The object of this study is the assessment process in digital learning environments, where multimedia educational materials such as textual documents and instructional images and require efficient methods for automatic question generation. The main problem investigated is the difficulty in generating high-quality, relevant questions from various multimedia learning materials (text and images). A multimedia question-generation system is proposed that integrates the text-to-text transfer transformer (T5) model with classification based on Bloom's taxonomy. A preprocessing workflow has been created that extracts and combines textual representations from text and images using optical character recognition (OCR) for data tokenization and performs named entity recognition (NER). The question generator application can generate various question types, including multiple-choice, short-answer, and essay questions. These questions are classified according to Bloom's taxonomy. The generated questions were evaluated using bilingual evaluation understudy (BLEU) and recall-oriented understudy for gisting evaluation (ROUGE). Experimental results demonstrated strong performance, with average scores of BLEU-1 = 0.86, BLEU-2 = 0.79, ROUGE-1 = 0.88, and ROUGE-2 = 0.81. Evaluation scores indicate that the multimodal quiz generator application produces high-quality and contextually relevant questions. Evaluation scores show similarities between reference questions and generated questions, with structured questions receiving higher scores than essay questions. The system allows its use in education and intelligent tutoring systems to produce measurable, efficient assessments. The method proposed in this study is limited to multimedia input consisting of text and images

Keywords: automated question generation, multimedia, T5 transformer, bloom taxonomy, BLEU, ROUGE

UDC 004.55
DOI: 10.15587/1729-4061.2026.355867

DEVELOPMENT OF AN AUTOMATED QUESTION GENERATION SYSTEM FROM MULTIMEDIA CONTENT USING TEXT-TO-TEXT TRANSFER TRANSFORMER WITH BLOOM TAXONOMY CLASSIFICATION

Marvin Chandra Wijaya
Corresponding Author
Philosophy Doctor of Information Communication Technology, Associate Professor*
E-mail: marvin.cw@eng.maranatha.edu
ORCID: https://orcid.org/0000-0001-5920-4348
Markus Tanubrata
Senior Lecturer*
E-mail: marvin.cw@eng.maranatha.edu
ORCID: https://orcid.org/0009-0006-0129-6749
*Department of Computer Engineering
Maranatha Christian University
Suria Sumantri str., 65, Bandung, Indonesia, 40164

Received 23.02.2026
Received in revised form 30.04.2026
Accepted date 11.05.2026
Published date 30.06.2026

How to Cite: *Wijaya, M. C., Tanubrata, M. (2026). Development of an automated question generation system from multimedia content using text-to-text transfer transformer with bloom taxonomy classification. Eastern-European Journal of Enterprise Technologies, 3 (2 (141)), 61–70. https://doi.org/10.15587/1729-4061.2026.355867*

1. Introduction

Digital learning platforms have changed the way learners worldwide access and engage with educational content. Increased access and engagement have made academic assessment significantly more burdensome than before, due to higher workloads from the need to create high-quality examinations. Constructing effective assessments requires significant effort to produce accurate, unambiguous, and appropriately challenging questions, which can only be done by educators with a high degree of subject-matter knowledge and teaching experience [1].

Automated question generation (AQG) has emerged as a possible solution to this issue; AQG systems generate questions from source content by applying pre-defined rules across content types. Early-generation systems used rule-based transformation techniques and shallow syntactic parsing to produce non-diverse items that are difficult to generalize across subject areas [2]. With the introduction of large-scale pretrained language models, it has been shown that the fluency and contextual accuracy of machine-generated questions can be significantly improved. However, most AQG systems remain limited by treating only plain text as valid input, while ignoring the rich semantic information in images and video, which together make up much of the instructional material used in educational contexts [3].

Another limitation of AQG (automatic question generation) study at the present time is the inability to assess the cognitive quality of generated questions. Assessment instruments are products of an evaluation process, and their effectiveness can be enhanced by providing cognitive complexity across item levels. The revised Bloom taxonomy identifies six hierarchical levels of cognitive demand, ranging from remembering to creating [4]. It establishes the traditional way of aligning assessment with learning objectives in both formal education and training. The types of questions generated

by AQG systems do not align with any cognitive framework, and as a result, their outputs are not well-balanced cognitively. Henceforth, these outputs will have limited value for the development of structured assessments [5].

Therefore, studies devoted to the development of automated question-generation systems from multimedia content are highly relevant to science. This is due to the increasing use of heterogeneous digital learning materials in modern educational environments. Multimedia-based data processing is a field of study aimed at improving the quality, scalability, and pedagogical value of automated question-generation systems.

2. Literature review and problem statement

Paper [6] presents preliminary study on automatic question generation based on rule-based transformation techniques. The system is implemented on declarative sentences. In this study, questions can be generated by rearranging syntactic structures such as subject-verb-object and applying them to predefined templates. However, it suffers from a lack of flexibility and limited generalizability across domains. The system is unable to handle the wide variety of sentences that need to be extracted. The lack of flexibility stems from its reliance on linguistic processing. The generated question forms in this study are limited by the similarity of the input provided.

Paper [7] discusses the use of a sequence-to-sequence neural model in automated question generation. This study focused on improving the relevance between questions. A weakness of this system was the interconnectedness between questions. The initial model was unable to capture the relevance between questions. To address this, a more contextual coding mechanism was needed.

Paper [8] focused on improving neural question generation by combining copying mechanisms and paragraph-level context. Existing paragraphs enhanced the relevance of questions by forming interrelated sequences. However, the relevance between paragraphs was not yet able to be interconnected. Paper [9] presented the use of a large-scale pre-trained transformer model for automated question generation. Training in this study focused on long sentences. This processing of long sentences weakened the accuracy of sentence extraction and decreased translation accuracy. Therefore, to solve this problem, it is necessary to have a translation and sentence-extraction method specifically designed for long sentences.

The papers [10] present the use of bidirectional encoder representations from transformers (BERT) and its variants as encoding-based architectures for natural language processing tasks. These models are shown to achieve strong performance in text understanding through contextual representations. However, unresolved issues included their inability to generate text and the high computational cost of processing [11]. The reason may be that BERT is an encoder-only architecture, limiting its applicability to generation tasks and increasing processing complexity. To overcome these difficulties, one can use encoder–decoder transformer models for efficient text generation.

Paper [12] highlights the importance of efficient multimedia processing, especially when handling large data sets, such as images, in learning environments. It shows that modern education systems use multiple formats, including text and visual materials, to convey information. Artificial intelligence methods can be used to extract text from multimedia content [13]. Artificial intelligence (AI) techniques can transform visual information into textual representations, but with various limitations [14]. Limitations in integrating interrelated text and the presence of multiple media types can degrade the quality of generated questions. The complexity of image media is a fundamental drawback of this method. The extraction quality of complex images requires improved OCR methods.

Paper [15] discusses advances in OCR technology for extracting text from images as a solution to complex media extraction. OCR methods can achieve high accuracy and effectiveness. These methods can focus on generating high-quality questions [16]. However, there are still unresolved issues in their practical application. The differences in the purpose of question generation and the generated questions do not match perfectly. To address this issue, integration of the OCR system with the user interface of the system is necessary.

In paper [17], the focus is on using Bloom's taxonomy as a framework for classifying cognitive levels for each generated question. This paper shows that effective evaluation is crucial and necessary. Evaluation must be at every processing level. The evaluation can use AI so that the evaluation process can run automatically and adaptively [18]. Based on the results, problems remain in applying this method to cognitive-level classification. Therefore, it is necessary to integrate automatic classification models to support the generation of higher-quality questions.

The paper [19] presents the use of machine learning classifiers, such as support vector machines and logistic regression, for text classification tasks. These methods achieve moderate accuracy using features such as bag-of-words and TF-IDF. However, there were unresolved issues related to limited performance across diverse datasets. The reason for this may be the inability of traditional models to capture deep contextual information. A way to overcome these difficulties is to use fine-tuned transformer-based models, such as BERT, which provide higher classification accuracy.

The above review identifies several interrelated gaps that motivate this study. First, most existing automated question-generation systems are limited to processing textual input, which is typically in contrast to e-learning materials presented in visual formats, such as images and slide presentations. Second, there is a lack of integrated preprocessing mechanisms capable of extracting and transforming textual information from heterogeneous multimedia sources into unified representations suitable for neural-based question generation. Third, although Bloom's taxonomy and automated question generation have been extensively studied, they are generally treated as separate components without a unified framework, thereby limiting their effectiveness in supporting structured, pedagogically meaningful assessments.

3. The aim and objectives of the study

The aim of this study is to develop an automated question generation system from multimedia educational content using a text-to-text transfer transformer (T5) integrated with Bloom's taxonomy classification. This will enable the automated generation of high-quality, structured exam questions from text and image media. High-quality questions can contribute to and improve the education system.

The following objectives are required:

– to propose an automation question generator structure.

– to design a multimedia preprocessing workflow by extracting media sources;

– to build a text-to-text transfer transformer model on the dataset to generate exam questions relevant to the multimedia content;

– to implement a Bloom's taxonomy classifier automatically on each generated question;

– to evaluate the generation results using BLEU and ROUGE metrics.

4. Materials and methods

4. 1. The object and hypothesis of the study

The object of this study is the assessment process in digital learning environments, where multimedia educational materials such as textual documents and instructional images and require efficient methods for automatic question generation. The primary hypothesis of this study is that integrating multimedia data preprocessing with a transformer-based model and Bloom's taxonomy classification can improve the quality of generated questions.

This study assumes that the data extracted from text and images is sufficiently accurate to generate high-quality questions. The existing data must be structured in such a way that it can be processed properly in subsequent stages. Therefore, it is necessary to limit the scope of this study. These simplifications include the use of a predetermined dataset, appropriate media source data, and good OCR accuracy.

4. 2. Proposed methodological framework

This section describes the methodological framework used to develop the automated question generation system. This framework consists of four sequential stages: system design, multimedia preprocessing, T5-based question generation, and question evaluation. The first stage defines the overall system workflow and the interactions between system components. The second stage transforms heterogeneous multimedia input into a unified textual representation. The third stage generates candidate questions using the refined T5 model. The final stage evaluates the generated questions using quantitative metrics.

4. 3. Preprocessing

The preprocessing stage transforms multimedia input (text and images) into a structured textual representation. This stage integrates the text and OCR extraction results, which will be processed using NLP. The NLP process consists of tokenization, named entity recognition (NER), and syntactic parsing. The multimedia input sources, both text and image extractions, are combined into a textual form. The integration steps are processed using the following formula

$$T = T_{text} \cup OCR(I), \tag{1}$$

T_{text} – the original text input, and $OCR(I)$ – the text extracted from image I. The result T is used as input for subsequent preprocessing steps. The next process is tokenization to divide the input text T into a series of tokens as follows

$$T = \{w_1, w_2, w_3, \ldots, w_n\}. \tag{2}$$

In transformer-based models like T5, subword tokenization is performed using the Byte-Pair Encoding (BPE) algorithm. This method is useful for handling difficult-to-understand words. The next step is named entity recognition (NER), which identifies and classifies the generated tokens. Given tokens $\{w_1, w_2, w_3, \ldots, w_n\}$. The task of NER is to assign y_i to each token

$$y_i \in \{PER, LOC, ORG, DATE, O\}. \tag{3}$$

The formula can be modeled as a sequence labeling problem

$$P(Y \mid X) = \prod_{i=1}^{n} P(y_i \mid x), \tag{4}$$

X – the input token sequence, and Y – the corresponding label sequence. In modern implementations, models such as Bidirectional Long Short-Term Memory-Conditional Random Field (BiLSTM-CRF) or transformer-based architectures are used. The conditional random field (CRF) layer globally optimizes the label sequence

$$y^* = \arg\max_{y} \sum_{i=1}^{n} \left(s_i(y_i) + t(y_{i-1}, y_i)\right), \tag{5}$$

s_i – the emission score and t – the transition score. Named entity recognition (NER) enriches the semantic content of the input, enabling the system to focus on key concepts for question generation. Syntactic parsing analyzes the grammatical structure of sentences to determine relationships between tokens. One commonly used approach is dependency parsing, which represents a sentence as a tree

$$G = (V, E), \tag{6}$$

V – the set of tokens, and E represents dependency relations between them. The objective is to find the optimal tree structure

$$T^* = \arg\max_{T \in \tau(x)} Score(T). \tag{7}$$

Parsing algorithms analyze the syntactic structure of an input based on specific grammatical rules. The aim of these algorithms is to transform raw data into a structured form. The results of these parsing algorithms provide basic information used to generate correct queries.

4. 4. Natural language processing model: fine-tuned text-to-text transfer transformer

The study uses a text-to-text transfer transformer (T5) as its foundational automated question generation NLP model (T5). T5 is a transformer-based architecture that creates a unified text-to-text framework for reformulating all types of natural language processing (NLP) tasks into text sequence representations for both input and output, allowing it to perform these tasks flexibly and consistently (Fig. 1). This approach simplifies the overall system design and allows the model to generate different types of questions from the same input flexibly.

The T5 architecture is designed using an encoder-decoder architecture. The encoder processes an input sequence and converts it into a form that captures its semantic meaning, context, etc. The encoder captures the context of each token

in the input sequence as a "hidden representation". The hidden representation provides the basis for decoding the sequence of generating a series of outputs through the decoder. For example, given an input sequence

$$H = Encoder\left(\{x_1, x_2, \ldots, x_n\}\right). \quad (8)$$

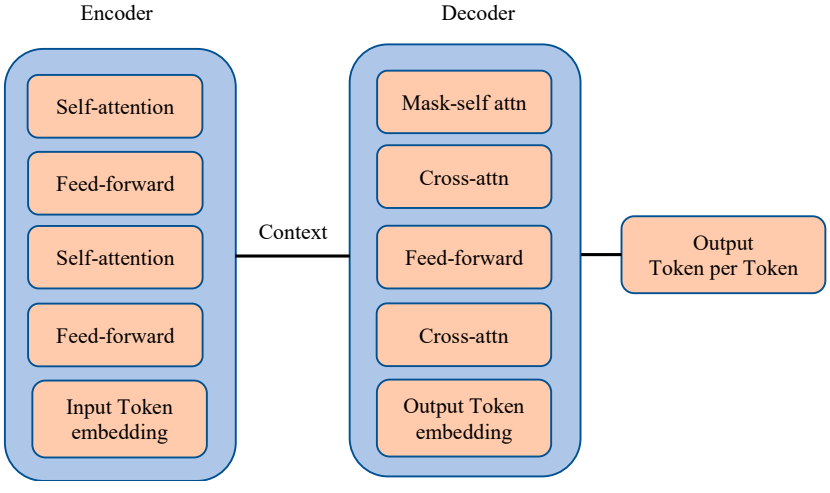


Fig. 1. Encoder-decoder structure of transformer T5

Then, through decodable generation, the decoder would produce an output sequence (Y) for each hidden representation and all previously produced outputs (Y). The generation of an output (Y) is serial in nature: each word produced (Y) depends upon its predecessors and the context of the input sequence.

In this study, the question generation task is formulated as a text transformation problem. The input is structured using a prompt-based format

$$X = \text{"}generate\ question\text{: "} + context. \quad (9)$$

Context is obtained from the preprocessing stage, including text and OCR results from images. To support different cognitive levels, the input is further extended using Bloom's taxonomy labels

$$X = \text{"}generate\ C_k\ question\text{: "} + context, \quad (10)$$

where $C_k \in \{C_1, C_2, \ldots, C_6\}$. This allows the model to generate questions that correspond to specific levels of difficulty, from simple recall (C_1) to higher-order thinking (C_6).

Through a fine-tuning process, the T5 model, which has been previously trained, can now be adapted for a particular purpose – generating questions. Fine-tuning consists of training the model on a dataset containing both inputs and outputs. The following describes the process of fine-tuning a T5 Model:

1) start with a pre-trained T5;

2) prepare the data by combining the context, bloom level, and task implementation into a single input sequence;

3) feed the input into the encoder to get contextual representations;

4) decode the output sequence token by token;

5) compare the generated output to the reference question to calculate the loss;

6) update model parameters with gradient descent;

7) repeat the step 3 to 6 until convergence occurs.

4. 5. Question generator with Bloom taxonomy

Question generator is a module that automatically generates assessment-style questions from the context of the input text, using a fine-tuned T5 model. The question-generation system produces three types of questions: multiple choice question (MCQ), fill-in-the-blank, and essay. These three types of questions are generated using a controlled text-generation approach, with the questions being constructed according to predetermined templates and adhering to strict semantic constraints. Given an input context X and a target Bloom level C_k, the question generation process can be formulated as

$$Q = f(X, C_k), \quad (11)$$

where Q represents the generated question and *fff* denotes the trained T5 model. The model produces a sequence of tokens

$$Q = \{q_1, q_2, \ldots, q_n\}. \quad (12)$$

To support different question types, the input is augmented with a task-specific prefix with $task\_type \in$ {"*MCQ*", "*Fill-Blank*", "*Essay*"}

$$X' = task\_type + \text{" "} + context. \quad (13)$$

MCQs consist of a question stem, one correct answer, and several distractors. The generation process involves two stages:

1. Question stem generation.

The model generates a question based on the input context

$$Q\_stem = f(X', C_k). \quad (14)$$

2. Answer and distractor generation.

Let the correct answer be $a+$, extracted from key entities or phrases in the context. Distractors $D = d, d_2, d_3$ are generated based on semantic similarity

$$D = g\,(a^+), \quad (15)$$

where function g generates distractors that are contextually relevant but incorrect. In practice, distractors can be generated using similarity-based ranking

$$\text{score}(d_i) = \text{sim}(d_i, a^+). \quad (16)$$

The final MCQ structure is:

– Question: Q_stem;

– Options: $\{a^+, d_1, d_2, d_3\}$.

Fill-in-the-blank questions are generated by masking important tokens in a sentence. A key token w_k is selected based on importance then replaced with a blank

$$S' = \{w_1, \ldots, \_\_\_, \ldots, w_n\}. \quad (17)$$

The removed token becomes the correct $answer = w\_k$. The selection of w_k can be guided by $w\_k = argmax\ importance(w_i)$, where $importance(w_i)$ is computed based on NER labels or term frequency.

Essay questions are designed to assess higher-order thinking skills (C3–C6 in Bloom's Taxonomy). The model generates open-ended questions directly

$$Q\_essay = f(X', C_k),\ C_k \geq C3. \quad (18)$$

These questions involve the words "explain," "analyze," or "evaluate," to ask students for explanations on the answers to questions generated by the system.

Algorithm. Question generation:

1. Input: context X, Bloom level C_k, task type.
2. Construct input: $X' = tasktype + context$.
3. Generate question: $Q = f(X', C_k)$.
4. If task = MCQ:
– extract correct answer a^+;
– generate distractors D.
5. Else if task = Fill-in-the-Blank:
– select key token w_k;
– replace with blank.
6. Else if task = Essay:
– output generated question.
7. Return final question.

4. 6. Question quality evaluation (bilingual evaluation understudy and recall-oriented understudy for gisting evaluation)

Evaluation of generated questions using the BLEU and ROUGE formulas. The scores of generated questions (Q_{gen}) and reference questions (Q_{ref}) are compared. The BLEU and ROUGE formulas are as follows:

$$Score_{avg} = \frac{1}{N}\sum_{i=1}^{N} Metric\left(Q_{gen}^{i}, Q_{ref}^{i}\right), \quad (19)$$

$$ROUGE - N = \frac{\sum_{gram \in Q_{ref}} Count_{match}\left(gram\right)}{\sum_{gram \in Q_{ref}} Count\left(gram\right)}. \quad (20)$$

N – the total number of predetermined samples. The ROUGE metric score estimates the number of questions generated based on the existing sequence. The BLEU metric score measures the number of exact matches using the generated question data. However, because the ROUGE metric emphasizes the recall process, these two measures together provide a balanced evaluation [20].

5. Research results of the question generation system

5. 1. Proposed automation question generator structure

The proposed automatic question generation (AQG) consists of several stages, as shown in Fig. 2.

The AQG system receives input sources in the form of text and images (multimedia) for further processing using various natural language processing (NLP) algorithms. These processes consist of tokenization, named entity recognition (NER), and syntactic parsing. These NLP techniques are used to clean, structure, and identify patterns in multimedia input sources. The next step is to process the input sources using an NLP model (T5) to obtain candidate questions for each source. All candidate questions generated by the NLP model will be classified into Bloom's taxonomy categories. These candidate questions will be categorized according to Bloom's Taxonomy levels C1–C6. Next, the candidate questions will be selected to ensure an even distribution of C1, C2, C3, C4, C5, and C6. After the questions are selected, their quality will be evaluated using automatic metrics (BLEU and ROUGE).

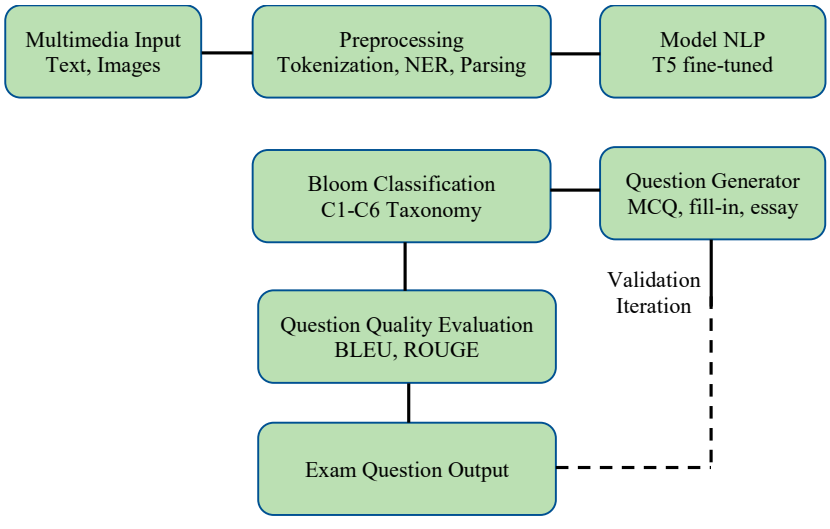


Fig. 2. The proposed automation question generator

Implementation of the proposed automated question generation system using several software and libraries. The proposed system is implemented using the Python programming language. The NLP and deep learning modules are developed using the PyTorch framework (version 2.0). Tokenization and syntactic processing use the Natural Language Toolkit (NLTK version 3.8). The optical character recognition (OCR) process uses Tesseract OCR (version 5.3). The BLEU and ROUGE metric evaluation processes utilize the NLTK evaluation package and the rouge-score library.

5. 2. Multimedia preprocessing results

The data sources in the experiment used multimedia data in the form of text and images. The image data was processed using optical character recognition (OCR) to generate text content for the images. The two data sets were combined to create a single data source. These data sources were interconnected to maintain the integrity of the data source. Table 1 shows an example of input from text and image data sources. One text and three images form a unified meaning within a single learning resource.

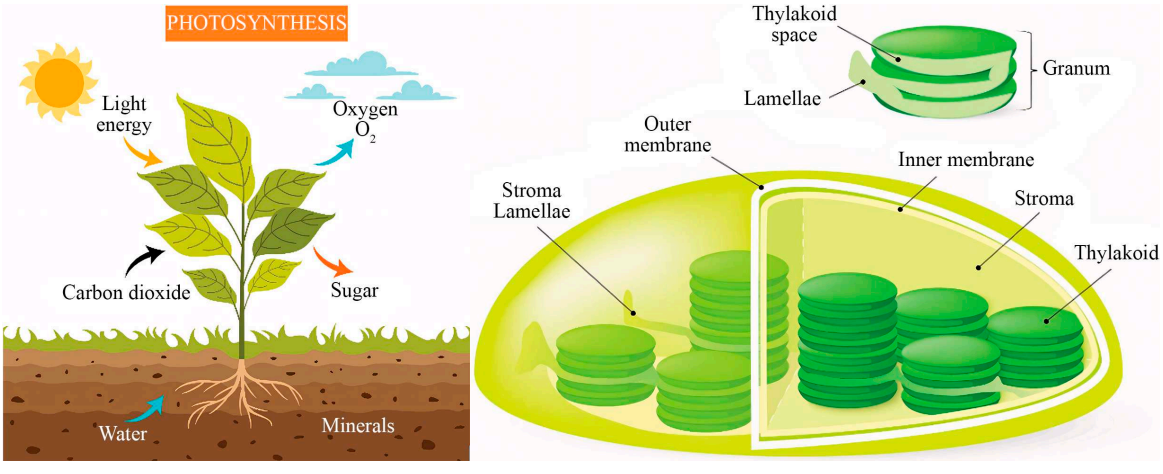
By combining the textual and OCR language input from all images, the system produces its most complete source of information. Together, these two forms of representing the same information help create a richer context for the NLP model and enable more accurate, appropriate question generation. Fig. 3, *a*, *b*, 4 show examples of diagrams of learning material sources in the form of photosynthesis.

The combined multimedia input undergoes preprocessing, converting it into a structured format suitable for processing by the natural language processing (NLP) model. Among other things, this process involves breaking each word or group of words into individual tokens (tokenization), identifying named entities (people and places) in the text (named entity recognition [NER]), and analyzing how words relate to each other grammatically (syntactic parsing).

Table 1

Sample of multimedia input

No.	Input type	Content description	Extracted text (OCR result)
1	Text	Photosynthesis is the process by which plants convert sunlight into chemical energy. This process occurs in the chloroplasts and is essential for plant growth	–
2	Image 1	Diagram illustrating the photosynthesis process including sunlight, carbon dioxide, water, glucose, and oxygen (Fig. 3, *a*)	Sunlight + Carbon Dioxide + Water → Glucose + Oxygen
3	Image 2	Diagram of chloroplast structure showing internal components such as thylakoids and stroma (Fig. 3, *b*)	Chloroplasts contain thylakoids and stroma, where photosynthesis occurs
4	Image 3	Diagram presenting the chemical equation of photosynthesis (Fig. 4)	$6CO_2 + 6H_2O$ + Light Energy → $C_6H_{12}O_6 + 6O_2$



a *b*

Fig. 3. Media diagram: *a* – image 1; *b* – image 2

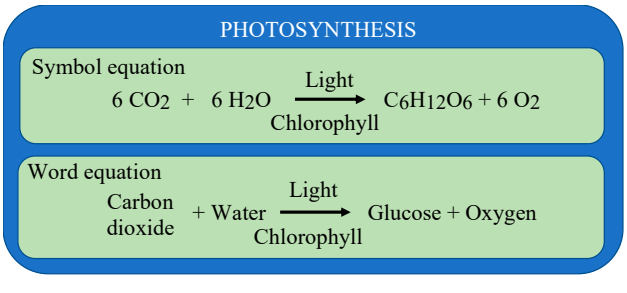


Fig. 4. Media diagram Image 3

The first step in this process is to create a document containing the original text and any OCR output from Table 1, and to normalize it into a single continuous text that can then be processed by the NLP engine. The following is an example of that output: "Photosynthesis is the process by which plants convert sunlight into chemical energy. This process occurs in the chloroplasts and is essential for plant growth. Sunlight, carbon dioxide, and water produce glucose and oxygen. Chloroplasts contain thylakoids and stroma. $6CO_2 + 6H_2O$ + light energy produce $C_6H_{12}O_6 + 6O_2$."

The combined text is divided into a sequence of tokens

[Photosynthesis, is, the, process, by, which, plants, convert, sunlight, into, chemical, energy, ...].

Important entities are identified and classified into predefined categories. The results are summarized in Table 2. The NER process highlights key concepts that are important for generating meaningful questions.

Table 2

Named entity recognition (NER) results

Token	Entity type
Photosynthesis	PROCESS
Plants	ORGANISM
Sunlight	ENERGY
Carbon dioxide	CHEMICAL
Water	CHEMICAL
Glucose	CHEMICAL
Oxygen	CHEMICAL
Chloroplast	ORGANELLE
Thylakoids	ORGANELLE
Stroma	ORGANELLE

Syntactic parsing identifies grammatical relationships within sentences. A simplified example is shown below:

– subject: photosynthesis;
– verb: is;
– object: process.

Another example:

– subject: plants;
– verb: convert;
– object: sunlight into chemical energy.

These structures help the system understand sentence relationships and improve the quality of generated questions.

5. 3. Text-to-text transfer transformer question generation results

Below presents the output of the refined T5 model for generating candidate questions from multimedia input. The aim

of this stage is to transform the structured textual representations obtained from the preprocessing stage into candidate natural language questions. The input to the T5 model consists of text that has been processed to generate contextually relevant questions at varying levels of cognitive complexity. Examples of input prompts are shown below:

– C1 (remembering):

"generate C1 question: photosynthesis is the process by which plants convert sunlight into chemical energy";

– C2 (understanding):

"generate C2 question: photosynthesis uses sunlight, carbon dioxide, and water to produce glucose and oxygen";

– C4 (analyzing):

"generate C4 question: chloroplast contains thylakoids and stroma where photosynthesis occurs".

The model generates question candidates in natural language form. Examples of generated outputs are presented in Table 3.

Table 3

Output examples

No.	Input prompt	Generated output
1	Generate C1 question: Photosynthesis is the process by which plants convert sunlight into chemical energy	What is photosynthesis?
2	Generate C2 question: Photosynthesis uses sunlight, carbon dioxide, and water to produce glucose and oxygen	How does photosynthesis produce glucose and oxygen?
3	Generate C4 question: Chloroplast contains thylakoids and stroma where photosynthesis occurs	Why are thylakoids and stroma important in photosynthesis?

The T5 model produced grammatically correct, contextually appropriate questions based on the input. The types of questions produced follow the specified levels of Bloom's taxonomy. These levels are:

– Low levels (C1–C2) produce very simple, factual, descriptive questions;

– High levels (C3–C4) produce more analytical and explanatory questions.

5. 4. Bloom classification and final question results

Below presents the classification of the generated questions into Bloom's taxonomy levels and their final transformation. Each question from the T5 process was assigned to the appropriate cognitive level (C1–C6) according to Bloom's taxonomy. This process ensures that the generated questions are pedagogically aligned with the various learning objectives. As detailed in the input and the T5 output, examples of the generated questions for each format are shown in Table 4.

According to this study's results, the system generated different types of questions from a single piece of content. Some examples of multiple-choice questions (MCQs) were created by identifying key concepts, such as "photosynthesis," and using related terms to generate good answers to the question. To create fill-in-the-blank type questions, the key were removed from the text, creating blanks in the middle of the sentence. Essay-type questions were created to promote deeper learning and are generally more cognitively complex. The generated questions maintain coherence and are contextually relevant to the input data, whether from textual sources or images.

Table 4

Question generation results

Bloom level	Question type	Output
C1	MCQ	What is the main function of photosynthesis? a. Produce oxygen b. Convert sunlight into chemical energy. c. Absorb minerals. d. Release carbon dioxide. Answer: B
C2	Fill-in-the-Blank	Photosynthesis is the process by which plants convert _____ into chemical energy. Answer: sunlight
C4	Essay	Explain the process of photosynthesis and its role in plant growth

5. 5. Quantitative evaluation

To measure how well the generated and reference questions match, it is possible to use BLEU and ROUGE scores to assess how similar the two groups are. BLEU and ROUGE are computed with respect to n-grams (sets of contiguous items in a sequence), so they can identify both simple and more complex word overlaps. BLEU measures how many of the individual words in the generated question were found in the reference question. Therefore, this score represents how many of the generated question's words matched up to the individual words of the reference question. BLEU-2 represents an amount for how both fluent (using two consecutive words) and local (using both of the two-word combinations) the word order in the generated question is to that of the reference question. ROUGE measures how much of the content of the generated question overlaps with that of the reference question; thus, it reflects how much of the reference question's content is present in the generated question. ROUGE captures the degree of uniqueness of phrases between each of the reference and generated questions.

Based on Table 5, the proposed question-generation framework produces question outputs similar to the reference questions across all evaluation methods.

Table 5

Detailed quantitative evaluation results

Question type	BLEU-1	BLEU-2	ROUGE-1	ROUGE-2
MCQ	0.85	0.78	0.81	0.76
Fill blank	0.88	0.82	0.84	0.80
Essay	0.78	0.71	0.78	0.72

Fill-in-the-blank questions consistently achieve the highest evaluation scores across all evaluation methods; this is directly related to their proximity to the sentence's origin, which provides strong overall word and phrase overlap with the reference. While MCQs perform just below fill-in-the-

blank questions, the performance across both BLEU and ROUGE evaluation metrics is consistently very similar; due to the diverse number of ways to write the MCQ's and multiple answer options available, the amount of *n*-gram overlap is somewhat lower, particularly when comparing bigrams. The essay questions produced the fewest similar outputs, which is expected given their open-ended format.

Overall, as measured by both BLEU and ROUGE, the question generation framework effectively produces structured, contextually accurate questions and performs particularly well in generating objective formats.

6. Discussion of results on multimedia-based question generation

The analysis is explained based on the effectiveness of each stage in the method proposed in this study. As illustrated in Fig. 1, this system integrates preprocessing, transformer-based generation, and cognitive classification. The preprocessing stage transforms multimedia input, both text and images, into structured textual representations. The effectiveness of the preprocessing stage is shown in Table 2. named entity recognition (NER) successfully extracts key domain concepts from the textual representations. These structured representations enhance and extract keywords from the multimedia input.

These textual representations are processed to generate questions in various formats, including multiple-choice, short-answer, and essay. The textual representations in Table 2 successfully generated questions such as those in Tables 3, 4. The use of the T5 model successfully generated grammatically correct questions. The resulting questions are contextually aligned and classified according to Bloom's taxonomy level. The quantitative evaluation results are presented in Table 5, including BLEU and ROUGE scores. Table 5 shows that structured questions, such as fill-in-the-blank and multiple-choice questions, achieve higher BLEU and ROUGE scores than essay questions.

The proposed method demonstrates several features that differ from existing studies. Traditional rule-based systems generate questions using fixed syntactic transformations, reducing flexibility [6]. Neural approaches based on sequence-to-sequence models improve fluency but lack the ability to process multimedia data sources [7, 8]. The proposed method uses a transformer-based encoder-decoder architecture (T5), enabling flexible text generation. This study has the advantage of integrating multimedia preprocessing in textual and visual forms, as shown in Table 1, Fig. 3, 4. In this study, the incorporation of Bloom's taxonomy classification successfully improved the quality of question generation compared to previous studies [17, 19].

The results obtained demonstrate several distinctive features of the proposed method. First, the system integrates multimodal preprocessing, transformer-based question generation, and Bloom's taxonomy classification. Second, unlike previous AQG approaches that primarily rely on plain text input, the proposed method incorporates OCR-based extraction of instructional images [6–11]. Third, the direct integration of cognitive level classification into the question generation flow enables the system to generate questions across Bloom's levels C1–C6.

This study has several limitations, including reliance on OCR methods to generate accurate textual representations from visual input. Another limitation is that the system was evaluated on a controlled dataset (Table 1), limiting the generalizability of the results. The reproducibility of the results depends on the availability of annotated datasets.

This study has the disadvantage that the distractor-generation mechanism remains relatively simple. The proposed method relies on basic semantic similarity, resulting in less challenging multiple-choice questions. Future study suggests incorporating knowledge-based methods or advanced semantic models to generate more challenging multiple-choice questions. Future development of this study could involve extending the system to handle more complex multimedia formats, such as video and audio data. However, integrating richer multimedia data sources requires more sophisticated alignment techniques. Furthermore, processing richer multimedia formats will increase model complexity and speed up computation.

7. Conclusion

1. The proposed structure consists of four integrated modules: multimedia input processing, multimodal preprocessing, T5-based question generation, and Bloom's Taxonomy classification. These four modules form an integrated framework for generating cognitively structured questions from heterogeneous educational materials.

2. A multimedia preprocessing workflow designed to extract and integrate textual representations from instructional text and images has been successfully implemented. Structured representations of the source data were created using optical character recognition (OCR) on images. The data was preprocessed using tokenization, named entity recognition, and syntactic parsing. This methodology facilitates the effective extraction of critical information from heterogeneous data sources and improves data quality for subsequent processing.

3. A Text-to-text transfer transformer (T5) model has also been built and fine-tuned to generate examination questions from integrated multimedia-based textual representations. The T5 model has demonstrated a strong ability to generate meaningful, grammatically correct question candidates by transforming contextual input into satisfactory question forms. The use of a unified text-to-text framework in the T5 model enables to handle diverse input types with flexibility, thereby supporting the generation of diverse question types.

4. The T5 model incorporates Bloom's taxonomy by associating the generated questions with the cognitive levels (C1-C6), allowing for the creation of questions that are diverse in terms of difficulty and assist in achieving the goals of education.

5. BLEU and ROUGE were used to assess system performance by comparing generated questions with reference questions. The experimental results showed that the proposed system performed effectively, achieving BLEU-1 = 0.86, BLEU-2 = 0.79, ROUGE-1 = 0.88, and ROUGE-2 = 0.81 on average. The experimental results showed that the proposed system performed effectively. Positive results were obtained for structured question types, such as multiple-choice and short-answer questions.

Conflict of interest

The authors declare that they have no conflict of interest in relation to this study, whether financial, personal, authorship, or otherwise, that could affect the study and its results presented in this paper.

Financing

The study was performed without financial support.

Data availability

Manuscript has no associated data.

Use of artificial intelligence

Artificial intelligence tools (SCITE) were used to search relevant literature during the literature review stage. Authors manually verified sources by cross-checking them against reputable academic publications to ensure accuracy and relevance.

Authors' contributions

Marvin Chandra Wijaya: Conceptualization, Investigation, Methodology, Validation, Visualization, Writing – original draft, Validation, Writing – review & editing; **Markus Tanubrata:** Investigation, Formal Analysis, Writing – original draft, Writing – review & editing.

References

1. Kurdi, G., Leo, J., Parsia, B., Sattler, U., Al-Emari, S. (2019). 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
2. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M. et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21. https://doi.org/10.48550/arXiv.1910.10683
3. Zhou, L., Hu, J., Zhang, S., Du, X., Song, M., Zhang, X., Feng, Z. (2024). DenseSAM: Semantic Enhance SAM for Efficient Dense Object Segmentation. Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence, 7994–8002. https://doi.org/10.24963/ijcai.2024/889
4. Adhikari, Y. (2024). A Review of Revised Bloom's Taxonomy of Educational Objectives. Education Review Journal, 1, 115–126. https://doi.org/10.3126/erj.v1i1.82852
5. Du, X., Shao, J., Cardie, C. (2017). Learning to Ask: Neural Question Generation for Reading Comprehension. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1342–1352. https://doi.org/10.18653/v1/p17-1123
6. Rus, V., Wyse, B., Piwek, P., Lintean, M., Stoyanchev, S., Moldovan, C. (2010). The first question generation shared task evaluation challenge. in Proceedings of the 6th International Natural Language Generation Conference, 251–257. Available at: https://aclanthology.org/W10-4234/
7. Sun, X., Liu, J., Lyu, Y., He, W., Ma, Y., Wang, S. (2018). Answer-focused and Position-aware Neural Question Generation. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 3930–3939. https://doi.org/10.18653/v1/d18-1427
8. Zhao, Y., Ni, X., Ding, Y., Ke, Q. (2018). Paragraph-level Neural Question Generation with Maxout Pointer and Gated Self-attention Networks. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 3901–3910. https://doi.org/10.18653/v1/d18-1424
9. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North. https://doi.org/10.18653/v1/n19-1423
10. Wijaya, M. C. (2021). Automatic Short Answer Grading System in Indonesian Language Using BERT Machine Learning. Revue d'Intelligence Artificielle, 35 (6), 503–509. https://doi.org/10.18280/ria.350609
11. Alzubi, J. A., Jain, R., Singh, A., Parwekar, P., Gupta, M. (2021). COBERT: COVID-19 Question Answering System Using BERT. Arabian Journal for Science and Engineering, 48 (8), 11003–11013. https://doi.org/10.1007/s13369-021-05810-5
12. Wijaya, M. C. (2025). Development of adaptive congestion control mechanism for real-time multimedia streaming in variable network condition. Eastern-European Journal of Enterprise Technologies, 5 (9 (137)), 54–63. https://doi.org/10.15587/1729-4061.2025.339985
13. Oliinyk, V., Biziuk, A., Deineko, Z., Chelombitko, V. (2025). Formalization of text prompts to artificial intelligence systems. Eastern-European Journal of Enterprise Technologies, 5 (2 (137)), 84–97. https://doi.org/10.15587/1729-4061.2025.335473
14. Omarkhanova, D., Oralbekova, Z. (2024). Interpretation of georadar data based on machine learning technologies. EUREKA: Physics and Engineering, 4, 193–204. https://doi.org/10.21303/2461-4262.2024.003289
15. Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S. et al. (2019). What Is Wrong With Scene Text Recognition Model Comparisons? Dataset and Model Analysis. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 4714–4722. https://doi.org/10.1109/iccv.2019.00481

16. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., Parikh, D. (2015). VQA: Visual Question Answering. 2015 IEEE International Conference on Computer Vision (ICCV), 2425–2433. https://doi.org/10.1109/iccv.2015.279
17. Gummineni, M. (2020). Implementing Bloom's Taxonomy Tool for Better Learning Outcomes of PLC and Robotics Course. International Journal of Emerging Technologies in Learning (iJET), 15 (05), 184. https://doi.org/10.3991/ijet.v15i05.12173
18. Voloshchuk, Y., Mitsa, O. (2025). Determining the effectiveness of GPT-4.1-mini for multiclass text categorization. Eastern-European Journal of Enterprise Technologies, 5 (2 (137)), 98–106. https://doi.org/10.15587/1729-4061.2025.340492
19. Gani, M. O., Ayyasamy, R. K., Sangodiah, A., Fui, Y. T. (2023). Bloom's Taxonomy-based exam question classification: The outcome of CNN and optimal pre-trained word embedding technique. Education and Information Technologies, 28 (12), 15893–15914. https://doi.org/10.1007/s10639-023-11842-1
20. Chatoui, H., Ata, O. (2021). Automated Evaluation of the Virtual Assistant in Bleu and Rouge Scores. 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), 1–6. https://doi.org/10.1109/hora52670.2021.9461351