IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 2, April 2026, pp. 1804~1813
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i2.pp1804-1813

IndoBERT for educational assessment: comparative analysis of transformer models in Indonesian question generation

Handaru Jati[1], Yuniar Indrihapsari[1], Pradana Setialana[1], Danang Wijaya[2], Satya Adhiyaksa Ardy[3], Dhista Dwi Nur Ardiansyah[1]
[1]Department of Electronics and Informatics Engineering Education, Universitas Negeri Yogyakarta, Sleman, Indonesia
[2]Department of Computer Science and Information Engineering, National Central University, Taoyuan, Taiwan
[3]Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan

Article Info

Article history:

Received Sep 26, 2025
Revised Jan 5, 2026
Accepted Jan 22, 2026

Keywords:

Automatic question generation
BERT-large
Educational assessment
IndoBERT
Indonesian NLP
Monolingual pretraining
Multilingual BERT

ABSTRACT

This study asks whether a monolingual encoder can realistically outperform multilingual and larger transformer models for Indonesian automatic question generation (AQG) when all models share the same training budget. We compare Indonesian bidirectional encoder representations from transformers (IndoBERT), multilingual BERT (mBERT), and BERT-large using a single fine-tuning pipeline with answer highlighting, applied to an Indonesian version of TyDiQA-GoldP and a 20,000 translated subset of SQuAD 2.0. We evaluate model quality using bilingual evaluation understudy score n-gram 4 (BLEU-4), metric for evaluation of translation with explicit ordering (METEOR), and ROUGE-Lincoln (ROUGE-L). IndoBERT consistently achieves the best scores on both datasets (e.g., BLEU-4 of 19.69 on TyDiQA-GoldP and 3.79 on the SQuAD 2.0 subset) while requiring less computation than mBERT and BERT-large. Our results show that language-specific pretraining gives clear advantages for Indonesian AQG, yielding higher accuracy at lower computational cost than multilingual or larger encoders. The work closes a gap in Indonesian AQG benchmarking by providing the first head-to-head comparison of IndoBERT, mBERT, and BERT-large under a shared fine-tuning and evaluation protocol. For educational assessment, the findings offer a practical recipe for building deployable AQG systems on mid-range GPUs that generate higher-quality questions without prohibitive training or inference budgets.

Corresponding Author:

Handaru Jati
Department of Electronics and Informatics Engineering Education, Universitas Negeri Yogyakarta
Depok, Sleman, Yogyakarta 55281, Indonesia
Email: handaru@uny.ac.id

1. INTRODUCTION

A standard method of assessing student understanding through targeted questioning has remained a core concept of educational assessment and a valuable tool for understanding, provided the questions are aligned with appropriate learning objectives [1]. In this context, creating high-quality questions for assessing student understanding is a demanding task, even for experienced teachers. The items must be valid, coherent, and varied enough to elicit more than simple recall [2]–[6].

Automatic question generation (AQG) closes this gap by automatically generating items from text and/or databases, thereby improving coverage and variety with less authoring time [7], [8]. Carefully designed AQG has the potential to improve assessment efficiency and refocus attention from remembering answers to reasoning and understanding. The evolution of AQG can be traced back to its origins in 1976 [9].

From a methodological perspective, the evolution of AQG methods began with rule-based pipelines that included semantic/syntactic matching and handwritten templates. AQG has expanded beyond the classroom and found applications in commercial sectors such as chatbot development [7] and healthcare services [10]. It progressed to neural models that learn question-answering directly from data [11]–[15]. The process requires systems to accurately interpret text and generate well-structured, meaningful questions [16]–[18]. Encoder-decoder models and cross-lingual transfer have made them applicable in low-resource environments as well [19], [20], and pre-training goals such as masked language modeling (MLM) have enhanced linguistic capabilities [21], [22]. The recurrent neural networks (RNNs) enable natural language processing (NLP) tasks by considering the sequential context. However, RNNs face challenges like vanishing gradients and long dependencies. The long short-term memory (LSTM) networks solve the above challenges by incorporating memory cells to store data for long sequences [23]–[26].

The use of transformer models has further advanced the field of AQG. They efficiently exploring long-range dependencies. Currently, NLP is widely applying models like bidirectional encoder representations from transformers (BERT), generative pre-trained transformer (GPT), and text-to-text transfer transformer (T5) [27]–[40].

For Indonesian, Indonesian BERT (IndoBERT) provides a monolingual encoder pre-trained on Indonesian text [41]. Previous studies have investigated monolingual encoders for Indonesian AQGs using sequence-to-sequence architectures (open neural machine translation (OpenNMT) using bidirectional gated recurrent unit (BiGRU), bidirectional long short-term memory (Bi-LSTM), or transformer) that achieve competitive performance on SQuAD 2.0 translation and TyDiQA-style datasets [41]–[44]. However, there has been no comprehensive evaluation that simultaneously considers the effectiveness and cost efficiency of encoders for general AQGs. In this paper, IndoBERT is compared for the first time to mBERT and BERT-large in terms of their performance on identical downstream datasets, the Indonesian SQuAD 2.0 (20,000 subset) and TyDiQA-GoldP, using bilingual evaluation understudy score n-gram 4 (BLEU-4), metric for evaluation of translation with explicit ordering (METEOR), and ROUGE-Lincoln (ROUGE-L) scores, respectively, as evaluation criteria for AQG models.

There are three main contributions from this study. First, to the best of our knowledge, this study offers the first comprehensive comparison of IndoBERT, multilingual BERT (mBERT), and BERT-large for the Indonesian AQG task under equal training and testing conditions. The second contribution is a comprehensive fine-tuning pipeline that leverages the answer-highlighting mechanism via the tags `[HL]…[/HL]` to ensure the model maintains attention on the relevant answer portion and enables the generation of genuine questions. The third contribution is an analysis of efficiency aspects, including the number of parameters, the time per epoch, GPU memory consumption, and processing rate, in a practical setting on mid-range graphics processing unit (GPU) models. Together, these contributions show that language-specific pretraining can offer a practical advantage over multilingual and larger encoders for Indonesian AQG. Beyond this task, the same insight is relevant for other Indonesian NLP applications, such as summarization, translation, and adaptive learning systems, where resource-efficient yet accurate models are essential [41]–[44].

2. METHOD

We adopt a research and development methodology to build and evaluate Indonesian AQG models using transformer-based techniques. Each of the three pre-trained BERT variants (IndoBERT, mBERT, and BERT-large) was fine-tuned on our Indonesian question answering (QA) datasets under a consistent training regimen [45], [46]. Fine-tuning allows the models to adapt their pre-trained language understanding to the specific task of question generation. To ensure a fair comparison, we maintained the same data preprocessing, training schedule, and optimization parameters for all models, adjusting only the model-specific components such as the tokenizer and pre-trained weights. All experiments were implemented in Python using the PyTorch deep learning framework, leveraging its facilities for transformer models and sequence generation. This uniform methodology enables a direct performance comparison to determine the optimal model for Indonesian AQG.

2.1. Procedure

The overall procedure for developing the Indonesian question generation models was identical for IndoBERT, mBERT, and BERT-large. We followed three main stages for each model: dataset preparation, model fine-tuning (training), and evaluation. This supervised learning pipeline adheres to standard practices in machine learning model development, ensuring that each model undergoes the same sequence of steps. By keeping the procedure consistent, we can attribute performance differences to the models themselves rather than to any variation in processing. The orchestration of these processes is facilitated through a Python environment and using PyTorch.

2.2. Data preparation

We use SQuAD 2.0 [47] and TyDiQA-GoldP [48] parallel sources of context-answer-question triples. We first translate both corpora into Indonesian and then split them into training, validation, and test sets. Because SQuAD 2.0 is substantially larger than TyDiQA-GoldP, we sample 20,000 QA pairs from the SQuAD 2.0 training split. This choice keeps training time manageable and yields a training set that is comparable in size to TyDiQA-GoldP. All three models are trained on the same set of instances to ensure a fair comparison. The resulting Indonesian paragraphs, answers, and reference questions provide a shared supervision signal across encoders. To balance training time and maintain a comparable computational budget on a single 16 GB GPU, we cap the SQuAD 2.0 portion at 20,000 instances while retaining the full TyDiQA-GoldP split. As a robustness check, we retrain IndoBERT on disjoint 20,000 SQuAD samples and observe no qualitative change in the model ranking or conclusions.

2.3. Model fine-tuning

We frame the Indonesian answer-aware AQG as a conditional sequence generation with transformer encoders. During training, each model receives a passage in which the target answer span is marked with [HL]…[/HL] and learns to predict the next question token at [MASK] given the highlighted context and the previously generated tokens (cf. BERT-based question generation (QG) [49]–[51] and the IndoBERT setting [41]). The [HL] tags act as a soft pointer to the answer span, helping the model focus on the intended content and reducing off-target questions. We apply the same fine-tuning recipe to IndoBERT, mBERT, and BERT-large and keep the core training hyperparameters fixed to preserve protocol parity across models. Table 1 summarizes the learning rate, batch configuration, sequence length, optimizer, precision, and early-stopping settings for each encoder.

Table 1. Hyperparameters used across models

Model	Learning rate	Batch size	Gradient accumulation	Effective batch	Max seq length	Optimizer	Precision	Early stopping	Epoch
IndoBERT	5×10^{-5}	8	4	32	128	AdamW	16	Y	3
mBERT	5×10^{-5}	8	4	32	128	AdamW	16	Y	3
BERT-large	5×10^{-5}	8	8	64	96	AdamW	16	Y	3

We train all three encoders with AdamW at a fixed learning rate of 5×10^{-5} and a per-device batch size of 8, using FP16 mixed precision. To achieve larger effective batch sizes without exceeding 16 GB of GPU memory, we use gradient accumulation: IndoBERT and mBERT accumulate 4 steps (effective batch size 32), whereas BERT-large accumulates 8 steps (effective batch size 64) to stabilize training for the deeper 24-layer architecture. We set the maximum sequence length to 128 tokens for IndoBERT and mBERT, and 96 tokens for BERT-large, to balance contextual coverage with memory usage. Each model trains for up to 3 epochs with early stopping based on validation loss, providing a fair, computationally comparable setup across encoders.

2.4. Evaluation

We evaluate the generated questions with BLEU-4 [52], METEOR [53], and ROUGE-L [54], with additional background in text summarization [55]. BLEU-4 measures n-gram precision with a brevity penalty, METEOR emphasizes recall using synonym and stem matching, and ROUGE-L computes overlap via the longest common subsequence. We use reference implementations with default settings. We acknowledge that these n-gram overlap metrics do not fully reflect pedagogical quality or semantic adequacy; we revisit their limitations in the discussion and provide complete formulas and notation.

2.5. Model architecture

Figure 1 illustrates the IndoBERT-AQG pipeline. We first mark the answer span in the passage with [HL]…[/HL], to obtain the highlighted context C'. The highlighted context and the partial question tokens are then tokenized, and encoded with IndoBERT. At each decoding step i, the model predicts the next question token at [MASK] given C' and the previously generated tokens $\hat{q}_{1:i-1}$; the new token is appended until the full question (q). For motivation and prior work on [HL] tagging in answer-aware question generation, we refer the reader to subsection 2.5 and in studies [41], [49], [50].

Formally, we follow the BER-based hierarchical label-aware sentence question generation (BERT-HLSQG) formulation [44]. IndoBERT has the BERT-base architecture but is pre-trained on Indonesian. At decoding step *i*, the model receives the input sequence in (1).

$$X_i = ([\text{CLS}], C, [\text{SEP}], \hat{q}_1, \ldots, \hat{q}_i, [\text{MASK}]) \quad (1)$$

Here, $\hat{q}_{1=i-1}$ are the previously generated question tokens, and [MASK] marks the position to predict next. The next-token distribution is produced from the final hidden state at [MASK] via an affine layer and softmax in (2), and the token choice is made by argmax in (3).

$$\Pr(w|X_i) = softmax\left(h_{[\text{mask}]}.W_{HLSQG} + b_{HLSQG}\right) \quad (2)$$

$$\hat{q}_i = argmax_w \Pr(w|X_i) \quad (3)$$

Together, (1) to (3) formalize the loop depicted in Figure 1: highlight the answer span, encode the context with IndoBERT, predict the next token at [MASK], and append it to the partial question until [SEP].

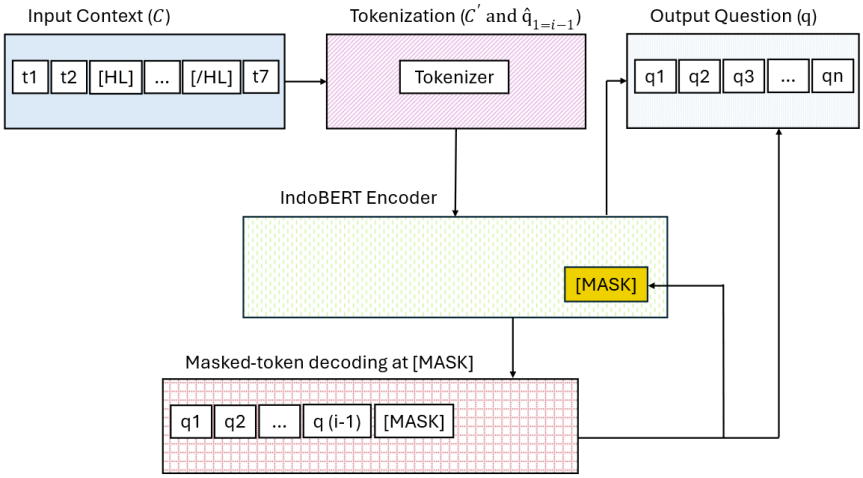

Figure 1. IndoBERT-AQG pipeline

2.6. Data analysis techniques

Data analysis encompassed the evaluation of each model configuration, which underwent the training stage utilizing standard automatic evaluation metrics to compare performance results. These metrics assessed and determined the model configuration with the most optimal performance during the training process. Graph and plotting tools were adeptly utilized to visually interpret and comprehend the performance trends in analyzing the newly developed QG model.

3. RESULTS AND DISCUSSION

3.1. Data preparation

We selected the translated SQuAD 2.0 and TyDiQA-GoldP as our datasets. We follow the original splits from both SQuAD 2.0 and TyDiQA-GoldP. For training, we use only 20,000 instances from the SQuAD 2.0 training set, while we use the entire TyDiQA-GoldP dataset.

3.2. Experimental setting

We apply the same fine-tuning setup to IndoBERT, mBERT, and BERT-large. IndoBERT and mBERT follow the 12-layer BERT-base configuration, whereas BERT-large uses the 24-layer variant. All experiments run on a single NVIDIA RTX 4060 Ti GPU (16 GB VRAM). We use AdamW with a learning rate of 5×10^{-5}, the maximum sequence length specified in Table 1, FP16 mixed precision, and early stopping. For IndoBERT and mBERT, we use an adequate batch size of 32, and for BERT-large, we rely on a smaller per-step batch with gradient accumulation to fit within 16 GB of VRAM; the smaller batch also adds regularization under limited data. We match the total number of optimization steps and the learning rate schedule across models so that observed differences reflect architectural and pre-training effects rather than the training protocol. To maintain computational-budget parity in this single-GPU setting, we cap the

SQuAD 2.0 training portion at 20,000 instances (with a robustness re-sampling check yielding the same qualitative ranking) while retaining the full TyDiQA-GoldP split.

3.3. Evaluation

The performance of IndoBERT, mBERT, and BERT-large on the two Indonesian QA datasets is summarized in Tables 2 and 3. IndoBERT achieved the highest scores across all evaluation metrics on both datasets, while mBERT showed the second-best performance and BERT-large the lowest. This ranking is consistent for each metric (BLEU-1 through BLEU-4, METEOR, and ROUGE-L), reflecting the advantages of a language-specific model over a multilingual model and the drawbacks of using a large model not pre-trained in the target language.

Table 2. Performance of the three models on the TyDiQA-GoldP dataset

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
IndoBERT	54.81	38.68	26.86	19.69	31.62	58.83
mBERT	47.67	30.83	19.82	13.69	25.64	52.15
BERT-large	46.15	27.78	17.25	11.12	24.34	50.64

On TyDiQA-GoldP, IndoBERT's BLEU-4 score of 19.69 is about 6 points higher than mBERT's 13.69 and 8.5 points higher than BERT-large's 11.12. Likewise, IndoBERT leads substantially in METEOR and ROUGE-L, indicating it generates questions that not only match the reference wording more closely but also capture more of the reference content. mBERT's scores, while lower than IndoBERT's, are clearly above those of BERT-large. Notably, mBERT outperforms BERT-large by around 2-3 points on most metrics, demonstrating the benefit of multilingual pre-training that includes Indonesian: even though mBERT is a smaller model than BERT-large, its familiarity with Indonesian gives it an edge. BERT-large's underperformance on TyDiQA-GoldP suggests that its large capacity remains underutilized due to the model's lack of prior Indonesian knowledge and the limited fine-tuning data available. In practical terms, the IndoBERT model shows strength in handling the TyDiQA-GoldP material, likely leveraging its pre-trained understanding of Indonesian nuances to produce more accurate and fluent questions.

Table 3. Performance of the three models on the SQuAD 2.0 (20,000 subset) dataset

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
IndoBERT	33.75	16.24	7.32	3.79	16.25	37.45
mBERT	28.23	10.31	3.64	1.51	12.63	32.77
BERT-large	25.30	8.72	3.19	1.46	11.22	30.32

On the SQuAD 2.0 subset, the overall scores are lower for all models, reflecting the increased difficulty of this dataset and the smaller training sample. IndoBERT still holds the highest scores by a notable margin. For instance, IndoBERT's BLEU-4 (3.79) is more than double that of mBERT (1.51) or BERT-large (1.46). This dramatic gap underscores IndoBERT's efficiency in learning from a limited dataset. mBERT and BERT-large both struggle on this dataset, but mBERT maintains a slight advantage over BERT-large on every metric. The differences between mBERT and BERT-large, though small in absolute terms here, consistently favor mBERT (ROUGE-L of 32.77 vs. 30.32), reinforcing that knowing the language is crucial for performance. The BERT-large model, despite having over three times the parameters of the others, fails to outperform the base models in this low-resource scenario. We attribute this to its inability to generalize from such a limited fine-tuning set, its large capacity cannot be effectively used without far more data. IndoBERT, by contrast, avoids this pitfall thanks to its prior Indonesian pre-training, which allows it to generalize better from the same small sample.

The results show that IndoBERT is the best-performing model for Indonesian question generation in our experiments, excelling especially in scenarios with limited training data. The multilingual mBERT provides decent performance and can be considered a strong baseline for Indonesian AQG, but it consistently lags behind IndoBERT. The large-capacity BERT-large model did not yield any performance benefit in this context; on the contrary, it underperformed even the smaller mBERT. From a practical perspective, these findings suggest that for Indonesian-language AQG tasks, one should favor a model that has been pre-trained on Indonesian rather than simply opting for a model with greater size or general multilingual training. Additionally, IndoBERT's superior performance, coupled with its relatively smaller size and thus faster inference and training, makes it an attractive choice for real-world applications where computational

resources and training data may be limited. Meanwhile, mBERT's results indicate that if a multi-language solution is required, it can handle Indonesian QG reasonably well, though with some loss in question quality. BERT-large, given its resource demands and low payoff here, would likely only be justified if significantly more Indonesian training data were available or if an Indonesian-specific large model were pre-trained.

3.4. Efficiency and resource usage

We compare model size, time per epoch, peak VRAM usage, and inference throughput, and summarize the results in Table 4. As shown in Table 4, IndoBERT is the most efficient model: it uses fewer parameters (124 M), trains faster (1 h/epoch), requires less VRAM (4.8 GB), and reaches the highest throughput (5,200 tokens/s). mBERT sits in the middle (179 M; 3 h/epoch; 11 GB; 3,200 tokens/s), whereas BERT-large is the most resource-demanding with the lowest throughput (340 M; 4 h/epoch; 11 GB; 1,600 tokens/s). Taken together, Table 4 shows that IndoBERT offers the best trade-off between accuracy and efficiency for Indonesian AQG under realistic resource budgets.

Table 4. Efficiency summary of IndoBERT, mBERT, and BERT-large (parameters (M), time per epoch (h), peak GPU memory (GB), throughput (tokens/s))

Model	Parameter (M)	Time per epoch (h)	Peak GPU memory (GB)	Throughput (token/s)
IndoBERT	124	1	4.8	5,200
mBERT	179	3	11	3,200
BERT-large	340	4	11	1,600

These efficiency results carry direct system implications. IndoBERT is a practical backbone for Indonesian AQG on mid-range GPUs (≈5 GB VRAM), enabling fast retraining (≈1 h/epoch) and high throughput (5,200 tokens/s). If multilingual support is required, mBERT is a reasonable fallback at a higher cost, whereas BERT-large is unlikely to be justified without substantially more Indonesian data and compute.

3.5. Limitations and future works

Our evaluation relies on automatic overlap metrics (BLEU-4, METEOR, and ROUGE-L). However, these scores correlate only imperfectly with pedagogical usefulness and perceived question quality, and we did not include human-rater studies with teachers or subject-matter experts. The Indonesian SQuAD 2.0 portion uses a 20,000 translated subset, which may contain translation artifacts and therefore does not fully mirror classroom discourse. At the same time, TyDiQA-GoldP focuses on information-seeking questions rather than curriculum-aligned materials. Working on a single 16 GB GPU, we used smaller effective batches for BERT-large and could not explore a vast hyperparameter space; even though we matched the number of optimization steps across models and ran a re-sampling robustness check on SQuAD 2.0, some confounding factors may persist. We also limit our comparison to encoder-only transformers and do not benchmark decoder or instruction-tuned large language models. In future work, we plan to add expert and teacher ratings, curriculum-specific Indonesian datasets, and a broader set of model families.

4. CONCLUSION

This study evaluated IndoBERT, mBERT, and BERT-large for Indonesian AQG and found that IndoBERT consistently produced the highest-quality questions while using resources most efficiently. mBERT remains a viable multilingual alternative, albeit with a modest accuracy trade-off, whereas BERT-large offers no clear advantage in this setting and requires a greater computational budget. We therefore recommend IndoBERT as the default backbone for Indonesian AQG in educational applications and deployments on mid-range GPUs. For developers and educators, a practical approach is to fine-tune IndoBERT with [HL] tagging at a learning rate of 5×10^{-5} (batch size 16, or 8 for larger models) and utilize FP16/mixed precision with early stopping. We recommend evaluating systems using BLEU, METEOR, and ROUGE-L. For deployment, enable batching/caching and consider FP16 or 8-bit inference to meet typical latency and memory budgets on mid-range (~5 GB VRAM) GPUs.

ACKNOWLEDGMENTS

We used AI tools solely to improve the clarity, grammar, and consistency of the manuscript text. All technical content, analyses, figures, and conclusions were produced and verified by the authors. No data were generated, modified, or synthesized using AI tools.

FUNDING INFORMATION

This research was supported by Universitas Negeri Yogyakarta (UNY) under Contract No. T/10.12/UN34.15/PT.01.02/2023. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Handaru Jati	✓	✓		✓	✓	✓		✓	✓	✓		✓	✓	✓
Yuniar Indrihapsari	✓	✓			✓	✓	✓		✓	✓			✓	
Pradana Setialana	✓		✓	✓		✓		✓	✓		✓			
Danang Wijaya		✓	✓	✓	✓			✓	✓	✓				
Satya Adhiyaksa Ardy		✓		✓		✓	✓		✓	✓	✓			
Dhista Dwi Nur Ardiansyah		✓		✓		✓	✓		✓	✓	✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The author declares that there are no known conflicts of interest associated with this publication. There are no financial or personal relationships that could inappropriately influence or bias the content of this work.

INFORMED CONSENT

Not applicable. This study did not involve human participants, human data, or any personally identifiable information. All data used were either publicly available, fully anonymized, or derived from non-human sources, and therefore no informed consent was required from individuals.

ETHICAL APPROVAL

Not applicable. This research did not involve human subjects, human biological materials, or experimental procedures on animals.

DATA AVAILABILITY

The data that support the findings of this study are openly available in the Github.com at https://github.com/google-research-datasets/tydiqa.

REFERENCES

[1] A. K. Salmon and M. X. Barrera, "Intentional questioning to promote thinking and learning," *Thinking Skills and Creativity*, vol. 40, Jun. 2021, doi: 10.1016/j.tsc.2021.100822.
[2] R. G. W. Mueller, "Examining teachers' development and implementation of compelling questions," *Social Studies Research and Practice*, vol. 13, no. 1, pp. 1–15, May 2018, doi: 10.1108/SSRP-08-2017-0042.
[3] S. Saarinen, S. Krishnamurthi, K. Fisler, and P. T. Wilson, "Harnessing the wisdom of the classes," in *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, Feb. 2019, pp. 606–612, doi: 10.1145/3287324.3287504.
[4] T. Tofade, J. Elsner, and S. T. Haines, "Best practice strategies for effective use of questions as a teaching tool," *American Journal of Pharmaceutical Education*, vol. 77, no. 7, Sep. 2013, doi: 10.5688/ajpe777155.
[5] C. Chin and D. E. Brown, "Student-generated questions: a meaningful aspect of learning in science," *International Journal of Science Education*, vol. 24, no. 5, pp. 521–549, May 2002, doi: 10.1080/09500690110095249.
[6] M. Gottlieb, J. Bailitz, M. Fix, E. Shappell, and M. J. Wagner, "Educator's blueprint: a how-to guide for developing high-quality multiple-choice questions," *AEM Education and Training*, vol. 7, no. 1, Feb. 2023, doi: 10.1002/aet2.10836.

[7] S. Soni, P. Kumar, and A. Saha, "Automatic question generation: a systematic review," *SSRN Electronic Journal*, 2019, doi: 10.2139/ssrn.3403926.

[8] K. P. Kumar *et al.*, "Automated question paper generator using LLM," *International Journal of Research and Innovation in Applied Science*, vol. X, no. IV, pp. 266–275, 2025, doi: 10.51584/IJRIAS.2025.10040020.

[9] J. H. Wolfe, "Automatic question generation from text - an aid to independent study," in *Proceedings of the ACM SIGCSE-SIGCUE technical symposium on Computer science and education*, 1976, pp. 104–112, doi: 10.1145/800107.803459.

[10] P. Tanwar, K. Bansal, A. Sharma, and N. Dagar, "AI based chatbot for healthcare using machine learning," in *2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)*, Nov. 2023, pp. 751–755, doi: 10.1109/ICAICCIT60255.2023.10465728.

[11] M. Flor and B. Riordan, "A semantic role-based approach to open-domain automatic question generation," in *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 2018, pp. 254–263, doi: 10.18653/v1/W18-0530.

[12] A. Gashkov and M. Eltsova, "Automatization of answers to questions by matching syntactic graphs," in *Proceedings of the X International Conference "Word, Utterance, Text: Cognitive, Pragmatic and Cultural Aspects" (WUT 2020)*, Aug. 2020, pp. 419–425, doi: 10.15405/epsbs.2020.08.49.

[13] E. Sneiders, "Automated question answering using question templates that cover the conceptual model of the database," in *Natural Language Processing and Information Systems (NLDB 2002)*, 2002, pp. 235–239, doi: 10.1007/3-540-36271-1_24.

[14] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, "A systematic review of automatic question generation for educational purposes," *International Journal of Artificial Intelligence in Education*, vol. 30, no. 1, pp. 121–204, Mar. 2020, doi: 10.1007/s40593-019-00186-y.

[15] M. Last and G. Danon, "Automatic question generation," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 6, Nov. 2020, doi: 10.1002/widm.1382.

[16] M. Essam, M. A. Deif, and R. Elgohary, "Deciphering Arabic question: a dedicated survey on Arabic question analysis methods, challenges, limitations and future pathways," *Artificial Intelligence Review*, vol. 57, no. 9, Aug. 2024, doi: 10.1007/s10462-024-10880-6.

[17] I. E. Fattoh, A. E. Aboutabl, and M. H. Haggag, "Tapping into the power of automatic question generation," *International Journal of Computer Applications*, vol. 103, no. 1, pp. 1–6, Oct. 2014, doi: 10.5120/18035-7636.

[18] A. Asadi and R. Safabakhsh, "The encoder-decoder framework and its applications," in *Deep Learning: Concepts and Architectures*, Springer, Cham, 2020, pp. 133–167. doi: 10.1007/978-3-030-31756-0_5.

[19] C. Escolano, M. R. C.-jussà, and J. A. R. Fonollosa, "Multilingual machine translation: deep analysis of language-specific encoder-decoders," *Journal of Artificial Intelligence Research*, vol. 73, pp. 1535–1552, Apr. 2022, doi: 10.1613/jair.1.12699.

[20] J. Yu, S. Wang, and J. Yin, "Adaptive cross-lingual question generation with minimal resources," *The Computer Journal*, vol. 64, no. 7, pp. 1056–1068, Aug. 2021, doi: 10.1093/comjnl/bxab106.

[21] D. Luitel, S. Hassani, and M. Sabetzadeh, "Improving requirements completeness: automated assistance through large language models," *Requirements Engineering*, vol. 29, no. 1, pp. 73–95, Mar. 2024, doi: 10.1007/s00766-024-00416-3.

[22] Y. Zhao, R. Cao, J. Bai, W. Ma, and H. Shinnou, "Determining the logical relation between two sentences by using the masked language model of BERT," in *2020 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, Dec. 2020, pp. 228–231, doi: 10.1109/TAAI51410.2020.00049.

[23] K. M. Tarwani and S. Edem, "Survey on recurrent neural network in natural language processing," *International Journal of Engineering Trends and Technology*, vol. 48, no. 6, pp. 301–304, Jun. 2017, doi: 10.14445/22315381/IJETT-V48P253.

[24] A. Graves, "Long short-term memory," in *Supervised Sequence Labelling with Recurrent Neural Networks*, Springer, Berlin, Heidelberg, 2012, pp. 37–45, doi: 10.1007/978-3-642-24797-2_4.

[25] N. Zucchet and A. Orvieto, "Recurrent neural networks: vanishing and exploding gradients are not the end of the story," in *Proceedings of the 38th International Conference on Neural Information Processing System*, 2024, pp. 139402–139443.

[26] P. Nerella, D. Pittu, S. Undrakonda, S. Chennamsetty, V. P. Kumar S, and V. K. Kishore K, "An efficient Seq2Seq model to predict question and answer response system," in *2024 Second International Conference on Advances in Information Technology (ICAIT)*, Jul. 2024, pp. 1–6, doi: 10.1109/ICAIT61638.2024.10690343.

[27] Y. Sun, "The evolution of transformer models from unidirectional to bidirectional in natural language processing," *Applied and Computational Engineering*, vol. 42, no. 1, pp. 281–289, Feb. 2024, doi: 10.54254/2755-2721/42/20230794.

[28] N. Patwardhan, S. Marrone, and C. Sansone, "Transformers in the real world: a survey on NLP applications," *Information*, vol. 14, no. 4, Apr. 2023, doi: 10.3390/info14040242.

[29] M. Arsalan, "Transformers in natural language processing: a comprehensive review," *International Journal for Research in Applied Science and Engineering Technology*, vol. 12, no. 5, pp. 5591–5597, May 2024, doi: 10.22214/ijraset.2024.62863.

[30] J. Park, M. R. Babaei, S. A. Munoz, A. N. Venkat, and J. D. Hedengren, "Simultaneous multistep transformer architecture for model predictive control," *Computers & Chemical Engineering*, vol. 178, Oct. 2023, doi: 10.1016/j.compchemeng.2023.108396.

[31] M. Zaheer *et al.*, "Big bird: transformers for longer sequences," in *Proceedings of the 34th International Conference on Neural Information Processing Syste*, 2020, pp. 17283–17297.

[32] B. Zhuang, J. Liu, Z. Pan, H. He, Y. Weng, and C. Shen, "A survey on efficient training of transformers," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, Aug. 2023, pp. 6823–6831, doi: 10.24963/ijcai.2023/764.

[33] A. Bhattacharyya, "Revolutionizing knowledge retrieval: a comprehensive study of question answering system," *SSRN Electronic Journal*, 2023, doi: 10.2139/ssrn.4649585.

[34] 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*, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[35] M. Suguna and K. S. S. Prabha, "Reciprocating encoder portrayal from reliable transformer dependent bidirectional long short-term memory for question and answering text classification," *IEEE Access*, vol. 12, pp. 117800–117811, 2024, doi: 10.1109/ACCESS.2024.3426604.

[36] N. Fatima, S. M. Daudpota, Z. Kastrati, A. S. Imran, S. Hassan, and N. S. Elmitwally, "Improving news headline text generation quality through frequent POS-Tag patterns analysis," *Engineering Applications of Artificial Intelligence*, vol. 125, Oct. 2023, doi: 10.1016/j.engappai.2023.106718.

[37] A. Karak, K. Kunal, N. Darapaneni, and A. R. Paduri, "Implementation of GPT models for text generation in healthcare domain," *EAI Endorsed Transactions on AI and Robotics*, vol. 3, Apr. 2024, doi: 10.4108/airo.4082.

[38] D. Jyoti, J. Srivastava, and D. P. Mahato, "Implementing T5 for text summarization: an algorithmic approach," in *2025 International Conference on Information Networking (ICOIN)*, Jan. 2025, pp. 648–652, doi: 10.1109/ICOIN63865.2025.10992766.

[39] J. Ranganathan and G. Abuka, "Text summarization using transformer model," in *2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, Nov. 2022, pp. 1–5, doi: 10.1109/SNAMS58071.2022.10062698.
[40] T. T, M. Borah, P. Dadure, and P. Pakray, "Comparative analysis of T5 model for abstractive text summarization on different datasets," *SSRN Electronic Journal*, 2022, doi: 10.2139/ssrn.4096413.
[41] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: a benchmark dataset and pre-trained language model for Indonesian NLP," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 757–770, doi: 10.18653/v1/2020.coling-main.66.
[42] K. Vincentio and D. Suhartono, "Automatic question generation monolingual multilingual pre-trained models using RNN and transformer in low resource Indonesian language," *Informatica*, vol. 46, no. 7, Nov. 2022, doi: 10.31449/inf.v46i7.4236.
[43] F. J. Muis and A. Purwarianti, "Sequence-to-sequence learning for Indonesian automatic question generator," in *2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA)*, Sep. 2020, pp. 1–6, doi: 10.1109/ICAICTA49861.2020.9429032.
[44] M. M. Henry, G. N. Elwirehardja, and B. Pardamean, "Automatic question generation for bahasa Indonesia examination using copynet," *Procedia Computer Science*, vol. 245, pp. 953–962, 2024, doi: 10.1016/j.procs.2024.10.323.
[45] Y.-H. Chan and Y.-C. Fan, "BERT for question generation," in *Proceedings of the 12th International Conference on Natural Language Generation*, 2019, pp. 173–177. doi: 10.18653/v1/W19-8624.
[46] Y.-H. Chan and Y.-C. Fan, "A recurrent BERT-based model for question generation," in *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, 2019, pp. 154–162. doi: 10.18653/v1/D19-5821.
[47] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: unanswerable questions for SQuAD," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 784–789, doi: 10.18653/v1/P18-2124.
[48] J. H. Clark *et al.*, "TyDi QA: a benchmark for information-seeking question answering in typologically diverse languages," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 454–470, Dec. 2020, doi: 10.1162/tacl_a_00317.
[49] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: pre-trained language models for Vietnamese," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 1037–1042, doi: 10.18653/v1/2020.findings-emnlp.92.
[50] H.-G. Zhao, X.-Z. Li, and X. Kang, "Development of an artificial intelligence curriculum design for children in Taiwan and its impact on learning outcomes," *Humanities and Social Sciences Communications*, vol. 11, no. 1, Oct. 2024, doi: 10.1057/s41599-024-03839-z.
[51] W. Zhang, X. Li, Y. Yang, and R. Dong, "Pre-training on mixed data for low-resource neural machine translation," *Information*, vol. 12, no. 3, Mar. 2021, doi: 10.3390/info12030133.
[52] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 2001, pp. 311–318, doi: 10.3115/1073083.1073135.
[53] S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, pp. 65–72.
[54] C.-Y. Lin, "ROUGE: a package for automatic evaluation of summaries," in *Text Summarization Branches Out*, Barcelona: Association for Computational Linguistics, 2004, pp. 74–81.
[55] D. Suleiman and A. Awajan, "Deep learning based abstractive text summarization: approaches, datasets, evaluation measures, and challenges," *Mathematical Problems in Engineering*, vol. 2020, pp. 1–29, Aug. 2020, doi: 10.1155/2020/9365340.

BIOGRAPHIES OF AUTHORS

Handaru Jati received his Ph.D. degree in Computer and Information Science from Universiti Teknologi Petronas, Malaysia. He is currently an associate professor at the Department of Electronics and Informatics Engineering Education, Universitas Negeri Yogyakarta, Indonesia. His research interests include machine learning, artificial intelligence, decision support systems, data mining, software development, and vocational education. He has published numerous papers in international journals and conferences, particularly in the areas of AI applications for education and intelligent systems. He has also been involved in several collaborative research projects and community service activities focusing on digital learning innovation. In this paper, he contributed to conceptualization, methodology design, and supervision. He can be contacted at email: handaru@uny.ac.id.

Yuniar Indrihapsari is an assistant professor at Universitas Negeri Yogyakarta (UNY), Indonesia, and a Ph.D. student at National Taiwan University of Science and Technology, Taiwan. She received her Bachelor's degree in Electrical Engineering and Master's degree in Information Technology from Universitas Gadjah Mada, Indonesia. Her research interests include social network analysis, e-learning, human-computer interaction, and educational technology. In this paper, she contributed to conceptualization, writing – original draft preparation, and formal analysis. She can be contacted at email: yuniar@uny.ac.id.

Pradana Setialana is an assistant professor in the Department of Electronics and Informatics Engineering Education at Universitas Negeri Yogyakarta, Indonesia. He received his Bachelor degree in Informatics Engineering Education from Universitas Negeri Yogyakarta and his Master degree in Information Technology from Universitas Gadjah Mada, Indonesia. His research interests include software engineering, natural language processing, mobile and cloud computing, and database systems. In this paper, he contributed to methodology, software development, and data curation. He can be contacted at email: pradana.setialana@uny.ac.id.

Danang Wijaya received his Bachelor degree in Informatics Engineering Education from Universitas Negeri Yogyakarta, Indonesia, and his Master degree from the International Master's program in Artificial Intelligence, National Central University, Taiwan. His research interests include artificial intelligence and natural language processing. In this paper, he contributed to investigation, validation, and visualization. He can be contacted at email: danangwijaya750@gmail.com.

Satya Adhiyaksa Ardy received his Bachelor degree in Information Technology from Universitas Negeri Yogyakarta, Indonesia. He is currently a Master's student in the Department of Electronic and Computer Engineering at National Taiwan University of Science and Technology, Taiwan. His research interests include artificial intelligence, natural language processing, and educational technology. In this paper, he contributed to writing – review and editing, resources, and project administration. He can be contacted at email: satyaadhiyaksa@gmail.com.

Dhista Dwi Nur Ardiansyah holds a Bachelor's degree in Informatics Engineering Education from Universitas Negeri Yogyakarta (UNY), Indonesia. He is part of the Digital Transformation Directorate at University Nahdatul Ulama Yogyakarta, where he supports institutional digitalization initiatives. His research interests focus on machine learning and its applications. In this paper, he contributed to investigation, data curation, software/processing scripts, visualization, and writing, review, and editing. He can be contacted at email: dhistadna@gmail.com.