



Optimizing Transformer Model FlanT5 for Multi-Question Answering with Context-Aware Learning Rate

Tri Lathif Mardi Suryanto^{1,2}, Aji Prasetya Wibawa², Hariyono³, Andrew Nafalski⁴

^{1,2}Department of Electrical Engineering and Informatics, Faculty of Engineering, Universitas Negeri Malang

¹Information System, Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jawa Timur

³Department of History, Faculty of Social Sciences, Universitas Negeri Malang

⁴UniSA Education Futures, School of Engineering, University of South Australia

¹²trilathif@gmail.com, ^{1*}aji.prasetya.ft@um.ac.id, ³hariyono.fis@um.ac.id, ⁴nafalski@unisa.edu.au

Abstract

This study investigates the performance of FlanT5-based transformer models in handling Multiple-Question Answering (M-QA) tasks, in which multiple semantically related questions must be addressed with a single cohesive answer. Unlike traditional QA systems that focus on one-to-one question-answer pairs, the M-QA approach challenges the model to understand contextual relationships across several questions tied to the same topic. A custom dataset was developed with shared context, grouped questions, and a unified answer to train and evaluate the model. The FlanT5 architecture was fine-tuned using different learning rates (0.0001, 0.0002, 0.0003) to explore the effect of training configurations on model performance. The evaluation was conducted using the ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum metrics. The results indicate that a learning rate of 0.0003 provides the optimal performance, achieving a ROUGE-Lsum score of 0.7390. This study confirms the capability of instruction-tuned transformers to manage complex summarization scenarios that require contextual coherence. The findings are relevant for real-world applications such as intelligent digital assistants, clinical decision support, and educational chatbots. Furthermore, this study emphasizes the importance of hyperparameter tuning in improving the effectiveness of question-driven summarization systems for scalable and efficient deployment.

Keywords: context-aware learning rate; flanT5; NLP; questions answers system; transformer

How to Cite: T. L. M. Suryanto, A. P. Wibawa, H. Hariyono, and A. Nafalski, "Optimizing Transformer Model FlanT5 for Multi-Question Answering with Context-Aware Learning Rate", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 9, no. 6, pp. 1517 - 1525, Jan. 2026.
Permalink/DOI: <https://doi.org/10.29207/resti.v9i6.6985>

Received: July 21, 2025

Accepted: December 16, 2025

Available Online: January 8, 2026

This is an open-access article under the CC BY 4.0 License

Published by Ikatan Ahli Informatika Indonesia

1. Introduction

Recent advancements in Natural Language Processing (NLP), particularly with the advent of Transformer-based models, have substantially enhanced the capabilities of various applications, including chatbots and question-answering (QA) systems. These models have revolutionized language understanding, enabling improved contextual interpretation and response generation in tasks like question answering [1], [2]. Despite the advancements made, a substantial portion of research involving Transformer models continues to focus on simpler frameworks that address single-question, single-answer tasks.

However, studies on more complex scenarios such as Multiple-Question Answering (M-QA), including research on UnifiedQA [3] and PEGASUS [4], highlight the challenges and solutions to handling

multiple related questions within a single context. Additionally, recent developments in instruction-tuned models have proven effective in enhancing question answering capabilities, as seen in models like T5 and FlanT5.

These studies underscore the gap in the current literature concerning M-QA tasks, which this research aims to address, while Transformer models such as T5 [5] - [8] existing benchmarks such as UnifiedQA (48%) [3] and PEGASUS (57%) [4] provide strong baselines for question answering, focusing primarily on single-question scenarios, whereas this study targets multiple related questions requiring a cohesive answer. have shown remarkable success in handling basic QA systems

The ongoing challenge within Transformer-based approaches is managing more complex scenarios, such as providing a cohesive response to multiple

interrelated questions that share the same answer, Multiple-Question Answering (M-QA) is essential to distinguish it from single-question QA, where each question is answered independently. In M-QA, multiple questions sharing the same context are linked to generate a single, cohesive answer.

For example, questions about Goddess Durga such as "What does Goddess Durga symbolize?" and "Who is the goddess representing strength?" are answered consistently as Durga is a symbol of strength. This issue, which concerns maintaining semantic coherence across multiple questions, remains unresolved. Despite the widespread application of Transformers for individual question-answer pairs, the scalability of these models for M-QA tasks, which require a deeper understanding of inter-question relationships, remains an open area of research [9]. This research aims to fill this gap by exploring Transformer models' effectiveness in answering multiple related questions with a unified response.

The importance of this research lies in its potential to enhance the efficiency and accuracy of NLP systems, especially in complex, real-world applications such as customer [10] - [13], educational platform [14] - [16] and healthcare [17] - [19]. By examining how

Transformer models handle semantic relationships across multiple questions, this study will contribute to the development of more advanced and practical QA systems.

The focus of this study is to assess various Transformer models FlanT5 in M-QA tasks. This will involve a comparative analysis of their effectiveness in understanding the relationships between questions and providing accurate, consistent answers, thereby offering insights into model training strategies, architecture enhancements, and dataset diversification for improved performance. Using evaluation metrics such as ROUGE, this study aims to provide valuable insights into optimizing these models for complex question-answering scenarios, ultimately advancing the development of sophisticated AI chatbot systems.

2. Methods

This section outlines the methodologies employed in the design experiment and evaluation of the proposed model, detailing the data preparation process, model architecture, hyperparameters, and the evaluation metrics used to assess its performance.

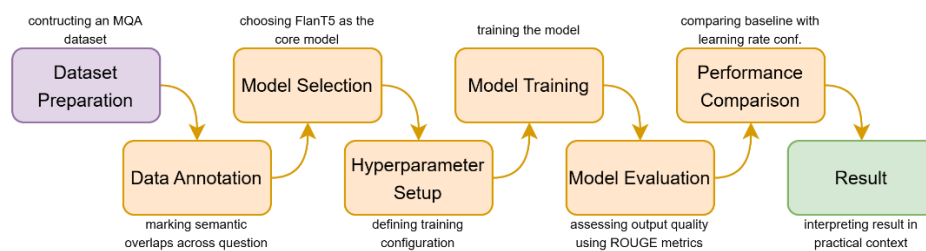


Figure 1. Research Method

This research follows a structured pipeline starting with the preparation of an M-QA dataset containing multiple semantically related questions and a unified answer. Data annotation highlights overlapping semantics to guide model understanding. FlanT5 is selected as the core model. The rationale for using FlanT5 with learning rate optimization is present but weakly argued as novel. Context-Aware Learning Rate (CALR) has been shown to improve model convergence and stability by dynamically adjusting the learning rate based on the context of training data [20], [21].

This approach is expected to outperform traditional fine-tuning methods by reducing the likelihood of overfitting and accelerating learning, particularly in tasks that require deep semantic understanding, followed by hyperparameter setup including learning rate, batch size, and optimizer. The model is trained for 60 epochs and evaluated using ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum. Various learning rates are compared to identify optimal performance.

2.1 Dataset Preparation

Optimising the data generation process to produce large-scale data sets involves utilising methodologies that improve efficiency and effectiveness. Various approaches, including generative modelling, simulation techniques, and advanced data processing methods, can significantly improve the quality and quantity of data set. In Figure 2, suggestion model experiment dataset

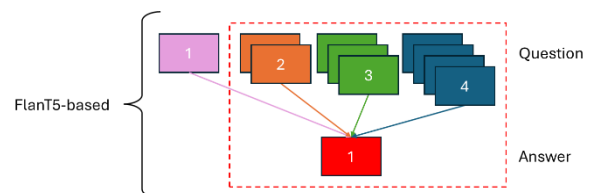


Figure 2. Convergent Questioning Mapping

The coloured boxes labelled 2, 3, and 4 represent different but related questions that are linked to the same topic. All these questions point to one unified answer, shown in red at the bottom. This setup helps the model understand different ways a question can be asked and still generate one clear and accurate response.

It trains the model to connect multiple questions to the same correct answer using the given context.

This study's Multiple-Question Answering (M-QA) dataset was constructed by gathering context-specific questions and coherent and consistent replies. This dataset has 500 questions and 2,000 answer separated into four categories: symbolism, physical traits, history, and mythology. Manual annotation was used to verify each question and response by two experts for contextual relevance and correctness; to prevent overfitting, techniques such as data augmentation and regularization methods were employed, ensuring that the model generalizes well despite the relatively small dataset size.

Cross-checking and error checking were used to ensure that the responses were relevant and appropriate to the questions posed. Data quality analysis demonstrates

that this dataset accurately represents the themes investigated, with an equitable distribution of questions across each category and strong answer consistency.

In Table 1. It is explained that each word that has been marked becomes knowledge of the answers to all questions which are then summarised to become answers, which questions are marked more, it raises the possibility of a greater answer match, for example in the dur_1 context with the answer 'Durga is a symbol of strength', question 2 and question 4 have a chance of questions that will be quickly responded to this is marked with yellow, purple, and blue which are included in the mark of the answer. Likewise with dur_2 and dur_3, the number of similar questions is one of the factors that provide faster answers.

Table 1. Creation Dataset Multiple Question Answer (M-QA)

Context	Question 1	Question 2	Question 3	Question 4	Answer
dur-1	apa yang dilambangkan oleh dewi durga "what does goddess durga symbolize"	siapa dewi dengan simbol kekuatan di hindu "who is the goddess with the hindu symbol of power"	durga adalah perlambangan dari simbol apa "durga is the epitome of what symbol"	dewi apa yang melambangkan kekuatan "what goddess symbolizes power"	durga adalah simbol kekuatan. durga is a "symbol of strength"
dur-2	bagaimana mata dewi durga "how is the eye of goddess durga"	seperti apa mata dewi durga "what does the eye of goddess durga look like"	mata durga berukuran apa "what size are durga's eyes"	apa ciri mata dewi durga "what is the meaning of goddess durga"	durga memiliki ukuran mata yang besar. "durga has large eye size"
dur-3	siapa musuh yang dikalahkan oleh dewi durga "who was the enemy defeated by goddess durga"	dewi durga mengalahkan musuh siapa "goddess durga defeated whose enemy"	durga mengalahkan musuh yang mana "durga defeated which enemy"	musuh durga yang mana yang dikalahkan "which enemy of durga was defeated"	musuh durga yang dikalahkan adalah mahishasura. "durga's defeated enemy was mahishasura"

Table 2 give an explanation the hyperparameters used in the training of this model are designed to ensure the stability, efficiency, and generalisability of the model in handling complex M-QA tasks. A learning rate of 0.0002 was chosen to provide stable training and prevent large jumps in optimisation, so that the model can achieve convergence gradually without losing stability. The selection of learning rates (0.0001, 0.0002, 0.0003) was based on the principles of context-aware [20], [22], [23], learning rate [24], [25], which dynamically adjusts the learning rate based on the model's context during training. This research suggests that moderate learning rates are optimal for maintaining gradient stability and improving convergence speed, especially in tasks with complex interdependencies, like M-QA. The batch size for training and evaluation was set at 24, an ideal compromise to ensure gradient stability while utilising GPU capacity efficiently. The consistency of batch size between training and evaluation ensures that the evaluation results reflect the true performance of the trained model.

The seed was set at 42 to guarantee deterministic experiments, which is important to ensure the results can be replicated and support the validity of the study.

The use of AdamW as the optimiser with parameters `betas=(0.9, 0.999)` and epsilon of `1e-08` provides adaptive capabilities in optimising the gradient, controlling momentum, and preventing overfitting through the weight decay mechanism. Linear learning rate scheduler is applied to gradually decrease the learning rate during training, allowing the model to avoid overfitting in the final epoch while still learning optimally.

Table 2. Result Implementing Hyperparameter

Hyperparameter	Nilai
learning_rate	0.0002
train_batch_size	24
eval_batch_size	24
seed	42
optimizer	adamw_torch with betas=(0.9, 0.999) epsilon=1e-08
lr_scheduler_type	Linear
num_epochs	60

2.2 Hyperparameter

In this study, a careful selection of hyperparameters is made to ensure the model can be trained optimally and produce maximum performance on M-QA tasks.

Learning rate is chosen to provide a balance between convergence speed [26] - [28] and training stability [25], [29], [30] so that the model can learn gradually without the risk of losing convergence. Train batch size and eval batch size to maximise GPU computational efficiency while maintaining gradient stability during training [31]. This study employs the AdamW optimizer to update model parameters during training. The update rule, shown in Equation 1, enables adaptive learning rates while decoupling weight decay to improve training stability and regularization.

$$n_t = n_0 \times \max \left(0, \frac{1-t/T}{1} \right) \quad (1)$$

In Equation 1, θ_t denotes the model parameters at training step t , α represents the learning rate, $\hat{m}_t$ and $\hat{v}_t$ correspond to the bias-corrected first and second moment estimates of the gradients, respectively, ϵ is a small constant added for numerical stability, and λ indicates the weight decay coefficient used to prevent overfitting.

The choice of AdamW is particularly strategic, as it has demonstrated excellent adaptive capabilities for optimizing gradients [32], consistent performance across various tasks [33], and strong regularization properties to prevent overfitting [34], especially in models with a large number of parameters.

To quantitatively evaluate the lexical overlap between generated answers and reference answers in the M-QA setting, this study employs the ROUGE metric. Specifically, ROUGE- n is computed as defined in Equation 2, measuring the recall of overlapping n -grams between system outputs and gold-standard summaries.

$$\begin{aligned} m_t &= \beta_1 m_t - 1 + (1 - \beta_1) \nabla \theta_t \mathcal{L} \\ v_t &= \beta_2 v_t - 1 + (1 - \beta_2) (\nabla \theta_t \mathcal{L})^2 \\ \hat{m}_t &= \frac{m_t}{1 - \beta_1^t}, \hat{v}_t = \frac{v_t}{1 - \beta_2^t} \\ \theta_t &= \theta_{t-1} - n_t = \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \end{aligned} \quad (2)$$

In Equation 2, n denotes the n -gram length, $\text{Count}_{\text{match}}(n)$ represents the number of overlapping n -grams between the generated answer and the reference answer, while $\text{Count}_{\text{ref}}(n)$ indicates the total number of n -grams in the reference text.

ROUGE assesses the similarity between generated text and reference text by analyzing n -grams, phrase lengths, and lexical-level information. Studies [5], [35] have demonstrated that ROUGE is highly effective in evaluating outputs from generative models, particularly in assessing accuracy and contextual relevance in tasks like open-domain QA and sarcasm generation. However, these studies also highlight ROUGE's limitations, such as its inability to capture more complex semantic relationships, suggesting the need to combine ROUGE with manual evaluations or supplementary metrics for more comprehensive results.

[36] introduced ROUGE-SS, a new variant that is more sensitive to semantic structures, making it more suitable for evaluating complex text tasks.

Additionally, research by [37] - [39] emphasized that while ROUGE remains a dominant metric, integrating it with human evaluations or graph-based statistical methods can improve the accuracy of assessing domain-specific outputs. ROUGE-L (Equation 3) is used to evaluate structural coherence via the longest common subsequence between generated and reference answers.

$$\text{ROUGE} - L_{\text{sum}} = \frac{1}{m} \sum_{i=1}^m \frac{\text{LCS}(S_i, R_i)}{\text{Length}(R_i)} \quad (3)$$

In Equation 3, $\text{LCS}(X, Y)$ denotes the length of the longest common subsequence between the generated text X and the reference text Y . The precision and recall components are computed based on the LCS length relative to the lengths of X and Y , enabling balanced evaluation of both content coverage and structural alignment.

In this study, unlike conventional one-to-one question answering, M-QA requires evaluation metrics that reflect both lexical accuracy and consistency across multiple semantically linked questions.

3. Results and Discussions

In this chapter, the results obtained from the study are presented and analyzed in detail. The discussion will focus on interpreting these findings, relating them to the research objectives, and comparing them with existing literature. This section aims to provide new insights into M-QA, examining how the findings align with or differ from previous studies, and exploring the implications of these results for both theory and practice in the field in figure.

Figure 3 illustrates the performance of the FlanT5-based baseline model over 30 epochs of training. The upper plot shows a significant decrease in both training and validation loss during the initial epochs, indicating rapid model adaptation. As training progresses, the losses continue to decline at a slower pace and remain closely aligned, suggesting that the model learns effectively without overfitting. By the final epoch, the model achieves approximately 0.73 in ROUGE-1, 0.71 in ROUGE-2, and 0.73 in ROUGE-L, indicating strong summarization performance. These results confirm that the FlanT5-based model effectively learns to produce high-quality, coherent, and contextually accurate summaries over time.

Figure 4 presents the training performance of the FlanT5-based model with a learning rate of 0.0001. The top graph displays the training and validation loss over 30 epochs. Unlike the smoother convergence seen in the baseline model, the loss curves here show more fluctuation, particularly in the training loss, which experiences abrupt drops and spikes. This irregular pattern suggests that the learning rate might be too high for stable gradient descent, leading to oscillations rather

than steady convergence. Despite these instabilities, the validation loss continues to decrease and remains relatively lower, indicating that the model is still improving in its generalization performance. The bottom graph shows the evolution of ROUGE scores, where ROUGE-1, ROUGE-2, and ROUGE-L steadily

increase throughout training. However, the improvement appears to plateau earlier compared to the baseline. ROUGE-1 and ROUGE-L approach values around 0.43, while ROUGE-2 stabilizes just under 0.30, indicating moderate summarization performance.

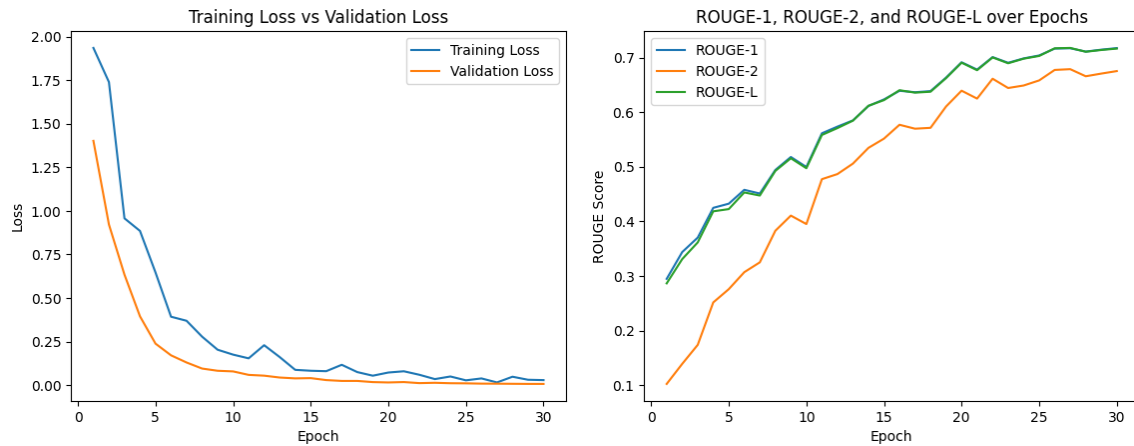


Figure 3. FlanT5-based

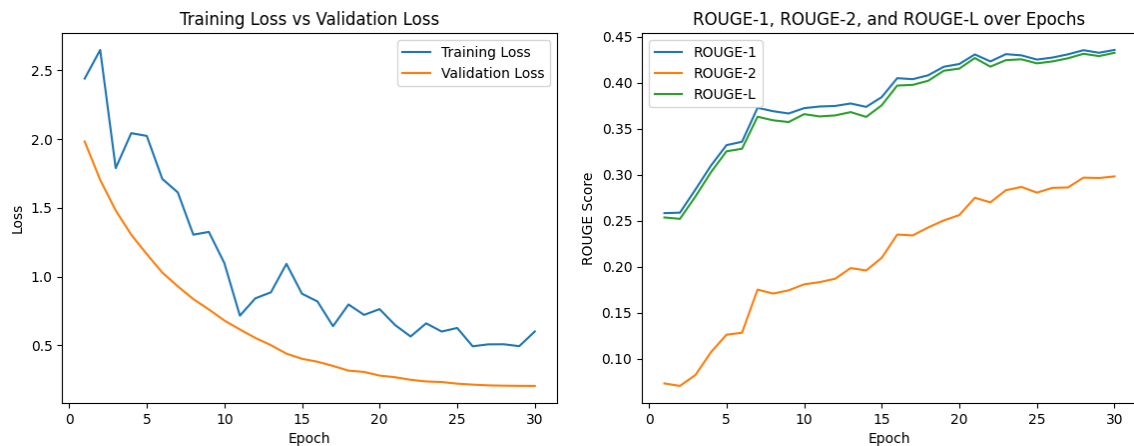


Figure 4. FlanT5-based+LR1

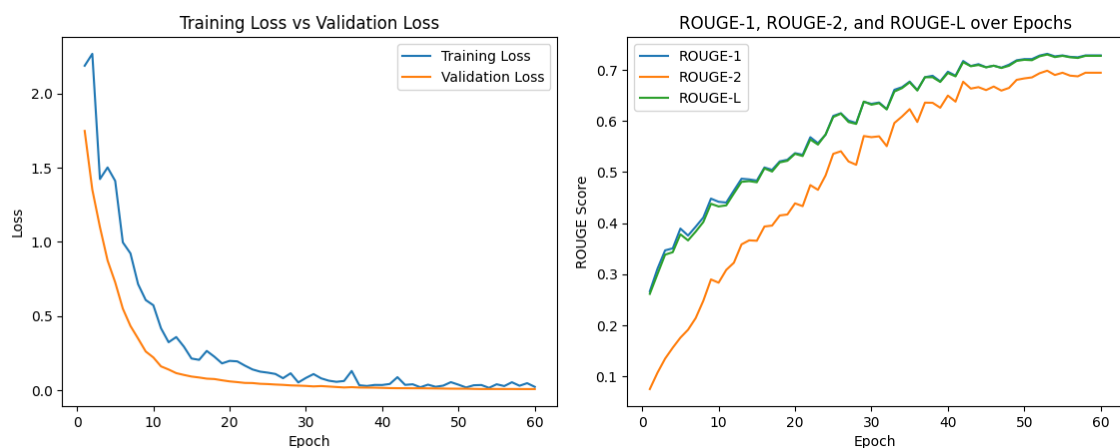


Figure 5. FlanT5-based+LR2

Figure 5 displays the performance of the FlanT5-based model trained with a learning rate of 0.0002 across 60 epochs. After this initial phase, the losses continue to decrease at a slower pace and converge toward zero, suggesting the model successfully minimizes error

without signs of overfitting, as the validation loss closely follows the training loss throughout. Notably, ROUGE-L reaches slightly above 0.72, while ROUGE-1 and ROUGE-2 approach approximately 0.73 and 0.69 respectively by the final epoch. These results indicate

that the learning rate of 0.0002 supports more stable and sustained learning compared to lower rates, enabling

the model to generate high-quality, semantically accurate, and structurally coherent summaries.

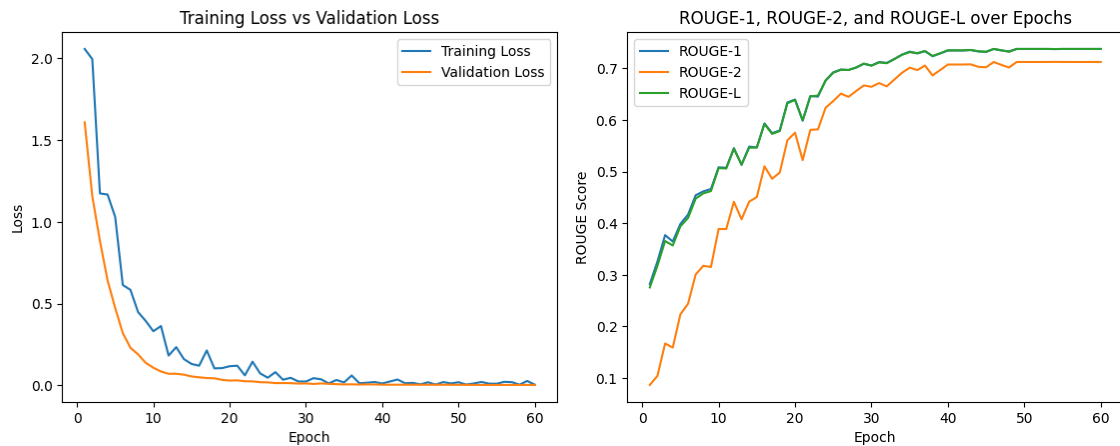


Figure 6. FlanT5-based+LR3

Figure 6 presents the training dynamics of the FlanT5-based model with a learning rate of 0.0003 over 60 epochs. The top plot illustrates a smooth and rapid decrease in both training and validation loss during the first 20 epochs, followed by a gradual decline and convergence towards near-zero values in later epochs. The close alignment between training and validation loss across epochs reflects consistent model generalization and the absence of overfitting, even at this higher learning rate. The bottom graph in Figure 6 shows the ROUGE score progression throughout training. ROUGE-1, ROUGE-2, and ROUGE-L all demonstrate strong, steady growth, with no sharp drops or instability. By the final epochs, ROUGE-1 and ROUGE-L scores stabilize around 0.7379, while ROUGE-2 peaks at 0.7128. This plateau in high performance metrics suggests that the learning rate of 0.0003 allows the model to reach its optimal summarization capability efficiently and effectively. Overall, this configuration achieves the best convergence and highest ROUGE scores, making it the most effective setting for training the FlanT5-based summarization model.

The ROUGE metrics in Table 3, this table compares the performance of FlanT5-baseline with three model variants trained using different learning rates.

Table 3 presents a comprehensive evaluation of the FlanT5-based model's summarization performance using ROUGE metrics, comparing the baseline configuration with variations using different learning rates. The baseline FlanT5 model already shows strong results with ROUGE-1 at 0.7173, ROUGE-2 at 0.6752, and ROUGE-L at 0.7164, indicating solid performance in capturing unigrams, bigrams, and sentence-level structures in the generated summaries. However, when the model is fine-tuned with a learning rate of 0.0001, performance significantly drops across all metrics (e.g., ROUGE-1 = 0.4357), suggesting underfitting due to overly conservative updates [40], who noted that small learning rates can inhibit transformer adaptation during fine-tuning.

Table 3. Model Performance Based on ROUGE

Model	Rouge1	Rouge2	RougeL	RougeLsum
FlanT5-based	0.7173	0.6752	0.7164	0.7174
FlanT5+LR 0.0001	0.4357	0.2982	0.4326	0.4327
FlanT5+LR 0.0002	0.7279	0.6940	0.7270	0.7260
FlanT5+LR 0.0003	0.7378	0.7126	0.7379	0.7390

Conversely, increasing the learning rate to 0.0002 leads to a noticeable improvement across all scores, with ROUGE-1 reaching 0.7279 and ROUGE-2 reaching 0.6940. This aligns with prior work [41], which emphasizes the importance of optimizing learning rates for achieving effective convergence and generalization. The best performance is obtained at a learning rate of 0.0003, where the model attains its highest scores—ROUGE-1 at 0.7378, ROUGE-2 at 0.7126, and ROUGE-Lsum at 0.7390—demonstrating that a slightly more aggressive learning rate can facilitate better learning dynamics in FlanT5.

This trend is consistent with observations [4] reported that modest increases in learning rate often led to faster convergence and higher performance in encoder-decoder models when carefully regularized. Overall, the results underscore the critical role of learning rate tuning in maximizing the performance of QA-based summarization systems built on transformer.

This research is critical because it targets the intersection of summarization and question answering a hybrid use case increasingly demanded in knowledge-intensive domains such as healthcare, law, policy, and education. Unlike generic summarization, question-driven summarization requires models not only to compress information but also to understand and highlight context-relevant content in response to specific queries.

The FlanT5-based architecture, enhanced through careful hyperparameter tuning as demonstrated in this study, offers a compelling solution for such tasks by

integrating instruction tuning and transfer learning capabilities [42]. The empirical results in this study provide strong evidence that even minor architectural adjustments—such as tuning the learning rate—can significantly improve the quality of generated answers or summaries, thus enhancing user trust and task efficiency.

Which abstractive summarization was explored in task-specific contexts, and it was demonstrated that domain-aware summarization consistently outperforms general-purpose models when fine-tuned appropriately. Additionally [4], [43] on retrieval-augmented QA systems has shown that models trained with context relevance in mind produce more accurate, human-aligned responses, reinforcing the value of fine-grained learning strategies like those applied in our experiments.

This study significantly enhances the FlanT5 model via a targeted fine-tuning method for many-Question Answering (M-QA) tasks, allowing the model to provide coherent responses to many interconnected questions inside a single context. This method contrasts with earlier models that concentrated only on one-to-one responses, using semantic connections between questions to provide a more holistic answer.

Experiments on hyperparameter optimisation, specifically on the learning rate, indicate that a rate of 0.0003 produces optimal results, achieving a ROUGE-Lsum of 0.7390, hence enhancing both convergence time and output quality relative to standard configurations. The value 0.0003 is the optimal solution for datasets with specific characteristics, namely 20,000 questions with 5,000 answers. Other datasets will certainly require different learning rate characteristics.

The results indicate that FlanT5 performs well in most M-QA tasks, with optimal learning rates leading to significant improvements in contextual coherence. However, despite these positive outcomes, the model faced challenges in handling questions that were highly ambiguous or required complex reasoning. Further error analysis revealed that questions with overlapping meanings, but subtle differences led to inconsistent answers, highlighting the need for further refinements. These limitations should be addressed in future work by incorporating techniques such as more advanced context modelling or error correction algorithms.

FlanT5's applications span from educational chatbots to virtual healthcare guidance and intelligent digital assistants owing to its prior efficiency and adaptability over earlier models. It can handle complex interlinked queries, reduce the need for human input, and operate on low-resource devices. The results from this study suggest that hyperparameter optimization as well as model-specific adjustments of FlanT5 greatly enhance the quality of M-QA, thus revealing a notable potential for tackling complex and multi-layered text data in real-world scenarios.

Moreover, in the context of cultural heritage and historical texts, the proposed M-QA setting is particularly relevant due to the presence of symbolic meanings, repetitive narratives, and semantically overlapping questions. The experimental results indicate that FlanT5 can maintain contextual coherence when generating unified answers across culturally grounded questions, such as those found in mythology and historical narratives. This demonstrates the model's suitability for cultural heritage and history-oriented question-driven summarization tasks.

This lightweight efficiency makes the model deployable on edge devices or within cloud-constrained environments, broadening its accessibility. As such, the model is not only technically robust but also socio-technically inclusive, offering a scalable solution that democratizes access to advanced natural language processing capabilities. Its real-world utility extends to enhancing productivity, reducing cognitive workload, and supporting more equitable access to AI-enhanced decision support across diverse sectors and user groups.

4. Conclusions

This study confirms that fine-tuning the FlanT5-based model with carefully selected learning rates significantly enhances its performance in question-driven summarization tasks. Among the tested configurations, a learning rate of 0.0003 yielded the highest ROUGE scores across the board, demonstrating the model's ability to generate concise, coherent, and context-aware summaries that align well with user intent.

These findings emphasize the value of targeted hyperparameter optimization in unlocking the full potential of instruction-tuned transformer architectures like FlanT5. Given its high accuracy and efficiency, the model holds strong potential for deployment in real-world applications across knowledge-intensive sectors such as healthcare, legal services, and digital public administration.

However, several limitations should be acknowledged. The study's evaluation relied exclusively on automatic metrics such as ROUGE, which, while informative, may not fully reflect the semantic depth, fluency, or factual accuracy of the generated summaries. Moreover, the experiments were limited to a single dataset and task domain, which may constrain the generalizability of the findings. Other influential hyperparameters and architectural modifications—such as batch size, optimizer variants, or attention heads—were not explored in this research. Additionally, aspects related to real-world deployment, such as model latency, interpretability, and user feedback, were not measured.

For future research, it is recommended to incorporate human-centered evaluations to assess the contextual and semantic quality of the summaries more holistically. Expanding the experiments to include

multilingual datasets and few-shot learning scenarios could improve the model's adaptability. Further optimization strategies, such as dynamic learning rate schedules, curriculum learning, or reinforcement learning with human feedback, should also be explored. Finally, integrating the model into end-user applications and assessing its practical impact on usability, task efficiency, and decision-making outcomes would provide valuable insights into its real-world utility

Acknowledgements

This research was carried out with institutional support from Universitas Negeri Malang (UM) and Universitas Pembangunan Nasional "Veteran" Jawa Timur (UPNVJT), as well as the B-26 Research Group, whose collaborative academic environment contributed to the completion of this study. The authors declare no conflict of interest.

References

- [1] S. S. Lakkimsetty, S. V. Latchireddy, S. M. Lakkoju, G. R. Manukonda, and R. V. V. M. Krishna, "Fine-Tuned Transformer Models for Question Answering," in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, IEEE, Jul. 2023, pp. 1–5. doi: 10.1109/ICCCNT56998.2023.10307046.
- [2] A. Rawat and S. Singh Samant, "Comparative Analysis of Transformer based Models for Question Answering," in *2022 2nd International Conference on Innovative Sustainable Computational Technologies (CISCT)*, IEEE, Dec. 2022, pp. 1–6. doi: 10.1109/CISCT55310.2022.10046525.
- [3] D. Khashabi et al., "UNIFIEDQA: Crossing format boundaries with a single QA system," *Find. Assoc. Comput. Linguist. Find. ACL EMNLP 2020*, pp. 1896–1907, 2020, doi: 10.18653/v1/2020.findings-emnlp.171.
- [4] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "PEGASUS: Pre-Training with extracted gap-sentences for abstractive summarization," *37th Int. Conf. Mach. Learn. ICML 2020*, vol. PartF16814, pp. 11265–11276, 2020.
- [5] P. Tummala and C. K. Roa, "Exploring T5 and RGAN for Enhanced Sarcasm Generation in NLP," *IEEE Access*, vol. 12, pp. 88642–88657, 2024, doi: 10.1109/ACCESS.2024.3416692.
- [6] B. Guan, X. Zhu, and S. Yuan, "A T5-based interpretable reading comprehension model with more accurate evidence training," *Inf. Process. Manag.*, vol. 61, no. 2, p. 103584, Mar. 2024, doi: 10.1016/j.ipm.2023.103584.
- [7] S. Soltanmohammadi, Z. U. Shamszaman, G. Shatha, and M. Jamila, "Job-Matching Chatbots Powered by T5: A Comparative Performance Study with GPT-2," *Proc. - Int. Conf. Dev. eSystems Eng. DeSE*, pp. 325–330, Mar. 2024, doi: 10.1109/DESE60595.2023.10468896.
- [8] T. L. M. Suryanto, A. P. Wibawa, H. Hariyono, and A. Nafalski, "Comparative Performance of Transformer Models for Cultural Heritage in NLP Tasks," *Adv. Sustain. Sci. Eng. Technol.*, vol. 7, no. 1, p. 0250115, Jan. 2025, doi: 10.26877/asset.v7i1.12111.
- [9] S. Wiegrefe, O. Taffjord, Y. Belinkov, H. Hajishirzi, and A. Sabharwal, "Answer, Assemble, Ace: Understanding How Lms Answer Multiple Choice Questions," *ICLR 2025*, pp. 1–25, 2025.
- [10] I. D. Raharjo and Egia Rosi Subhiyako, "Implementing Long Short Term Memory (LSTM) in Chatbots for Multi Usaha Raya," *Adv. Sustain. Sci. Eng. Technol.*, vol. 6, no. 4, p. 02404018, Oct. 2024, doi: 10.26877/asset.v6i4.934.
- [11] A. Janson, "How to leverage anthropomorphism for chatbot service interfaces: The interplay of communication style and personification," *Comput. Human Behav.*, p. 107954, 2023, doi: 10.1016/j.chb.2023.107954.
- [12] P. D. Larasati, A. Irawan, S. Anwar, M. F. Mulya, M. A. Dewi, and I. Nurfatima, "Chatbot helpdesk design for digital customer service." Accessed: Dec. 20, 2024. [Online]. Available: <https://pubs2.ascee.org/index.php/aet/article/view/684>
- [13] C. V. Misischia, F. Poecze, and C. Strauss, "Chatbots in customer service: Their relevance and impact on service quality," *Procedia Comput. Sci.*, vol. 201, no. C, pp. 421–428, 2022, doi: 10.1016/j.procs.2022.03.055.
- [14] S. Kesarwani, Titiksha, and S. Juneja, "Student Chatbot System: A Review on Educational Chatbot," in *2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)*, IEEE, Apr. 2023, pp. 1578–1583. doi: 10.1109/ICOEI56765.2023.10125876.
- [15] D. P. P. Villanueva and I. Aguilar-Alonso, "A Chatbot as a Support System for Educational Institutions," in *2021 62nd International Scientific Conference on Information Technology and Management Science of Riga Technical University (ITMS)*, IEEE, Oct. 2021, pp. 1–6. doi: 10.1109/ITMS52826.2021.9615271.
- [16] S. Hussain, S. H. Al-Hashmi, M. H. Malik, and S. I. Ali Kazmi, "Chatbot in E-learning," *SHS Web Conf.*, vol. 156, p. 01002, Jan. 2023, doi: 10.1051/shsconf/202315601002.
- [17] H. Yu, C. Yu, Z. Wang, D. Zou, and H. Qin, "Enhancing Healthcare through Large Language Models: A Study on Medical Question Answering," 2024, [Online]. Available: <http://arxiv.org/abs/2408.04138>
- [18] Y. W. Sung, D. S. Park, and C. G. Kim, "A Study of BERT-Based Classification Performance of Text-Based Health Counseling Data," 2022, doi: 10.32604/cmcs.2022.022465.
- [19] A. Babu and S. B. Boddu, "BERT-Based Medical Chatbot: Enhancing Healthcare Communication through Natural Language Understanding," *Explor. Res. Clin. Soc. Pharm.*, vol. 13, p. 100419, Mar. 2024, doi: 10.1016/j.jrcsop.2024.100419.
- [20] N. Nascimento, P. Alencar, C. Lucena, and D. Cowan, "A Context-Aware Machine Learning-based Approach A Context-Aware Machine Learning-based Approach," *CASCON'2018, Oct. 2018, Markham, Ontario Canada*, no. July 2020, 2018.
- [21] T. L. M. Suryanto, A. P. Wibawa, H. Hariyono, and H. Shili, "Contextual Relevance-Driven Question Answering Generation: Experimental Insights Using Transformer-Based Models," *Int. J. Eng. Sci. Inf. Technol.*, vol. 5, no. 4, pp. 103–113, Jul. 2025, doi: 10.52088/ijesty.v5i4.989.
- [22] P. K. Rachabatuni, F. Principi, P. Mazzanti, and M. Bertini, "Context-aware chatbot using MLLMs for Cultural Heritage," in *Proceedings of the ACM Multimedia Systems Conference 2024 on ZZZ*, New York, NY, USA: ACM, Apr. 2024, pp. 459–463. doi: 10.1145/3625468.3652193.
- [23] D. A. Sulisty, D. D. Prasetya, F. A. Ahda, and A. P. Wibawa, "Pivoted Low Resource Multilingual Translation with NER Optimization," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 24, no. 5, pp. 1–16, May 2025, doi: 10.1145/3727876.
- [24] Y. Wu and L. Liu, "Selecting and Composing Learning Rate Policies for Deep Neural Networks," *ACM Trans. Intell. Syst. Technol.*, vol. 14, no. 2, Feb. 2023, doi: 10.1145/3570508/ASSET/CC4BF25D-3B4F-4EAF-BD15-33E5CBB999E0/ASSETS/GRAPHIC/TIST-2020-12-0363-F07.JPG.
- [25] C. Xu, S. Liu, Z. Yang, Y. Huang, and K. K. Wong, "Learning Rate Optimization for Federated Learning Exploiting Over-the-Air Computation," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3742–3756, Dec. 2021, doi: 10.1109/JSAC.2021.3118402.
- [26] G. Retsinas, G. Sfikas, P. P. Filntisis, and P. Maragos, "Newton-Based Trainable Learning Rate," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Jun. 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10094576.
- [27] A. Mah, C. E. M. Golden, and C. M. Constantinople, "Dopamine transients encode reward prediction errors independent of learning rates," *Cell Rep.*, vol. 43, no. 10, p. 114840, Oct. 2024, doi: 10.1016/j.celrep.2024.114840.
- [28] K. Lv, H. Yan, Q. Guo, H. Lv, and X. Qiu, "AdaLomo: Low-memory Optimization with Adaptive Learning Rate," Oct. 2023, Accessed: Dec. 13, 2024. [Online]. Available:

- <https://arxiv.org/abs/2310.10195v3>
- [29] M. A. Hannan *et al.*, "SOC Estimation of Li-ion Batteries with Learning Rate-Optimized Deep Fully Convolutional Network," *IEEE Trans. Power Electron.*, vol. 36, no. 7, pp. 7349–7353, Jul. 2021, doi: 10.1109/TPEL.2020.3041876.
- [30] Y. Wu *et al.*, "Demystifying Learning Rate Policies for High Accuracy Training of Deep Neural Networks," in *2019 IEEE International Conference on Big Data (Big Data)*, IEEE, Dec. 2019, pp. 1971–1980. doi: 10.1109/BigData47090.2019.9006104.
- [31] M. Choi, "An Empirical Study on the Optimal Batch Size for the Deep Q-Network," 2019, pp. 73–81. doi: 10.1007/978-3-319-78452-6_8.
- [32] Z. Zhuang, M. Liu, A. Cutkosky, and F. Orabona, "Understanding AdamW through Proximal Methods and Scale-Freeness," no. 2019, 2022, [Online]. Available: <http://arxiv.org/abs/2202.00089>
- [33] R. Llusi, S. El Yacoubi, A. Fontaine, and P. Lupera, "Comparison between Adam, AdaMax and Adam W optimizers to implement a Weather Forecast based on Neural Networks for the Andean city of Quito," in *2021 IEEE Fifth Ecuador Technical Chapters Meeting (ETCM)*, IEEE, Oct. 2021, pp. 1–6. doi: 10.1109/ETCM53643.2021.9590681.
- [34] Shem L. Gonzales, "Enhancing Image Classification Performance: A Comparative Analysis of Optimization Algorithms," *Int. J. Adv. Res. Sci. Commun. Technol.*, pp. 641–645, Jun. 2023, doi: 10.48175/IJARSCT-11918.
- [35] S. Sanniboina, S. Trivedi, and S. Vijayaraghavan, "LoRE: Logit-Ranked Retriever Ensemble for Enhancing Open-Domain Question Answering," Oct. 2024, Accessed: Dec. 12, 2024. [Online]. Available: <https://arxiv.org/abs/2410.10042v1>
- [36] S. Kumar and A. Solanki, "ROUGE-SS: A New ROUGE Variant for Evaluation of Text Summarization," *Authorea Prepr.*, Jul. 2023, doi: 10.22541/AU.168984209.92955863/V1.
- [37] H. S. Ali, L. M. Kefali, S. Parida, and S. R. Dash, "Amharic ATS - A Comparison Between Graph Based and Statistical Based Approach using Rouge Metric and Human Evaluation," *2022 OPJU Int. Technol. Conf. Emerg. Technol. Sustain. Dev. OTCON* 2022, 2023, doi: 10.1109/OTCON56053.2023.10114029.
- [38] A. Auriemma Citarella, M. Barbella, M. G. Ciobanu, F. De Marco, L. Di Biasi, and G. Tortora, "Rouge Metric Evaluation for Text Summarization Techniques." 2024. doi: 10.2139/ssrn.4753220.
- [39] M. Barbella and G. Tortora, "Rouge Metric Evaluation for Text Summarization Techniques," *SSRN Electron. J.*, May 2022, doi: 10.2139/SSRN.4120317.
- [40] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," no. 1, 2019, [Online]. Available: <http://arxiv.org/abs/1907.11692>
- [41] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 1–67, 2020.
- [42] H. W. Chung *et al.*, "Scaling Instruction-Finetuned Language Models," pp. 1–54, 2022, [Online]. Available: <http://arxiv.org/abs/2210.11416>
- [43] S. Narayan *et al.*, "Conditional Generation with a Question-Answering Blueprint," *Trans. Assoc. Comput. Linguist.*, vol. 11, pp. 974–996, Dec. 2023, doi: 10.1162/tacl_a_00583.