International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 21s, 2025
https://www.theaspd.com/ijes.php

# Adaptive Question Generation and Answer Evaluation with Dynamic Text Complexity Adjustment

[1]Minidu Tissera, [2]Bhagya Peramuna, [3]Muditha Jayawickrema, [4]Pasindu Serasinghe, [5]Manori Gamage, [6]Karthiga Rajendran
[1,2,3,4,5,6]Sri Lanka Institute of Information Technology
[1]miniduspc@gmail.com, [2]bhagyadewmi17@gmail.com, [3]mudithapradeeptha@gmail.com, [4]pasinduserasinghe@gmail.com, [5]manori.g@sliit.lk, [6]karthiga.r@sliit.lk

***Abstract*—** *The ability to read and comprehend a non-native language should be developed from the early childhood. The education system in Sri Lanka requires students to learn English language which is a dominant universal language as the secondary language. Still, the individual attention provided in the classroom is not enough for the proper grasp of language skills by the students. And it is time consuming for the teachers to generate multiple questions and to evaluate the students individually (PROBLEM). The system proposed as solution facilitates enhancing language comprehension of students within the age group of 8-11 years, in a personalized manner (SOLUTION). The student is provided with an interactive and eye-catching web application interface built using React and Tailwind. System allows the teachers to add comprehension passages related to the student's age. The system can challenge the students with questions of different difficulty levels that are individually catered to the student's performance level, which is an exceptional feature of this system. To achieve this, a question-generation model is created and trained using a custom-made dataset created from domain-specific textbooks, customizing the T5 pre-trained model, incorporating style embeddings, and custom methods. In parallel answer evaluation is done using a pre-trained model deepset/roberta-base-squad2. The difficulty level of the provided text passage is calculated based on Flesch Reading Ease score. The pain points in learning and teaching comprehension skills are reduced through this solution due to the personalized and automatic question generation, and answer evaluation features. The educators are given the privilege to add relevant text passages and keep in track with the student's engagement and scoring with ease. The students are given personalized attention to their performance and relevant passages are provided.*

**Index Terms—** *difficulty level adjustment, domain-specific question generation, English reading and understanding skills evaluation, question and answer generator.*

## I. INTRODUCTION

The worldwide educational systems and educational approaches have developed over time. During the era of digital learning, the nature of adaptive learning systems has emerged drastically. The different solutions built for adaptive learning address the diverse needs of learners; it also helps in personalized learning experiences for students. In the last few years, with the advancement of natural language processing (NLP) and deep learning (DL), it is now possible to develop sophisticated question generation (QG) and answer evaluation frameworks. These systems fundamentally help the revolutionary learning in early language learning by dynamically tailoring learning materials for one's performance level. This proposed system focuses on creating an adaptive learning platform for Sri Lankan students under age eleven, where the system provides passages adjusted accordingly to the performance level of the student while generating questions and evaluating the answers provided.

Even though the students need constant guidance over a subject like English, the schools lack the capability to provide enough attention to all the students in Sri Lanka. So implementing a system that automates the teaching and learning process together with facilitating the teacher and providing constant overlook on the student holds a quite significance. Current research trends in educational technology underscore the importance of automated QG. In Particularly, QG models such as T5 have shown the world promising results in generating relevant and appropriate questions [1]. Also, models in transformer-based question answering, such as deepset/Roberta-base-squad2, have evolved greatly in the field of evaluating student responses [2]. Even though QG and answer evaluation areas have been developed over time, the attention given to text passage simplification and complexification relative to difficulty levels is a bit low [3].

The fundamental aim of this research is to develop an adaptive learning framework concentrating the English subject reading and understanding skill development. The system provided an automated QG and answer evaluation mechanism with the feature of simplifying and complexifying the text passages according to the scoring levels. For each functionality, some pre-trained models are used, together with freezing some layers and adding custom layers to enhance the functionality. Considering the QG functionality, two types of questions are generated: short answer questions (SAQ) and jumbled sentence questions (JSQ). Then, from the answer evaluating model, the answers are checked, and a score is produced. The next passage is provided, adjusting the difficulty level based on the student's scoring level. The adaptive approach aims to improve the reading and understanding of a non-native language, English, by providing the students with relevant subject matter in a more tailored way. Furthermore, the integration of automatic QG and answer evaluation models facilitates the tutors by reducing the manual effort that needs to be put into individual attention and personalization of learning subjects.

## II. RELATED WORK

QG and answer evaluation are areas that have been discussed and improved over time. When considering the area of QG studies, Gumaste et al. [4] have studied a question generation system that is developed and implemented using diverse natural language processing (NLP) libraries. This includes using Spacy and NLTK but is not limited to them. After generating the question from the input text, the text is divided into tokens, which are the basis for analyzing semantic and syntactic features. The main aim of this study was to generate questions that schools and colleges use in educational scenarios rather than only automatically generating question papers for examinations. There is also a part of speech tagging that would be used in order to assign meaning to words as to how they are located within the sentence structure of a normal sentence. Another method used in text processing is Named Entity Recognition (NER), which is used for classifying such as names, dates, and places. It employs part-of-speech (POS) tagging combined with chunking in order to group the sequences of words into coherent phrases like noun phrases and applies a multitude of regular expression patterns to create questions from the phrases. The most powerful aspect of this system is its ability to automatically generate “wh” questions necessary to assess reading comprehension level mastery. The authors also clearly define how such a system can aid teachers in improving question papers or aid students in self-assessment. This system is useful for simple comprehension question generation, which is based on syntactic patterns. This restricts more advanced question-generation authorities such as those based on inferential or valuative thinking skills of higher orders of Bloom's Taxonomy. In addition, no provision is made for an adaptive learning model to enable the system to learn how to adapt to the level of difficulty that the learner is able to handle. In short, the research presented a good starting point for the development of basic automated question generation, but in a sense, the work could not take advantage of an actual learning adaptive model or working with more advanced question types.

The study by Riza et al. [5] describes a system that was developed to generate short answers. The answers are produced on reading comprehension using complex natural language processing (NLP) and kNN, or the k-nearest neighbor algorithm. This system pulls out plain sentences from the input text. It picks out the sentences from reliable resources and formulates who, what, when, where, why, and how. This process follows several steps to construct question sentences for a given text passage and comparison with existing International English Language Testing System (IELTS) training data. The authors recommend this approach as it applies kNN, which helps the system to create questions that are very similar to the questions found in the IELTS database. Other elements used for NLP include POS tagging, which helps the system handle all sorts of sentence structures. Also, this system uses an elementary sentence extraction system (ESES) to reduce the complexity of the text before question generation. Also, removing the pronouns, which makes the sentence clear, and applying grammar checks along with replacing words with synonyms to make the question harder helps to output more precise questions. The study uses three evaluation metrics. They are grammatical correctness with an average of 59.52% and answer existence 95. 24% and the difficulty index score set to 34. The integrated measures have shown that the system has high accuracy for answer existence, but there are possible advancements that can be made in formulating more complex questions as well as in grammatical accuracy. Unfortunately, the authors observe that the

system mainly produces elementary factual short answer questions suitable for some standardized tests like IELTS but not for other educational purposes. The study highlights the ability to reduce the effort needed in the preparation of questions in learning institutions. However, since the usability is based on datasets from proficiency exams like IELTS the younger students or non-standardized studies will not be able to showcase their skills via the system. Also, this system lacks the ability to develop inferential questions needed for easy understanding. Also, this system does not provide questions for difficulty adjustments.

Alkhuzaey et al. [6] studied the practices towards the use of automatic methods to identify the difficulty level of Text-Based Questions (TBSQs). The study has focused on using Machine Learning(ML) and NLP approaches. The authors also have considered the aspect of question difficulty, for which predictive models occupy a unique position when designing educational assessments. In earlier days, question difficulty was determined either by subjective assessment of experts or by pre-testing, which are time-consuming and costly. This study highlights the diverging path toward data-driven and automatic solutions to overcome the limitations in TBSQs' difficulty level identification. The presented review helps to explore the types of supervised ML models used for difficulty prediction. This paper mainly focuses on the aspects of syntactic and semantic features as contributing to the difficulty level of questions. Simple readability formulas and complex readability formulas, such as word embedding and transformer-like techniques are reviewed in the paper. The paper also focuses attention on how syntactical and semantic features affect the difficulty level of questions. As gaps are found by the authors, the paper states the lack of publicly shared and standardized datasets that can be used in training the models and testing. Also, the lack of stable and diverse databases is noted. In conclusion, it is stated that there is a continuous search for more approaches to improve algorithms and feature extraction to improve and generalize educational systems depending on the learning environment.

Neural Question Generation(NQG) approaches that were proposed in the period between 2016 and the early half of 2022 in the education domain are discussed in the work by Al Faraby et al. [7]. NQG is a type of Automatic Question Generation (AQG), that uses a neural network model for generating questions. These questions are fluent, valid, and also in compliance with learning objectives. According to the study, neural global models like sequence-to-sequence and Transformers outperform rule-based systems, particularly at a large scale. But it also notes certain limitations, including the lack of educational contexts, the need for datasets of adequate quality for question generation tasks, and the challenges of writing higher-order questions that stimulate higher-order thinking skills. The authors recommend the continuation of studies aiming at creating and regulating the NGQ system, which will function at varying difficulty levels. Also, the system should be sensitive in the area of question formulation for specified learning goals. They also emphasize the generation of questions, with the interaction with external information, which will improve the quality and educational value of the generated questions.

Suhartono et al. [8] studied the creation of an AQG system in the domain of the Indonesian language. They have used pre-trained models, namely, the Indonesian version of the Bidirectional Encoder Representations from Transformer (BERT) model (IndoBERT), and a state-of-the-art (SOTA) language model for Indonesian based on the Generative Pre-training Transformer (GPT) model IndoGPT. The study aimed to remove the burden of creating good and structured tests from educators by automating the creation of questions for Reading Comprehension (RC). The study focuses on developing the sequence-to-sequence learning framework, using models such as Bidirectional Gated Recurrent Unit (BiGRU), Bidirectional Long Short-Term Memory (BiLSTM), Transformer, BERT, and GPT. It also discusses the pros of the usage of these models for creating text. With the help of contextual embeddings, POS tagging, and Named Entity Recognition (NER) to improve the quality of the questions. As novelties, the system introduces two architectures IndoBERT, which includes BERT as the encoder and a transformer as a decoder, and IndoTransGPT, which employs a transformer only for encoding. The study focuses on and emphasizes the effectiveness of these models in producing proper and relevant questions from the input texts. Bilingual Evaluation Understudy (BLEU), and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) were used as parameters to measure the outcomes of the study. They showed that the IndoBERT former model generates higher-quality questions with better efficiency and

better context. Considering this, the authors pointed out the difficulties in a system's scalability for the wider multilingual needs due to the insufficient funding for the Bhasha Indonesia language.

The paper published by Xu et al. [9] proposes and implements a Question Answering (QA) model for answering questions about Nanjing Yunjin digital resources, which reflects the traditional craft of silk weaving. Knowledge management is done by the Neo4J graph database to improve the understanding of the Nanjing Yunjin culture among civil servants. In intent recognition, BERT was used to categorize the user queries and entity recognition. The proposed model is BERT, BiGRU, and Case Report Form (CRF), which helps the system reply in natural language. The results from the study reveal that precision, recall, and the F1 score in the system validate the efficiency of the system responses to the user's queries. This is not only for the retrieval and use of Yunjin knowledge but also the framework for constructing similar systems for other fields, making the knowledge graphs vital for cultural conservation and education.

Another study conducted by Xu et al. [10] put forward a Nanjing Yunjin Intelligent QA System that is based on the Knowledge Graphs (KGs) and utilizes the Retrieval Augmented Generation (RAG) tactics to overcome the identified issues, which are associated with KGs-based systems. In the proposed system, the textual information is vectorized using the Roberta model. This approach helps the system in performing semantic matching of the user's queries with the information stored in the system, which aids in improving the interpretability and trustworthiness of the system's answers. Adding the Large Language Model (LLM) goes a level deeper by improving the natural language answer generation. The problems of graph update, type difference of the entities, and even semantic analysis is well taken care of here. The ease and difficulty in the stratum of the processed information and the queried information provide evidence of the increased efficiency of the system and its scope as the crucial framework of the significant preservation and promotion of the relevant intangible cultural heritage with the aid of Artificial Intelligence (AI) technologies.

Y. Tomikawa et al. [11] studied a system with adaptive QG and incorporated difficulty control using Item Response Theory(IRT). The system uses pre-trained transformer models. The goal of this research was to generate question and answer pairs with difficulty levels tailored to learners' abilities. Also, this dynamically estimates their proficiency. The system used BERT for answer extraction and T5 for QG component implementation, together with embedding difficulty values into the models. Moreover, a Computer Adaptive Testing (CAT) -inspired framework was used to extract difficulty selection based on learners' responses. The results stated positive results on the study achieving the research goals. A notable limitation of the study was the use of SAQuAD dataset, which assesses surface level comprehension. The study concludes that using more complex datasets and reading passage selection can improve adaptive learning outcomes.

According to the above literature review we can observe the improvement of automatic QG and answer evaluation area. Still we can find some open gaps in the research area. In forming basic factual questions, both rule-based systems and even kNN techniques work relatively well, however traditional NLP techniques struggle with the generation of more advanced inferential questions and do not pay attention to the needs of individual learners. Iterative response theory as a technique adapts flexibly according to a user's features, but requires well defined dataset that unfortunately remain superficial at best. On the other hand, neural model T5 and BERT do have the ability to produce questions that are context-sensitive and fluent, but like the other models previously mentioned, they are limited in scope with resource deficient contexts and situations. These shortcomings suggest the urgent need to develop a single efficient system that integrates all technologies and methods available to solve various educational problems and adapt to different learning environments.

## III. METHODOLOGY

This research consists of three main parts. Automatic passage difficulty adjustment, question generation, and answer evaluation. Under this section, the technologies, algorithms, and architectures used in developing, model training, and integrating the research components are discussed, together with data collection and processing. Figure 1 gives a brief idea on the overall system.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 21s, 2025
https://www.theaspd.com/ijes.php

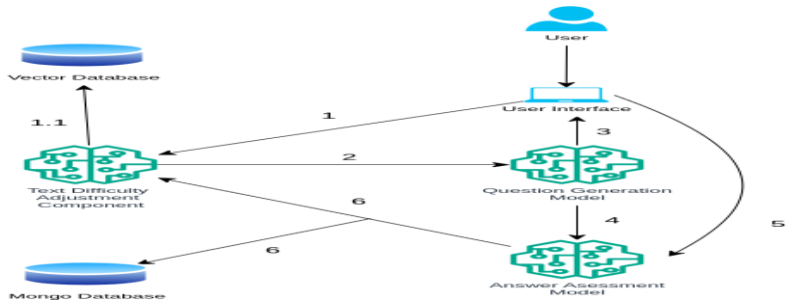


**Figure 1. High level system diagram.**

*A. Question Generation (QG) Component*

This section explains the development of a style-aware question generation module that uses the provided text passages to generate JSQ and SAQ. In this regard, all classes are taught using the English language curriculum for grades 3–5 in Sri Lanka, utilizing a customized model that is based on pre-trained T5-small but incorporates some style-specific enhancements.

1) *Data Preprocessing and Preparation:* The collection of data and preparation of the data for model training are discussed under this section.

a) *Raw Data and Task-Specific Transformation:* One of the major strengths of this research is the small domain-specific dataset it produced [12], which was meticulously chosen from English language textbooks for grades three to five in Sri Lanka [13] [14] [15]. The dataset is a custom-made dataset focused on the Sri Lankan English syllabus. Nearly 4000 examples that precisely reflect Sri Lankan primary school education terminology are included in the dataset. The dataset is stored as a JSON file where each sample includes an "input," "output," and "task" field. Processing for SAQ and JSQ tasks involves tagging input text with style markers ([SAQ] or [JSQ]) and using numeric identifiers (0 for SAQ and 1 for JSQ). The explicit style annotation directs both the preprocessing and generation stages. A fixed random seed is used to shuffle the transformed dataset to guarantee reproducibility. The dataset is split into a training portion of 70%, a validation portion of 20%, and a test portion of 10%, which are then saved individually as JSON files.

b) *Tokanization and Tensor Preparation:* T5 tokenizer (t5-small) processes both inputs and outputs to tokenize them into a standardized format. The tokenizer processes each sample by transforming the text into input IDs, attention masks, and decoder input IDs while applying necessary padding and truncation. The tensor representation includes the style identifier. The datasets tokenized with PyTorch (.pt) files allow for efficient training data loading

*2) Model Architecture:* Architectures followed by the used models are discussed in this section.

a) *Base T5 Framework:* We propose the custom model, which extends existing state-of-the-art T5-small based architecture using robust pre-trained language representation. It still relies on a shared embedding layer of the encoder and decoder made of deep copies of the original T5 architecture to benefit from the strengths of T5 for learning high-level language. This is further strengthened by novel, style-specific adaptation of SOTA style embeddings and an adapter network that allows the model to produce questions at the level of precise, task-specific nuances. [16].

*b) Incorporation of Style Embeddings:* The main innovation is a new learnable embedding layer to the model, which is dedicated to capturing style information. Since there are two tasks (SAQ and JSQ), the Embedding is created with two entries (indexed as 0 and 1) and a specified embedding size. The new layer allows the model to recognize and adjust to various question formats by providing two embedding elements that correspond to the SAQ and JSQ tasks [17].

c) *Adapter Network for Fusion:* The style embeddings enter the decoder through a small feed-forward adapter network, which combines them with hidden state outputs. The style information becomes integrated by this adapter through its two linear centric layers with ReLU activations to provide proper stylistic representation in generated questions. Other parameter-efficient transfer learning tasks have shown successful results from using adapter networks, according to Houlsby et al. [18].

*d) Custom Decoding Procedure:* A custom decoding loop to complete this procedure is done as the process generator. Starting with padding on a token, it iterates through token possibilities one by one till the sequence's end or until its maximum length. Accordingly, the decoding method encodes across the

two parameters while preserving the legal syntax and the format of the query to which it will be asked. This custom decoding loop makes sure that questions are syntactically generated in a style similar to the intended format.

*3) Training and Evaluation:* The training and evaluation process is carefully planned to maximize the style-aware question-generating model while maintaining reliable and repeatable results on the customized dataset. A dedicated QGDataset class that reads preprocessed JSON files and performs on stand tokenization over coming data in a dynamic fashion, while data is, again, batched efficiently with PyTorch DataLoaders. After the encoding, these are moved from the input IDs, attention masks, labels, and style identifiers to the computing device at each training iteration. The model then calculates the Cross-Entropy Loss and masks padding tokens by replacing them with -100 to focus on predicting the valid tokens. Furthermore, style-specific layers are integrated and monitored, and updates are used to track the gradients of the adapter network and dedicated style embeddings in a very close manner. To stabilize the training, gradient clipping is used, and the parameters are updated using the AdamW optimizer. A ReduceLROnPlateau scheduler is also used, besides adding validation loss and simply stopping training whenever it remains unchanged for a period of time.

*B. Question Answer Evaluation Component*

For the question-answering. It runs on the fine-tuned version of RoBERTa [19], a robust question answering model that has been pre-trained on the SQuAD2.0 set using a pre-trained model deepset/roberta-base-squad2. The data pipeline is then created by feeding in the QA model and its associated tokenizer so that the system can handle a passage and a corresponding question as input. During inference, the pipeline will pick the most relevant answer span from the passage and return with a confidence score. By ensuring modular integration, this method resolves the implementation and optimizes the whole solution, giving the QA module to work with the custom question generation module while maintaining good precision and reliability. This design choice leverages state-of-the-art pre-trained models, aligning with current best practices in natural language processing research 20].

*C. Text Passage Difficult Adjustment Component*

This part of the system enables users to have a customized learning experience by dynamically changing text difficulty depending on real-time user performance. The system uses a transparent linear regression model to perform data-driven correction that combines text analysis techniques with vector-based retrieval of passages. This architecture works to provide adaptive content delivery by enabling system speed in collecting passages that match the specified Flesch score. The system calculates the adjustment value for the Flesch score using a linear regression model with ridge regularization and polynomial features that are trained using a synthetically generated dataset that simulates user interactions [21]. This dataset comprises three features: Current difficulty (based on Flesch Reading Ease score) [22], Answer similarity (continuous value between 0 and 1 representing the mean difference between a asked QA's pipeline and a user's response), and User reward (1 for strong similarity, -1 for weak similarity, and 0 for neutral performance).

When the prediction is positive, it means it should be simpler (increase a Flesch score), otherwise, it should be harder (decreasing a Flesch score). The synthetic dataset has Gaussian noise added and has extreme examples. The polynomial degree and regularization strength are tuned on a grid search procedure using 5-fold cross-validation. This approach moves beyond fixed, rule-based methods by learning directly from simulated user performance, providing a transparent and interpretable mapping from user behavior to content difficulty adjustment [23]. Once trained, the linear regression model is integrated into the system's adaptive framework to dynamically adjust text difficulty in real time.

In the initial step, the system requires teachers to upload CSV files with text passages and their relevant grade levels as the first step of operation. Then, the Textstat library enables the system to evaluate Flesch Reading Ease scores, which serve as standard measures for text readability. The uploaded CSV is processed using the state-of-the-art SentenceTransformer model('all-MiniLM-L6-v2') to embed high-quality vectors for each of the defined passages in the CSV. Qdrant, a cutting-edge vector database, stores these embeddings along with original text, reading ease scores, and grade info. With this SOTA model for semantic embeddings, we ensure the system's speed and accuracy for retrieving the proper text passages through advanced metadata filtering and semantic search. When the user starts to answer the questions

in the inference stage, the system retrieves passages from the Qdrant vector database indexed with target Flesch Reading Ease range and uses them to the custom question generation model to generate questions. The QA pipeline is then used to compute answer similarity and assign a user reward as people interact with the system. A QA pipeline tries to evaluate the responses of the users, and the answer similarity is calculated for a user reward. The input feature vector for the regression model is then formed by these metrics, as well as the current difficulty, and the adjustment value is predicted. This prediction directly influences whether the difficulty of succeeding passages should be adjusted to contribute to raising or lowering the content continuously according to the needs of each student.

This data driven and feedback centric methodology ensures that the adaptive learning system is flexible and customized, and at the same time it devices for a successful connection between user performance and content complexity.

## IV. RESULT

The research focused on developing a web-based system that integrates automated Question Generation (QG), answer evaluation, and dynamic text complexity adjustment. The system was able to address key educational challenges effectively:

- Student Progress Tracking: The system monitors students' learning progress individually.
- Automated QG: Questions are generated dynamically from the passage content.
- Answer Evaluation: Answers are automatically graded with high accuracy.
- Text Adjustment: The level of difficulty of the input passage is tailored based on user performance.

Key aspects and outcomes are as follows:

- QG Accuracy: The QG model (T5) showed an increase in accuracy by 5.56% after fine tuning the model over 50 epochs, while validation accuracy remained consistent, as seen in Figure 2 and Figure 3.

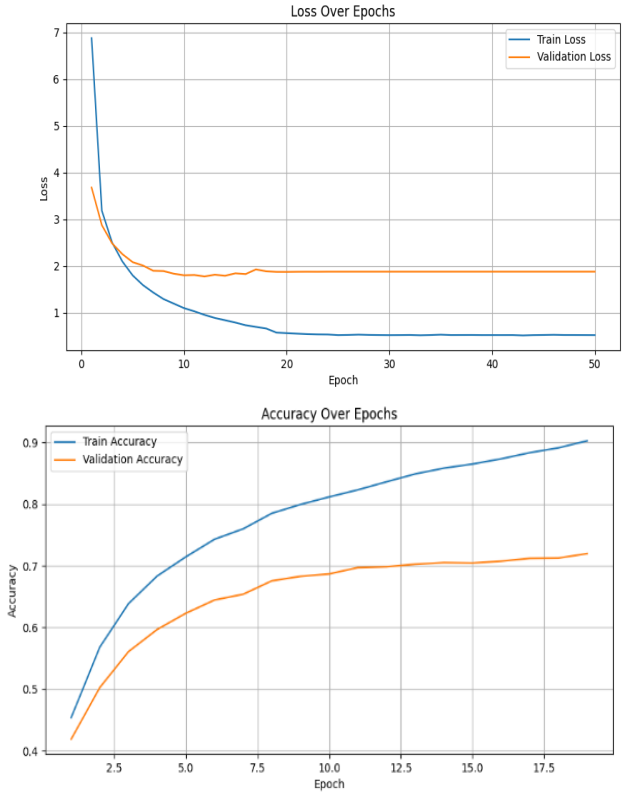


**Figure 2.Training script visualization of accuracy in QG model for 50 epochs before fine tuning**

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 21s, 2025
https://www.theaspd.com/ijes.php

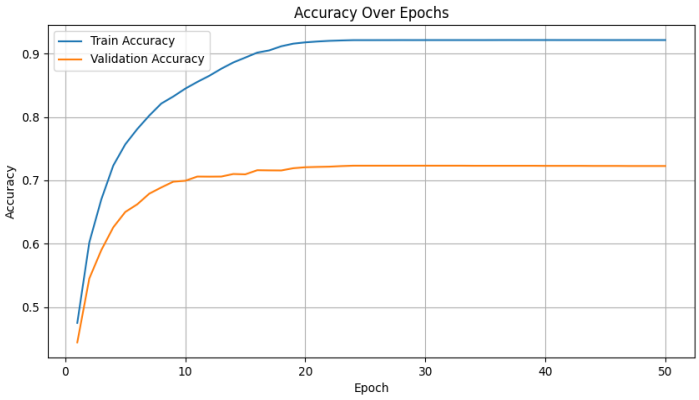


**Figure 3. Training script visualization of accuracy in QG model for 50 epochs after fine tuning**

• QG Loss Behavior: As can be observed in figure 4 and figure 5, train loss was enhanced by 20% after the QG model fine tuning, which means the model learned well on training data, though validation loss did not improve, indicating generalization problems.

**Figure 4. Training script visualization of loss over epochs in QG model before fine tuning.**

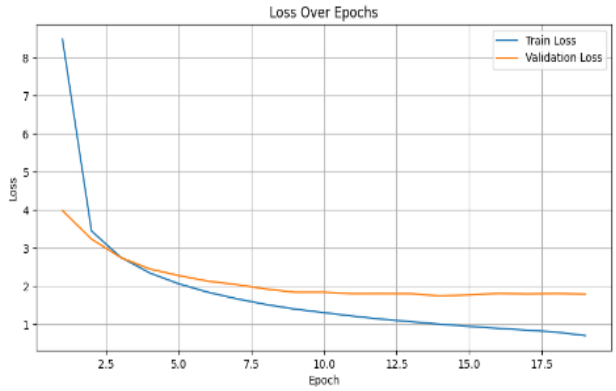


**Figure 5. Training script visualization of loss over epochs in QG model after fine tuning.**

• The system was capable of generating both SAQs and JSQs, which were effective in measuring student understanding and promoting deeper engagement. These questions were displayed to the user with a user friendly web interface.

## V. DISCUSSION

The application of T5 to QG and RoBERTa to answer evaluation was effective in generating relevant questions and evaluating answers appropriately. Adding a real-time difficulty adjustment function enhanced the system's responsiveness to individual learners. However, certain limitations and areas for improvement were observed:

• Overfitting of QG Model: Even though the training loss had decreased, inability to notice improvement in validation loss suggests overfitting, and the model had not generalized that well on unseen data.

• Synthetic Data Bias: Training from synthetically generated data may have introduced biases which could impact the performance of how well the model would do if trained on actual data.

• Simple Difficulty Model: While efficient, the current regression-based difficulty adaptation scheme can also be optimized further with the aid of deep learning-based personalization.

• Limited Question Types: Currently, support is restricted to SAQs and JSQs. Facilitating other diverse question types can facilitate learner measurement and improved engagement.

## VI. CONCLUSION

This research was able to effectively demonstrate the effectiveness and viability of an adaptive QG and assessment system with text complexity adjustment at runtime. The system was successful in achieving its

goals of automating material generation, evaluating student answers, and adjusting the difficulty level of learning content. The results have some issues with generalizability and adaptability with the existing models but are promising and have a great deal of potential for real-world application in education. Future studies should focus on:

- Using multiple, actual datasets to achieve greater generalizability.
- Enriching the difficulty adjustment module with deep learning methodologies.
- Providing more varieties of questions and broader coverage. The figure 6 and figure 7 displays the current web user interface with the generated short answer question and jumble sentence question.

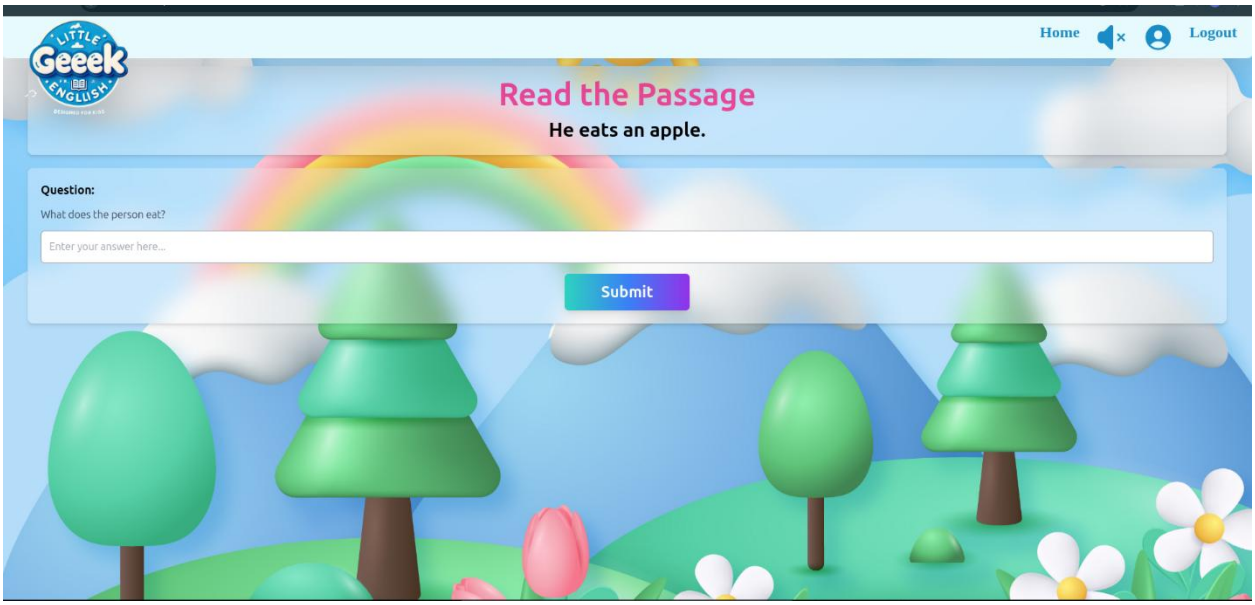


**Figure 6. Generated short answer question**

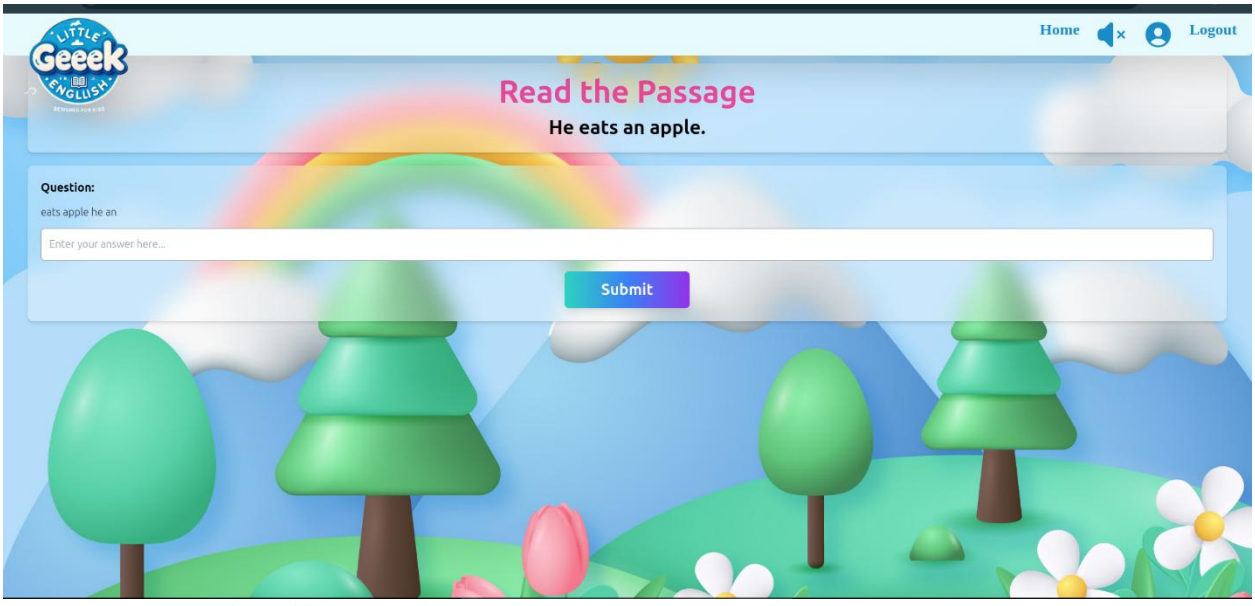


**Figure 7. Generated jumble sentence question**

By addressing these factors, the system can be made more effective and inclusive as an instructional resource, overcoming gaps of traditional learning modes and minimizing manual intervention by teachers.

**ACKNOWLEDGMENT**

The authors highly appreciate the continuous guidance and encouragement provided by the Sri Lanka Institute of Information Technology lecturers, colleagues, and family members. Special thanks is given to Mrs. Sureshinie Fernando, who helped us as an external supervisor to gather Sri Lankan curriculum data in the English language

## REFERENCES

[1] K. Grover, K. Kaur, K. Tiwari, N. Rupali, and P. Kumar, "Deep learning based question generation using T5 transformer," in Communications in computer and information science, 2021, pp. 243–255. doi: 10.1007/978-981-16-0401-0_18.

[2] X. Xue, J. Zhang, and Y. Chen, "Question-answering framework for building codes using fine-tuned and distilled pre-trained transformer models," Automation in Construction, vol. 168, p. 105730, Sep. 2024, doi:

10.1016/j.autcon.2024.105730.

[3] I. Rets, L. Astruc, T. Coughlan, and U. Stickler, "Approaches to simplifying academic texts in English: English teachers' views and practices," English for Specific Purposes, vol. 68, pp. 31–46, Jun. 2022, doi: 10.1016/j.esp.2022.06.001.

[4] P. Gumaste, S. Joshi, S. Khadpekar, and S. Mali, "Automated question generator system using NLP libraries," *Int. Res. J. Eng. Technol.*, vol. 7, no. 6, pp. 4568, Jun. 2020. e-ISSN: 2395-0056, p-ISSN: 2395-0072.

[5] L. S. Riza, Y. Firdaus, R. A. Sukamto, N. Wahyudin, and K. A. F. A. Samah, "Automatic generation of short-answer questions in reading comprehension using NLP and KNN," *Multimedia Tools and Applications*, vol. 82, no. 27, pp. 41913–41940, Apr. 2023, doi: 10.1007/s11042-023-15191-6.

[6] S. AlKhuzaey, F. Grasso, T. R. Payne, and V. Tamma, "Text-based question difficulty Prediction: A Systematic review of automatic approaches," *International Journal of Artificial Intelligence in Education*, vol. 34, no. 3, pp. 862–914, Sep. 2023, doi: 10.1007/s40593-023-00362-1.

[7] S. A. Faraby, A. Adiwijaya, and A. Romadhony, "Review on neural question Generation for education purposes," *International Journal of Artificial Intelligence in Education*, vol. 34, no. 3, pp. 1008–1045, Oct. 2023, doi: 10.1007/s40593-023-00374-x.

[8] D. Suhartono, M. R. N. Majiid, and R. Fredyan, "Towards automatic question generation using pre-trained model in academic field for Bahasa Indonesia," *Education and Information Technologies*, vol. 29, no. 16, pp. 21295–21330, Apr. 2024, doi: 10.1007/s10639-024-12717-9.

[9] L. Xu, L. Lu, M. Liu, C. Song, and L. Wu, "Nanjing Yunjin intelligent question-answering system based on knowledge graphs and retrieval augmented generation technology," *Heritage Science*, vol. 12, no. 1, Apr. 2024, doi: 10.1186/s40494-024-01231-3.

[10] L. Xu, L. Lu, and M. Liu, "Construction and application of a knowledge graph-based question answering system for Nanjing Yunjin digital resources," *Heritage Science*, vol. 11, no. 1, Oct. 2023, doi: 10.1186/s40494-023-01068-2.

[11] Y. Tomikawa, A. Suzuki, and M. Uto, "Adaptive Question–Answer Generation With Difficulty Control Using Item Response Theory and Pretrained Transformer Models," *IEEE Trans. Learn. Technol.*, vol. 17, pp. 2186-2198, 2024. doi: 10.1109/TLT.2024.3491801.

[12] Z. Dong, Q. Ding, W. Zhai, and M. Zhou, "A Speech Recognition Method Based on Domain-Specific Datasets and Confidence Decision Networks," *Sensors*, vol. 23, no. 13, p. 6036, 2023. doi: 10.3390/s23136036. [Online]. Available: https://www.researchgate.net/publication/372010682_A_Speech_Recognition_Method_Based_on_Domain-Specific_Datasets_and_Confidence_Decision_Networks

[13] Educational Publications Department, *English: Pupil's Book, Grade 3*. 3rd ed. Sri Lanka: Educational Publications Department, 2019. [Online]. Available: http://www.edupub.gov.lk/Administrator/English/3/english%20PB%20G-3/english%20PB%20G-3.pdf.

[14] Educational Publications Department, *English: Pupil's Book, Grade 4*. 2nd ed. Sri Lanka: Educational Publications Department, 2019. [Online]. Available: http://www.edupub.gov.lk/Administrator/English/4/en%20pb%20G-4/english%20PB%20G-4.pdf.
[15] Educational Publications Department, *English: Pupil's Book, Grade 5*. First ed. Sri Lanka: Educational Publications Department, 2019. [Online]. Available: http://www.edupub.gov.lk/Administrator/English/5/english%20pb%20G-5/English%20PB%20G-5.pdf.
[16] G. C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 140:1–140:67, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:204838007.
[17] Q. Gou, Z. Xia, B. Yu, H. Yu, F. Huang, Y. Li, and C.-T. Nguyen, "Diversify Question Generation with Retrieval-Augmented Style Transfer," in *Proc. EMNLP*, 2023, pp. 1677-1690. doi: 10.18653/v1/2023.emnlp-main.104. [Online]. Available: https://www.researchgate.net/publication/376392973_Diversify_Question_Generation_with_Retrieval-Augmented_Style_Transfer.
[18] N. Houlsby et al., "Parameter-Efficient Transfer Learning for NLP," *ArXiv*, vol. abs/1902.00751, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:59599816.
[19] J. Briskilal and C. N. Subalalitha, "An ensemble model for classifying idioms and literal texts using BERT and RoBERTa," *Information Processing & Management*, vol. 59, no. 1, p. 102756, Sep. 2021, doi: 10.1016/j.ipm.2021.102756.
[20] P. Rajpurkar, R. Jia, and P. Liang, "Know What You Don't Know: Unanswerable Questions for SQuAD," in *Proc. ACL*, 2018, pp. 784-789. doi: 10.18653/v1/P18-2124. [Online]. Available: https://www.researchgate.net/publication/334116183_Know_What_You_Don't_Know_Unanswerable_Questions_for_SQuAD.
[21] W. Warwick, R. Ford, and M. Funke, "Using synthetic datasets to hone intuitions within an adaptive learning environment," in *Lecture notes in computer science*, 2021, pp. 491–500. doi: 10.1007/978-3-030-90328-2_33.
[22] D. Rooein, P. Röttger, A. Shaitarova, and D. Hovy, "Beyond Flesch-Kincaid: Prompt-based metrics improve difficulty classification of educational texts," 2024. https://www.semanticscholar.org/paper/Beyond-Flesch-Kincaid%3A-Prompt-based-Metrics-Improve-Rooein-R%C3%B6ttger/d3338b08ade2a6ce0a1f54616936b3da0c44cd34
[23] D. Maulud and A. M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 1, pp. 140–147, 2020. doi: 10.38094/jastt1457.