identical in semantic content but differ in language. The two datasets were then concatenated to ensure balanced exposure of Indonesian and English data during model training. Data integration was performed by merging datasets from different information contexts into a single unified dataset using the pandas pd.concat() function. The integrated dataset was subsequently exported into CSV (Comma-Separated Values) format using the to_csv() function. The distribution of the dataset after transformation, augmentation, and integration is shown in Table IV.

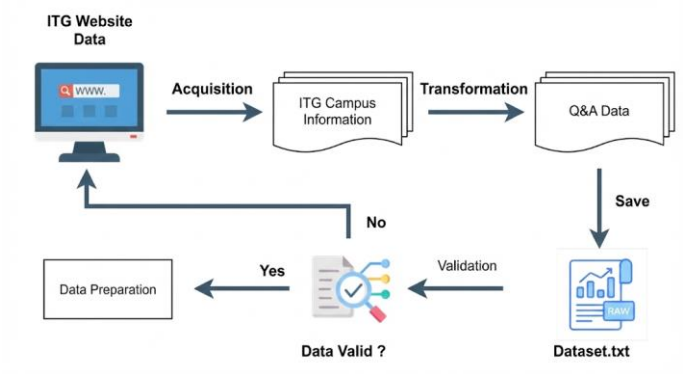

Fig. 2 Data collection workflow

TABLE III
TRANSFORMED DATASET

Question	Answer
Assalamualaikum	Waalaikumsalam! How can I help you?
What is the registration fee for the regular admission track at ITG?	The registration fee for the regular admission track at ITG is IDR 225,000.
Who is the Head of the Computer Science Department at ITG?	The Head of the Computer Science Department at ITG is Dr. Dede Kurniadi, S.Kom., M.Kom.

As shown in Table IV, the dataset consists of 10,184 question–answer pairs, with an equal distribution of 5,092 instances in Indonesian and English. After data cleaning, which involved removing empty or invalid entries, six instances were excluded, resulting in a final dataset of 10,178 valid pairs. The dataset was then partitioned into a training set (8,142 instances, 80%) and an independent testing set (2,036 instances, 20%) using the train_test_split function from the scikit-learn library. To prevent data leakage during model optimization, a validation subset was internally extracted from the training data to fine-tune hyperparameters such as batch size and epoch count. This ensures that the final model performance is verified against a strictly unseen test set. The balanced distribution across both languages enables the bert-base-multilingual-cased (mBERT) model to learn from each language in proportion, reducing language bias and enhancing multilingual chatbot performance.

3) Data Analysis: This stage involves analyzing and preprocessing the training dataset derived from the data preparation phase. To construct a balanced bilingual dataset, Indonesian-language data were augmented through automatic translation into English using the deep_translator library, producing a semantically equivalent parallel corpus. Preprocessing begins with data type standardization, in which both the Question and Answer columns are converted to string format. For the Question column, text normalization is applied, including the removal of special characters, filtering of non-alphabetic symbols, and conversion to lowercase to reduce noise and ensure consistent semantic representation. The Answer column is preserved in its original form to maintain readability in chatbot responses. Each text sequence is then formatted for mBERT input by adding the [CLS] token at the beginning and the [SEP] token at the end. Tokenization is performed using Hugging Face's BertTokenizer, converting text into numerical Token IDs. The tokenizer is configured with return_tensors='pt', truncation=True, padding=True, and max_length=128 to ensure uniform input length and efficient processing. This transformation enables mBERT to generate accurate semantic embeddings for question–answer matching.

4) Model

The training stage represents the core process in developing the bilingual academic chatbot. In this stage, the mBERT model is trained on a bilingual dataset of question–answer pairs in Indonesian (ID) and English (EN), enabling the model to capture contextual and semantic relationships across both languages. Training is performed using paired inputs consisting of anchors, namely user queries, and positive passages, namely relevant answers. A Siamese architecture is employed, in which both inputs are processed using the same mBERT encoder with shared weights. This shared-encoder mechanism ensures consistent sentence representation while reducing computational complexity by avoiding separate models for queries and answers. The architecture of the Siamese mBERT model used for bilingual semantic retrieval training is illustrated in Fig. 3.

TABLE IV
DATASET DISTRIBUTION BY INFORMATION FEATURE

No	Information Feature	Total_id	Total_en
1	New Student Admissions 2024	848	848
2	ITG Institutional Profile	449	449
3	Academic Calendar	241	241
4	ITG Lecturers and Staff	3,379	3,379
5	Chatbot Information	172	172
	Total	**5,092**	**5,092**

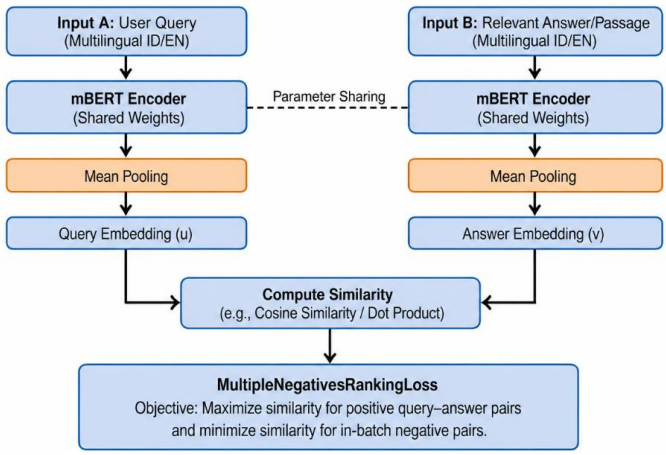

Fig. 3 Siamese mBERT Architecture for bilingual semantic retrieval training

As shown in Fig. 3, the training process uses two parallel input branches: one for the user query and one for the relevant answer or passage. Both inputs are encoded using the same mBERT encoder, and the resulting contextualized token embeddings are aggregated using a mean pooling layer to generate fixed-dimensional sentence embeddings. The query embedding and answer embedding are then compared using a similarity function, such as cosine similarity, to measure their semantic closeness. To optimize the semantic embedding space, this study applies the MultipleNegativesRankingLoss function. This loss function maximizes the similarity between each query and its corresponding positive answer while minimizing its similarity with other non-relevant answers within the same batch. The loss function is formulated in (1).

$$L = -\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{sim(u_i,v_i)}}{\sum_{j=1}^{N} e^{sim(u_i,v_j)}} \quad (1)$$

where L denotes the MultipleNegativesRankingLoss value, N represents the number of samples in a batch, u_i denotes the embedding vector of the i-th user query, v_i denotes the embedding vector of the corresponding positive answer, v_j denotes the embedding vector of the j-th answer candidate in the batch, and $sim(u_i, v_j)$ represents the similarity score between the query and answer embeddings.

C. Evaluate Model

The first evaluation employs cosine similarity to measure the semantic closeness between informal user queries and questions stored in the knowledge base. The model generates embeddings for each query and dataset entry, and similarity is computed by comparing the query embedding with all question–answer (QA) embeddings. Cosine similarity is the normalized dot product of two vectors, yielding values between 0 and 1, with 1 indicating perfect similarity. Each query is then matched with the dataset question yielding the highest similarity score. The results are presented in Tables V and VI.

The evaluation results for Indonesian are presented in Table V. Based on these results, the model demonstrates a reasonable ability to recognize semantic similarity despite variations in wording and sentence structure. For example, the informal query *"Siapakah yang menjadi ketua jurusan Ilmu Komputer?"* is correctly matched with the dataset question *"Siapa ketua jurusan Ilmu Komputer di ITG?"* with a similarity score of 0.5050. In addition, the query *"Berapasaja NIDN pak Dedi Sadudin Taptajani?"* with spelling variations is successfully recognized, with a similarity score of 0.6476. These results indicate that the model is sufficiently robust to handle everyday Indonesian language use, including both formal and informal expressions.

TABLE V
INDONESIAN LANGUAGE SIMILARITY RESULTS

Informal Question	Dataset Question	Dataset Answer	Similarity Score
Assalamualaikum	*Assalamualaikum!*	*Waalaikumsalam! Ada yang bisa CHATIT bantu?*	0.8120
Berapa biaya pendaftaran reguler di ITG?	*Berapa biaya pendaftaran untuk jalur reguler di ITG…*	*Biaya pendaftaran untuk jalur reguler di ITG adalah…*	0.6886
Kapan wisuda untuk Program Sarjana III dijadwalkan?	*Kapan wisuda untuk program Sarjana III dijadwalkan?*	*Wisuda Sarjana III dijadwalkan pada bulan November…*	0.3974
Siapakah yang menjadi ketua jurusan Ilmu Komputer?	*Siapa ketua jurusan Ilmu Komputer di ITG?*	*Ketua Jurusan Ilmu Komputer di ITG adalah Dr…*	0.5050
Berapasaja NIDN pak Dedi Sadudin Taptajani?	*Berapa NIDN Pak Dedi Sa'dudin Taptajani?*	*NIDN dengan nama dosen Dedi Sa'dudin Taptajani…*	0.6476

TABLE VI
ENGLISH LANGUAGE SIMILARITY RESULTS

Informal Question	Similar Dataset Question	Dataset Answer	Similarity Score
Good Morning	Good morning!	Good morning! How can CHAT IT help you?	0.8008
How much is the regular registration fee?	What is the registration fee for the regular path at ITG…	The registration fee for the regular path at ITG is…	0.8568
When's the undergrad graduation?	When was the graduation for the Bachelor of III program scheduled?	Graduation Bachelor III is scheduled in…	0.5185
Who is the head of the Computer Science department?	Who is the Head of the Computer Science Department at the Garut Institute of Technology?	… is the Head of the Computer Science Department at the Garut Institute of Technology.	0.9165

The evaluation results for English, presented in Table VI, show higher and more consistent cosine similarity scores compared to Indonesian, indicating effective handling of informal English queries, such as mapping *"Who is the head of the Computer Science department?"* with a similarity score of 0.9165 and *"Good Morning"* with 0.8008. To further assess response quality, BERTScore was employed to provide a semantic-level evaluation beyond surface-level word overlap.

The selection of BERTScore as the primary evaluation metric is justified by its ability to capture contextual and semantic nuances that traditional lexical overlap metrics, such as BLEU or ROUGE, often overlook. Unlike n-gram-based metrics, BERTScore leverages contextual embeddings from pre-trained transformer models to measure semantic similarity between candidate and reference texts, resulting in a higher correlation with human judgment [17]. In a semantic retrieval-based chatbot, ensuring that the retrieved response is not merely a keyword match but a semantically coherent answer is essential. Therefore, BERTScore provides a more robust, semantically meaningful evaluation by leveraging mBERT's contextual embeddings to compare the system's output with human-verified reference answers.

The fine-tuned mBERT model (exp_4) demonstrates consistent improvements across all metrics for both languages, with median precision increasing from 0.7985 to 0.8641 for Indonesian and from 0.7976 to 0.8338 for English, and median recall improving from 0.7256 to 0.8335 and from 0.7236 to 0.8036, respectively (Fig. 4). Consequently, the F1 Score shows the most substantial overall gains, with the mean increasing from 0.7638 to 0.8152 for Indonesian and from 0.7556 to 0.8005 for English, accompanied by narrower score distributions that indicate more stable and consistent model performance after fine-tuning.

To complement the visual analysis, the quantitative comparison of BERTScore metrics is summarized in Table VII.

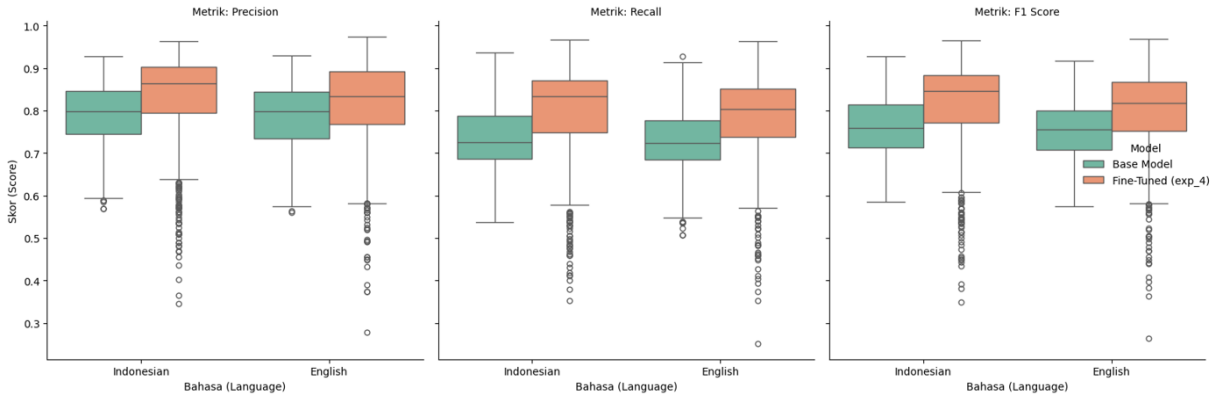

Fig. 4 Presents a boxplot comparison of two models-the base model and the fine-tuned model (exp_4)-across three evaluation metrics-precision, recall, and f1 score-for both Indonesian and English.

TABLE VII
BERTSCORE METRIC COMPARISON

Language	Model	Precision	Recall	F1 Score
Indonesian	Base Model	0.7914	0.7405	0.7638
Indonesian	Fine-Tuned (exp_4)	0.8325	0.8002	0.8152
English	Base Model	0.7685	0.7308	0.7556
English	Fine-Tuned (exp_4)	0.8173	0.7855	0.8005

Overall, the boxplot in Fig. 4 and the results in Table VII show that the fine-tuned model (exp_4) consistently outperforms the base model across all metrics and both languages. Higher mean and median scores indicate improved semantic relevance prediction, with slightly better performance in Indonesian, likely due to greater data availability or domain alignment. The most significant improvements are observed in Recall and F1 Score, suggesting that fine-tuning enhances the model's ability to retrieve relevant answers while maintaining a balanced precision–recall trade-off. These results confirm that fine-tuning mBERT improves semantic understanding and response quality in bilingual academic chatbot systems.

D. Optimize Model

The model optimization stage was conducted after fine-tuning and initial evaluation to enhance performance and efficiency before deployment. This stage aimed to identify the optimal hyperparameter configuration for maximizing mBERT's ability to capture semantic relationships between questions and answers. Systematic hyperparameter tuning experiments were performed by varying the number of training epochs and batch size. All experiments used the MultipleNegativesRankingLoss function for its effectiveness in semantic retrieval tasks, along with the AdamW optimizer from the sentence-transformers library. All fine-tuning experiments were run on a Fedora Linux system with a GPU to accelerate training. The model was trained with a learning rate of 2e-5, and the average training time per configuration was approximately 45 minutes. Model performance was evaluated using cosine similarity on the test dataset to measure semantic matching accuracy. The results of these optimization experiments are summarized in Table VIII.

According to Table VIII, hyperparameter tuning significantly improves model performance. Increasing the number of epochs from 2 to 3 consistently improves cosine similarity scores, indicating better semantic learning without evident overfitting. Reducing the batch size from 32 to 16 further enhances performance, likely because more frequent weight updates improve convergence and generalization. As shown in Fig. 5, experiment *exp_4* achieves the highest cosine similarity score (0.8446), confirming that a batch size of 16 and 3 epochs is the optimal configuration for this task.

TABLE VIII.
MODEL OPTIMIZATION EXPERIMENTS

Experiment	Batch Size	Epochs	Loss Function	Cosine Similarity Score
exp_1	32	2	MultipleNegativesRankingLoss	0.8192
exp_2	32	3	MultipleNegativesRankingLoss	0.8315
exp_3	16	2	MultipleNegativesRankingLoss	0.8342
exp_4	16	3	MultipleNegativesRankingLoss	0.8446

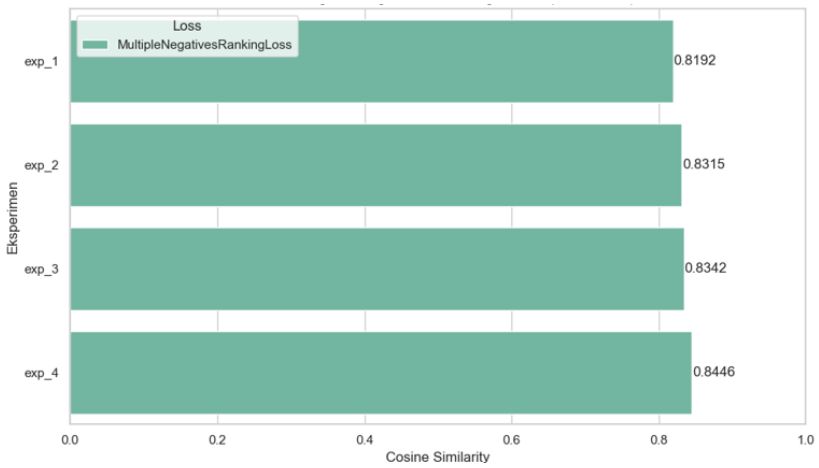

Fig. 5 Comparison of mBERT fine-tuning configurations

E. Production System

After training, evaluation, and optimization, the model is deployed into a web-based prototype of the bilingual mBERT chatbot. The implementation uses Streamlit to enable rapid development and seamless integration with the model. The optimized model is connected to the backend, allowing users to submit queries and receive real-time responses in Indonesian or English. The system interface is illustrated in Fig. 6.

The results of this study demonstrate that fine-tuning Multilingual BERT (mBERT) with a semantic retrieval approach significantly improves the performance of a bilingual academic chatbot. Across all evaluation stages, the fine-tuned model consistently outperforms the base model in terms of semantic relevance, stability, and robustness for both Indonesian and English queries. The cosine similarity evaluation confirms the model's ability to match informal user queries with semantically equivalent questions stored in the knowledge base, despite variations in wording, spelling, and sentence structure, indicating strong generalization capability. While standard Information Retrieval (IR) metrics, such as Top-k accuracy and Mean Reciprocal Rank (MRR), are commonly used to assess retrieval performance, this study emphasizes the semantic quality of responses. The high mean cosine similarity scores serve as a proxy for retrieval effectiveness, suggesting that the model reliably identifies the most semantically relevant question–answer pairs. However, a more comprehensive evaluation involving a broader set of unseen queries and formal IR metrics would further strengthen the validation of the system's robustness in large-scale environments.

Performance in English is observed to be slightly higher and more consistent than in Indonesian, which may be attributed to the broader availability of English-language training corpora and the inherent advantages of pre-trained multilingual models in high-resource languages. The use of BERTScore further reinforces these findings by providing a semantic-level evaluation of response quality. The fine-tuned model demonstrates notable improvements in precision, recall, and F1 across both languages, with particularly strong gains in recall and F1. This indicates that the optimized model not only reduces irrelevant responses but also enhances its ability to retrieve a wider range of contextually appropriate answers. In addition, the narrower score distributions suggest improved consistency and reliability across diverse query types.

The hyperparameter optimization results highlight the importance of training configuration in semantic retrieval tasks. Specifically, smaller batch sizes combined with longer training durations lead to more frequent weight updates and deeper semantic learning, thereby improving model performance. The optimal configuration (batch size = 16, epochs = 3) effectively balances learning capacity and generalization, making it well-suited for bilingual academic datasets. From a system perspective, the successful integration of the optimized model into a web-based prototype demonstrates its practical applicability. The chatbot can deliver real-time academic information in both Indonesian and English, supporting digital transformation initiatives in higher education by improving accessibility, efficiency, and user experience.

Despite these promising results, several limitations should be acknowledged. The dataset is derived from a single institution, which may limit the model's generalizability to other academic contexts. In addition, the current evaluation is conducted in an offline experimental setting and has not yet been validated through real-world user interactions or user-based studies. Consequently, the practical effectiveness and usability of the chatbot in real academic service environments remain to be further investigated. Future work should include user-centered evaluations involving students and academic staff, as well as deployment in real institutional settings to assess system performance under actual usage conditions. Furthermore, the chatbot currently operates in a single-turn question–answer setting and does not yet support multi-turn conversational context or reasoning over multiple knowledge sources. Another limitation arises from the use of automated translation to construct the English portion of the dataset, which may introduce linguistic bias. Although this approach ensures a parallel corpus, it may lack the natural variability and idiomatic nuances of authentic English queries. This limitation may affect the model's generalization capability in real-world English interactions and likely contributes to the slightly better performance observed in Indonesian. Future research may address these limitations by expanding the dataset across multiple institutions, incorporating dialogue context modeling, and exploring hybrid retrieval–generation approaches to enhance response quality and adaptability further.

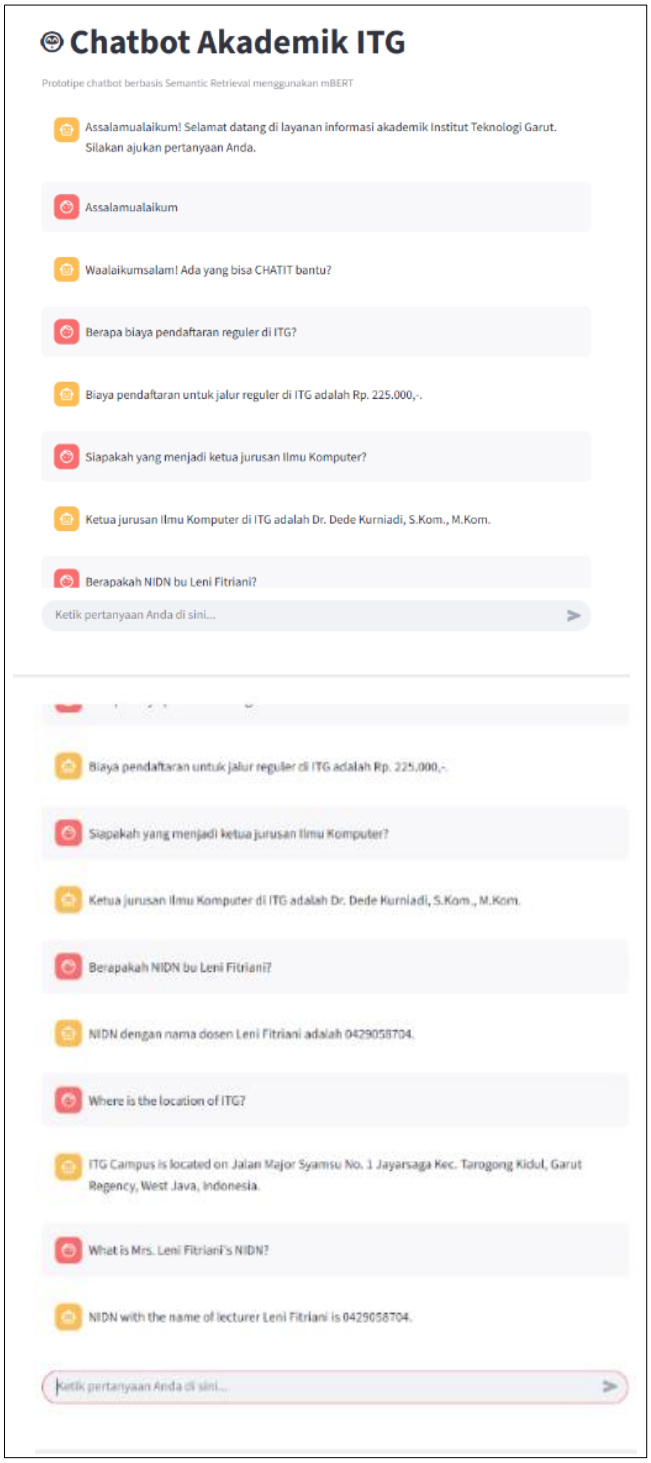

Fig. 6 Academic chatbot ITG

IV. CONCLUSION

This study presents the development and evaluation of a bilingual academic chatbot based on a semantic retrieval approach using Multilingual BERT (mBERT). The proposed system was designed to address limitations in conventional academic information services by providing fast, accurate, and bilingual responses to user queries in Indonesian and English. Experimental results demonstrate that fine-tuning mBERT significantly improves semantic retrieval performance. The optimized model achieves higher cosine similarity scores and consistently outperforms the base model across all BERTScore metrics, including precision, recall, and F1 score, for both languages. These improvements indicate enhanced semantic understanding, increased relevance of retrieved answers, and more stable performance across diverse query types. Hyperparameter optimization further shows that smaller batch sizes combined with longer training durations yield superior results, with the configuration using a batch size of 16 and 3 training epochs identified as the most effective. The successful deployment of the optimized model into a web-based prototype confirms its practical applicability in real-world academic environments. The chatbot can deliver real-time, bilingual academic information, thereby supporting digital transformation initiatives and improving the efficiency and accessibility of academic services in higher education institutions. Despite these promising results, this study has limitations. The dataset is limited to a single institution, and the chatbot operates in a single-turn interaction setting. Future research may extend this work by incorporating multi-institutional datasets, supporting multi-turn conversational context, and integrating hybrid retrieval–generation techniques to enhance system adaptability and response quality further. Overall, this research demonstrates that combining mBERT with semantic retrieval and systematic optimization provides an effective approach for developing bilingual academic chatbots and contributes to advancing intelligent academic service systems in higher education.

REFERENCES

[1] Kompas, “21 Negara dengan Pengguna ChatGPT Terbanyak, Ada Indonesia.” [Online]. Available: https://tekno.kompas.com/read/2024/09/02/07020037/21-negara-dengan-pengguna-chatgpt-terbanyak-ada-indonesia

[2] J. Salmi and A. A. Setiyanti, “Persepsi Mahasiswa Terhadap Penggunaan ChatGPT di Era Pendidikan 4.0,” *J. Ilm. Wahana Pendidik.*, vol. 9, no. 19, pp. 399–406, Oct. 2023, doi: 10.5281/zenodo.8403233.

[3] S. S. Babalola and C. A. Genga, “Managing Digital Transformation in African Higher Education Institutions: Challenges and Opportunities: Managing Digital Transformation,” *Int. J. E-Learn. Distance Educ. Int. E-Learn. Form. À Distance*, vol. 39, no. 1, 2024, doi: https://doi.org/10.55667/10.55667/ijede.2024.v39.i1.1333.

[4] C. Chaka, T. Shange, T. Nkhobo, and V. Hlatshwayo, "An Environmental Review of the Generative Artificial Intelligence Policies and Guidelines of South African Higher Education Institutions: A Content Analysis," *Int. J. Learn. Teach. Educ. Res.*, vol. 23, no. 12, pp. 487–511, 2024, doi: https://doi.org/10.26803/ijlter.23.12.25.

[5] L. Fitriani, M. L. Khodra, and K. Surendro, "A conceptual framework for AI adoption in business architecture with case studies in higher education and government," *Discov. Artif. Intell.*, vol. 5, no. 1, p. 409, Nov. 2025, doi: 10.1007/s44163-025-00673-3.

[6] S. Hadid, U. Ramadhani, S. Dian Suari, and A. G. Eka Putri, "Analisis Dampak Penggunaan Chatbot AI Dalam Pembelajaran Di Kalangan Mahasiswa PGSD Universitas Jambi," *J. Pendidik. Terap.*, vol. 02, no. September, pp. 160–166, 2024, doi: 10.61255/jupiter.v2i3.461.

[7] T. N. Alruqi and S. M. Alzahrani, "Evaluation of an Arabic chatbot based on extractive question-answering transfer learning and language transformers," *AI*, vol. 4, no. 3, pp. 667–691, 2023, doi: https://doi.org/10.3390/ai4030035.

[8] H. Fan, B. Han, and W. Gao, "(Im)Balanced customer-oriented behaviors and AI chatbots' Efficiency–Flexibility performance: The moderating role of customers' rational choices," *J. Retail. Consum. Serv.*, vol. 66, p. 102937, May 2022, doi: 10.1016/j.jretconser.2022.102937.

[9] R. H. Hafizh, "Pengembangan Chatbot Berbasis Jaringan Saraf Tranformer untuk Layanan Informasi Akademik dan Keuangan Mahasiswa di Universitas Muhammadiyah Sukabumi," vol. 12, no. 3, 2024, doi: 10.23960/jitet.v12i3.5002.

[10] A. B. Permadi, N. S. H, L. Handayani, and Yusra, "Implementasi Question Answering System Tafsir Al-Azhar Menggunakan Langchain Dan Large Language Model Berbasis Chatbot Telegram," *J. Teknoif Tek. Inform. Inst. Teknol. Padang*, vol. 12, no. 1, pp. 62–69, 2024, doi: 10.21063/jtif.2024.v12.1.62-69.

[11] G. Z. Nabiilah, I. N. Alam, E. S. Purwanto, and M. F. Hidayat, "Indonesian Multilabel Classification Using IndoBERT Embedding and MBERT Classification," *Int. J. Electr. Comput. Eng.*, vol. 14, no. 1, pp. 1071–1078, 2024, doi: 10.11591/ijece.v14i1.pp1071-1078.

[12] D. C. Febrianto, M. A. Fitriani, M. Afrad, and M. A. Khadija, "Aspect Based Sentiment Analysis Menggunakan Indobert Model Terhadap Review Pengunjung Objek Wisata Baturraden," *Melek IT Inf. Technol. J.*, vol. 10, no. 2, pp. 157–166, 2024, doi: https://doi.org/10.30742/melekitjournal.v10i2.358.

[13] F. T. Sabilillah, S. Winarno, and R. B. Abiyyi, "Implementasi BERT dan Cosine Similarity untuk Rekomendasi Dosen Pembimbing berdasarkan Judul Tugas Akhir," vol. 8, no. 2, pp. 585–594, 2024, doi: 10.29408/edumatic.v8i2.27791.

[14] M. Maskey, R. Ramachandran, I. Gurung, B. Freitag, J.J. Miller, M. Ramasubramanian, D. Bollinger, R. Mestre, D. Cecil, A. Molthan, and C*l*. Hain, "Machine learning lifecycle for earth science application: a practical insight into production deployment," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 2019, pp. 10043–10046. doi: https://doi.org/10.1109/IGARSS.2019.8899031.

[15] L. Fitriani, D. Tresnawati, and M. B. Sukriyansah, "Image Classification On Garutan Batik Using Convolutional Neural Network with Data Augmentation," *JUITA J Inf.*, vol. 11, no. 1, p. 107, 2023, doi: https://doi.org/10.30595/juita.v11i1.16166.

[16] L. Fitriani, A. Sanusi, R. Rismala, and D. Tresnawati, "Transformer-Based Detection Model for Number Recognition on Electric kWh Meters," *JUITA J. Inform.*, vol. 13, no. 2, pp. 135–143, 2025, doi: https://doi.org/10.30595/juita.v11i1.16166.

[17] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," Feb. 24, 2020, *arXiv*: arXiv:1904.09675. doi: 10.48550/arXiv.1904.09675.