Pertanika J. Sci. & Technol. 34 (3): 1493 - 1523 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Automated Generation of Calculation-based Physics Questions Using Text-to-Text Transfer Transformer (T5) for Malaysian Upper Secondary Education

Tuong Kiet Ngang* and Chih How Bong

Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia

ABSTRACT

Creating assessment questions to evaluate student achievement is a time-intensive task, especially in specialised subjects like Physics, particularly for structured, calculation-based problems in Science, Technology, Engineering, and Mathematics (STEM) disciplines. This study presents a web-based Automatic Question Generation (AQG) system that generates Physics questions for Malaysian upper secondary levels (Form 4 and 5). Covering topics such as Force and Motion, Heat, Light and Optics (Form 4), Pressure, and Electricity (Form 5), the system leverages transfer learning through the fine-tuning of the Text-to-Text Transfer Transformer (T5) model, a state-of-the-art natural language processing (NLP) technique. The methodology encompasses data construction, pre-processing, dataset generation, fine-tuning T5 model, evaluation, and inference. Performance was assessed through system experiments, ROUGE-L automatic evaluations, and human evaluations by expert educators, focusing on relevance, correctness, usefulness, and variety. The high ROUGE-L scores (0.82–0.85) indicate strong alignment with reference questions, while human evaluations demonstrate that the system generates contextually relevant and high-quality questions. The results from this study show that the AQG system matches the template approach for quality, but it is far more flexible and saves teachers a lot of manual work. It can also be scaled easily should more questions are needed. A comparative analysis with ChatGPT-4 was conducted, revealing the edge that a purpose-built, structured generator has over a broad and open-ended one. In short, deep-learning NLP can automate domain-specific question writing and make large-scale assessment design much simpler. These findings should interest researchers in computational linguistics, AI, and test automation.

Keywords: Artificial intelligence, automatic question generation, deep learning, natural language processing, question generation, Text-to-Text Transfer Transformer (T5), transfer learning

ARTICLE INFO

Article history:
Received: 27 July 2025
Accepted: 04 May 2026
Published: 19 June 2026

DOI: https://doi.org/10.47836/pjst.34.3.10

E-mail addresses:
tkngang@gmail.com (Tuong Kiet Ngang)
chbong@unimas.my (Chih How Bong)
* Corresponding author

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

INTRODUCTION

Question generation (QG) is defined by Rus et al. (2008) as "the task of automatically generating questions from various inputs such as raw text, database, or semantic representation" (p. 1). This definition gives academics and researchers a clear starting point for choosing the right questions and inputs. There are three important automatic question generation (AQG) components, namely the input, the output and the relationship between the input and the output. This is shown in Figure 1.

Automatic question generation (AQG), on the other hand, converts written text into interrogatives that are logically warranted by the source. This approach has attracted much attention since MOOCs put large-scale assessment on the agenda (Goldbach & Hamza-Lup, 2017). In interactive applications like conversational agents, Dialogue Policy Learning (DPL) manages multi-turn flow (Kwan et al., 2023). However, a fundamental and distinct challenge for AQG is crafting a single, high-quality question from a context. This core task is crucial to AQG's expanding role in tutoring and health-literacy tools, confirming its rising status (Mishra et al., 2020).

Despite this progress, producing physics questions for Malaysia's upper-secondary curriculum remains a stubborn challenge. Manual writing is slow, teachers have limited assessment training, and pupils receive too few and often repetitive practice tasks, undermining motivation and timely feedback (Erinosho, 2013; Saleh, 2014). Most existing AQG tools were built for reading comprehension and falter on numerical reasoning. These limitations highlight the need for domain-specific systems (Kurdi et al., 2020).

To address this need, the present study introduces a T5-based generator to automate the creation of physics questions for Malaysia's upper-secondary curriculum. Through this generator, it aims to lighten teachers' workload and sustain learner engagement by matching items to current understanding (Chew & Cerbin, 2021). The paper details the framework (dataset preparation, system architecture and evaluation metrics), the experimental setup with test designs and quality criteria, the results with interpretation, and finally the principal findings alongside directions for future research.

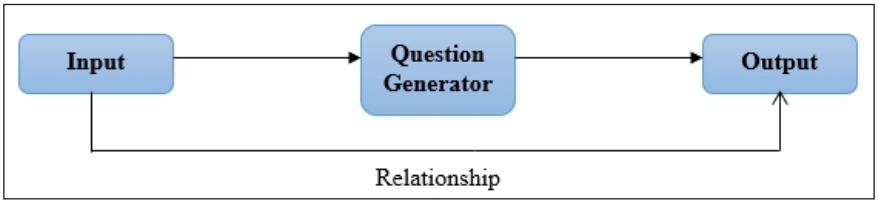

Figure 1. The components of AQG

PROPOSED METHODS

The methodology emphasises practical implementation and domain adaptation of the T5 model for physics question generation. The selection of the T5 architecture for the AQG system is justified by its strong performance in structured, domain-specific natural language generation tasks. Although newer foundational models have emerged, recent comparative studies suggest that the performance gap between fine-tuned T5 models and newer large language models is often marginal in educational contexts (Molina et al., 2024). For example, a study comparing GPT-3.5 Turbo, Llama 2-Chat 13B, and Flan-T5 XXL in generating questions from university lecture slides reported that although the newer LLMs slightly outperformed T5 in terms of clarity and question–answer alignment, all models exhibited comparable performance in relevance and question difficulty (Molina et al., 2024).

Furthermore, comparative studies Maity et al. (2024) report that T5 generally achieves stronger performance than BART and GPT-based models on automated evaluation metrics (e.g., ROUGE, BLEU) in educational NLP tasks. This is attributed to its unified text-to-text framework, which is well-suited for mapping structured inputs to natural language outputs in AQG.

Proposed Architecture of AQG

The proposed AQG architecture consists of six main processes: (1) Data Construction, (2) Data Pre-processing, (3) Dataset Generation, (4) Fine-tuning T5 Model, (5) Model Evaluation and Decoding, and (6) Inference. This architecture is illustrated in Figure 2. The T5 model, a transformer-based language model by Google, uses a "text-to-text" approach, enabling fine-tuning across tasks without task-specific architecture, reducing training data needs.

The architecture integrates template-based data construction with neural generalisation, enabling the T5 model to learn beyond fixed templates.

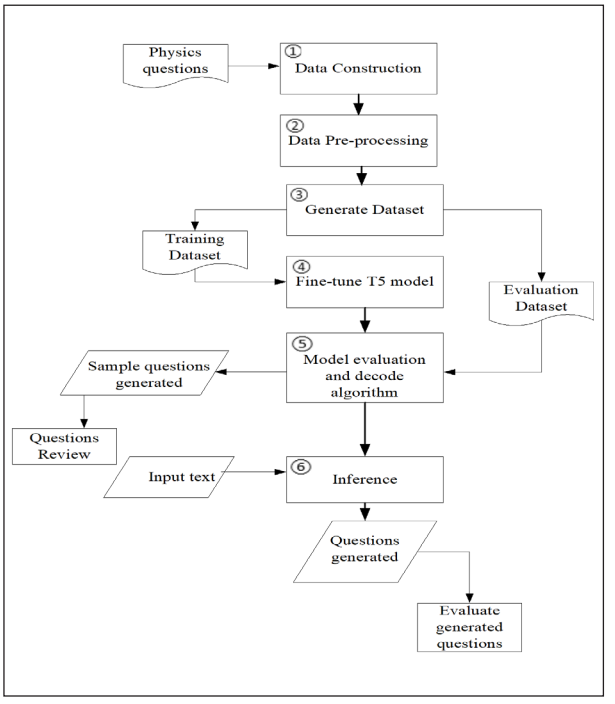

Figure 2. Proposed AQG architecture

Data Construction

In creating this generator, questions were exclusively derived from upper secondary Physics sources, including Form 4 and Form 5 textbooks (Choy et al., 2019; Chuan et al., 2020), the educational website Bumi Gemilang (https://www.bumigemilang.com), and a teacher-provided question bank. Sample questions underwent careful selection for clarity, accuracy, and relevance, following a structured procedure to ensure quality and appropriate difficulty levels. These questions align with curricular objectives and practical scenarios, with selection criteria prioritising relevance, coverage of content, and difficulty appropriate to each topic and subtopic. Three experienced Physics teachers reviewed and discussed each question to ensure consistency and educational alignment.

To introduce diversity and to ensure the system assesses computational skills rather than surface-level numeric substitution, sample questions were manually modified into eight distinct formats: "What is," "Determine," "Find," "Calculate," "What is the amount," "What is the value," "Figure out," and "How much." Each question was rephrased or restructured accordingly while retaining the central concept. For instance, "What is the acceleration of an object?" was adapted to "Determine the acceleration of the object" or "What is the value of acceleration," as demonstrated in Table 1. This variety requires students to actively decode the narrative context to identify physical variables.

In this study, a total of 4 modified question types were determined for different topics or subtopics. The primary emphasis was on the eight specific question types mentioned earlier.

Table 1
New question types were modified from the original question type

Original Question Type	New Modified Question Types
What is the acceleration of an object?	1. 1. **What** is the value of acceleration when a luggage with mass 15 kg is pulled by Elizabeth along a smooth horizontal surface with a force of 71 N?
	2. Elizabeth pulls a luggage of mass 15 kg along a smooth horizontal surface by a force of 71 N. **What** is the acceleration of the luggage?
	3. A piece of luggage with mass 15 kg is pulled by Elizabeth along a smooth horizontal surface. The force applied to the luggage is 71N. **Determine** the acceleration of the luggage.
	4. 4. Elizabeth pulls a luggage of mass 15 kg along a smooth horizontal surface by a force of 71 N. **Find** the acceleration of the luggage?

Data Pre-processing

Data pre-processing generates varied question types by randomly replacing values within set ranges, such as mass and force in acceleration questions. This process, automated using Python in the Anaconda Spyder environment, enhances scalability, maintains consistency, and minimises human error. This approach also saves time and ensures consistent randomisation across the dataset. Table 2 provides a summary of key variables and their ranges for various physics topics in Form 4 and 5.

Table 2
Variables and their ranges for selected Physics topics

No.	Form 4		
	Topic (question)	**Variables**	**Range of Values**
1	Topic: Linear Motion (displacement)	initial velocity	30 – 40 m/s
		duration	0.12 – 0.18 minutes (precision up to 2 decimal points)
		final velocity	50 – 60 m/s
2	Topic: Force (force)	mass	20 – 30 kg
		acceleration	0.5 – 2.0 m/s2 (precision up to 2 decimal points)
3	Topic: Force (acceleration)	force	60 – 80 N
		mass	10 – 20 kg
4	Topic: Force (mass)	force	60 – 70 N
		duration	5 – 9 s
		velocity	30 – 40 m/s
5	Topic: Specific Heat Capacity (specific heat capacity)	mass	0.2 – 0.6 kg (precision up to 2 decimal points)
		power	30 – 60 W
		initial temperature	18 – 21 oC
		final temperature	40 – 46 oC
		duration	80 – 100 s
6	Topic: Specific Latent Heat (specific latent of fusion)	power	450 – 500 W
		mass	0.22 – 0.30 kg (precision up to 2 decimal points
		duration	2 – 4 minutes
7	Topic: Gas Laws (temperature of air)	initial pressure	200 – 230 kPa
		initial temperature	24 – 27 oC
		final pressure	250 – 270 kPa
8	Topic: Refraction of Light (apparent distance)	depth	6 – 10 cm
		height	50 – 70 cm
9	Topic: Thin Lens Formula (size of image)	height	6 – 10 cm
		distance	50 – 70 cm
		focal length	12 – 18 cm

Table 2 (*continued*)

No.	Form 5		
	Topic (question)	**Variables**	**Range of Values**
1	Topic: Resultant Force (friction force)	force	70 – 80 N
		mass	5 – 10 kg
		acceleration	5 – 10 m/s2
2	Topic: Elasticity (elastic potential energy)	spring constant	20 – 30 N/cm
		mass	80 – 100 N
3	Topic: Pressure in Liquids (pressure in liquids)	depth	0.5 – 2.0 m (precision up to 2 decimal points)
		density	900 – 1200 kg/m3
4	Topic: Archimedes Principle (buoyant force)	volume	0.002 – 0.005 m3 (precision up to 3 decimal points)
		density of water	997 – 1003 kg/m3
5	Topic: Current and Potential Difference (current flow)	electric charge	500 – 550 C
		duration	3 – 7 minutes
6	Topic: Electrical Energy and Power (electrical energy)	potential difference	5 – 10 V
		current	0.5 – 0.9 A (precision up to 2 decimal points)
		duration	2 – 4 minutes

Dataset Generation

In this process, a modified question type is structured within a template, as shown in Table 3. Using a template format enhances organisation, clarity, and adaptability for deep learning analysis. Physics question templates cover various topics and difficulty levels, with each topic having a main template for core questions and three sub-templates introducing variations in problem-solving, conceptual understanding, and applied scenarios. In total, there are 15 main templates and 45 sub-templates, ensuring comprehensive and diverse question generation.

Table 3
Template to generate a dataset for acceleration

Object	Name	Question Type for Acceleration
Cable box	Maria	**1.** **Elizabeth** pulls a **NN** of mass **x** kg along a smooth horizontal surface by a force of **y** N. What is the acceleration of the **NN**?
Wooden door	Andrew	
Luggage	Terry	
Cement bag	Tommy	
Bench	Isabella	
Cupboard	Leslie	

Table 3 illustrates a sample question template with placeholders for objects and names. In an acceleration question, for example, placeholders (e.g., "NN" for objects and "PERSON" for names) are substituted with specific items and human names (like "Elizabeth") to increase question diversity. This substitution involves selecting an object from a list and using a function with regular expressions to replace placeholders efficiently. Named Entity Recognition (NER), via libraries like SpaCy, was used to identify and substitute human names to enhance the dataset variety for physics questions.

During dataset generation, Named Entity Recognition (NER) is employed to identify and tag human names within question templates. For example, in a template involving acceleration, names like "Elizabeth" are automatically identified by the NER model and labelled as "PERSON" to enable systematic substitution. This process also extends to other placeholders: the object placeholder "NN" is replaced with items from a predefined list (e.g., "table," "luggage"), while the numerical variables x (mass) and y (force) are replaced with values sampled from specified numerical ranges. This structured approach, which leverages NER for entity replacement and value ranges for numerical variability. Table 4 illustrates this NER-based process.

Using NLP techniques like Named Entity Recognition (NER) and object substitution, original question templates were customised with various names and objects to generate diverse scenarios. Python programmes further varied questions by substituting variables with random numbers, creating a robust dataset of 6,000 questions across 15 physics topics of varying difficulty. The dataset was split 70% (4,200) for training and 30% (1,800) for evaluation, covering 13 main topics structured into six chapters from the Form 4 and 5 syllabi, with four question types per topic. Table 5 presents examples of the "What is" question type, showing "target_text," "input_text," and "prefix" columns.

Table 4
NER process in dataset generation

Entity Type	Original Placeholder	NER Identification	Example of Substitution
PERSON	Elizabeth	Identified as PERSON	Andrew, Tommy

Table 5
Generated dataset for acceleration using questions type What is

target_text	input_text	prefix
Maria pulls a cable box of mass 10 kg along a smooth horizontal surface by a force of 68 N. **What is the acceleration of the cable box?**	{'mass', 'force', 'cable box'}	ask_question
Andrew pulls a luggage of mass 14 kg along a smooth horizontal surface by a force of 62 N. **What is the acceleration of the luggage?**	{'force', 'mass', 'luggage'}	ask_question

Table 5 (*continued*)

target_text	input_text	prefix
Terry pulls a cement bag of mass 20 kg along a smooth horizontal surface by a force of 76 N. **What is the acceleration of the cement bag?**	{'cement bag', 'mass', 'force'}	ask_question

Table 5 shows variations of similar questions generated by replacing values for mass, force, person names, and objects, producing a diverse dataset. For instance, mass and force vary across examples (e.g., 10 kg and 68 N with "Maria" and "cable box"; 14 kg and 62 N with "Andrew" and "luggage"; 20 kg and 76 N with "Terry" and "cement bag"). Each dataset pair includes an input_text containing key variables, sometimes with corrupted inputs, and a target_text that frames a complete question. This structure enables machine learning models to learn question generation patterns. The dataset, covering selected Form 4 and Form 5 Physics topics, was split into training and evaluation sets to train and assess a T5 model.

Although templates are used to structure the dataset, the diversity of templates, linguistic reformulations, and variable substitutions enables the model to learn generalisable question generation patterns rather than memorising fixed structures.

Fine-tune T5 Model

The T5 base model was selected due to its strength in generative tasks (Xue et al., 2021), such as those highlighted in modern e-learning applications (Patil et al., 2022). This choice is further supported by its proven effectiveness in generating structured educational content in STEM domains, such as the use of the multilingual T5 (mT5) model for creating math word problems (Gao, 2023). The text-to-text framework of T5 is particularly advantageous as it allows diverse NLP tasks, including question generation, to be modelled uniformly as generating target text from a source text prompt.

The model applies the Masked Language Model (MLM) objective (Shi & Wolff, 2021), which involves predicting masked portions in the input, making it suitable for learning variable placement and structure in Physics questions. During fine-tuning, model weights are updated via backpropagation to reduce prediction errors. The final tuned weights are saved for evaluation and inference. Detailed fine-tuning setup, dataset configuration, and training parameters are discussed in the section “Fine-tuning T5 Model for AQG”.

Model Evaluation and Decoding

After fine-tuning, the AQG model was evaluated to assess the relevance, correctness, usefulness, and diversity of generated Physics questions. Each question was reviewed and compared to reference questions for quality. To generate coherent and natural-sounding

questions, decoding algorithms such as Top-k and Top-p sampling were used. These approaches filter word choices during generation to balance coherence and diversity (Tucker, 2020; von Platen, 2020). Decoding strategies (Top-k and Top-p) are employed to encourage variation in generated questions. Full evaluation configuration, sampling settings, and loss trends are described in the section "Evaluation Configuration and Decoding Settings".

Inference

The fine-tuned model was used for inference via a web-based Anvil interface linked to a Jupyter Notebook backend. Users entered Physics variables, and an object, and the system generated four diverse questions automatically. For example, with initial velocity, final velocity, and duration as inputs, and "car" as the object, the system produced four displacement questions. Four examples are shown in Table 6.

Table 6
Question generated for displacement during inference

Level:	Form 4
Chapter:	Force and Motion I
Topic:	Linear Motion
Question to be asked:	displacement
Input text (variables):	initial velocity, final velocity, duration
Object involved:	car
Generated Questions	
1.	Andrew drives a car on a straight path. The car accelerates uniformly from a velocity of 33 m/s to a velocity of 52 m/s in 0.13 minutes. **What is the total displacement** travelled by the car?
2.	Andrew drives a car on a straight path. The car accelerates uniformly from a velocity of 39 m/s to a velocity of 52 m/s in 0.15 minutes. **What is the total displacement** travelled by the car?
3.	As a car moves along a straight track, its velocity is 35 m/s. After 0.13 minutes, the car has reached 51 m/s. **What is its displacement?**
4.	A car is moving along a straight road with a velocity of 40 m/s. 0.16 minutes later, its velocity has increased to 56 m/s. **Determine the displacement of the car?**

EXPERIMENT SETUP

Dataset

In this research, the T5 model's input dataset was manually compiled from Physics sample questions across various topics and subtopics in Form 4 and Form 5. A total of 6 chapters and 13 topics were included (Form 4: 3 chapters, 7 topics; Form 5: 3 chapters, 6 topics), as shown in Table 7.

In addition, all labelling and question combinations were done manually to ensure accuracy, as detailed in Table 8. The target text represents the "questions set"

intended for generation or analysis, while the input text includes variables relevant to each question type. This provides context for meaningful question formulation. The prefix "ask_question" acts as a string identifier for executing the task of question generation or processing.

Table 7
List of chapters and topics

Form 4	**Topic (question)**
Chapter	
2	Topic: Linear Motion (displacement)
2	Topic: Force (force)
2	Topic: Force (acceleration)
2	Topic: Force (mass)
4	Topic: Specific Heat Capacity (specific heat capacity)
4	Topic: Specific Latent Heat (specific latent of fusion)
4	Topic: Gas Laws (temperature of air)
6	Topic: Refraction of Light (apparent distance)
6	Topic: Thin Lens Formula (size of image)
Form 5	
1	Topic: Resultant Force (friction force)
1	Topic: Elasticity (elastic potential energy)
2	Topic: Pressure in Liquids (pressure in liquids)
2	Topic: Archimedes Principle (buoyant force)
3	Topic: Current and Potential Difference (current flow)
3	Topic: Electrical Energy and Power (electrical energy)

Table 8
Dataset (combination of different Physics topics)

	target_text	**input_text**	**prefix**
1.	What is the value of acceleration when a luggage with mass 17 kg is pulled by Tommy along a smooth horizontal surface with a force of 74 N?	force, mass, luggage	ask_ question
2.	In a physics class, a student prepared a simple electrical circuit that connected a 6 V battery and a bulb. Calculate the electrical energy that is supplied for 2 minutes if the current is 0.71 A?	potential difference, bulb, current, duration	ask_ question
3.	What is the size of the image of a cup with a height of 6 cm, when it is placed at 63 cm from a concave lens with a focal length of 18 cm?	focal length, cup, distance, height	ask_ question

Table 8 (*continued*)

	target_text	input_text	prefix
4.	A steel block with a volume of 0.005 m3 is immersed in water. If the density of water is 997 kg/m3, what is the buoyant force experienced by the steel block? Gravitational acceleration is 9.81 m/s2.	steel block, density of water, volume	ask_ question
5.	A lorry is moving along a straight road with a velocity of 30 m/s. 0.14 minutes later, its velocity has increased to 53 m/s. Determine the displacement of the lorry?	initial velocity, lorry, duration, final velocity	ask_question

Implementation Details

Fine-tuning T5 Model for AQG

The T5-base pretrained model was fine-tuned using a dataset of 6,000 samples, split into 70% for training and 30% for evaluation (Table 9).

Table 9
Dataset distribution for model training

Dataset	Number of Samples
Training dataset	4200
Evaluation dataset	1800

Fine-tuning involved adjusting the model parameters for the AQG task using this custom dataset. The hyperparameters used are summarised in Table 10.

Table 10
Fine-tuning hyperparameters for the T5 model

Parameter	Value
Model	T5-base
Batch size	5
Number of epochs	3
Learning rate	1e-4
Maximum sequence length	80 tokens

These values were selected based on standard practices and hardware constraints. A batch size of 5 was used due to the relatively small dataset size and limited memory availability, enabling more frequent updates and reducing overfitting risk. The model was trained for 3 epochs, as training loss stabilised by then, minimising the chance of overtraining. A learning rate of 1e-4 allowed for gradual and stable updates, which is preferable for specialised datasets compared to the commonly used 0.001 rate (Raffel et al., 2020). The maximum sequence length of 80 tokens balanced processing efficiency and input completeness, compared to longer default lengths used in other Transformer-based models (Devlin et al., 2019). The AdaFactor optimiser (Shazeer & Stern, 2018) was used to reduce memory usage during training, which completed in approximately 15 minutes. The Weights and Biases (wandb) platform was used to log training metrics and visualise progress.

Figure 3 shows the total training time, while Figure 4 presents the training loss curve. The steadily decreasing loss indicates effective learning and convergence.

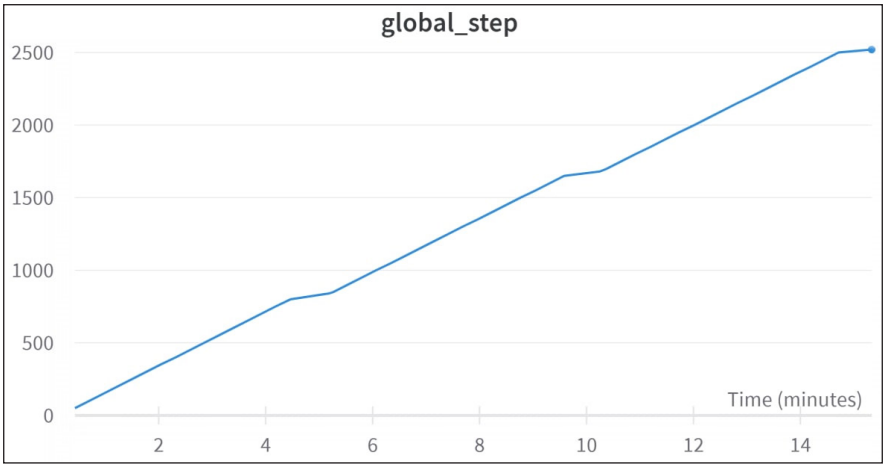

Figure 3. Total time taken for training the T5 model

Figure 4. Training loss curve

Evaluation Configuration and Decoding Settings

The AQG model was evaluated by generating questions from input text and comparing them to reference questions. Using parameters from Table 11, the evaluation ran in batches of 3, with 80 token sequences over one epoch, optimising memory and efficiency while preventing overfitting.

Table 11
Evaluation parameters for the AQG model

Parameter	Value
Batch size	3
Number of epochs	1
Maximum sequence length	80 tokens

The evaluation loss curve (Figure 5) shows a decrease over time, indicating the model's ability to generalise and learn effectively. Evaluation settings were adjusted based on early loss behaviour. A smaller batch size of 3 provided more frequent updates and improved learning precision for this generative task. Though smaller batches introduce noise, they can act as a regularise, enhancing generalisation.

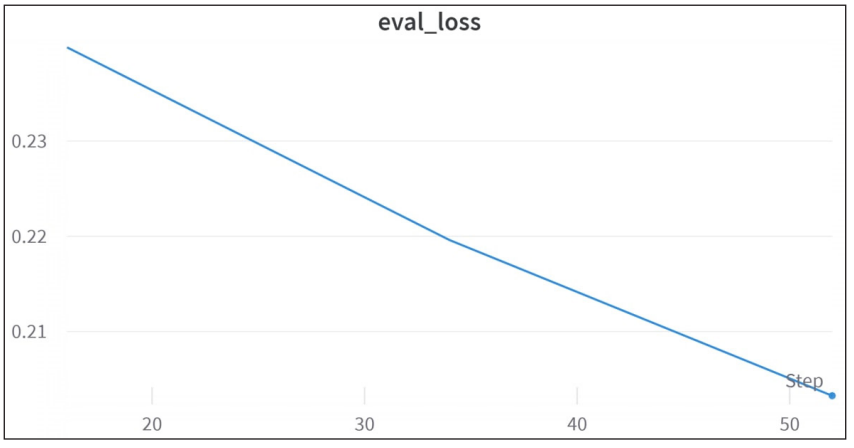

Figure 5. Graph of evaluation loss

The number of epochs was fixed at 1, as early experiments showed that additional epochs led to marginal improvement while increasing the risk of overfitting. A maximum sequence length of 80 was retained, balancing context capture and training efficiency. For decoding, this study used a combination of Top-k =50 and Top-p = 0.90 sampling, as shown in Table 12. These probabilistic methods filter out unlikely word candidates and dynamically adjust the word pool based on cumulative probability (Fan et al., 2018; Holtzman et al., 2020). This configuration consistently yielded coherent and contextually appropriate questions.

Table 12
Value of the sampling scheme

Parameter	Value
Top_k	50
Top_p	0.90

Four output sequences were generated for each target row in the evaluation dataset. All generated questions were stored in a .txt file. Samples are illustrated in Tables 14 and 16, based on real questions from Tables 13 and 15, respectively.

Table 13
Input text and real question for acceleration in the evaluation dataset

Input Text:	**{'mass', 'force', 'bench'}**
Real question:	What is the value of acceleration when a bench with mass 14 kg is pulled by Leslie along a smooth horizontal surface with a force of 63 N?

Table 14
Generated questions for acceleration

1.	Andrew pulls a bench of mass 18 kg along a smooth horizontal surface by a force of 80 N. What is the acceleration of the bench?
2.	Leslie pulls a bench of mass 19 kg along a smooth horizontal surface by a force of 68 N. Find the acceleration of the bench?
3.	A bench with mass 18 kg is pulled by Terry along a smooth horizontal surface. The force applied to the bench is 68 N. Determine the acceleration of the bench?
4.	What is the value of acceleration when a bench with mass 10 kg is pulled by Tommy along a smooth horizontal surface with a force of 63 N?

Table 15
Input text and a real question for specific heat capacity in the evaluation dataset

Input Text:	**{'bronze medal', 'mass', 'duration', 'temperature', 'power'}**
Real question:	What is the specific heat capacity of the bronze medal with mass 0.5 kg, when it is heated by an electric heater with a power of 57 W for 92 s and the temperature increases from 19 oC to 40 oC?

Table 16
Generated questions for specific heat capacity

1.	A bronze medal with mass 0.57 kg is heated by an electric heater with a power of 45 W for 85 s. The temperature of the bronze medal increases from 18 oC to 45 oC. Calculate the specific heat capacity of the bronze medal?
2.	A bronze medal with mass 0.53 kg is heated by an electric heater with a power of 46 W for 87 s. The temperature of the bronze medal increases from 20 oC to 40 oC. Determine the specific heat capacity of the bronze medal?
3.	What is the specific heat capacity of the bronze medal with mass 0.43 kg, when it is heated by an electric heater with a power of 39 W for 97 s and the temperature increases from 21 oC to 44 oC?
4.	A bronze medal with mass 0.35 kg is heated by an electric heater with a power of 39 W for 89 s. The temperature of the bronze medal increases from 18 oC to 42 oC. Determine the specific heat capacity of the bronze medal?

Inference Response Time

The AQG system measures response time using a Jupyter server and an Anvil web app. Users input data, which Anvil sends to Jupyter for question generation. The Jupyter notebook records response times, helping identify bottlenecks and optimise performance.

Experiment

Experiment 1: Question Generation with a New Object

This experiment tests the AQG model's ability to generate questions for unseen objects and achieve denoising using masked language modelling (MLM). Success is measured by the model's accuracy in predicting masked tokens and restoring sequences with minimal errors. This process improves question generation by training the model to handle incomplete inputs, enhancing its ability to produce coherent, relevant questions with high relevance and robustness.

This experiment focused on testing the concept of Force from the Form 4 chapter on Force and Motion I. Using the variables mass and acceleration, a new object, "trolley," was introduced to generate Force-related questions (Table 17).

Table 17
Experimental parameters for force-related question generation with a new object

Level:	Form 4
Chapter:	Force and Motion I
Topic:	Force
Question to be generated:	force
Input text (variables):	mass, acceleration
Object involved:	trolley

Table 18
Experimental parameters for pressure in liquids-related question generation with new object

Level:	Form 5
Chapter:	Pressure
Topic:	Pressure in Liquids
Question to be generated:	pressure in liquids
Input text (variables):	depth, density
Object involved:	prawn

The next topic in this experiment focused on pressure in liquids from the Form 5 pressure chapter. The question involved pressure in liquids, with "prawn" as the new object and depth and density as variables (Table 18).

The AQG model experiments used a Form 4 topic (force) and a Form 5 topic (pressure in liquids) to assess question generation. New objects like "trolley" (for force) and "prawn" (for pressure) tested the model's generalisation. These choices enhanced question diversity, reflecting real-world physics and challenging the model's contextual accuracy.

Experiment 2: Multiword Target

This experiment examines the AQG model's handling of multiword targets, as BERT and RoBERTa typically map single words to masks (Shi & Wolff, 2021). Using "steel pot" for specific heat capacity, the study assessed how well the model linked input variables to context. Table 19 presents the data. Success was measured by comparing generated questions with expected outputs, providing insights into the model's ability to process complex linguistic structures.

Other additional experiments are to generate questions for the topics Thin Lens Formula and Current and Potential Difference. These questions focus on calculating the size of an image (using focal length, distance, and height for a plastic bottle) and determining the current flow (using electric charge and duration for a copper block), exploring how the AQG model handles multiword targets.

Table 19
Experimental settings for AQG model handling of multiword targets for topic-specific heat capacity

Level:	Form 4
Chapter:	Heat
Topic:	Specific Heat Capacity
Question to be generated:	specific heat capacity
Input text (variables):	power, temperature, mass, duration
Object involved:	steel pot

Comparison Between ChatGPT-4 and AQG System

The emergence of Large Language Models (LLMs) has transformed education, particularly in automated question generation, by offering unprecedented generative capabilities (Kasneci et al., 2023). However, alongside these opportunities exist significant challenges concerning their reliability and suitability for specialised pedagogical tasks. This section addresses this gap by comparing the effectiveness of a purpose-built AQG system based on the T5 model with ChatGPT (GPT-4) in generating calculation-based physics questions. The evaluation focuses on four key dimensions: relevance, correctness, usefulness, and variety. These aspects are assessed by three experienced physics teachers, as detailed in the section "Evaluating Questions Generated by ChatGPT-4 and AQG System".

To ensure consistent experimental conditions, both systems were evaluated using identical input prompts. Nevertheless, the comparison reflects different modelling paradigms, as the T5-based AQG system is fine-tuned on domain-specific data, whereas ChatGPT-4 is evaluated using an instruction-guided zero-shot prompting approach. ChatGPT-4 was configured using custom instructions aligned with Table 20 to ensure consistent response structure, linguistic behaviour, and output constraints. In addition, both ChatGPT-4 and the AQG system were evaluated under a single-run generation setting, without repeated sampling, to maintain parity and control variability across systems.

Configuring Custom Instructions in ChatGPT-4

Custom instructions let users set preferences for ChatGPT-4, ensuring consistent responses without repetition. For this comparison, instructions are tailored to align with Table 20, ensuring relevant replies.

Table 20
Data and details for custom instruction alignment in ChatGPT-4

Custom Instructions	Data and Details
What would you like ChatGPT to know about you to provide better responses?	I am doing automatic question generation for physics calculation questions.
How would you like ChatGPT to respond?	Please help to generate a question, based on these details: 1. Question to be generated 2. The input 3. Object Condition: add value to the input Then, add diversity to the questions by using this format (What is, Determine, Find, Calculate, what is the amount, What is the value? Figure out, and How much? ** Note: Set the limit to generate 4 different types of questions only.

Evaluation Method

The evaluation assessed AQG-generated and template-based questions using human and automatic methods. Three Physics teachers evaluated AQG questions on relevance, correctness, usefulness, and variety, while template-based questions were assessed on relevance, correctness, and usefulness. The evaluators each possess over 15 years of experience in the Malaysian secondary education system and specialise in Physics teaching and curriculum, supporting the reliability and credibility of the human evaluation. For automatic evaluation, the nlg-eval Python package (Sharma et al., 2017) with the ROUGE-L metric was used to measure syntactic and semantic similarities by comparing the longest common subsequence (LCS) between generated and reference questions.

Human Evaluation

Human evaluation was conducted using four criteria: correctness, relevance, usefulness, and variety. A 1-4 scale adapted from previous work was used to ensure systematic and consistent assessment, where 1 indicated poor quality, and 4 indicated excellent quality (Amidei et al., 2018; Le & Pinkwart, 2015; Yao et al., 2012). Correctness evaluated grammatical accuracy and fluency, relevance assessed how well the question aligned with the input and the national curriculum, usefulness considered the question's potential to support meaningful learning (pedagogical value); and variety evaluated the diversity of question types, where a score of 1 indicated one type and 4 indicated four distinct types. The human evaluation involved three Physics teachers, who validated a total of

180 generated questions. Their evaluation particularly emphasised the criteria of usefulness and relevance to the national curriculum, thereby strengthening the credibility and educational alignment of the evaluation. The template-based approach used correctness, relevance, and usefulness

The reliability of the human evaluation was assessed using Randolph's Free-Marginal Kappa. This metric was chosen to provide a more accurate reflection of agreement than Fleiss' Kappa, which can be artificially deflated when rater agreement is exceptionally high (the 'Kappa Paradox'). The analysis was conducted on 180 evaluation units generated by the T5-based AQG model, excluding template-based results to focus strictly on the transformer model's performance.

Evaluation Setup of the AQG Model

A web page enabled AQG model evaluation, where users input data processed by a Jupyter server. Three Physics teachers generated four questions per click (Figure 6) and copied them to a Word template for review. Each evaluator reviewed 60 AQG-generated questions and 15 template-based questions. While the AQG system generated different questions for each evaluator, the template-based questions were identical, as there was only one set per template. This ensured a comprehensive and consistent assessment across both types of questions.

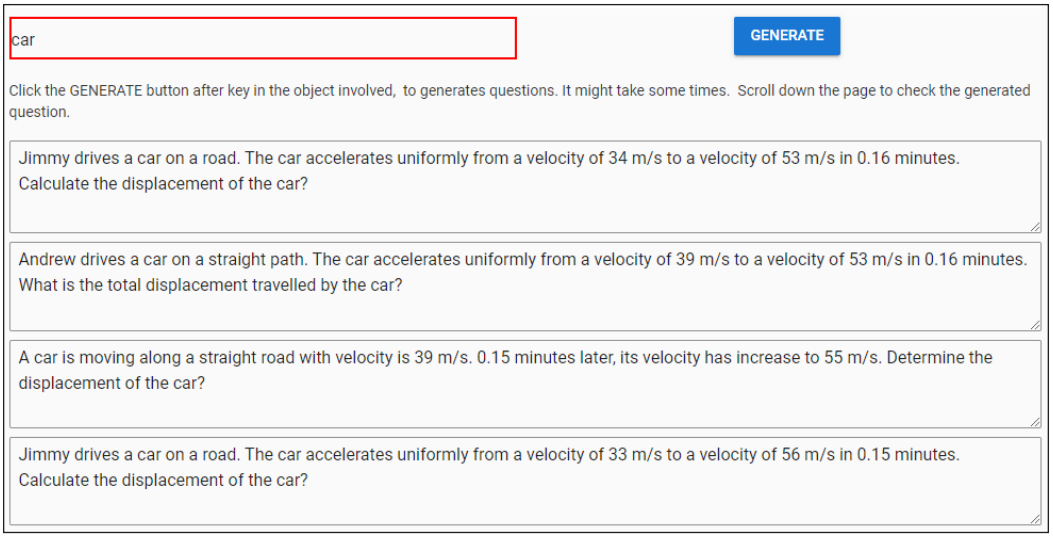

Figure 6. Questions generated using the AQG method

Evaluation Setup of Template-based Generated Questions

Template-based question evaluation followed a structured process. Question templates were created for various Physics topics, and a program allowed evaluators to select a topic and generate questions by filling template placeholders with relevant Physics data (Figure 7).

Automatic Evaluation

The nlg-eval tool evaluated generated questions by processing data from three evaluators, each generating 60 questions stored in separate files (hyp1.txt, hyp2.txt, hyp3.txt). These were then compared to a golden reference set (ref.txt), ensuring alignment by question type and topic. The nlg-eval command was executed for each hypothesis file against the reference, focusing on ROUGE-L, which measures the longest matching word sequence to assess accuracy and relevance (Figure 8).

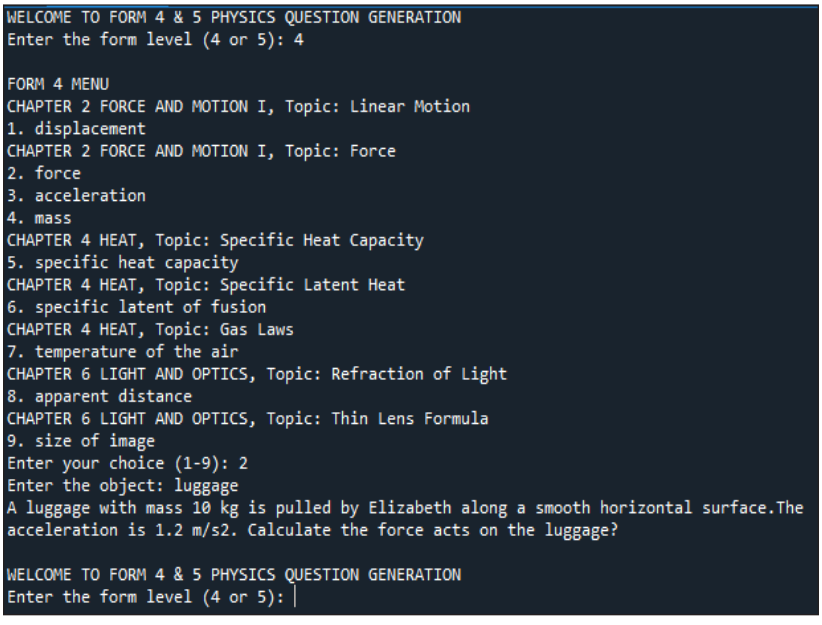

Figure 7. Questions generated using the template-based approach

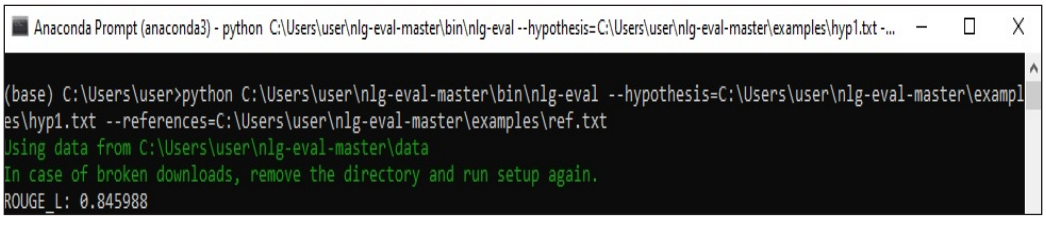

Figure 8. Evaluation of hypothesis files using the nlg-eval command

RESULT AND DISCUSSION

This section discusses experimental results and human evaluation findings to assess the quality of generated questions. The analysis offers insights into their effectiveness in educational question generation.

Inference Response Time Result

The response time test measured the time required to generate questions, with four questions generated per input to ensure consistent evaluation. The average number of words and generation time per question (in seconds) were recorded (Table 21). The longest processing time, 1.96 s, was observed for the Specific Heat Capacity topic, corresponding to the highest average word count (42.5). The overall mean response time per question was 1.89 s, indicating satisfactory response times for the AQG system.

Table 21
Result of the response time test

Form 4	Topic (question)	Average No. of Words per Question	Average Generation Time per Question (seconds)
Chapter			
2	Topic: Linear Motion (displacement)	36	1.94
2	Topic: force (force)	26.5	1.86
2	Topic: force (acceleration)	26	1.84
2	Topic: force (mass)	31	1.89
4	Topic: Specific Heat Capacity (specific heat capacity)	42.5	1.96
4	Topic: Specific Latent Heat (specific latent of fusion)	41.5	1.95
4	Topic: Gas Laws (temperature of air)	37	1.94
6	Topic: Refraction of Light (apparent distance)	40	1.88
6	Topic: Thin Lens Formula (size of image)	32	1.86
Form 5			
1	Topic: Resultant Force (friction force)	29	1.85
1	Topic: Elasticity (elastic potential energy)	33	1.86
2	Topic: Pressure in Liquids (pressure in liquids)	39	1.94
2	Topic: Archimedes Principle (buoyant force)	30	1.89
3	Topic: Current and Potential Difference (current flow)	19	1.77
3	Topic: Electrical Energy and Power (electrical energy)	34	1.90
	Mean average time per question		1.89

This performance can be evaluated against established Human–Computer Interaction (HCI) benchmarks and reported baseline latencies in educational AI workflows (Cornejo et al., 2024; Jauhiainen & Guerra, 2024), further supporting its suitability for interactive applications. Compared to general-purpose systems, which typically exhibit higher response times, the AQG system demonstrates improved efficiency for real-time educational use.

The observed latency of 1.89 seconds falls within the recommended response time for interactive systems, typically defined as between one and two seconds (Roth, 2013). Furthermore, based on the Seow framework as discussed in Cornejo et al. (2024), response times under five seconds are categorised as "continuous" interaction, indicating that system feedback remains sufficiently immediate to support user engagement. This level of responsiveness ensures that the question generation process does not disrupt the natural dialogue between teacher and system, thereby preserving the flow of thought essential for effective classroom pacing (Cornejo et al., 2024; Wang & Reani, 2017).

To provide a clearer rationale, the efficiency of the specialised T5 model was compared against reported baseline inference times of general-purpose Large Language Models (LLMs) used in educational contexts. While the underlying tasks and system configurations differ, prior studies consistently report higher per-interaction latency for such models. For example, Jauhiainen and Guerra (2024) report that GPT-4 requires an average of 18.51 seconds to process a single educational request. Similarly, Parker et al. (2024) indicate that GPT-4 responses for educational feedback tasks typically take approximately 10 seconds.

Although general-purpose models offer broader capabilities, their relatively high latency (approximately 10–18 seconds per interaction) may limit their suitability for time-sensitive, real-time educational applications. In contrast, the proposed specialised AQG system achieves an average response time of 1.89 seconds, suggesting a substantial improvement in responsiveness and making it more suitable for live, teacher-led instructional settings.

Experiment Result and Discussion

Experiment 1: Result Discussion

Experiment 1 (Force concept) demonstrates the AQG system's ability to generate questions for unseen objects, such as a "trolley". The system successfully incorporates the new object into generated questions in this controlled test case, as shown in Table 22, indicating its capability to generalise to unfamiliar inputs. This suggests that the model can adapt to new object instances by leveraging learned patterns from the training data rather than relying solely on memorised templates. However, this result should be interpreted as a case-specific observation rather than evidence of perfect performance. It reflects successful object handling under controlled conditions, while broader performance variations are observed in other experiments, as shown in Table 23.

Table 22
Experimental results on using a new object to generate questions for the topic Force

	Generated Questions
1.	What is the Force applied on a **trolley** with mass 19 kg, if it is pulled by David along a smooth horizontal surface with acceleration 1.29 m/s2?
2.	A **trolley** with mass 14 kg is pulled by David along a smooth horizontal surface. The acceleration is 0.86 m/s2. Calculate the force that acts on the **trolley**?
3.	A **trolley** with mass 19 kg is pulled by David along a smooth horizontal surface. The acceleration is 1.54 m/s2. Determine the Force of the **trolley**?
4.	A **trolley** with mass 12 kg is pulled by Tommy along a smooth horizontal surface. The acceleration is 0.63 m/s2. Calculate the force that acts on the **trolley**?

Table 23
Experimental results on using a new object to generate questions for the topic Pressure in Liquids

1.	A **shrimp** is at a depth of 1.57 m in a tank. The density of the water in the tank is 959 kg/m3. Find the pressure exerted by the water on the **fish**? g = 9.81 m/s2.
2.	When a **shrimp** is submerged at a depth of 1.85 m in a lake with a density of 1035 kg/m3, calculate the water pressure exerted on it? g=9.81 m/s2
3.	A **shrimp** is at a depth of 1.57 m in a tank. The density of the water in the tank is 959 kg/m3. Find the pressure exerted by the water on the **fish**? g = 9.81 m/s2.
4.	A **crab** is at a depth of 1.58 m in a lake. The density of water in the lake is 957 kg/m3. What is the pressure experienced by the crab caused by the water around it? g = 9.81 m/s2.

Table 23 presents the generated questions for topic pressure in liquids, which use "prawn" as the new object. The results show an overall inconsistency in object substitution, indicating a 0% strict accuracy under exact object preservation criteria. The system replaces the target object with different related terms such as “shrimp,” “crab,” and “fish,” reflecting both adaptive behaviour and variability in substitution patterns.

In particular, the first generated question, the object was substituted with "shrimp" (not in the dataset), reflecting the model's recognition of similarities between related terms. In question 4, "prawn" was replaced with "crab" (found in the dataset), showcasing the model's flexibility in adapting substitutions based on context or saliency. However, in questions 1 and 3, the substitution was inconsistent, with the first part referring to "shrimp" and the second to "fish," resulting in inaccuracies. Overall, these findings demonstrate that while the AQG system exhibits a degree of flexibility in handling unseen objects, its substitution mechanism is not fully stable across all cases.

Testing revealed that 5 out of 112 generated questions (approximately 4.5%) contained errors in object substitution, excluding the topic of Pressure in Liquids. While these substitutions did not entirely invalidate the questions, they introduced inaccuracies that could confuse students and reduce the educational value of the questions. Addressing these issues would improve the AQG system's effectiveness as a teaching tool. In a classroom context, such inconsistencies may distract students from the core physical concepts. Therefore, although the system effectively mitigates the cold-start problem in question generation, human-in-the-loop verification remains necessary to ensure pedagogical clarity and instructional reliability.

Experiment 2: Result Discussion

Tables 24 to 26 show the AQG model's ability to handle multiword targets in masked inputs using its stack of decoders. For instance, in questions 2 and 4 of Table 24, "steel pot" appears in three parts of the sentence. Similarly, "specific heat capacity" is treated as a multiword target. These results demonstrate the model's accuracy in multiword masking to generate coherent questions.

Table 24
Experimental results on using a multiword target in AQG for the topic Specific Heat Capacity

1.	What is the specific heat capacity of a **steel pot** with mass 0.6 kg, when it is heated by an electric heater with a power of 48 W for 93 s and the temperature increases from 19 oC to 42 oC?
2.	A **steel pot** with mass 0.42 kg is heated by an electric heater with a power of 44 W for 89 s. The temperature of the **steel pot** increases from 21 oC to 44 oC. Calculate the specific heat capacity of the **steel pot**?
3.	What is the amount of specific heat capacity required to raise the temperature of a **steel pot** with mass 0.34 kg from 21 oC to 45 oC, by using an electric heater with a power of 51 W for 80 s?
4.	A **steel pot** with mass 0.43 kg is heated by an electric heater with a power of 31 W for 84 s. The temperature of the **steel pot** increases from 19 oC to 45 oC. Calculate the specific heat capacity of the **steel pot**?

Table 25
Experimental results on using a multiword target in AQG for the topic Thin Lens Formula

1.	A plastic bottle of height 12 cm is positioned at 57 cm from a concave lens with a focal length of 17 cm. Determine the size of the image?
2.	A concave lens with a focal length of 18 cm is positioned at 55 cm from a plastic bottle with a height of 14 cm. Figure out the size of the image?
3.	A plastic bottle of height 8 cm is placed at 48 cm from a concave lens with a focal length of 17 cm. Find the size of the image?
4.	A plastic bottle of height 13 cm is positioned at 56 cm from a concave lens with a focal length of 18 cm. Find the size of the image?

Table 26
Experimental results on using a multiword target in AQG for the topic Current and Potential Difference

1.	If an electric charge of 536 C flows through a silver block for 4 minutes, what is the current flow?
2.	If an electric charge of 517 C flows through a silver block for 7 minutes, what is the current flow?
3.	What is the current flow if 506 C of electric charge passes through a silver block for 6 minutes?
4.	What is the current flow if 541 C of electric charge passes through a silver block for 5 minutes?

The model shows high accuracy and coherence in generating questions with both single-word and multiword targets. Despite the added complexity of multiword targets, such as maintaining context, the model handles them effectively. As shown in Tables 24 to 26, the model performs well with multiword targets, demonstrating similar accuracy and coherence as with single-word targets.

In Table 25, question 1 shows the model's ability to combine multiple multiword targets, such as "concave lens" and "size of image," within a coherent structure. The model effectively integrates scientific terminology and variable dependencies while maintaining coherence despite the complexity of multiword targets. Similarly, in Table 26, Question 1 demonstrates the model's versatility in incorporating multiword phrases like "current flow" and contextual objects like "silver block" to create relevant, meaningful questions.

The training dataset provides structured, specific input, such as "force, mass, luggage" or "electric charge, silver block, duration," with clearly defined variables and objects. This consistency helps the model accurately incorporate multiword targets into generated questions. For instance, in the "size of image" scenario, the input variables, namely, focal length, distance, and height, enable the model to generate a precise, coherent question. This well-defined context ensures the model understands the relationships between variables, leading to more accurate and coherent questions.

Despite the complexity of multiword targets, the model maintains high accuracy with structured input context. This is evident in topics like heat, light and optics, and electricity, where terms like "specific heat capacity," "size of image," and "current flow" are effectively used. These examples showcase the model's ability to integrate multiword targets across various physics topics, ensuring questions are relevant, contextually appropriate, and scientifically accurate.

AQG model's ability to handle complex terminology is reinforced by parallel success in STEM question generation. Gao (2023) achieved high performance by applying bidirectional training to the mT5 model on large, general datasets (Math23K, MAWPS; evaluated with 5-fold cross-validation). In contrast, our results demonstrate that standard fine-tuning of the T5-base model on a compact, highly-targeted dataset of 6,000 samples is equally effective for generating high-quality questions within a specific physics curriculum.

This contrast underscores the T5 architecture's versatility, proving that efficient, domain-focused fine-tuning can achieve robust results without the need for massive datasets or highly complex training schemes.

Evaluating Questions Generated by ChatGPT-4 and the AQG System

This section presents a comparative evaluation of the questions generated by ChatGPT-4 and the AQG system based on four criteria: relevance, correctness, usefulness, and variety. As summarised in Table 27, ChatGPT-4 excels in most criteria, especially relevance, correctness, and variety, while the AQG system slightly outperforms in usefulness, indicating its strength in generating practical questions.

Table 27
Performance comparison between ChatGPT-4 and the AQG system in question generation

Method	Relevance	Correctness	Usefulness	Variety
ChatGPT-4	4	3.89	3.65	4
AQG system	3.92	3.79	3.88	2.18

ChatGPT-4 generated more relevant and grammatically correct questions than the AQG system, scoring 4.00 for relevance and 3.89 for correctness. Its broad training and ability to follow precise prompts result in natural, well-structured questions. In contrast, the AQG system, though trained on physics-focused data, occasionally omits input values or uses awkward phrasing, leading to slightly lower scores (3.92 relevance, 3.79 correctness). However, the AQG system outperformed ChatGPT-4 in usefulness (3.88 vs. 3.65), as its questions align more closely with physics curricula. While ChatGPT-4 excels in generating varied formats (score 4.00), the AQG system is more limited (score 2.18), often producing fewer question types per topic. Overall, the AQG system offers better control and domain focus, whereas ChatGPT-4 provides greater linguistic fluency and diversity when guided properly. By using T5 as the primary generator and ChatGPT-4 as a benchmark, this study demonstrates that specialised fine-tuning is more effective than general prompting for producing the consistent, constrained, and curriculum-aligned assessments necessary for secondary-level STEM education.

Human Evaluation Results and Discussion

The human evaluation compared the AQG system with a template-based approach across four criteria: relevance, correctness, usefulness, and variety. While the template-based method scored slightly higher in relevance (4.00 vs. 3.92), correctness (3.80 vs. 3.79), and usefulness (3.91 vs. 3.88), the AQG system showed comparable performance, demonstrating its ability to produce grammatically accurate and generally relevant questions.

The template method's edge is likely due to its structured and predictable format, which closely aligns with expected question patterns. These results are shown in Table 28.

The AQG system's ability to automatically generate three distinct question types for topics such as latent heat and refraction contrasts with the inherently limited scope of template-based methods, which require manual creation of numerous templates for each scenario. This flexibility reflects the AQG system's ability to leverage learned patterns to produce diverse and adaptive questions. Although the AQG system had slightly lower average scores, it showed strong potential in creating effective and varied questions without requiring extensive manual template creation, making it a valuable tool for educators.

In addition to linguistic quality, the pedagogical value of the generated questions was examined in terms of cognitive learning outcomes. The questions target the Understand, Apply, and Analyse levels of Bloom's Taxonomy by requiring students to interpret contexts and perform multi-step reasoning. Specifically, they emphasise the Apply level, where learners must map verbal scenarios to physical laws and execute mathematical procedures. By demanding structured variable extraction from narratives, the system moves beyond rote substitution to support conceptual problem-solving. This ensures alignment with the Malaysian KSSM Physics syllabus and supports learning outcomes like conceptual transfer and structured physics reasoning.

Evidence of the AQG model's ability to move beyond original seed templates is reflected in its capacity to adapt phrasing based on the specific physics parameters provided in the input prompt. Experts observed consistent generation with linguistic variation, resulting in scores of 3.79 for Correctness and 3.88 for Usefulness. While some repetitive phrasing occurred, the questions remain pedagogically meaningful by requiring learners to interpret contexts and apply physical principles rather than relying on rote templates.

The domain-specific dataset, tailored to the Malaysian secondary Physics curriculum, ensures strong national alignment but may limit generalisability to other contexts without fine-tuning. This intentional design optimises accuracy and pedagogical relevance for the target curriculum. As noted by Deroy et al. (2024), such systems require a level of context-awareness that general-purpose models often lack. Consequently, adapting the system to other regions would require retraining on structured datasets to maintain its pedagogical and linguistic effectiveness.

Table 28
Human evaluation results

Method	Relevance	Correctness	Usefulness
AQG system	3.92	3.79	3.88
Template based	4.00	3.80	3.91

The inter-rater reliability of the T5-based AQG system was validated by analysing 180 evaluation units (representing 60 questions across three qualitative metrics: Relevance, Correctness, and Usefulness). With three expert Physics teachers providing 540 total ratings, the analysis yielded an observed agreement of 89%. Using Randolph's Free-Marginal Kappa to account for the high concentration of positive scores, the resulting coefficient of 0.85 indicates almost perfect agreement. This statistical evidence confirms that the high mean scores for question quality reflect strong expert consensus rather than chance agreement among raters.

Automatic Evaluation Results and Discussion

The ROUGE-L scores for all three evaluators exceed 0.82, closely aligning and indicating that the AQG system consistently generates questions similar to the reference ones (Table 29). The small score range (0.82-0.85) reflects strong evaluator agreement, highlighting the reliability and quality of the generated questions.

Table 29
ROUGE-L scores for AQG system questions based on evaluators' outputs compared to golden references

Generated by Evaluator	ROUGE-L Score
AQG System (Evaluator 1)	0.85
AQG System (Evaluator 2)	0.82
AQG System (Evaluator 3)	0.84

The high ROUGE-L scores show that the AQG system generates accurate questions that closely match the ideal phrasing. The slight difference between Evaluator 2 and the others may reflect minor variations in evaluation criteria. However, the consistently high scores across evaluators demonstrate the model's robust performance and alignment with educational or assessment expectations.

The ROUGE-L results demonstrate that the AQG system performs exceptionally well, producing questions closely aligned with the reference questions. Minor variations among evaluators highlight potential areas for refinement in ensuring consistency and alignment with reference questions. Overall, the high scores confirm the model's effectiveness in generating quality educational questions, achieving the primary objective of AQG tasks. However, ROUGE-L primarily measures surface-level textual overlap and may not fully capture conceptual correctness or pedagogical relevance, particularly in a template-driven context. To address this limitation, human evaluation was conducted by three experienced Physics teachers, assessing relevance, correctness, and usefulness value. Thus, ROUGE-L serves as a complementary metric, supported by qualitative evaluation to ensure overall question quality.

CONCLUSION

This study developed and evaluated a web-based Automatic Question Generation (AQG) system for Malaysian secondary-school Physics, leveraging the T5 model to generate contextually appropriate, curriculum-aligned questions within STEM education. By automating the generation of diverse and pedagogically sound assessment items, the system addresses a critical need for efficient educational tools. High ratings from human evaluators on relevance, correctness, and usefulness indicate the system's potential to enhance physics instruction and learning.

Despite these promising results, several limitations should be acknowledged. First, the evaluation was conducted using single-run generation, which does not capture variability in model outputs. Second, the study relied on a limited number of expert reviewers, and while pedagogical validity was established through expert judgement, empirical validation with students was not conducted. Third, although the system demonstrates linguistic variation, some generated questions remain partially constrained by template-like structures, indicating scope for improving contextual and semantic diversity. Finally, the model was trained on a domain-specific dataset aligned with the Malaysian secondary school curriculum, which may limit generalisability to other contexts without further adaptation.

Future work will address these limitations by incorporating multiple sampling runs and statistical testing to improve robustness, as well as expanding evaluation through classroom-based pilot studies to assess learning outcomes and student engagement. Additionally, benchmarking against a wider range of baseline and foundational models, such as Google Gemini and Claude, will provide a more comprehensive comparison across model architectures. Further enhancements may include enabling teacher-controlled difficulty levels, supporting multimodal inputs (e.g., images and text), and improving model performance through domain-specific pre-training and advanced data augmentation techniques. The integration of semantic-based evaluation metrics, such as BERTScore, is also proposed to complement lexical metrics and provide a more comprehensive assessment of question quality.

These developments would significantly enhance the system's versatility and context-awareness, yielding greater benefits for educators and students. Ultimately, this research establishes a strong foundation for the development of intelligent, curriculum-aligned AQG systems in STEM education.

ACKNOWLEDGEMENT

My tokens of appreciation go to the Centre for Graduate Studies, Universiti Malaysia Sarawak, for the advice and support given during the publication of this paper. There are no funding resources provided for the conduct of this research.

REFERENCES

Amidei, J., Piwek, P., & Willis, A. (2018). Evaluation methodologies in automatic question generation 2013-2018. In E. Krahmer, A. Gatt, & M. Goudbeek (Eds.), *Proceedings of the 11th International Conference on Natural Language Generation* (pp. 307-317). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6537

Chew, S. L., & Cerbin, W. J. (2021). The cognitive challenges of effective teaching. *The Journal of Economic Education, 52*(1), 17-40. https://doi.org/10.1080/00220485.2020.1845266

Choy, C. S., Chuan, K. K., Beng, O. H., Md Mustafa, M. K. A., & Ragavan, R. (2019). *Physics form 4*. Sasbadi Sdn. Bhd.

Chuan, K. K., Choy, C. S., Bongkek, N. R., Kasron, J., Md Mustafa, M. K. A., & Chakrabarty, P. K. (2020). *Physics form 5*. Penerbit Bestari Sdn. Bhd.

Cornejo, O., Briola, D., Micucci, D., Ginelli, D., Mariani, L., Santos Parrilla, A., & Juristo, N. (2024). A family of experiments about how developers perceive delayed system response time. *Software Quality Journal, 32*, 567-605. https://doi.org/10.1007/s11219-024-09660-w

Deroy, A., Maity, S., & Sarkar, S. (2024). *MIRROR: A novel approach for the automated evaluation of open-ended question generation* [Preprint]. arXiv.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Erinosho, S. Y. (2013). How do students perceive the difficulty of physics in secondary school? An exploratory study in Nigeria. *International Journal for Cross-Disciplinary Subjects in Education, 3*(Special Issue 3), 1510-1515. https://doi.org/10.20533/ijcdse.2042.6364.2013.0212

Fan, A., Lewis, M., & Dauphin, Y. (2018). Hierarchical neural story generation. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers* (pp. 889-898). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1082

Goldbach, I. R., & Hamza-Lup, F. G. (2017). Survey on e-learning implementation in Eastern Europe: Spotlight on Romania. In L. A. Ludovico & A. M. F. Yousef (Eds.), *Proceedings of the International Conference on Mobile, Hybrid, and Online Learning* (pp. 5-12). IARIA XPS Press.

Gao, K. (2023). Bidirectional training for generating math word problems using a pre-trained model and prompt. In *2023 20th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICCWAMTIP60502.2023.10387053

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text degeneration. In *Proceedings of the Eighth International Conference on Learning Representations*.

Jauhiainen, J. S., & Garagorry Guerra, A. (2024). *Evaluating students' open-ended written responses with LLMs: Using the RAG framework for GPT-3.5, GPT-4, Claude-3, and Mistral-Large* [Preprint]. arXiv. https://doi.org/10.54364/AAIML.2024.44177

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education, 30*(1), 121-204. https://doi.org/10.1007/s40593-019-00186-y

Kwan, W. C., Wang, H. R., Wang, H. M., & Wong, K. F. (2023). A survey on recent advances and challenges in reinforcement learning methods for task-oriented dialogue policy learning. *Machine Intelligence Research, 20*(3), 318-334. https://doi.org/10.1007/s11633-022-1347-y

Le, N.-T., & Pinkwart, N. (2015). Evaluation of a question generation approach using semantic web for supporting argumentation. *Research and Practice in Technology Enhanced Learning, 10*, Article 3. https://doi.org/10.1007/s41039-015-0003-3

Maity, S., Deroy, A., & Sarkar, S. (2024). *Exploring the capabilities of prompted large language models in educational and assessment applications* [Preprint]. arXiv.

Mishra, S. K., Goel, P., Sharma, A., Jagannatha, A., Jacobs, D. W., & Daumé, H., III. (2020). *Towards automatic generation of questions from long answers* [Preprint]. arXiv.

Molina, I. L., Švábenský, V., Minematsu, T., Chen, L., Okubo, F., & Shimada, A. (2024). Comparison of large language models for generating contextually relevant questions. In I. A. Chounta, P. D. Muñoz-Merino, V. Dimitrova, T. D. Leo, & I. M. Román-González (Eds.), *Towards hybrid human-AI learning technologies: Artificial intelligence supporting human intelligence: 19th European Conference on Technology Enhanced Learning, EC-TEL 2024* (pp. 237-250). Springer.

Parker, M. J., Anderson, C., Stone, C., & Oh, Y. (2024). A large language model approach to educational survey feedback analysis. *International Journal of Artificial Intelligence in Education, 35*(2), 444-481. https://doi.org/10.1007/s40593-024-00414-0

Patil, S., Chavan, L., Mukane, J., Vora, D., & Chitre, V. (2022). State-of-the-art approach to e-learning with cutting-edge NLP transformers: Implementing text summarisation, question and distractor generation, and question answering. *International Journal of Advanced Computer Science and Applications, 13*(1), 445-453. https://doi.org/10.14569/IJACSA.2022.0130155

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1-67.

Roth, R. E. (2013). Interactive maps: What we know and what we need to know. *Journal of Spatial Information Science, 6*, 59-115. https://doi.org/10.5311/JOSIS.2013.6.105

Rus, V., Cai, Z., & Graesser, A. (2008). Question generation: Example of a multi-year evaluation campaign. In V. Rus & A. Graesser (Eds.), *Online proceedings of the First Question Generation Workshop*. National Science Foundation.

Saleh, S. (2014). Malaysian students' motivation towards physics learning. *European Journal of Science and Mathematics Education, 2*(4), 223-232. https://doi.org/10.30935/scimath/9414

Sharma, S., El Asri, L., Schulz, H., & Zumer, J. (2017). *Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation* [Preprint]. arXiv.

Shazeer, N., & Stern, M. (2018). Adafactor: Adaptive learning rates with sublinear memory cost. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning* (*Proceedings of Machine Learning Research*, Vol. 80, pp. 4596-4604). PMLR.

Shi, H., & Wolff, P. (2021). What transformers might know about the physical world: T5 and the origins of knowledge. *Proceedings of the Annual Meeting of the Cognitive Science Society, 43*, 2218-2223.

Tucker, A. (2020). *Text generation with HuggingFace GPT-2* [Kaggle notebook]. Kaggle. https://www.kaggle.com/code/tuckerarrants/text-generation-with-huggingface-gpt2

von Platen, P. (2020, March 1). *How to generate text: Using different decoding methods for language generation with transformers*. Hugging Face. https://huggingface.co/blog/how-to-generate

Wang, W., & Reani, M. (2017). The rise of mobile computing for group decision support systems: A comparative evaluation of mobile and desktop. *International Journal of Human-Computer Studies, 104*, 16-35. https://doi.org/10.1016/j.ijhcs.2017.02.008

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, & Y. Zhou (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 483-498). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.41

Yao, X., Bouma, G., & Zhang, Y. (2012). Semantics-based question generation and implementation. *Dialogue & Discourse, 3*(2), 11-42. https://doi.org/10.5087/dad.2012.202