International Journal of Engineering, Science and Information Technology
Volume 5, No. 4 (2025) pp. 103-113
eISSN: 2775-2674
Website: http://ijesty.org/index.php/ijesty
DOI: https://doi.org/10.52088/ijesty.v5i4.989
Research Paper, Short Communication, Review, Technical Paper

Contextual Relevance-Driven Question Answering Generation: Experimental Insights Using Transformer-Based Models

Tri Lathif Mardi Suryanto[1,2], Aji Prasetya Wibawa[1*], Hariyono[3], Hechmi Shili[4]

[1]*Department of Electrical Engineering and Informatics, Faculty of Engineering, Universitas Negeri Malang, Malang, Indonesia*
[2]*Department of Information Systems, Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jawa Timur, Indonesia*
[3]*Department of History, Faculty of Social Sciences, Universitas Negeri Malang, Malang, Indonesia*
[4]*Department of Computer Science, Haql University College, University of Tabuk, Saudi Arabia*

**Corresponding author Email: aji.prasetya.ft@um.ac.id*

The manuscript was received on 24 February 2025, revised on 10 May 2025, and accepted on 23 August 2025, date of publication 4 November 2025

Abstract

This study investigates the impact of contextual relevance and hyperparameter tuning on the performance of Transformer-based models in Question-Answer Generation (QAG). Utilising the FlanT5 model, experiments were conducted on a domain-specific dataset to assess how variations in learning rate and training epochs affect model accuracy and generalisation. Six QAG models were developed (QAG-A to QAG-F), each evaluated using ROUGE metrics to measure the quality of generated question-answer pairs. Results show that QAG-F and QAG-D achieved the highest performance, with QAG-F reaching a ROUGE-LSum of 0.4985. The findings highlight that careful tuning of learning rates and training duration significantly improves model performance, enabling more accurate and contextually appropriate question generation. Furthermore, the ability to generate both questions and answers from a single input enhances the interactivity and utility of NLP systems, particularly in knowledge-intensive domains. This study underscores the importance of contextual modelling and hyperparameter optimisation in generative NLP tasks, offering practical insights for improving chatbot development, educational tools, and digital heritage applications.

Keywords: *Question Generation, Transformer Models, Hyperparameter Optimisation, Contextual Relevance, Cultural Heritage.*

1. Introduction

The rapid advancement of Natural Language Processing (NLP) has been largely propelled by the introduction of Transformer-based models such as T5, BERT [1] [2][3][4]. These models have transformed a wide range of NLP tasks, including text summarization [5][6], question answering [7][8] [9][10], and natural language generation [8][11] [12][13][14][15][16][17]. Despite their widespread success, fine-tuning these powerful models remains a significant challenge, primarily due to the sensitivity of hyperparameters such as the learning rate. Improper tuning can lead to poor generalisation or inefficient computation, limiting their practical effectiveness [18][19][20].

A core challenge in fine-tuning Transformer models lies in balancing convergence speed with model generalisation. Setting the learning rate too high risks training instability and suboptimal performance, while too low a learning rate can result in slow convergence and overfitting [5][21]. Various learning rate scheduling strategies—including warm-up, linear decay, cosine annealing, and adaptive gradient methods—have been proposed to address these issues [13][22], showing promise in enhancing model convergence and mitigating catastrophic forgetting [23][24]. However, a systematic and comprehensive evaluation of these learning rate techniques across multiple NLP tasks remains lacking.

This study discusses how context-based contributions to automated questions and answers, hence the need for fine-tuning techniques of Transformer-based NLP models. The key contributions of this work are: (1) a comprehensive evaluation of learning rate scheduling strategies tailored for Transformer-based NLP models; (2) an in-depth analysis of the influence of learning rate on model convergence and performance across multiple datasets; and (3) practical recommendations to guide researchers and practitioners in effectively fine-tuning Transformer models [12][25]. These contributions are especially relevant to domain-specific applications, such as cultural heritage research, where optimised NLP models can substantially improve information retrieval, knowledge extraction, and automated question-answering systems. By highlighting the crucial role of hyperparameter optimisation in Transformer fine-tuning, and provides actionable insights for enhancing NLP model performance. Ultimately, these findings aim to facilitate more accessible and effective NLP solutions across diverse research domains.

2. Methods

This section details the research methodology employed to investigate the impact of learning rate scheduling on fine-tuning Transformer-based models for natural language processing (NLP) tasks.

2.1. Data Preparation and Scenario Experiment

In Figure 1, the contextual relevance-driven question answer generation model development pipeline, the process starts with the selection of the base model (Model Selection), then is trained using a dataset of question-answer pairs (Dataset Usage). Fine-tuning is then performed to adapt the model to the data domain.

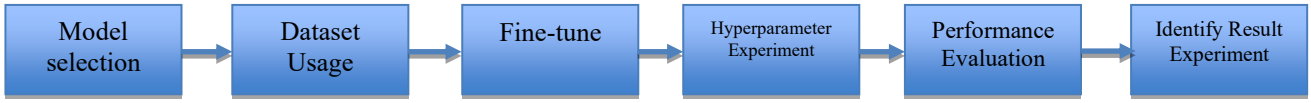
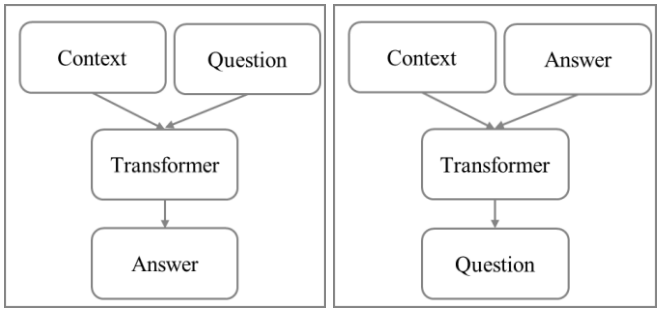


Fig 1. Methods Research

The Hyperparameter Experiment stage aims to find the optimal parameter configuration. The model is evaluated in the Performance Evaluation stage, and the experimental results are analysed in the last stage (Identify Result Experiment) to determine the effectiveness of the model and the direction of further development.

In most research in the field of Question Answer Generation (QAG), context only acts as a static input that supports the generation of questions or answers, so the model is less able to capture the dynamics and changes in context in actual conversations. As a result, the chatbot's ability to provide relevant and adaptive responses to the evolving context is limited.



(a) (b)

Fig 2. Question Generation Conventional, (a) Modelling Patterns to Generate Answers, (b) Model Pattern for Generating Answers.

In general, both patterns have limitations in building a responsive and adaptive context-based chatbot. Pattern (a) focuses more on matching questions and context to answer, but lacks flexibility in understanding the context of the conversation holistically. Pattern (b) is useful in training data generation, but less efficient for direct interaction with users. While conventional approaches often rely on manually annotated datasets and limited linguistic flexibility, modern Transformer-based models provide superior generalisation by capturing broader contexts and enabling multi-task learning.

The study builds on recent advances in Transformer-based question generation models, with a focus on generating both questions and answers simultaneously to improve contextual relevance and semantic coherence. Unlike traditional models that generate either questions or answers independently (Figure 1), the proposed method leverages a unified architecture inspired by recent pre-trained models such as T5 and its instruction-tuned variant, FlanT5 [26][27][28].

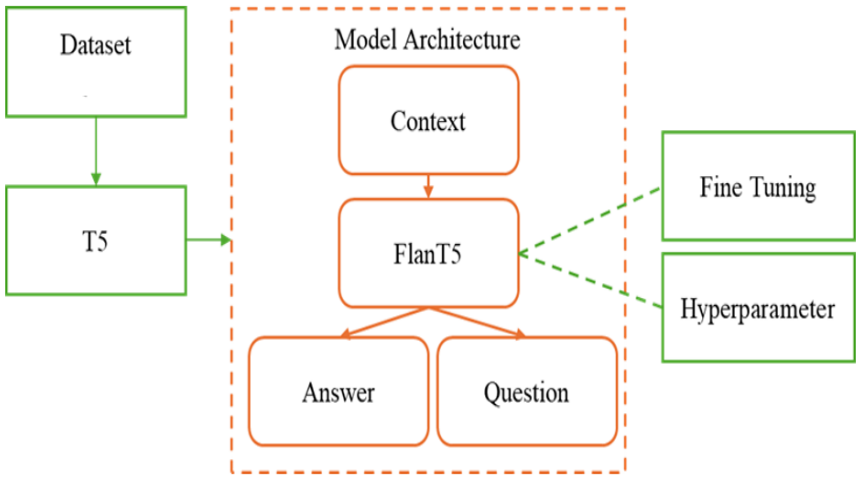


Fig 3. Question-Answer Contextual Relevance Model Architecture.

In Figure 3, This latest image shows the architecture of a more advanced and comprehensive model using FlanT5 as the core of the context-based question and answer generation system. This model has some significant advantages over the simple pattern shown in the

previous Figures (a) and (b) in Figure 2. Firstly, FlanT5 is a generative transformer model that has undergone pre-training and fine-tuning, so it can process context, questions and answers simultaneously and flexibly. Unlike the models in Figures (a) and (b), which only process one direction (question to answer or vice versa), FlanT5 can learn the complex relationship between conversational context, question, and answer in more depth. Secondly, the fine-tuning process and explicit setting of hyperparameters in this architecture provide better optimisation space, so that the model can be customised to specific datasets and application domains. This makes the context-based chatbot more accurate, relevant and adaptive to variations in user input.
Thirdly, the model can generate two outputs at once, namely answers and questions from a single context input. This allows for applications in two-way learning, such as additional data generation, question validation, and more natural interactive dialogue, which is an improvement over the models in Figures (a) and (b) in Figure 2, which only produce one output. Fourthly, the use of live datasets as input to FlanT5 shows that the system learns thoroughly from real data, in contrast to previous models that are simpler and less supportive of deep learning. This improves the chatbot's ability to understand the broad context and nuances of complex conversations.

2.1. Optimisation Methods and Model Training

So, the core model architecture in Figure 3 utilised is FlanT5, an instruction-tuned extension of the T5 model that improves stability, generalisation, and performance on diverse NLP tasks. FlanT5 employs an encoder-decoder Transformer design, enabling it to simultaneously generate questions and answers from input contexts effectively, Fine-tuning procedures involve training the FlanT5 model variants on the Question-Answer Pairs Goddess Durga dataset and settings hyperparameters explored Cross-Entropy Loss [11][29], Objective Function [11][30] [31], Gradient Accumulation [28][32][33] [34][35], Optimizer Step – AdamW [36][37], include:

Cross-Entropy Loss

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}^{target}.\log(y_{ij}^{pred}) \quad (1)$$

Definition 2.1:
N is the sample size.
M is the length of the token in the target sequence.
y_{ij}^{target} is the target token to-j from data to-i.
y_{ij}^{target} is the probability of the predicted token to-j from data to-i.

Objective Function

$$\theta = \arg\min_{\theta} \mathcal{L}\,(f(X,\theta),Y) \quad (2)$$

Definition 2.2:
θ is parameter model.
$f(X,\theta)$ is the model's prediction of the input X.

Y is the sequence target.
$\mathcal{L}$ is the loss function (Cross-Entropy Loss).

Gradient Accumulation

$$\Delta\theta = \frac{1}{G}\sum_{a=1}^{G}\eta\,.\frac{\partial\mathcal{L}_g}{\partial\theta} \quad (3)$$

Definition 2.3:
G is the number of gradient accumulation steps (gradient_accumulation_steps = 4).
η is the learning rate (7×10^{-5}).
$\mathcal{L}_a$ is the loss of the batch to-g.

Optimizer Step – AdamW

$$\theta_{t+1} = \theta_t - \eta.\frac{\partial\mathcal{L}}{\partial\theta_t} \quad (4)$$

Definition 2.4:
θ_{t+1} is the updated parameter.
η is the learning rate.

Hyperparameter optimization is vital for efficient training and better performance in deep learning models, especially for complex tasks like question answering, especially for complex tasks like question answering. Studies show that settings like learning rate and batch size

greatly impact accuracy [4][38]. For the T5 model, fine-tuning with optimal hyperparameters improves NLP task results by accelerating convergence and enhancing effectiveness.

Table 1. Research Experiments

Model Name	Learning Rate	Num Epochs	Optimizer
QAG-CR1	0.0003	60	Adam w_torch (betas=(0.9, 0.999), epsilon=1e-08, no additional optimizer arguments)
QAG-CR2	0.0003	90	
QAG-CR3	0.0003	120	
QAG-CR4	5,00E-05	120	
QAG-CR5	1,00E-05	120	
QAG-CR6	3,00E-05	120	

The experiments systematically evaluated QAG models across different hyperparameters, focusing on learning rates, epochs, and optimiser settings. QAG-60 and QAG-90 were trained for 60 and 90 epochs, respectively, at a fixed learning rate of 0.0003, while QAG-120 variants tested various learning rates with 120 epochs. Consistent batch sizes (24) and a fixed seed (42) ensured reproducibility. Using the Adam W optimiser contributed to stable training. Results aim to assess if lower learning rates improved convergence and generalisation, and whether longer training caused overfitting or performance gains.

ROUGE metrics were used to evaluate generated text quality by comparing overlaps with reference summaries. Widely adopted in NLP, ROUGE has proven effective for benchmarking models in tasks like question answering [7][33][39][40] and text summarisation [6] [41][42][43][44] [45].

$$ROUGE - Lsum_{recall} = \frac{LCS(X,Y)}{m}$$
$$ROUGE - Lsum_{precision} = \frac{LCS(X,Y)}{n} \quad (5)$$
$$ROUGE - Lsum_{F1} = \frac{(1+\beta^2).ROUGE - Lsum_{precision} \,.\, ROUGE - Lsum_{recall}}{\beta^2.ROUGE - Lsum_{precision} \,.\, ROUGE - Lsum_{recall}}$$

Definition 2.5:

$LCSsum(X,Y)$ is the length of the longest common subsequence computed across all sentences between the reference summary (X) and candidate summary (Y).

m is the total number of words in the reference summary.

n is the total number of words in the candidate summary.

β controls the weighting between recall and precision (typically set to 1 for equal weight).

An easy way to comply with the paper formatting requirements is to use this document as a template and simply type your text into it.

3. Results and Discussions

This section presents experimental findings on how different training configurations affect accuracy, convergence speed, and computational efficiency when fine-tuning Transformer-based NLP models. It highlights the trade-offs between performance and resource usage, especially for tasks like question answering.

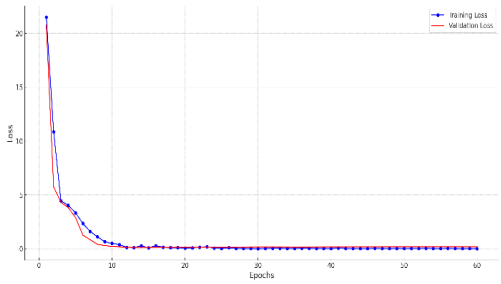


Fig 4. Model Experiment QAG-A

For the QAG-A in Figure 4, the training loss begins at approximately 21.5, indicating the model starts with significant room for optimisation. Throughout the 60 epochs, the training loss steadily declines to ~0.19, demonstrating effective learning within the given duration. This trend closely mirrors the validation loss, which also stabilises around ~0.19 by the end of the training. The alignment between training and validation losses reflects the model's ability to generalise effectively to unseen data, with minimal overfitting. However, the shorter training duration of 60 epochs appears to constrain the model's potential to reach the lower loss levels observed in models trained for 90 or 120 epochs. This suggests that additional epochs might allow the model to fine-tune its parameters further, resulting in even better convergence and potentially improved downstream performance.

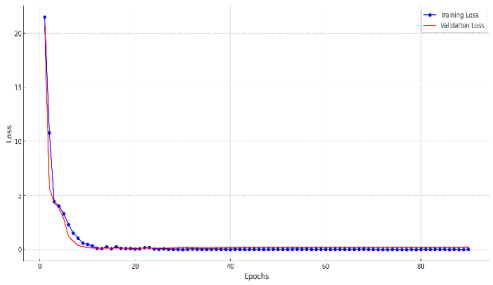


Fig 5. Model Experiment QAG-B

Figure 5 for QAG-B, the training loss exhibits a significant decline, starting at approximately 21.5 and dropping to ~0.22 by the end of 90 epochs. This steady reduction indicates effective learning throughout the training process, demonstrating that the model was able to capture the underlying patterns in the data within the given epoch limit. The validation loss stabilises around ~0.22, showing a slightly higher value than the training loss, which is expected as it reflects the model's performance on unseen data. This consistency between training and validation losses suggests that the model generalises reasonably well without significant overfitting.

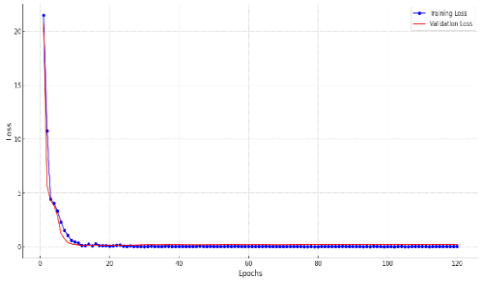


Fig 6. Model Experiment QAG-C

For the QAG-C model in Figure 6, the training loss starts at approximately 21.5 and steadily decreases to around 0.20 by the end of 120 epochs, demonstrating consistent learning throughout the training process. The validation loss follows a similar trend, stabilising at approximately 0.20, which is slightly higher than the final training loss. This indicates that while the model is learning effectively, it may not be fully optimised for generalisation compared to some of the other models, such as 120d and 120b, which achieve lower final losses.

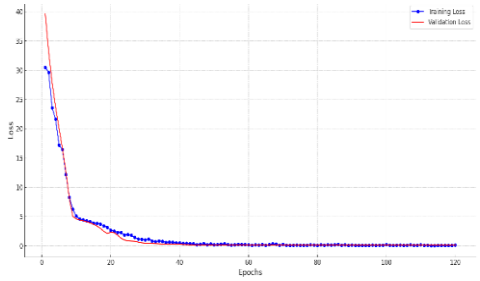


Fig 7. Model Experiment QAG-D

In Figure 7 QAG-D model, the QAG-D model training loss exhibits a rapid decline from approximately 30.5 at the start to a remarkably low 0.13 by the end of 120 epochs. This significant reduction in training loss reflects the model's ability to effectively learn the task as it progresses through the training iterations. Similarly, the validation loss mirrors this behaviour, decreasing steadily and stabilising around 0.13, closely aligning with the training loss. The QAG-D model demonstrates strong learning performance, achieving low final loss values. This outcome suggests an optimal balance between model complexity and the duration of training, allowing the model to achieve high task-specific performance without overfitting or underfitting. This model configuration could serve as a benchmark for efficient and effective training strategies.

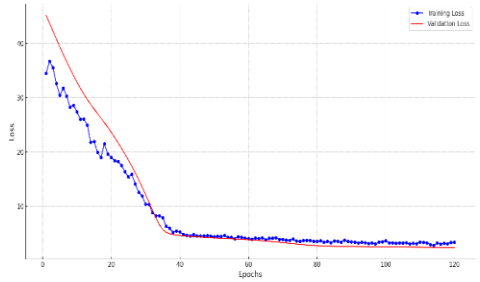


Fig 8. Model Experiment QAG-E

The QAG-E model in Figure 8 exhibits a noticeable decline in training loss, starting at approximately 34.5 and decreasing to around 2.39 by the end of 120 epochs. While this indicates that the model is learning to some extent, the training loss remains significantly higher compared to the corresponding loss in models like 120d, which suggests that the learning process is less efficient. Similarly, the validation loss stabilises at approximately 2.38, mirroring the pattern of the training loss.
However, the relatively high final loss values for both training and validation highlight potential issues with the model's configuration or learning process. One plausible explanation is suboptimal hyperparameters, such as the lower learning rate used in this model, which might slow down the convergence or limit the depth of learning. Additionally, the higher loss values might also reflect challenges in adapting to the dataset or capturing the nuances of the question generation task. Compared to 120d, the learning appears less effective, suggesting that improvements in parameter tuning or optimisation strategy could significantly enhance the model's performance.

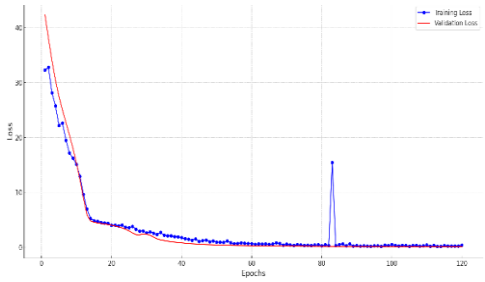


Fig 9. Model Experiment QAG-F

For the QAG-F model in Figure 9, the training loss begins at a high value of approximately 32, reflecting the initial state of the model's parameters and its lack of knowledge about the task. Throughout 120 epochs, the training loss steadily decreases, eventually converging to a low value of ~0.12. This consistent decline indicates that the model effectively learns the patterns and relationships in the training data, optimising its parameters progressively throughout the training process.

Table 2. Research Experiments

Model	ROUGE1	ROUGE2	ROUGEL	ROUGELSum
QAG-A	0.4027	0.0936	0.3317	0.3595
QAG-B	0.2213	0.0920	0.2137	0.2180
QAG-C	0.3778	0.1232	0.3460	0.3447
QAG-D	0.4965	0.1956	0.4702	0.4882
QAG-E	0.2935	0.2208	0.2920	0.2959
QAG-F	0.5056	0.2117	0.4942	0.4985

Among the evaluated models, QAG-F and QAG-D emerge as the top performers, achieving ROUGE-1 scores of 0.5056 and 0.4965, respectively, alongside strong ROUGE-2 and ROUGE-L results. These scores reflect a high level of lexical and semantic overlap with reference data, indicating that these models effectively capture contextual information and generate fluent, relevant questions. In contrast, QAG-A and QAG-C demonstrate moderate performance with ROUGE-1 scores around 0.40 and 0.37, respectively. Although their ROUGE-2 and ROUGE-L scores fall below those of QAG-D and QAG-F, they still suggest reasonable question generation quality, albeit with less precision in capturing longer or more complex n-gram patterns. On the lower end, QAG-B and QAG-E perform significantly worse, with QAG-B recording the lowest ROUGE-1 score of 0.2213 and a poor ROUGE-L of 0.2137, indicating weaker lexical overlap and contextual relevance. While QAG-E shows a relatively higher ROUGE-2 score of 0.2208, its ROUGE-1 and ROUGE-L scores remain low, near 0.29, suggesting challenges in overall lexical similarity and coherence.

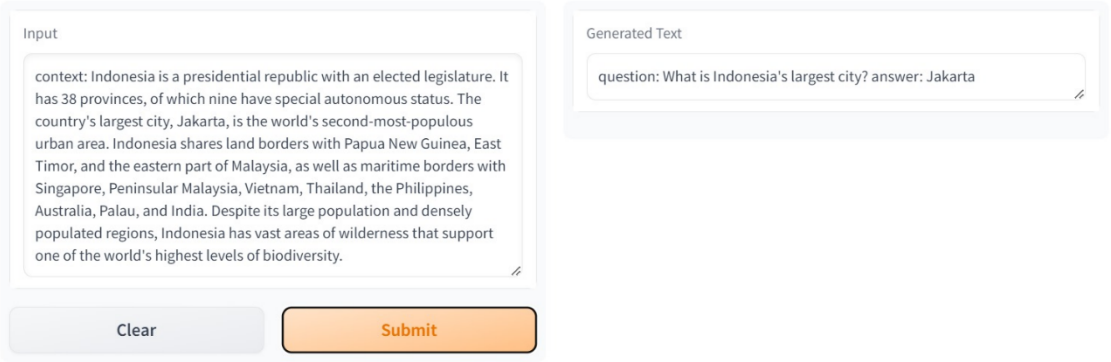


Fig 10. Result Experiment Output Text

The generated text output is as shown in Figure 10. The success in generating questions and answers from a given context, commonly referred to as Question Answer Generation, is a key indicator in the development of natural language understanding systems. The outcome of this process is a coherent and meaningful question-answer pair that can be utilised for various purposes, such as training question answering models, developing chatbots, or assessing reading comprehension. The effectiveness of QAG is strongly influenced by the quality of the input context, the model's ability to understand sentence structure and meaning, and the accuracy of the generated question-answer pairs.

It shows that from the same context, the system can generate different pairs of questions and answers in each iteration. This capability provides flexibility in information exploration, where a variety of questions can explore different aspects of a single text source. In Natural Language Processing (NLP) based research, especially in specific domains such as cultural heritage, this approach allows exploring deeper insights into a cultural heritage by automatically generating questions that cover historical, social, and anthropological perspectives.

This study systematically evaluated the performance of six Transformer-based Question-Answer Generation (QAG) models—QAG-A through QAG-F—focusing on the effects of hyperparameter tuning on accuracy, convergence speed, and computational efficiency. The key findings revealed that models QAG-F and QAG-D outperformed others, achieving the highest ROUGE-1 scores of 0.5056 and 0.4965, respectively, indicating strong lexical and semantic overlap with the reference data. These results suggest that precise hyperparameter tuning, particularly in terms of learning rate, number of training epochs, and fine-tuning strategy, significantly enhances model performance.

The results align with the initial hypothesis that appropriate hyperparameter optimisation plays a critical role in model efficacy. The top-performing models not only demonstrated substantial reductions in training loss but also maintained a close alignment between training and validation loss, reflecting strong generalisation capabilities. These findings are consistent with prior studies [4][21], which underscore the importance of hyperparameter configuration in optimising Transformer-based models.

Compared to earlier research, such as [13], which explored adaptive learning techniques, this study contributes practically by demonstrating that even with minimal fine-tuning, strategically selected configurations can achieve competitive results. The findings support the theoretical foundations of attention mechanisms, emphasising their effectiveness in capturing token-level dependencies and generating semantically coherent questions. In practical terms, the implications of this research are significant, particularly for applications in intelligent question-answering systems, educational chatbots, and natural language processing (NLP) tasks related to cultural heritage. As illustrated in Figure 10, the model's ability to generate diverse question-answer pairs from a single context allows for flexible exploration of information. This capability is especially valuable in domains that require multi-perspective analysis, such as historical or anthropological content.

From a theoretical perspective, the findings reinforce the importance of deep contextual modelling in generative NLP tasks. The study confirms that architectural elements such as attention mechanisms and contextual embeddings are critical in producing coherent and contextually appropriate outputs. This opens avenues for future exploration of hybrid architectures that integrate retrieval-based approaches with generative models to further enhance relevance and specificity.

Nevertheless, the study is not without limitations. First, the experiments were conducted using a domain-specific dataset (Goddess Durga), which may constrain the generalizability of the results across other textual domains. Second, the evaluation relied heavily on ROUGE metrics, which, while useful, primarily capture surface-level similarities and may not fully reflect semantic quality. Third, the scope of hyperparameter optimisation was limited; advanced strategies such as learning rate warm-up and decay methods were not explored. Future studies should consider expanding the dataset to include multi-domain corpora to assess the adaptability of the models in more complex contexts. Additionally, incorporating human evaluation alongside automated metrics would provide a more nuanced assessment of question quality. Exploring hybrid architectures that combine generative capabilities with retrieval mechanisms, and applying reinforcement learning to optimise question relevance based on user feedback, also represent promising directions for further research.

4. Conclusions

This study systematically provides compelling and insightful findings by evaluating the performance of six distinct Question-Answer Generation (QAG) models—QAG-A, QAG-B, QAG-C, QAG-D, QAG-E, and QAG-F—using a straightforward fine-tuning approach. The evaluation employed the RPUGE Score as a robust metric to assess the models' effectiveness in generating questions that are contextually relevant and semantically coherent. Notably, the QAG-F model achieved the highest RPUGE score of 0.498, demonstrating superior capability in generalising complex contextual information into meaningful and accurate question-answer pairs. This strong performance reflects the model's advanced understanding of contextual nuances and its ability to maintain natural language fluency, underscoring the crucial role of sophisticated architectural components such as mindfulness mechanisms and contextual embeddings that are finely optimised during the training refinement phase.

These findings substantiate that the architectural design choices, particularly the adoption of attention-based Transformer mechanisms coupled with enhanced input retrieval techniques, are pivotal to the success of QAG models. Such designs facilitate the generation of queries that are not only relevant but also coherent and appropriately aligned with the provided context. While some models, like QAG-D and QAG-F, showed relatively modest performance scores above 0.7 in other evaluation metrics, their results nevertheless open avenues for more nuanced analysis and model improvement. This indicates that even models with moderate scores can contribute valuable insights and serve as platforms for further experimentation.

Looking forward, future research directions could involve the exploration of hybrid architectures that integrate generative models with retrieval-based methods to leverage the strengths of both paradigms, thereby enhancing the relevance and precision of generated questions. Additionally, incorporating reinforcement learning techniques could optimise query generation dynamically based on continuous user feedback, enabling adaptive and user-centric QA systems. Such advancements would help address current limitations and propel the development of more intelligent and context-aware question generation frameworks. Overall, this research highlights the critical importance of effective context modelling within the question generation process, establishing a foundational benchmark for subsequent innovations. By emphasising architectural sophistication and optimisation strategies, the study contributes to the broader goal of developing intelligent QA systems capable of understanding and generating meaning-sensitive queries, ultimately improving human-computer interactions and information retrieval effectiveness.

Acknowledgement

Thank you to DPPM, which provided funding assistance Number: 098/C3/DT.05.00/PL/2025 so that this research can run, also thank you to Universitas Pembangunan Nasional Veteran Jawa Timur, which supported this research Number: 10/UN63.8/LT-Kontrak/VI/2025.

And thanks to Universitas Negeri Malang for providing direction in carrying out this research. The authors declare no conflict of interest in this research.

References

[1] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," *Adv. Neural Inf. Process. Syst.*, vol. 2020-December, May 2020, Accessed: Dec. 19, 2024. [Online]. Available: https://arxiv.org/abs/2005.14165v4

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[3] M. Lewis *et al.*, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 7871–7880, Oct. 2019, doi: 10.18653/v1/2020.acl-main.703.

[4] 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.

[5] Y. Liu and M. Lapata, "Text Summarization with Pretrained Encoders," *EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, pp. 3730–3740, Aug. 2019, doi: 10.18653/v1/d19-1387.

[6] A. See, P. J. Liu, and C. D. Manning, "Get To The Point: Summarization with Pointer-Generator Networks," *ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.*, vol. 1, pp. 1073–1083, Apr. 2017, doi: 10.18653/v1/P17-1099.

[7] M. M. Henry, G. N. Elwirehardja, and B. Pardamean, "Automatic question generation for bahasa indonesia examination using copynet," *Procedia Comput. Sci.*, vol. 245, no. C, pp. 953–962, 2024, doi: 10.1016/j.procs.2024.10.323.

[8] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, "A systematic review of automatic question generation for educational purposes," *Int. J. Artif. Intell. Educ.*, vol. 30, no. 1, pp. 121–204, Mar. 2020, doi: 10.1007/s40593-019-00186-y.

[9] N. Mulla and P. Gharpure, "Genetic Algorithm Optimized Topic-aware Transformer-Based Framework for Conversational Question Generation," *Procedia Comput. Sci.*, vol. 230, no. 2023, pp. 914–922, 2023, doi: 10.1016/j.procs.2023.12.041.

[10] M. Zhang and X. Shang, "Chinese Short Text Classification by ERNIE Based on LTC_Block," *Wirel. Commun. Mob. Comput.*, vol. 2022, 2022, doi: 10.1155/2022/1411744.

[11] V. Kumar, G. Ramakrishnan, and Y. F. Li, "Putting the horse before the cart: A generator-evaluator framework for question generation from text," *CoNLL 2019 - 23rd Conf. Comput. Nat. Lang. Learn. Proc. Conf.*, pp. 812–821, 2019, doi: 10.18653/v1/k19-1076.

[12] J. Ling and M. Afzaal, "Automatic question-answer pairs generation using pre-trained large language models in higher education," *Comput. Educ. Artif. Intell.*, vol. 6, no. April, p. 100252, 2024, doi: 10.1016/j.caeai.2024.100252.

[13] L. Murakhovs'ka, C. S. Wu, P. Laban, T. Niu, W. Liu, and C. Xiong, "MixQG: Neural Question Generation with Mixed Answer Types," *Find. Assoc. Comput. Linguist. NAACL 2022 - Find.*, pp. 1486–1497, 2022, doi: 10.18653/V1/2022.FINDINGS-NAACL.111.

[14] C. Patil and M. Patwardhan, "Visual question generation: the state of the art," *ACM Comput. Surv.*, vol. 53, no. 3, May 2020, doi: 10.1145/3383465.

[15] S. Shen *et al.*, "On the generation of medical question-answer pairs," *AAAI 2020 - 34th AAAI Conf. Artif. Intell.*, pp. 8822–8829, 2020, doi: 10.1609/AAAI.V34I05.6410.

[16] H. C. Wang, M. Maslim, and C. H. Kan, "A question–answer generation system for an asynchronous distance learning platform," *Educ. Inf. Technol.*, vol. 28, no. 9, pp. 12059–12088, Sep. 2023, doi: 10.1007/S10639-023-11675-Y.

[17] Q. Zaman, S. Safwandi, and F. Fajriana, "Supporting Application Fast Learning of Kitab Kuning for Santri' Ula Using Natural Language Processing Methods," *Int. J. Eng. Sci. Inf. Technol.*, vol. 5, no. 1, pp. 278–289, Jan. 2025, doi: 10.52088/ijesty.v5i1.713.

[18] M. N. Dorabati, R. Ramezani, and M. A. Nematbakhsh, "Research of LSTM Additions on Top of SQuAD BERT Hidden Transform Layers," *2022 12th Int. Conf. Comput. Knowl. Eng. ICCKE 2022*, pp. 415–422, 2022, doi: 10.1109/ICCKE57176.2022.9960031.

[19] W. Wang *et al.*, "Understanding and Improving Sequence-to-Sequence Pretraining for Neural Machine Translation," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, vol. 1, pp. 2591–2600, Mar. 2022, doi: 10.18653/v1/2022.acl-long.185.

[20] L. S. Hartono, E. I. Setiawan, and V. Singh, "Retrieval Augmented Generation-Based Chatbot for Prospective and Current University Students," *Int. J. Eng. Sci. Inf. Technol.*, vol. 5, no. 3, pp. 268–277, Jun. 2025, doi: 10.52088/ijesty.v5i3.951.

[21] A. Ushio, F. Alva-Manchego, and J. Camacho-Collados, "An Empirical Comparison of LM-based Question and Answer Generation Methods," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 14262–14272, May 2023, doi: 10.18653/v1/2023.findings-acl.899.

[22] P. M. Patil, R. P. Bhavsar, and B. V. Pawar, "A Review on Natural Language Processing based Automatic Question Generation," *2022 Int. Conf. Augment. Intell. Sustain. Syst.*, 2022, doi: 10.1109/ICAISS55157.2022.10010799.

[23] S. Indurthi, D. Raghu, M. M. Khapra, and S. Joshi, "Generating natural language question-answer pairs from a knowledge graph using a RNN based question generation model," *15th Conf. Eur. Chapter Assoc. Comput. Linguist. EACL 2017 - Proc. Conf.*, vol. 1, pp. 376–385, 2017, doi: 10.18653/V1/E17-1036.

[24] S. Rao and H. Daumé, "Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information," *ACL 2018 - 56th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.*, vol. 1, pp. 2737–2746, 2018, doi: 10.18653/v1/p18-1255.

[25] B. Das, M. Majumder, S. Phadikar, and A. A. Sekh, "Automatic question generation and answer assessment: a survey," *Res. Pr. Technol. Enhanc. Learn.*, vol. 16, no. 1, Dec. 2021, doi: 10.1186/s41039-021-00151-1.

[26] M. Bahani, A. El Ouaazizi, and K. Maalmi, "The effectiveness of T5, GPT-2, and BERT on text-to-image generation task," *Pattern Recognit. Lett.*, vol. 173, pp. 57–63, Sep. 2023, doi: 10.1016/j.patrec.2023.08.001.

[27] M. Fuadi, A. D. Wibawa, and S. Sumpeno, "idT5 : Indonesian Version of Multilingual T5 Transformer," 2023, [Online]. Available: https://doi.org/10.48550/arXiv.2302.00856

[28] 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.
[29] T. Ji, C. Lyu, G. Jones, L. Zhou, and Y. Graham, "QAScore—an unsupervised unreferenced metric for the question generation evaluation," *Entropy*, vol. 24, no. 11, p. 1514, Nov. 2022, doi: 10.3390/e24111514.
[30] S. Snekha and N. Ayyanathan, "An Educational CRM Chatbot for Learning Management System," *Shanlax Int. J. Educ.*, vol. 11, no. 4, pp. 58–62, Sep. 2023, doi: 10.34293/education.v11i4.6360.
[31] J. C. Sánchez-Prieto, V. Izquierdo-álvarez, M. T. Del Moral-Marcos, and F. Martínez-Abad, "Generative artificial intelligence for self-learning in higher education: Design and validation of an example machine," *RIED-Revista Iberoam. Educ. a Distancia*, vol. 28, no. 1, pp. 59–81, Jan. 2025, doi: 10.5944/RIED.28.1.41548.
[32] W. Zhong *et al.*, "ProQA: Structural Prompt-based Pre-training for Unified Question Answering," *NAACL 2022 - 2022 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Conf.*, pp. 4230–4243, 2022, doi: 10.18653/v1/2022.naacl-main.313.
[33] R. Rodriguez-Torrealba, E. Garcia-Lopez, and A. Garcia-Cabot, "End-to-End generation of Multiple-Choice questions using Text-to-Text transfer Transformer models," *Expert Syst. Appl.*, vol. 208, p. 118258, Dec. 2022, doi: 10.1016/J.ESWA.2022.118258.
[34] H. Yen, T. Gao, J. Lee, and D. Chen, "MoQA: Benchmarking Multi-Type Open-Domain Question Answering," pp. 8–29, 2023, Accessed: Mar. 27, 2025. [Online]. Available: https://github.com/princeton-nlp/MoQA
[35] B. Weng, "Navigating the Landscape of Large Language Models: A Comprehensive Review and Analysis of Paradigms and Fine-Tuning Strategies," Apr. 2024, Accessed: Dec. 19, 2024. [Online]. Available: https://arxiv.org/abs/2404.09022v1
[36] 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
[37] 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
[38] L. Xue *et al.*, "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 483–498. doi: 10.18653/v1/2021.naacl-main.41.
[39] A. Mohammadshahi *et al.*, "RQUGE: Reference-Free Metric for Evaluating Question Generation by Answering the Question," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 6845–6867, 2023, doi: 10.18653/v1/2023.findings-acl.428.
[40] R. Rodriguez-Torrealba, E. Garcia-Lopez, and A. Garcia-Cabot, "End-to-End generation of Multiple-Choice questions using Text-to-Text transfer Transformer models," *Expert Syst. Appl.*, vol. 208, Dec. 2022, doi: 10.1016/j.eswa.2022.118258.
[41] M. Barbella and G. Tortora, "Rouge Metric Evaluation for Text Summarization Techniques," *SSRN Electron. J.*, May 2022, doi: 10.2139/SSRN.4120317.
[42] M. Mieskes and U. Padó, "Summarization Evaluation meets Short-Answer Grading," *Proc. 8th Work. NLP Comput. Assist. Lang. Learn.*, no. Nlp4call, pp. 79–85, 2019, [Online]. Available: https://www.aclweb.org/anthology/W19-6308
[43] 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.
[44] 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.
[45] 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.v7i1.1211.