JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)
VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8295

TRANSFORMER-BASED GENERATIVE CHATBOT FOR HIGHER EDUCATION INFORMATION SERVICES WITH HYPERPARAMETER OPTIMIZATION AND MODEL EVALUATION

Leni Fitriani[1*]; Endang Prayoga Hidayatulloh[1]; Dede Kurniadi[1]; Muhammad Rikza Nashrulloh[1]

Department of Computer Science[1]
Institut Teknologi Garut, Garut, Indonesia[1]
https://www.itg.ac.id/[1]
leni.fitriani@itg.ac.id*, 2006189@itg.ac.id, dede.kurniadi@itg.ac.id, rikza@sttgarut.ac.id

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *Generative AI has revolutionized the creation of realistic multimedia content, including chatbots that generate human-like responses. This technology significantly improves higher education by enabling universities to provide fast, accurate, and efficient information services. Generative chatbots can handle multiple users simultaneously and operate 24/7, increasing productivity and accessibility. The model was developed using Machine Learning Lifecycle (MLLC) with deep learning algorithm and Transformer architecture. The dataset used consists of 5,403 question-answer pairs from Institut Teknologi Garut (ITG), which are divided into 5,089 pairs for training and 314 pairs for testing. From 12 hyperparameter configurations, the best combination (maxlen 80, num_layers 2, batch_size 128, embedding_dim 256, fully_connected_dim 256, num_heads 2, positional_encoding_length 512, learning_rate 0.0002, and epoch 100) achieved a BLEU score of 71.03% on the ITG dataset. Evaluation using ROUGE and METEOR also shows consistent performance, indicating good content coverage and semantic similarity. Retraining with another dataset using the same approach resulted in a slightly higher BLEU score of 72.05%, with a different optimal learning rate of 0.00025. The results of this study indicate that Transformer-based generative chatbots can support higher education services and highlight the importance of adjusting hyperparameters based on dataset characteristics. This research also provides opportunities to develop similar models in other universities by adapting datasets and exploring more advanced methods to improve performance in broader educational contexts.*

Keywords: *Chatbot, Deep Learning, Generative AI, Higher Education, Transformer.*

Intisari— *Kecerdasan Buatan Generatif (Generative AI) telah merevolusi pembuatan konten multimedia realistis, termasuk chatbot yang menghasilkan respons mirip manusia. Teknologi ini secara signifikan meningkatkan pendidikan tinggi dengan memungkinkan universitas menyediakan layanan informasi yang cepat, akurat, dan efisien. Chatbot generatif dapat menangani beberapa pengguna secara bersamaan dan beroperasi 24/7, meningkatkan produktivitas dan aksesibilitas. Model ini dikembangkan menggunakan Machine Learning Lifecycle (MLLC) dengan algoritma deep learning dan arsitektur Transformer. Data yang digunakan terdiri dari 5.403 pasangan pertanyaan-jawaban dari Institut Teknologi Garut (ITG), yang dibagi menjadi 5.089 pasangan untuk pelatihan dan 314 pasangan untuk pengujian. Dari 12 konfigurasi hyperparameter, kombinasi terbaik (maxlen 80, num_layers 2, batch_size 128, embedding_dim 256, fully_connected_dim 256, num_heads 2, positional_encoding_length 512, learning_rate 0.0002, dan epoch 100) mencapai skor BLEU sebesar 71,03% pada dataset ITG. Evaluasi menggunakan ROUGE dan METEOR juga menunjukkan kinerja yang konsisten, yang mengindikasikan cakupan konten dan kesamaan semantik yang baik. Pelatihan ulang dengan dataset lain menggunakan pendekatan yang sama menghasilkan skor BLEU yang sedikit lebih tinggi yaitu 72,05%, dengan laju pembelajaran optimal yang berbeda yaitu 0,00025. Hasil penelitian ini menunjukkan bahwa chatbot generatif berbasis Transformer dapat mendukung layanan*

pendidikan tinggi dan menyoroti pentingnya menyesuaikan hyperparameter berdasarkan karakteristik dataset. Penelitian ini juga memberikan peluang untuk mengembangkan model serupa di universitas lain dengan mengadaptasi dataset dan mengeksplorasi metode yang lebih canggih untuk meningkatkan kinerja dalam konteks pendidikan yang lebih luas.

***Kata Kunci**: Chatbot, Pembelajaran Mendalam, Kecerdasan Buatan Generatif, Pendidikan Tinggi, Transformer.*

INTRODUCTION

Generative Artificial Intelligence (AI), particularly through large language models (LLMs), has significantly advanced natural language processing by enabling systems to generate coherent, context-aware, and human-like text. These models, primarily based on the Transformer architecture, have demonstrated strong performance across a wide range of language tasks, including conversational systems and question–answering applications [1], [2].

Key Generative AI algorithms, such as Generative Adversarial Networks (GAN) and Transformers, have been widely applied in scientific domains, including synthetic data generation. Generative chatbots such as ChatGPT have attracted significant interest and are utilized across sectors for research, marketing, customer interaction, and decision-making [3].

Generative chatbots, powered by artificial intelligence and machine learning, can process data and generate responses to various user requests. These capabilities have enabled their application across multiple sectors, including higher education [4], [5]. In this context, generative chatbots such as ChatGPT can support adaptive learning, provide personalized feedback, assist research and data analysis, automate administrative services, and enable innovative assessment methods [6].

Higher education has continuously evolved alongside advancements in Information and Communication Technology (ICT) [7]. Despite these developments, universities still face challenges in providing fast, accessible, and accurate information services [8], [9]. Artificial Intelligence (AI), particularly generative chatbots, offers a promising solution by delivering relevant information, handling multiple users simultaneously, and operating continuously. In higher education, chatbots can enhance efficiency by enabling quick and easy access to academic and non-academic information [10], [11].

Previous research demonstrated the effectiveness of a generative chatbot for academic administrative services using a sequence-to-sequence architecture with a Bi-LSTM network [12]. Using various experimental settings, the model achieved a training accuracy of 0.9964 and a validation accuracy of 0.8756, with the best configuration at a 90:10 data split, Adam optimizer (learning rate 0.01), and batch size of 16.

The model also obtained a BLEU score of 0.7724, indicating high relevance of generated responses. However, despite these promising results, sequence-based architectures still have limitations in capturing long-range dependencies and maintaining contextual coherence in conversational tasks.

This limitation becomes more significant when dealing with complex or diverse user queries. In terms of user acceptance, the chatbot obtained a CUQ score of 62.2, which falls into the “Good” category on the SUS rating scale. Although these results demonstrate strong performance, the approach still relies on sequential architectures that may struggle to capture long-range dependencies and complex contextual relationships in conversational data. However,

Recent studies have demonstrated the potential of Transformer architectures for generative chatbot development. By leveraging self-attention and parallel sequence processing, Transformer-based models can effectively capture contextual relationships in conversational data and have achieved promising performance on generative dialogue tasks [13]. Therefore, this study adopts the Transformer architecture for developing the proposed higher-education generative chatbot.

Therefore, this study adopts the Transformer architecture due to its effectiveness and adaptability to small datasets. Nevertheless, most existing studies focus on general datasets and do not specifically investigate the impact of hyperparameter configurations in domain-specific applications, particularly in higher education environments. Therefore, this research aims to implement a Transformer-based generative chatbot model and analyze the effect of hyperparameter tuning on performance using a domain-specific dataset.

In line with this objective, this research focuses on the implementation of a Generative Intelligence Chatbot based on the Transformer model for higher education. The Transformer model is a neural network architecture that uses self-

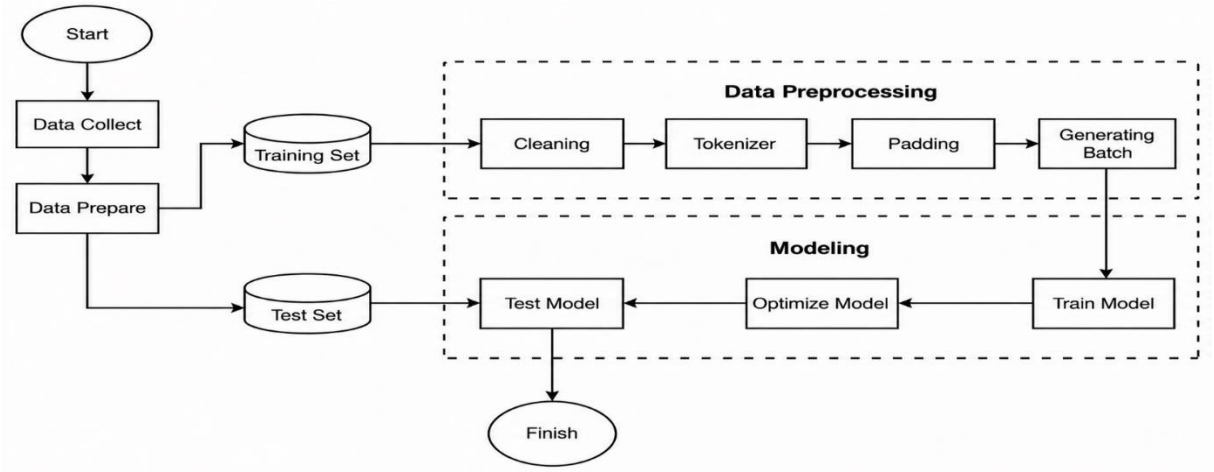
Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8295

attention to capture contextual relationships in sequential data, enabling efficient handling of long dependencies and parallel processing [14]. The dataset was collected from the official website of Garut Institute of Technology and converted into an Indonesian question–answer format, with validation conducted by campus staff to ensure accuracy. This study aims to develop a generative chatbot capable of providing relevant information and effectively assisting students, prospective students, lecturers, and staff in academic and administrative activities.

MATERIALS AND METHODS

This section presents the research methodology used in this study. A generative chatbot model for higher education is developed using machine learning and deep learning techniques based on the Transformer architecture, as shown in Figure 1. The research consists of several stages, including data collection, data preparation, preprocessing, model training with hyperparameter optimization, and evaluation.



Source: (Research Results, 2025)

Figure 1. The study Methodology

Each subsequent stage is carried out systematically, from preparing and preprocessing the data into a suitable format for the Transformer model, to training and evaluating the model's performance in handling diverse user inputs.

Data Collection

This study collects data by compiling conversational information related to Institut Teknologi Garut (ITG). The dataset was constructed manually using information sourced from the official ITG website, covering various features such as lecturer and staff data, new student admissions, the general campus profile, the academic calendar for the 2023/2024 academic year, and additional chatbot-related information. To enrich the dataset structure, the development process also referred to existing question–answer dataset patterns from similar publicly available sources, while all content was manually adapted and rewritten to ensure relevance and contextual suitability for ITG.

The collected information was then transformed into question–answer (QA) pairs designed to reflect realistic user queries. This transformation process was carried out manually by considering the most likely questions that users might ask, resulting in a diverse and comprehensive set of QA pairs. The dataset was designed with different characteristics for model learning and evaluation purposes. The training data emphasizes structured, clear, and consistent QA patterns to support effective learning, while the evaluation data uses more natural, varied, and less predictable language.

To ensure data quality, the constructed dataset underwent a validation process involving the institution. This step ensures the accuracy, consistency, and reliability of the information. Only validated data were used in the subsequent modeling process.

Data Preparation

The collected data are processed through data transformation and integration. The data are first transformed from plain text into a dataframe format using the pandas library in Python. Since the collected data are categorized based on different information contexts, a data integration step is

performed to unify them into a single dataset, ensuring consistency, completeness, and accuracy. The final dataset consists of 5,403 question–answer pairs, divided into 5,089 for training and 314 for testing. The majority of the training data (3,379 pairs) relate to information about ITG lecturers and staff. The testing dataset reflects similar attributes to the training data. Examples of the original question–answer pairs are presented in Table 1. These examples are intentionally presented in their original Indonesian language to preserve the authenticity of the dataset and accurately represent the linguistic characteristics of the data used for model training and evaluation.

Table 1. Examples of the Original Indonesian Question–Answer Pairs in the Dataset

No	Question	Answer
1	Berapa biaya pendaftaran untuk jalur reguler di ITG?	Biaya pendaftaran untuk jalur reguler di ITG adalah Rp. 225.000,-.
2	Kapan wisuda untuk program Sarjana III dijadwalkan pada tahun 2024?	Wisuda Sarjana III dijadwalkan pada bulan November 2024.
3	Kapan masa heregistrasi dimulai untuk Semester Genap tahun akademik 2023/2024 di ITG?	Masa heregistrasi untuk Semester Genap tahun akademik 2023/2024 di ITG dimulai pada tanggal 5 Februari 2024.

Source: (Research Results, 2025)

Data Preprocessing

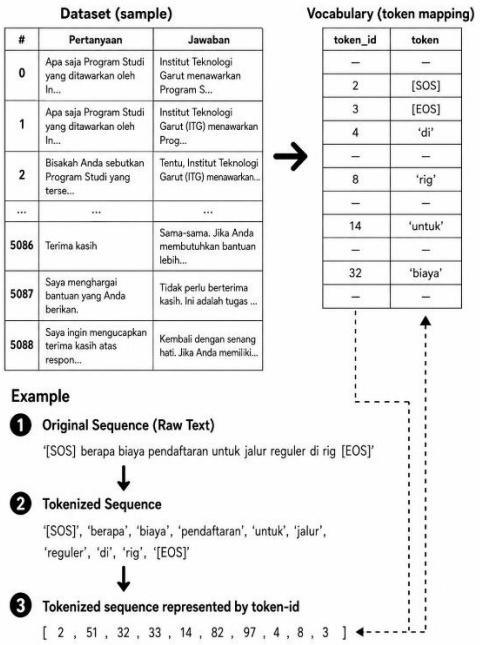
After data preparation, the dataset is processed for modeling through four main stages: data cleaning, tokenization, padding, and batch generation. Data cleaning is applied to the Question (input) column by removing special characters, retaining only alphabetic characters, and converting all text to lowercase to ensure consistency. The Answer column is not fully cleaned to preserve its natural language structure for generating readable outputs. Next, special tokens, Start of Sentence ([SOS]) and End of Sentence ([EOS]), are added to both Question and Answer sequences to indicate sentence boundaries and support the training process. The resulting dataset after cleaning and token addition is then used for further preprocessing steps.

Table 2. Dataset Conditions After Data Cleaning and Token Addition for the Indonesian Language

No	Question	Answer
1	[SOS] berapa biaya pendaftaran untuk jalur reguler di itg [EOS]	[SOS] Biaya pendaftaran untuk jalur reguler di ITG adalah Rp. 225.000,-. [EOS]
2	[SOS] kapan wisuda untuk program sarjana iii dijadwalkan pada tahun 2024 [EOS]	[SOS] Wisuda Sarjana III dijadwalkan pada bulan November 2024. [EOS]
3	[SOS] kapan masa heregistrasi dimulai untuk semester genap tahun akademik 2023/2024 di itg [EOS]	[SOS] Masa heregistrasi untuk Semester Genap tahun akademik 2023/2024 di ITG dimulai pada tanggal 5 Februari 2024. [EOS]

Source: (Research Results, 2025)

The examples in Table 2 are intentionally presented in their original Indonesian language to preserve the authenticity of the dataset and the linguistic characteristics of the data used in the preprocessing and modeling stages. After cleaning the dataset and adding [SOS] and [EOS] tokens, the dataset was tokenized to convert the text into numbers. Each number represents a word (special token) in each vocabulary. This is important because machine learning models can only handle numeric data. An illustration of the tokenization process on the dataset can be seen in Figure 2.



Source: (Research Results, 2025)

Figure 2. Illustration of Tokenization Process

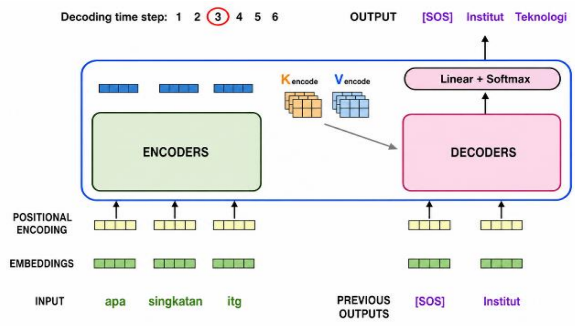
The next step is determining the maximum sequence length for the encoder and decoder, followed by padding the sequences to ensure uniform input size. In this study, both are limited to a maximum length of 80, based on the longest sentence in the dataset. The padded data are then converted into the appropriate data type and finalized for training. Finally, the dataset is divided into smaller batches to improve computational efficiency, with an initial batch size of 32.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8295

Train Model

This stage includes selecting the algorithm or architecture, and training the model with the prepared data. In this modeling, the deep learning algorithm is implemented and the Transformer architecture is applied.



Source: (Research Results, 2025)

Figure 3. Model Transformer Architecture for Chatbot

Figure 3 illustrates the Transformer architecture used to develop the chatbot model, consisting of Encoder and Decoder components, along with Embedding, Positional Encoding, and Linear + Softmax layers. The Embedding layer represents tokens as vectors, while Positional Encoding provides information about word order, which is not inherently captured by the Transformer architecture. The Encoder processes input sequences into contextual representations using self-attention and feed-forward layers, supported by layer normalization and residual connections [15]. These representations capture both individual token meaning and overall context. The Decoder then generates output sequences based on the Encoder output and previous tokens, utilizing masked self-attention, multi-head attention, and feed-forward layers. The final output is passed through linear and softmax layers to predict the next word in the sequence.

Evaluate Model

The evaluation method used to assess the chatbot model in this study primarily employs the BLEU score, which has been widely used as a standard metric for evaluating text generation tasks [16], [17]. The BLEU score evaluates modified n-gram precision by comparing n-grams in candidate sentences with those in reference texts. The modified precision $p_n$is calculated as the ratio of matching n-grams to the total number of candidate n-grams. The final BLEU score is obtained by computing the geometric mean of these precisions up to length N, weighted by w_n, and adjusted by a brevity penalty (BP).

$$Pn = \frac{\sum_{C\in\{candidate\}} \sum_{n-gram\in C} Count_{clip}(n-gram)}{\sum_{C_0\in\{candidate\}} \sum_{n-gram_0\in C_0} Count(n-gram_0)} \quad (1)$$

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} w_n \log p_n) \quad (2)$$

$$Where, BP = \begin{cases} 1, & if\ c > r \\ e^{\left(1-\frac{r}{c}\right)}, & if\ c \le r \end{cases} \quad (3)$$

In addition to BLEU, this study also employs ROUGE and METEOR as complementary evaluation metrics to provide a more comprehensive assessment. ROUGE focuses on recall-based matching by measuring how much of the reference text is captured in the generated response, making it suitable for evaluating content coverage. Variants such as ROUGE-1, ROUGE-2, and ROUGE-L are commonly used to capture unigram overlap, bigram overlap, and longest common subsequence similarity. Recent studies show that ROUGE remains effective for evaluating summarization and text generation tasks due to its ability to measure information completeness [17], [18].

Meanwhile, METEOR evaluates generated text by considering not only exact word matches but also semantic similarity through stemming, synonym matching, and alignment. This allows METEOR to better capture linguistic variations and meaning preservation compared to purely n-gram-based metrics. Recent research highlights that METEOR provides better correlation with human judgment in evaluating generated text, especially in conversational AI and natural language generation tasks [17], [19].

Optimize Model

The optimization stage aims to improve model performance by tuning hyperparameters. The parameters adjusted include encoder–decoder max length, number of layers, batch size, embedding dimension, fully connected dimension, number of heads, positional encoding length, learning rate, and number of epochs. A total of 12 experiments were conducted, grouped into three sets, each containing four different hyperparameter configurations, as detailed in Table 3.

Table 3. Transformer Model Hyperparameter Pairs

Hyperparameter	File 1	File 2	File 3
Maxlen	80	80	80
num_layers	2	2	2
batch_size	32	64	128
embedding_dim	64	128	256
fully_connected_dim	64	128	256
num_heads	2	2	2
positional_encoding_length	128	256	512

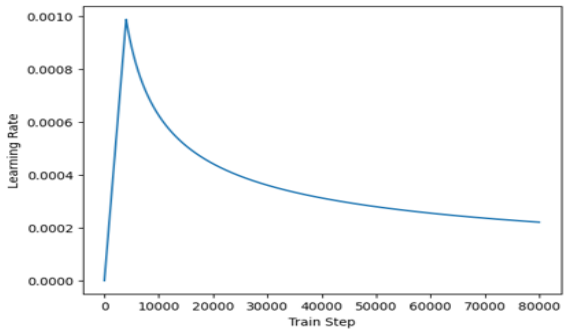
Hyperparameter	File 1	File 2	File 3
learning_rate	0.0002, 0.00025	0.0002, 0.00025	0.0002, 0.00025
epoch	80, 100	80, 100	80, 100

Source: (Research Results, 2025)

RESULTS AND DISCUSSION

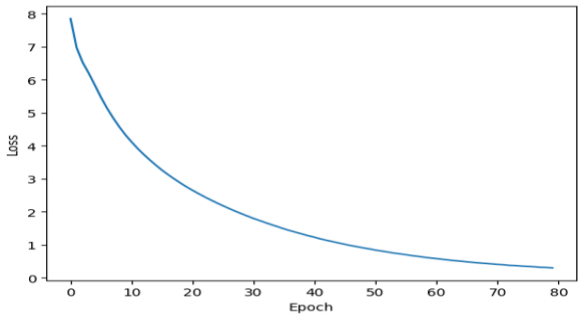
Experimental Result

To create a generative chatbot model for Indonesian language colleges with the proposed method, this research uses the python programming language and utilizes the TensorFlow library. The model will be trained using deep learning algorithms and applying the Transformer model architecture, the model will also be trained using training datasets that have been prepared through data stages. In the training process we utilize free support from google colab and use a T4 GPU. In the first experiment, this research sets a batch size of 32 with a number of iterations of 80. To optimize the model, this research uses the adam optimizer with learning rate parameters that will be searched by learning schedule, as well as $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\in = 1e-9$.



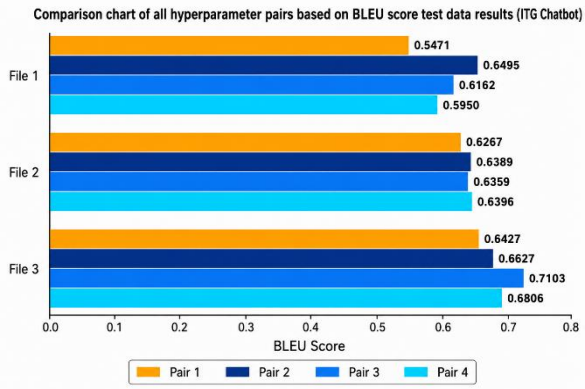
Source: (Research Results, 2025)
Figure 4. Sample Learning Rate Curve For 80000 Training Steps

Figure 4 shows the learning rate curve obtained from the learning schedule over 80,000 training steps, resulting in an optimal learning rate of 0.0002. The model is trained using sparse categorical cross-entropy with a masked loss function to ignore padding tokens. The initial experiment uses a question–answer dataset from Institut Teknologi Garut (ITG), with baseline hyperparameters set as follows: maxlen 80, 2 layers, batch size 64, embedding dimension 128, fully connected dimension 128, 2 attention heads, positional encoding length 256, learning rate 0.0002, and 80 epochs. Additional configurations are tested as outlined in Table 3.



Source: (Research Results, 2025)
Figure 5. Graph of Changes in Loss Values in the Initial Model Training

During training, the loss value decreases rapidly in the early epochs and gradually stabilizes, indicating effective learning. The final loss reaches 0.3371, with a training time of 409.70 seconds. Model performance is evaluated using 314 test samples, achieving a BLEU score of 0.5471 (54.71%), which falls within the "very high quality, adequate, and fluent" range [20]. To further improve performance, hyperparameter tuning is conducted, with the results of different configurations compared in Figure 6.



Source: (Research Results, 2025)
Figure 6. Comparison Diagram of Parameters Based on BLEU Score Test Results

The evaluation results for each hyperparameter configuration across all dataset files are presented as follows. The assessment uses BLEU as the primary metric, supported by ROUGE and METEOR.

Table 4. Evaluation Results for File 1

Hyperparameter	Pair 1	Pair 2	Pair 3	Pair 4
Maxlen	80	100	80	100
num_layers	2	2	2	2
batch_size	32	32	32	32
embedding_dim	64	64	64	64
fully_connected_dim	64	64	64	64
num_heads	2	2	2	2
positional_encoding_length	128	128	128	128

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8295

Hyperparameter	Pair 1	Pair 2	Pair 3	Pair 4
learning_rate	0.0002	0.0002	0.00025	0.00025
epoch	80	100	80	100
Avg BLEU score	54.71%	64.95%	61.62%	59.50%
ROUGE-1	0.8240	0.8336	0.8400	0.8294
ROUGE-2	0.7590	0.7708	0.7732	0.7605
ROUGE-L	0.7990	0.8094	0.8176	0.8097
METEOR	0.8075	0.8013	0.8169	0.8102
Training time (second)	~409.70	~434.88	~358.54	~466.67

Source: (Research Results, 2025)

Based on Table 4, Pair 2 achieves the highest BLEU score (64.95%), indicating the best alignment between generated responses and reference answers. Although Pair 3 shows slightly higher ROUGE and METEOR values, the difference is not significant, making Pair 2 the best configuration for File 1.

Table 5. Evaluation Results for File 2

Hyperparameter	Pair 1	Pair 2	Pair 3	Pair 4
Maxlen	80	100	80	100
num_layers	2	2	2	2
batch_size	64	64	64	64
embedding_dim	128	128	128	128
fully_connected_dim	128	128	128	128
num_heads	2	2	2	2
positional_encoding_length	256	256	256	256
learning_rate	0.0002	0.0002	0.00025	0.00025
epoch	80	100	80	100
Avg BLEU score	62.67%	63.89%	63.59%	63.96%
ROUGE-1	0.8458	0.8372	0.8520	0.8542
ROUGE-2	0.7799	0.7731	0.7948	0.7880
ROUGE-L	0.8264	0.8152	0.8304	0.8270
METEOR	0.8192	0.8172	0.8421	0.8298
Training time (second)	~357.92	~531.87	~380.23	~436.46

Source: (Research Results, 2025)

Based on Table 5, Pair 4 achieves the highest BLEU score (63.96%), indicating the best precision among all configurations in File 2. Although Pair 3 shows slightly higher ROUGE and METEOR values, Pair 4 provides a more balanced overall performance, making it the most reliable configuration for this dataset.

Table 6. Evaluation Results for File 3

Hyperparameter	Pair 1	Pair 2	Pair 3	Pair 4
Maxlen	80	100	80	100
num_layers	2	2	2	2
batch_size	128	128	128	128
embedding_dim	256	256	256	256
fully_connected_dim	256	256	256	256
num_heads	2	2	2	2
positional_encoding_length	512	512	512	512
learning_rate	0.0002	0.0002	0.00025	0.00025
epoch	80	100	80	100
Avg BLEU score	64.27%	66.27%	71.03%	68.06%
ROUGE-1	0.8527	0.8811	0.8738	0.8670
ROUGE-2	0.7894	0.8343	0.8182	0.8120
ROUGE-L	0.8369	0.8636	0.8557	0.8498
METEOR	0.8324	0.8719	0.8478	0.8396
Training time (second)	~787.42	~884.52	~688.31	~903.92

Source: (Research Results, 2025)

Based on Table 6, Pair 3 achieves the highest BLEU score (71.03%), indicating the best performance in generating accurate and contextually appropriate responses. Although Pair 2 records slightly higher ROUGE and METEOR values, Pair 3 offers a better balance between precision and overall response quality, making it the optimal configuration for File 3. Overall, the best result in this study is achieved by File 3 Pair 3, with the highest BLEU score of 71.03%. This demonstrates the model's ability to generate responses that closely align with reference answers while maintaining consistency across evaluation metrics, supported by strong ROUGE and METEOR scores. To further validate the proposed model, a comparative experiment was conducted using a baseline Sequence-to-Sequence (Seq2Seq) model with LSTM Encoder–Decoder and Bahdanau Attention, trained on the same dataset under similar conditions. The evaluation results of the baseline model are presented in the following tables.

Table 7. Evaluation Results of Sequence-to-Sequence Model (File 1)

Hyperparameter	Pair 1	Pair 2
batch_size	32	32
embedding_dim	64	64
units	128	128
epoch	80	100
Avg BLEU score	29.78%	34.17%
ROUGE-1	0.5411	0.5495
ROUGE-2	0.3897	0.3959
ROUGE-L	0.5100	0.5118
METEOR	0.4362	0.4396
Training time (second)	~1924.95	~2324.73

Source: (Research Results, 2025)

Table 8. Evaluation Results of Sequence-to-Sequence Model (File 2)

Hyperparameter	Pair 1	Pair 2
batch_size	64	64
embedding_dim	128	128
units	256	256
epoch	80	100
Avg BLEU score	30.28%	31.52%
ROUGE-1	0.5186	0.5452
ROUGE-2	0.3656	0.3875
ROUGE-L	0.4854	0.5075
METEOR	0.4217	0.4454
Training time (second)	~1615.52	~1998.13

Source: (Research Results, 2025)

Table 9. Evaluation Results of Sequence-to-Sequence Model (File 3)

Hyperparameter	Pair 1	Pair 2
batch_size	128	128
embedding_dim	256	256
units	512	512

Hyperparameter	Pair 1	Pair 2
epoch	80	100
Avg BLEU score	22.33%	30.14%
ROUGE-1	0.4528	0.5089
ROUGE-2	0.2975	0.3458
ROUGE-L	0.4258	0.4773
METEOR	0.3491	0.4113
Training time (second)	3599.84	4273.90

Source: (Research Results, 2025)

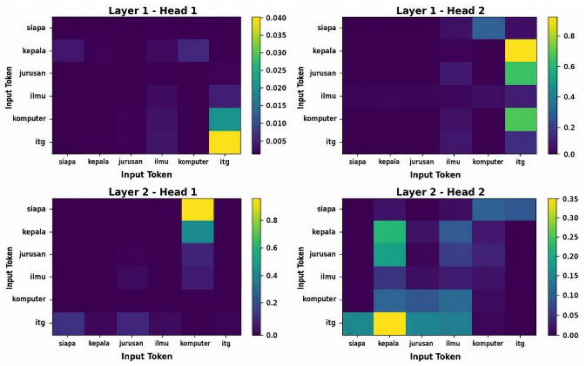
Based on Tables 7–9, the Seq2Seq model achieves BLEU scores between 22.33% and 34.17%, while the Transformer model reaches a best score of 71.03%. Similar trends are observed in ROUGE and METEOR, indicating differences in content coverage and semantic alignment. In terms of computational cost, the Transformer requires longer training time due to its more complex architecture, reflecting a trade-off between performance and efficiency. Within the scope of this study, the Transformer model demonstrates superior performance compared to the Seq2Seq baseline, although these results may not generalize beyond the experimental setting.

Table 10. Comparison of Transformer and Seq2Seq Models

Metric	Seq2Seq (LSTM + Attention)	Transformer
Best File-Pair	File 1 – Pair 2	File 3 – Pair 3
BLEU	34.17%	71.03%
ROUGE-1	0.5495	0.8738
ROUGE-2	0.3959	0.8182
ROUGE-L	0.5118	0.8557
METEOR	0.4396	0.8478
Training time (second)	~2324.73	~688.31

Source: (Research Results, 2025)

Based on Table 10, the Transformer model outperforms the Seq2Seq baseline across all evaluation metrics. It achieves a BLEU score of 71.03%, significantly higher than the 34.17% obtained by the Seq2Seq model, indicating improved response accuracy. This improvement is also reflected in ROUGE and METEOR scores, suggesting better content coverage and semantic alignment. However, the Transformer requires longer training time due to its more complex architecture, highlighting a trade-off between performance and computational cost. Overall, within the scope of this study, the Transformer demonstrates superior performance in generating accurate and contextually appropriate responses. To further analyze model behavior, the self-attention mechanism is examined by visualizing attention weights across encoder and decoder layers in the optimized model.

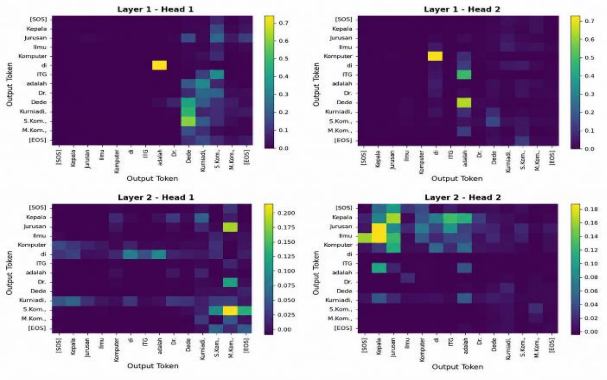


Source: (Research Results, 2025)

Figure 7. Visualization of Self-Attention Weights in the Encoder of the ITG Case Study Model

The self-attention visualization in encoder layers 1 and 2 illustrates how the model captures relationships between tokens and progressively refines its focus. In layer 1, Head 1 emphasizes the token "komputer" and its relation to "itg," indicating contextual relevance, while Head 2 concentrates primarily on "itg," suggesting a highly selective focus on key elements [21], [22].

In layer 2, attention becomes more abstract and contextualized. In Head 1, the model continues to emphasize the relationship between "komputer" and "itg," reinforcing the importance of the token "komputer." In Head 2, attention shifts to tokens such as "kepala" and "itg," indicating the model's ability to capture more complex contextual relationships. This reflects a transition from word-level focus to broader contextual understanding as layer depth increases.



Source: (Research Results, 2025)

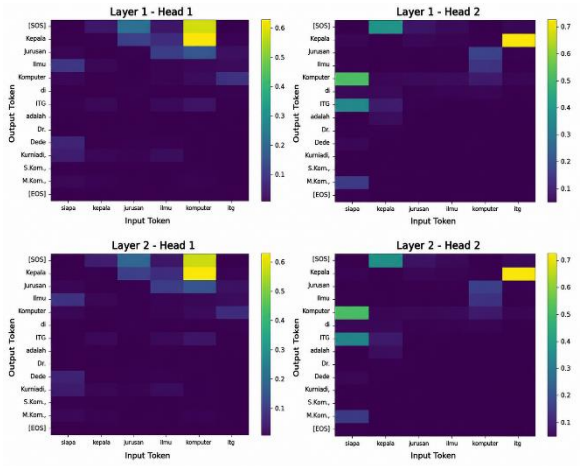
Figure 8. Visualization of Self-Attention Weights in the Decoder of the ITG Case Study Model

In the first decoder layer, the model begins generating output by focusing on relevant tokens in the partially generated sequence, such as "Kurniadi" and "M.Kom" (Head 1) and "adalah" and "itg" (Head 2), preserving key associations like names and titles.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8295

As processing moves to the second layer, attention becomes more refined and incorporates broader contextual information. Head 1 distributes attention more widely across tokens such as "jurusan" and "Dr.", while Head 2 emphasizes tokens like "Dede" and "adalah," improving contextual understanding. This refinement enables the decoder to generate more coherent and contextually relevant responses.



Source: (Research Results, 2025)

Figure 9. Visualization of Attention Weights in the Encoder-Decoder of the ITG Case Study Model

In the initial encoder–decoder attention layer, the model prioritizes key input tokens to guide output generation, ensuring relevance and alignment. Head 1 focuses on tokens such as "[SOS]" and "kepala," while Head 2 emphasizes "itg" and "kepala," indicating their importance in early decoding. In the second layer, attention becomes more complex, allowing the decoder to capture richer contextual information. Head 1 highlights tokens like "jurusan" and "komputer," while Head 2 focuses on "itg" and "siapa," reflecting a deeper understanding of sentence structure and relationships. Overall, these observations indicate that the self-attention mechanism enables Transformer models to integrate contextual information across tokens and layers. Encoder representations capture token-level contextual relationships, while decoder and encoder–decoder interactions contribute to context-aware prediction and source–target alignment during language generation [15].

Discussion

Experimental results on the ITG dataset demonstrate the effectiveness of the Transformer model in generating question–answer pairs with high BLEU scores. Hyperparameter tuning shows that configurations with larger embedding and fully connected dimensions tend to improve performance. This is further supported by consistent gains in ROUGE and METEOR scores, indicating strong precision, content coverage, and semantic alignment.

To further test the generalizability of the model, additional experiments were conducted using a dataset associated with the Campus X. The BLEU score results for the Campus X dataset showed similarities with the results from the ITG dataset, confirming that the Transformer model can be effectively applied to various academic question-answer datasets. However, there are differences in the optimal hyperparameter configuration, especially in terms of learning rate.

The model trained on the ITG dataset achieved the best performance with a learning rate of 0.0002, while the model trained on the Campus X dataset showed improved performance with a learning rate of 0.00025. This confirms the importance of hyperparameter adjustment based on the characteristics of the data set to optimize model performance. The highest BLEU score obtained for the Campus X dataset reaches 72.05%, while the ITG dataset achieves a slightly lower score of 71.03%. This difference indicates that although both datasets yield comparable performance, variations in data characteristics and optimal hyperparameter settings, particularly the learning rate, can influence the final model performance.

Compared to the Seq2Seq model with LSTM Encoder–Decoder and attention, the Transformer consistently achieves higher performance across all evaluation metrics. The Seq2Seq model records significantly lower BLEU, ROUGE, and METEOR scores, indicating weaker content coverage and semantic alignment. This suggests that the Transformer is more effective in capturing contextual relationships and generating accurate, coherent responses within the scope of this study.

The results also demonstrate that the proposed Transformer-based chatbot for Indonesian higher education, particularly at ITG, can provide relevant and context-aware responses. This aligns with previous studies showing that Transformer models outperform RNN and LSTM-based approaches in handling long-term dependencies and multi-turn conversations through self-attention mechanisms [23].

More broadly, these findings support prior research highlighting the effectiveness of Transformers in developing adaptive generative chatbots across domains, including higher education [24], [25]. While the model shows strong performance on the ITG dataset, its generalizability

may be affected by differences in data structure and terminology across institutions. Despite these promising results, several limitations remain. The dataset is limited to a single institution, the model has not been evaluated in multi-turn conversational scenarios, and potential data bias may influence response variation.

Future research should expand the dataset to include multiple universities in Indonesia to improve model generalization. Evaluating the model in more complex conversational scenarios and incorporating user feedback mechanisms can further enhance response quality. Additionally, advanced fine-tuning techniques may improve performance in specific domains. Future work may also explore frameworks such as LangChain to enhance contextual understanding and support longer interactions. By integrating external data sources and enabling more complex query processing, such approaches can produce more accurate, context-aware, and adaptable chatbot responses.

CONCLUSION

This study shows that careful hyperparameter tuning enables the Transformer model to achieve strong performance on both the ITG and Campus X datasets, with best BLEU scores of 71.03% and 72.05%, respectively, supported by consistent ROUGE and METEOR results. Differences in optimal learning rates highlight the importance of dataset-specific tuning. The findings confirm that no single hyperparameter configuration fits all datasets, emphasizing the role of tuning in optimizing performance. Comparative experiments also show that the Transformer outperforms the Seq2Seq model across all metrics, particularly in generating more accurate and contextually appropriate responses. However, several limitations remain, including limited dataset scope, evaluation restricted to single-turn interactions, and potential bias in manually constructed data. Future research should expand datasets across institutions, evaluate multi-turn dialogue scenarios, and integrate external knowledge or advanced frameworks to improve contextual understanding. Further exploration of optimization techniques and alternative architectures may also enhance performance and efficiency.

REFERENCE

[1] P. Kumar, “Large language models (LLMs): survey, technical frameworks, and future challenges,” *Artif. Intell. Rev.*, vol. 57, no. 10, p. 260, 2024, doi: 10.1007/s10462-024-10888-y.

[2] W. X. Zhao et al., "A Survey of Large Language Models," Frontiers of Computer Science, vol. 20, no. 12, Art. no. 2012627, pp. 1–144, 2026, doi: 10.1007/s11704-026-60308-3.

[3] C. Chaka, “Generative AI Chatbots - ChatGPT versus YouChat versus Chatsonic: Use Cases of Selected Areas of Applied English Language Studies,” *Int. J. Learn. Teach. Educ. Res.*, vol. 22, no. 6, pp. 1–19, 2023, doi: 10.26803/ijlter.22.6.1.

[4] T. Wang *et al.*, “Exploring the Potential Impact of Artificial Intelligence (AI) on International Students in Higher Education: Generative AI, Chatbots, Analytics, and International Student Success,” *Appl. Sci. Switz.*, vol. 13, no. 11, 2023, doi: 10.3390/app13116716.

[5] T. Adiguzel, M. H. Kaya, and F. K. Cansu, “Revolutionizing education with AI: Exploring the transformative potential of ChatGPT,” *Contemp. Educ. Technol.*, vol. 15, no. 3, 2023, doi: 10.30935/cedtech/13152.

[6] T. Rasul *et al.*, “The role of ChatGPT in higher education: Benefits, challenges, and future research directions,” *J. Appl. Learn. Teach.*, vol. 6, no. 1, pp. 41–56, 2023, doi: 10.37074/jalt.2023.6.1.29.

[7] T. Thamodharan and M. F. A. Ghani, “A proposed model of ICT facilities in the central zone vocational colleges, Malaysia,” *Int. J. Eval. Res. Educ.*, vol. 12, no. 2, pp. 845–858, 2023, doi: 10.11591/ijere.v12i2.24258.

[8] A. Aljanazrah, G. Yerousis, and G. Hamed, “Digital transformation in times of crisis : Challenges , attitudes , opportunities and lessons learned from students ’ and faculty members ’ perspectives,” no. October, pp. 1–14, 2022, doi: 10.3389/feduc.2022.1047035.

[9] A. Kharchenko et al., "Digital Technologies as a Factor of Transformation of Learning in the University Education," Revista Romaneasca pentru Educatie Multidimensionala, vol. 16, no. 4, pp. 97–126, Dec. 2024, doi: 10.18662/rrem/16.4/909.

[10] S. Janthakal, G. M. Reddy, Stevenson. P, K. S. Aqtar, and A. G. S, “AI-Based Chatbot for College Management System,” *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. 5, pp. 3633–3637, 2023, doi: 10.22214/ijraset.2023.52059.

[11] A. Sarkar et al., "Unlocking the microblogging potential for science and medicine," bioRxiv preprint bioRxiv:2022.04.22.488804, Apr. 2022, doi: 10.1101/2022.04.22.488804.

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8295

[12] Suryani and Mustakim, "An Intelligent Chatbot for Faculty Administration Using Bidirectional LSTM and Seq2Seq Architecture," in *2024 International Conference on Smart Computing, IoT and Machine Learning (SIML)*, 2024, pp. 226–231. doi: 10.1109/SIML61815.2024.10578161.

[13] N. Esfandiari, K. Kiani, and R. Rastgoo, "A Conditional Generative Chatbot using Transformer Model," Sep. 08, 2023, *arXiv*: arXiv:2306.02074. doi: 10.48550/arXiv.2306.02074.

[14] S. Islam *et al.*, "A comprehensive survey on applications of transformers for deep learning tasks," *Expert Syst. Appl.*, vol. 241, 2024, doi: 10.1016/j.eswa.2023.122666.

[15] J. Ferrando, G. I. Gállego, I. Tsiamas, and M. R. Costa-Jussà, "Explaining How Transformers Use Context to Build Predictions," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, vol. 1, pp. 5486–5513, 2023, doi: 10.18653/v1/2023.acl-long.301.

[16] Y. Heryanto and A. Triayudi, "Evaluating Text Quality of GPT Engine Davinci-003 and GPT Engine Davinci Generation Using BLEU Score," vol. 1, no. 4, pp. 121–129, 2023, doi: 10.58905/SAGA.vol1i4.213.

[17] A. Frummet and D. Elsweiler, "Decoding the Metrics Maze : Navigating the Landscape of Conversational Question Answering System Evaluation in Procedural Tasks," pp. 81–90, 2024.

[18] F. F. Kartamanah, A. R. Atmadja, and I. Budiman, "Analyzing PEGASUS Model Performance with ROUGE on Indonesian News Summarization," Sinkron: Jurnal dan Penelitian Teknik Informatika, vol. 9, no. 1, pp. 31–42, Jan. 2025, doi: 10.33395/sinkron.v9i1.14303.

[19] M. Lyu *et al.*, "Natural Language Generation in Healthcare: A Review of Methods and Applications," 2026, doi: 10.1016/j.jbi.2026.104997.

[20] G. Cloud, "Evaluating models | AutoML Translation Documentation." [Online]. Available: https://cloud.google.com/translate/automl/docs/evaluate

[21] Y. Guo, "Interpretability analysis in transformers based on attention visualization," *Appl. Comput. Eng.*, vol. 76, pp. 92–102, Jul. 2024, doi: 10.54254/2755-2721/76/20240571.

[22] R. Katırcı and H. Çelik, "Analyzing The Impact Of Attention Head Count On Transformer-Based Machine Translation," 2025, pp. 105–116.

[23] X. Ke, P. Hu, C. Yang, and R. Zhang, "Human–Machine Multi-Turn Language Dialogue Interaction Based on Deep Learning," Micromachines, vol. 13, no. 3, Art. no. 355, Feb. 2022, doi: 10.3390/mi13030355.

[24] Z. Lv, "Generative artificial intelligence in the metaverse era," *Cogn. Robot.*, vol. 3, no. May, pp. 208–217, 2023, doi: 10.1016/j.cogr.2023.06.001.

[25] R. Gozalo-Brizuela and E. C. Garrido-Merchán, "A Survey of Generative AI Applications," Journal of Computer Science, vol. 20, no. 8, pp. 801–818, Aug. 2024, doi: 10.3844/jcssp.2024.801.818.