Cognitively Aligned Assessment Item Generation with Open-Source LLMs: A Comprehensive Evaluation on LearningQ

Mahmoud Badry[1], Walaa Medhat[2], Shereen A. Taie[3], Asmaa Hashem Sweidan[4]
Faculty of Computers and Artificial Intelligence, Fayoum University, Fayoum, Egypt[1,3,4]
Faculty of Computers and Artificial Intelligence, Banha University, Qalyubiyya, Egypt[2]

***Abstract*—Automated generation of high-quality educational assessment items is still difficult, especially when it comes to higher-order cognitive skills. Although Large Language Models (LLMs) show promise, their structural validity and cognitive alignment are limited. This study systematically evaluates fine-tuning open-source LLMs using an enriched LearningQ dataset that includes Bloom's cognitive labels and evidence. The results show a clear performance contrast. Qwen2.5-3B-Instruct displays the best semantic reasoning, while Llama-3.2-3B shows better structural adherence, achieving a 94.9% validity rate and full compliance against answer leakage while maintaining high question validity. In contrast, older encoder-decoder models like FLAN-T5-XL do not generate valid questions. The study finds that small- to medium-sized instruction-tuned models, backed by strong data engineering, are successful at developing scalable, cognitively well-aligned assessment items.**

***Keywords*—Automated question generation; Large Language Models; educational assessment; Bloom's Taxonomy; Parameter-Efficient Fine-Tuning; LearningQ**

I. Introduction

The application of Large Language Models (LLMs) in educational technology has transformed automated content generation. This is particularly true for creating high-quality assessment items that assess not just factual recall but also higher-order thinking (cognitive) skills such as analysis, application, and synthesis. Traditional automated question generation (AQG) methods were based on rule-based systems or early neural models, which had limitations in scope and scalability. Instruction-tuned large language models can generate a wide range of effective questions at scale. However, models that are generally pre-trained often produce results that lack educational depth or fail to follow the strict structural rules required for standardized assessments.

Cognitive alignment refers to how well an assessment item measures the cognitive processes intended by a given learning outcome within the field of educational measurement theory. The idea of cognitive alignment is very similar to that of construct validity and alignment validity within the realm of test creation. Thus, assessing the cognitive alignment of test creation involves more than simply looking at linguistic similarities. In this study, an "assessment item" refers to a "single educational question" designed to evaluate a specific learning objective or cognitive level.

A major limitation in current AQG research is the reliance on datasets meant for machine reading comprehension, like SQuAD [1] or RACE [2]. Although these datasets are large, they mainly focus on extractive fact retrieval ("Remembering"), which does not capture the details of educational assessment. Assessment items need to evaluate specific cognitive levels, as defined by frameworks such as Bloom's Taxonomy. Additionally, valid assessment items must meet strict requirements. For example, the questions should be interrogative, the answer should not be included in the question, and premature explanations should not be given. General-purpose LLMs often violate these rules in zero-shot settings.

Beyond structural validity, high-quality educational items must also show a variety of assessment formats and Intended Learning Outcomes (ILOs). This includes generating items across multiple item types—such as conceptual, analytical, application-based, and synthesis-oriented prompts—and aligning each item with a specific learning objective rather than producing generic fact-recall items. Without enforcing this alignment, models often revert to simple cognitive tasks, which limits their usefulness in real teaching and evaluation situations.

To address these challenges, this study examines the ability of open-source LLMs to create standardized assessment items using LearningQ [3], a large dataset derived from real educational platforms, including TED-Ed and Khan Academy. Unlike datasets designed to find facts, LearningQ has many questions that require higher-order thinking. This quality makes it a suitable testing ground for developing questions that align with cognitive principles. This study focuses on assessing various open-source LLM models using a unified, cognitively aligned pipeline, rather than comparing them with existing state-of-the-art AQG systems. This is to establish the influence of models on cognitive alignment.

This study presents a detailed fine-tuning and evaluation framework to close the gap between raw educational content and usable assessment items. The proposed approach employs a comprehensive data enrichment pipeline that initially improves raw LearningQ data by adding synthetic cognitive labels and emphasized passage evidence, utilizing Qwen2.5-7B-Instruct [4]. Next, several models are adjusted using Parameter-Efficient Fine-Tuning (PEFT) through Low-Rank Adaptation (LoRA) [5]. This includes both contemporary decoder-only architectures, such as Llama-3.2-3B [6] and Qwen2.5-3B, and traditional encoder-decoder models like FLAN-T5-XL [7].

This research makes several significant contributions:

1) A Cognitive-Aligned Data Pipeline: that adds passage highlighting and Bloom's Taxonomy labels to educational datasets. This enables models to discover the relationship between cognitive targets and segments of the source text.
2) Comparative Model Analysis: which evaluates six different LLM architectures with 3 billion to 7 billion parameters. It looks at their performance not only on lexical overlap using BLEU and ROUGE but also on semantic preservation through BERTScore and pedagogical intent.
3) Insights for AQG: show that modern instruction-tuned LLMs, when combined with careful data engineering and fine-tuning, can consistently produce assessment items that align with cognitive goals, outperforming traditional architectures.

Our results indicate a distinct performance gap between contemporary instruction-tuned architectures and earlier models. This demonstrates that with proper data engineering and fine-tuning, small-to-medium LLMs can consistently function as tools for generating automated, cognitively aligned assessments. This may assist educators in developing standardized assessments.

The remainder of this study is organized as follows: Section II reviews related state-of-the-art literature and identifies the main challenges of current research. Section III describes the LearningQ dataset. Section IV presents the phases of the proposed approach. Section V and Section VI discuss the evaluation criteria and results based on the enriched LearningQ dataset. Section VII discusses the implications of the findings. Section VIII presents conclusions and discusses future work.

II. Related Work

This section presents related work on relevant datasets, methods for generating questions using LLMs, and applications of the LearningQ dataset [3].

A. Datasets for Question Generation

Significantly, the establishment of robust datasets has been critical in the development of machine reading comprehension and question generation. Early datasets such as SQuAD [1], which have over 100,000 crowd-sourced questions associated with corresponding passages on Wikipedia, with answers being restricted to a particular text span, have been instrumental. Although such datasets have been effective for extractive question answering, they have been primarily focused on factual recall. To address this issue, RACE [2], a dataset of multiple-choice questions created by experts and based on English examination papers for middle and high school students has been proposed. Unlike SQuAD, the answers in RACE cannot be extracted; they require a thorough level of reasoning and understanding of a passage.

The attempt to bridge QA studies and educational assessment has been realized in the EduQG dataset, released in 2023 [8], which contains a collection of 3,397 questions in multiple formats, all gathered from OpenStax textbooks. EduQG also introduced a new type of sentence-level annotation and included a subset of 903 questions annotated with their corresponding level of cognitive complexity according to Bloom's Taxonomy.

Following this trend, EduQuest, released in 2024 [9], has widened the field of application of question answering in education by collecting over 76,000 instructional documents and 68,000 questions from sources such as MIT OpenCourseWare, OpenStax, and Khan Academy. An important aspect of EduQuest is the addition of metadata labels, such as the corresponding cognitive complexity level and pedagogical labels, which are essential for applications that involve higher-level question generation beyond simple recognition and recall.

B. Educational Assessments and LLM-Based Question Generation

Recent advances in Large Language Models (LLMs) have motivated the use of these models to facilitate the automation of educational assessment design. Lamsiyah et al.'s (2024) [10] study presents an approach to automatic question generation using the RLLM-EduQG model, which fine-tunes the FLAN-T5 model on a hybrid objective function combining cross-entropy and reinforcement-learning rewards based on BLEU and semantic similarity scores. This approach addresses exposure bias and improves both syntactic and semantic generation performance. Another study on automatic question generation was proposed by Hasan et al. (2024) [11]. It presented an Automatic Question & Answer Generation (AQAG) framework based on the Llama-2 model and uses unsupervised prompts on the RACE corpus to generate multiple-choice and fill-in-the-blank question types.

Regarding cognitive alignment, Zhuge et al. (2025) [12] introduced TwinStar. It is a dual-LLM system in which one model generates questions and another. It was fine-tuned on ChatGPT-6B to assess alignment at cognitive levels. This refining process better preserved cognitive targets and ensured content relevance compared to models such as GPT-4 and Bard.

C. Applications of the LearningQ Dataset

LearningQ has gained popularity in recent research studies due to its scope, instructional value, and diverse questions. Focal (2023) [13], an end-to-end automated construction of an evaluation pipeline, utilized LearningQ to fine-tune its evaluator model. The study showed promising outcomes in enhancing the logical coherence and instructional value of questions generated by LLMs. Moreover, García-Méndez [14] (2025) emphasized LearningQ's value in training models for generating questions as well as automated evaluations, citing previous studies by Moore et al. (2022) and Bhat et al. (2022) on training models to evaluate the quality of questions generated by students with the aid of LearningQ.

In a study of multimodal learning, Stamatakis et al. (2025) [15] analyzed LearningQ, a subset of TED-Ed and Khan Academy, to assess Vision-Language Models (VLMs) such as Video-LLaMA and PG-Video-LLaVA. The study showed that when generating contextually relevant educational questions from video inputs, VLMs trained or fine-tuned on LearningQ achieved a significant improvement over zero-shot VLMs that generated irrelevant queries.

D. Cognitive Alignment and LLM Evaluation

Some studies also used cognitive models, such as Bloom's taxonomy, to evaluate LLM performance. Several LLMs, such as Llama-2, Mistral, and GPT-4, were assessed by Scaria et al. (2024) [16] using cognitive prompt techniques, such as the Chain of Thought. According to this study, LLM models typically perform well on lower-order thinking tasks such as "Remember" or "Understand", but struggle to produce reliable results on higher-order thinking tasks such as "Create". Another study by Huber and Niklaus (2025) [17] assessed well-known benchmark tests, such as MMLU and BIG-Bench Hard, and classified them according to Bloom's taxonomy levels. According to this study, there were very few instances of higher-order thinking tasks like "Evaluate" or "Create" in these benchmarks, while it found a high frequency of lower-order thinking tasks.

Most of the literature reviewed has focused on creating automated questions with the assessment AQG. These studies offer insights related to cognitive alignment. However, only a few studies investigate cognitive alignment directly. These have issues, including poor structural validity, poor cognitive alignment, and poor adherence to assessment constraints in previous AQG approaches. This study addresses these problems using the proposed approach.

III. Dataset: LearningQ

A. Dataset Description

LearningQ [3] is a large dataset created specifically to generate and analyze educational questions, with a special emphasis on learning- and assessment-oriented questions. This is unlike other datasets, such as SQuAD [1] and RACE [2], which have mostly focused on information retrieval questions. LearningQ comprises more than 230,000 document-question pairs across domains such as science, humanities, and mathematics.

LearningQ data was collected from two major online learning platforms: **1) TED-Ed** (https://ed.ted.com/) consists of questions created by instructors linked to educational video lectures, characterized by high quality and a well-structured educational framework. **2) Khan Academy** (https://www.khanacademy.org/) set includes questions prepared by learners and posted in the comments section of videos and articles. To reduce noise, the dataset creators used a convolutional neural network (CNN)-based classifier to filter out irrelevant content while retaining learning-relevant questions, achieving 80.5% accuracy.

LearningQ consists of videos and article documents coupled with educational questions. It features documents that are much longer than standard benchmarks and covers the revised Bloom's taxonomy levels with distributions of recall (31%), comprehension (36%), application (9%), analysis (11%), evaluation (2%), and creation (1%), along with nearly 10% of unknown/irrelevant content generated by the learner, as shown in Table I. For clarity, Table I presents the datasets by content type (video lectures and articles) obtained from the platforms TED-Ed and Khan Academy. To better understand the structural properties of LearningQ, the proposed framework conducted an in-depth statistical and schema-level inspection of the JSON files using Python utilities. The TED-Ed subset has a standardized, instructor-designed structure, unlike Khan Academy, where a high level of variability was observed, with a range of 1 to 42 questions per document, a mean of 12.44, and a standard deviation of approximately 8.71. LearningQ covers a modified version of Bloom's Taxonomy, covering a wide range of cognitive abilities from basic recall to complex thinking, offering a high level of cognitive and structural variation (see Table II).

TABLE I. Statistics of the LearningQ Dataset

Data Source	Content Type	# Docs	# Qs	Avg Q/Doc
TED-Ed	Video Lectures	1,102	7,612	6.91
Khan Academy	Video Lectures	7,924	201,273	25.40
Khan Academy	Articles	1,815	22,585	12.44
Total	**Mixed**	**10,841**	**231,470**	**21.3**

TABLE II. Question Examples of Different Bloom's Revised Taxonomy Levels in LearningQ Dataset.

Bloom Level	Example
Remembering	What is a negative and positive feedback in homeostasis?
Understanding	Why do some plankton migrate vertically?
Applying	If I double the area and take half of the fraction, do I get the same result?
Analyzing	Why are cities like London, Tokyo, and New York facing shortages in burial ground space?
Evaluating	Will all the cultures merge into one big culture, due to the fading genetic distinctions?
Creating	Can somebody please explain to me what marginal benefits are and give me some examples?

While LearningQ is presented as covering a range of cognitive levels, the downloaded dataset does not include Bloom's Taxonomy labels for individual questions. Thus, the provided distribution represents only a descriptive understanding of the data. Therefore, cognitive-level labeling has been newly applied in this study using a language model–based classifier.

IV. Methodology

This section describes the proposed framework, which applies a structured process to transform raw educational data into high-quality assessment items. As shown in Fig. 1, the framework includes four clear phases: 1) Data Preprocessing and Enrichment, which involves raw LearningQ data [3]; 2) Prompt Engineering, which organizes the data into instruction-tuned structures; 3) Parameter-Efficient Fine-Tuning (PEFT), which uses Low-Rank Adaptation (LoRA) to fine-tune LLMs for the educational domain; and 4) Comprehensive Evaluation, which includes both semantic measures and evaluation constraints.

A. Data Preprocessing and Enrichment

It plays an important role, significantly impacting the overall performance of subsequent processes. Consequently, the input data pass through a series of steps to clean, normalize, and semantically validate them. The preprocessing stage involves the following steps:

1) Dataset Cleaning: This step includes nested JSON object analysis, text normalization, handling of invalid

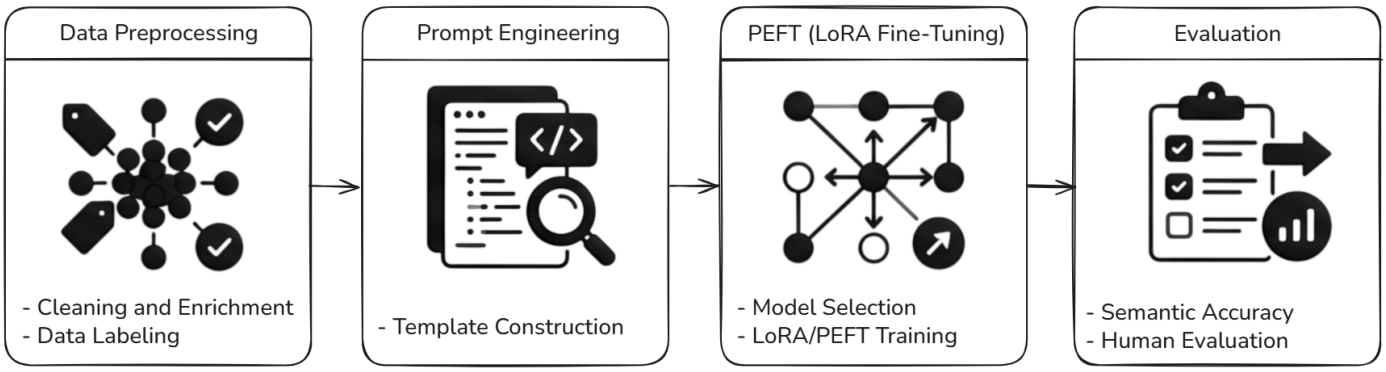

Fig. 1. Methodology pipeline

rows, and resolving inconsistencies between multiple sources. The output of this is a flat table, Table III. Specifically, any invalid entries (e.g., empty passages, questions, or malformed data) were filtered out. The text data was normalized by converting to lowercase and trimming, and any samples with invalid question-passage alignment were filtered.

2) Restructuring into One-Question-Per-Row: To ensure that the supervision during the training is consistent, the data is transformed into the following structure: "| ID | Title | Subject | Text | Question |".
3) Cognitive Level Classification: Since LearningQ does not provide cognitive labels, each question in LearningQ was labelled by following the taxonomy of Bloom. For this experiment, Qwen2.5-7B-Instruct [4] was used because it was observed to perform well at following instructions, was efficient in 8-bit deployment, and exhibited stability in differentiating between nearby cognitive levels. A new column, "Cognitive Level", was added to the schema.
4) Highlighting Relevant Passage Segments: A second LLM processing step (Qwen2.5-7B-Instruct) is utilized to pinpoint the relevant passage segments needed to answer the questions. These segments are marked with the HTML tag '<h>...</h>. The purpose of highlighting was to improve alignment between the passage and the question. This provides explicit contextual grounding and facilitates easier interpretation during evaluation.
5) Verify reliability: To verify the reliability of the supervision signals produced by the LLM for training purposes, a random set of 150 supervision signals was manually checked for their validity. The agreement between the two human evaluators for the Bloom-level labels was substantial, with a Cohen's κ of 0.79. Furthermore, the highlighted evidence segments were found to be 87% accurate according to human judgment.
6) Balancing Cognitive-Level Distribution: After labelling and highlighting, the distribution of classes at different Bloom levels was measured. To ensure a balanced distribution, classes that were overrepresented were downscaled, while underrepresented classes were upscaled.

TABLE III. DATASET CLEANING: INITIAL SHEET OUTPUT

Field	Description
ID	Source document identifier
Title	Document title
Subject	High-level semantic category
Text	Passage/content body
Questions	List of associated questions

Template Components

@@TASK@@
Generate exactly ONE assessment question.

@@SUBJECT@@
<Subject>

@@COGNITIVE_LEVEL@@
<Cognitive Level>

@@PASSAGE@@
<Text>

@@CONSTRAINTS@@
The question must be answerable using ONLY the passage
Do NOT include the answer
Do NOT include explanations
Output the question only

Fig. 2. Prompt template components.

B. Prompt Template Construction

To ensure consistent supervision during fine-tuning and minimize stylistic variability across training instances, each enriched record is converted into a unified instruction–response template Fig. 2, including the topic, knowledge level, and highlighted passage, with strict output constraints. Task specifications ("Create only one assessment question") guide the creation of a single item and establish a consistent pattern for fine-tuning. Metadata and <h>-tagged evidence provide context, while the output constraints — preventing explanations, avoiding answer leakage, and requiring only one question — ensure that assessment items are standardized, organized, and cognitively aligned.

C. Model Loading and LoRa Fine-Tuning

Pairs of prompt and response examples were used to fine-tune six open-source large language models using Parameter-Efficient Fine-Tuning (PEFT) [5]. To address resource constraints in educational natural language processing, the Low-Rank Adaptation (LoRA) technique was applied. This approach introduces a small number of trainable parameters into the main transformer layers while keeping the base model weights fixed, enabling efficient domain adaptation without full end-to-end training. The entire process is illustrated in Fig. 3.

1) Model Selection and Architectural Diversity: To ensure a comprehensive evaluation across architectures, six models were fine-tuned under identical experimental conditions: Llama-3.2-3B-Instruct [6], Qwen2.5-3B-Instruct [4], Mistral-7B-Instruct [18], MediPhi-Instruct [19], Gemma-3-4B [20], and FLAN-T5-XL [7]. These models span a spectrum of decoder-only, encoder–decoder, and multimodal architectures with 3B–7B parameters, allowing us to evaluate differences in model behavior under a uniform task formulation.

2) Quantized Model Loading: All models were loaded in 8-bit precision, consistent with the training script. This reduced VRAM usage by approximately 60%, enabling fine-tuning on modest GPU hardware such as Google Colab and Kaggle T4 instances. Despite the compression, 8-bit quantization preserves the representational depth required for reasoning over instructional prompts and highlighted passages.

3) Dataset Formulation and Preprocessing: The dataset was divided into subsets for training (80%), validation (10%), and testing (10%), subsets and organized into unified objects to facilitate batch processing. Input sequences were tokenized, using a structural separator (@@PASSAGE@@) to separate instructions from paragraph content, and the marker "### OUTPUT": to indicate the target response. Sequences were truncated to fit the model's context window, and selective masking of tokens (-100) was applied to the instructions, paragraph content, and padding to ensure that the loss function was optimized only for the generated question outputs.

4) Fine-Tuning Configuration and Execution: To optimize the pre-trained models for the target task while minimizing computational overhead, Low-Rank Adaptation (LoRA) was employed. The adapter configuration targeted all key linear layers within the transformer architecture, including the query, key, value, and output projections (q_proj, k_proj, v_proj, o_proj), as well as the feed-forward gate, up, and down projections. The hyperparameters were set to r=8, alpha=16, and a dropout rate of 0.05. This configuration introduces fewer than 0.3% additional trainable parameters, enabling efficient adaptation for task-specific behaviors—such as cognitive-level alignment and strict output formatting—without the resource intensity of full-model fine-tuning. Model training was performed with 8-bit quantization and mixed precision (FP16), a batch size of 1 per device, gradient accumulation every 4 steps, and a learning rate of 2e-4 for 3 epochs. The input sequences were limited to 2048 tokens, with a structured prompt format and masking, so that the loss was calculated only on the output tokens. The preprocessing included data normalization and filtering out invalid entries. Every 100 training steps, validation was used to evaluate the model's performance. During this period, the effectiveness of the network was evaluated through perplexity, defined as $e^{\text{eval\_loss}}$.

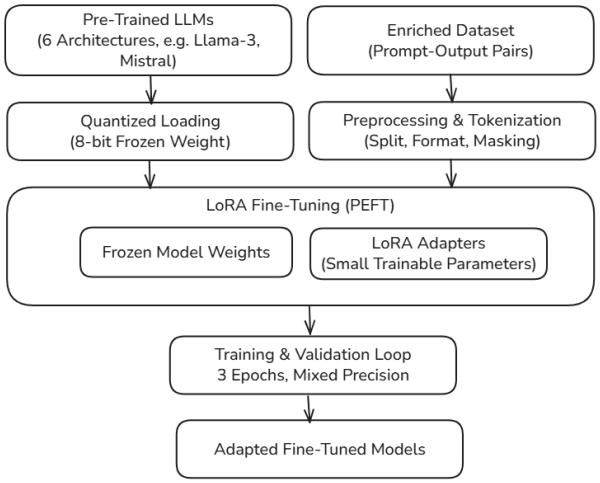

Fig. 3. The end-to-end LoRA fine-tuning framework across diverse model architectures.

D. Inference and Comprehensive Evaluation

Following fine-tuning, a series of evaluations was performed using a controlled inference process, that simulated the exact conditions under which the model was trained. In this process, the entire prompt template was recreated, including the task directive, subject, cognitive level, highlighted passage, and delimiter, to ensure no distributional shift occurred in the model's outputs. The outputs generated by a deterministic decoding (greedy, non-sampling) were then normalized to assess their structural validity, and ensure that only questions were evaluated.

To ensure reproducibility, experiments were performed with a fixed random number generator seed (42). The input sequences were limited to 2048 tokens, with the prompt and passage text truncated as needed. During inference, deterministic decoding was used (greedy decoding with a maximum of 64 tokens (max_new_tokens = 64)). The output was normalized, and the constraints were verified.

V. Evaluation Metrics

A thorough evaluation of the quality of generated questions in the education domain was conducted using multidimensional evaluation metrics. Unlike traditional text-generation applications, cognitive alignment is not an observable phenomenon that can be measured by a single automatic metric. Rather, it requires a multidimensional measurement approach, in which each measure reflects an element of the construct in question. Table IV shows how the components of cognitive alignment relate to the proxies used to measure them.

It is important to understand that the commonly used action verbs connected with Bloom's taxonomy ("define", "explain", "analyze") are very poor indicators of the cognitive process involved. Different cognitive processes can be represented by the same action verb based on their levels of difficulty. For example, "Explain what photosynthesis is" means the use of knowledge, while "Explain how the environmental changes

TABLE IV. MAPPING OF COGNITIVE ALIGNMENT COMPONENTS TO EVALUATION METRICS

Construct Component	Proxy Metric
Semantic Similarity	BERTScore, SBERT
Lexical and Structural	BLEU, ROUGE, METEOR
Structural Validity	is_question, length_ok
Assessment Integrity	no_answer_word, no_expl_word
Cognitive Intent Alignment	Human Evaluation

affect the rate of photosynthesis" requires analysis. Therefore, it is impossible to rely solely on keyword categorization to determine the cognitive process level.

A. Semantic Similarity Metrics

These metrics evaluate how well questions generated from a reference question preserve the original meaning and intent, despite possible variations in wording.

1) BERTScore (F1) [21]: It is used to evaluate semantic similarity based on the comparison of contextual word representations of generated candidates and reference texts. In contrast to n-gram-based metrics, BERTScore is robust to paraphrasing. Higher scores indicate greater semantic similarity between generated questions and the ground truth.

2) Sentence-BERT (SBERT) Cosine Similarity [22]: In the Sentence-BERT model, the cosine similarity is used to measure the similarity between the sentence embeddings of the generated output and the reference text.

B. Lexical and Structural Metrics

It measures Conventional n-gram-based metrics evaluate the generated text's language quality and structure in relation to the reference questions.

1) BLEU (Bilingual Evaluation Understudy) [23]: It assesses how similar the output and reference are in terms of n-gram overlap. In generative tasks, the BLEU metric is usually lower; however, a higher BLEU score denotes higher output precision.

2) ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) [24]: The F1 score of the ROUGE-L system has been used for the assessment of the longest common subsequence that exists between the generated candidate and the reference texts.

3) METEOR [25]: It uses stem, synonym, and word order similarities to measure similarities. This metric is generally more flexible with respect to linguistic variation than BLEU. It's also known to correlate strongly with human judgments of word similarity.

C. Task-Specific Constraint Validation

In addition to linguistic similarities, it is important that the generated output is functionally correct with respect to the assessment. Therefore, a constraint-checking module is implemented using heuristic algorithms.

1) Question Validity (is_question): The first heuristic evaluates if the output complies with the syntax of a question. It does this through the presence of a question mark and interrogative sentence structure.

2) Answer Leakage (no_answer_word): Another important failure mode in question generation involves including the answer within the answer (for example, "The answer is."). It ensures that there is no answer leakage within the generated question.

3) Explanation Prohibition (no_expl_word): The models are expected to produce questions only. There is no need for any explanation. The purpose of the check is to ensure the production of 'pure' questions without any explanation.

4) Length Constraints (length_ok): This metric checks whether the generated question falls within an appropriate range of characters or tokens, preventing the generation of either overly short or overly long questions.

D. Cognitive Intent Alignment

For the task of determining the relevance of the generated questions based on the specified Bloom's Taxonomy level, a human evaluation approach was employed. A random sample of questions generated was evaluated independently by two human annotators who had the passage, the required cognitive level, and the generated question available for consideration and were asked to check if the generated question matched the desired cognitive level. The basis of the evaluation is educational measurement theory, according to which an evaluation by human experts is the most reliable method for measuring cognitive processes and their alignment to learning goals. This human evaluation was conducted intentionally because automatic measurements alone do not capture all aspects of pedagogical value or cognitive alignment across Bloom's levels.

VI. RESULTS

A. Overall Performance Evaluation

A comprehensive set of semantic, lexical, and constraint-based metrics was used to assess six refined large language models (LLMs). Llama-3.2-3B, Qwen2.5-3B, Mistral-7B, MediPhi, Gemma-3-4B, and FLAN-T5-XL were the models evaluated. Their ability to create assessment items that meet cognitive expectations was evaluated. A summary of the performance comparison across all important metrics is presented in Table V. Table VI shows qualitative differences in model-generated assessment items.

B. Semantic Similarity Analysis

Embedding-based metrics are used to evaluate how well the generated questions maintain the ground truth's educational intent.

- Qwen2.5-3B achieved the highest BERTScore F1 (0.8275), indicating it was most effective at capturing the underlying semantic meaning of the reference questions, even when the specific phrasing differed.
- Mistral-7B demonstrated superior performance in Sentence-BERT (SBERT) cosine similarity (0.3272), suggesting a strong alignment in overall sentence-level intent.

- Llama-3.2-3B and MediPhi remained competitive with BERTScores above 0.80, whereas Gemma-3-4B significantly underperformed (0.447), indicating a failure to generate semantically relevant content.

C. Lexical and Structural Fidelity

In terms of exact lexical matching and structural coherence:

- Llama-3.2-3B outperformed all other models in BLEU (0.1131) and METEOR (0.0989) scores. This suggests that Llama-3.2-3B most frequently generated n-grams and stems that matched the ground truth references.
- Mistral-7B achieved the highest ROUGE-L F1 (0.0914), marginally outperforming Qwen2.5-3B (0.0904), indicating strong capabilities in preserving the longest common subsequences and structural phrasing of the assessment items.
- FLAN-T5-XL failed in these categories, scoring 0.0 across BLEU, ROUGE-L, and METEOR. Qualitative inspection revealed that this model often generated short, non-question phrases (e.g., "The need for rights in government") rather than valid interrogatives.

D. Task-Specific Constraint Compliance

The usefulness of the models for educational purposes was evaluated based on the models' compliance as follows:

- Question Validity (is_question): The highest validity rate of 94.9% was recorded by Llama-3.2-3B. MediPhi followed with a rate of 92.4%. This suggests that the generated text included question marks at the end and contained interrogative sentences. In contrast, Gemma-3-4B produced valid questions only 51.9% of the time. FLAN-T5-XL generated no valid questions, resulting in a 0% rate.
- Answer Leakage (no_answer_word): All models demonstrated high adherence to this constraint (>98%), ensuring that the answer key was not revealed within the question stem. Llama-3.2-3B achieved perfect compliance (100%).
- Explanation Prohibition (no_expl_word): Llama-3.2-3B was the most effective at excluding unwanted reasoning or explanations, achieving 67.1% compliance. Qwen2.5 followed with 60.8%. While FLAN-T5-XL shows 100% compliance here, this is an artifact of its failure to generate substantive text rather than successful instruction following.
- Length Constraints (length_ok): Qwen2.5-3B was the most successful at generating questions within the optimal length range (50.6%), whereas Llama-3.2-3B frequently generated output outside the target length parameters (10.1%)

E. Summary of Model Stability

Based on the "Output Stability" qualitative measure, Llama-3.2-3B and Qwen2.5-3B exhibited High stability, consistently following instructions and maintaining formatting. Mistral-7B and MediPhi showed Medium-to-High stability. Gemma-3-4B and FLAN-T5-XL were rated Low and Very Low, respectively, due to hallucinations and failure to adhere to the fundamental task of question generation. Qualitative samples confirm that while Llama and Qwen produced coherent, cognitively aligned questions (e.g., "Why did delegates think the Constitution needed limits on federal power?"), FLAN-T5-XL produced fragmented summaries.

To establish a solid basis for evaluation in educational measurement theory, a human assessment of cognitive alignment was performed on a random sample of 80 generated questions. Two human judges determined whether each question matched the corresponding Bloom level using a three-point correctness scale. The assessment yielded a Cohen's κ of 0.76, with 72% of the questions correctly aligned. This indicates that the improvements observed in the automatic metrics are genuine and reflect real enhancements in pedagogical validity. Statistical significance was assessed using paired t-tests on key metrics, confirming that the observed improvements are statistically significant ($p < 0.05$).

In an effort to determine how automatic metrics correlate with human assessment, a correlation analysis was performed using the evaluation metrics and human cognitive alignment assessment scores. The analysis shows a medium level of positive correlation between semantic similarity metrics, such as BERTScore and SBERT, and human assessment metrics. On the other hand, BLEU and ROUGE lexical metrics showed low correlation, owing to their focus on superficial similarities. This clearly shows that not all cognitive alignment can be captured by one evaluation metric. Instead, semantic metrics, combined with human assessment and validation using constraints, provide the best measures.

VII. Discussion

This study indicates that fine-tuned, instruction-aligned Large Language Models (LLMs) can reliably generate cognitively aligned educational assessment items when trained on enriched instructional data. By incorporating semantic similarity, lexical accuracy, and rule-based validation, this study provides an extensive understanding of model behaviour that goes beyond question generation. This section highlights the overall findings and their implications for educational assessment research.

First, the results clearly show a distinction between decoder-only and traditional encoder–decoder architectures. Specifically, Llama-3.2-3B and Qwen2.5-3B demonstrated superior performance compared to other models on both semantic preservation and output validity measures. Notably, Qwen2.5-3B achieved the highest F1 score on the BERTScore measure, indicating significant semantic preservation of the pedagogical intent encoded in the reference items. On the other hand, Mistral-7B achieved the highest score on the SBERT cosine similarity measure, indicating strong alignment at the sentence level.

Second, lexical-similarity metrics indicate that structural fidelity depends significantly on the model type. The best scores for the BLEU and METEOR metrics were achieved by the Llama-3.2-3B model, suggesting greater similarity in

TABLE V. PERFORMANCE COMPARISON OF FINE-TUNED LLMS

Category	Metric	Llama-3.2-3B	Qwen2.5-3B	Mistral-7B	MediPhi	Gemma-3-4B	FLAN-T5-XL
Semantic Similarity	BERTScore F1	0.8147	**0.8275**	0.8217	0.8068	0.447	0.7894
	SBERT Cosine	0.3029	0.3229	**0.3272**	0.3134	0.1909	0.0356
String-Matching Similarity	BLEU	**0.1131**	0.0607	0.0729	0.0677	0.0445	0
	ROUGE-L F1	0.069	0.0904	**0.0914**	0.0648	0.0096	0
	METEOR	**0.0989**	0.0718	0.0826	0.0758	0.0257	0
Question Validity	is_question	**94.90%**	91.10%	91.10%	92.40%	51.90%	0.00%
	length_ok	10.10%	**50.60%**	36.70%	27.80%	0.00%	0.00%
Constraint Compliance	no_answer_word	**100%**	98.70%	98.70%	98.70%	98.70%	100%
	no_expl_word	**67.10%**	60.80%	53.20%	55.70%	53.20%	100%

TABLE VI. QUALITATIVE COMPARISON OF GENERATED ASSESSMENT ITEMS ACROSS MODELS

Prompt (Shortened)	Ground Truth	Llama-3.2-3B	Qwen2.5-3B	Mistral-7B	MediPhi	Gemma-3-4B	FLAN-T5-XL
Humanities – Understand "Bill of Rights passage"	Why were the authors of the Federalist Papers against the Bill of Rights?	Why did some leaders at the Constitutional Convention believe a stronger federal government was necessary?	How did concerns about potential government tyranny influence the delegates' decision to add the Bill of Rights?	In what ways did fears about a powerful executive branch motivate the creation of the Bill of Rights?	Why were many delegates concerned that the new Constitution might threaten individual rights?	Why did delegates think the Constitution needed limits on federal power?	The need for rights in government.
Physics – Apply "Velocity graph passage"	How to calculate velocity and acceleration?	How would you determine the acceleration of an object using the slope of a velocity graph?	How can you calculate displacement using the area under a velocity graph?	How would you use slope and graph values to determine acceleration and displacement?	How do you apply the slope formula to find acceleration?	How can the slope of a velocity graph be used?	Using the graph to find acceleration.
Art History – Remember "Stonehenge passage"	(Example recall question)	What was the purpose of the Aubrey holes in the first phase of Stonehenge construction?	What significant feature characterized the first phase of Stonehenge construction?	When did the first phase of Stonehenge construction occur?	What was built during the first phase of Stonehenge?	What is Stonehenge?	Stonehenge construction phases.

phrasing and construction with human-generated texts. However, it is important to note that high lexical metric scores did not necessarily correlate with high cognitive alignment scores. This is also supported by some of Llama's outputs, which showed a tendency to paraphrase at the surface level rather than perform semantic-level transformations. This aligns with previous findings on LearningQ-based pipelines and indicates that it is not possible to generate higher-level questions solely through imitation of phrasing; instead, it requires cues and cognitive scaffolding incorporated into the instructional content. The relatively low performance on length constraints is due to the model's tendency to generate more descriptive questions.

Third, constraint evaluation for specific tasks is beneficial for understanding the feasibility of applying LLMs to standardized assessments. In this regard, metrics such as "is_question" and "no_answer_word" indicate that Llama-3.2-3B and MediPhi successfully constrained questions to maintain correct grammatical form and prevent answer words from appearing. In contrast, Qwen2.5-3B also scored highly on these metrics. Gemma-3-4B and FLAN-T5-XL, on the other hand, faced significant challenges with grammatical form and answer word inclusion. Moreover, while other models generated questions, FLAN-T5-XL failed to produce valid outputs. This suggests that model selection is critical when using AI to generate assessment questions. Finally, there is a critical observation regarding output stability, which is increasingly recognized as a primary requirement when applying LLMs to educational assessments. In this regard, Llama and Qwen scored highly on output stability. In contrast, while Mistral and MediPhi scored moderately, Gemma and FLAN-T5 scored very poorly. This is consistent with recent research suggesting that LoRA is most effective when applied to architectures explicitly pre-trained on instructional tasks.

Finally, the study used a data-enrichment pipeline. It included flattening the JSON structure of the raw data, using the Qwen2.5-7B model to generate cognitive labels, emphasizing key passage segments, and developing stringent templates for model outputs. The model's performance relied on this pipeline.

Despite these promising findings, there are also clear limitations: the experiments were conducted in a deterministic decoding setting, examined only models ranging from 3B to 7B in size, and used heuristic constraints that may not fully capture pedagogical quality. Overall, small- to medium-sized, instruction-tuned LLMs appear capable of generating cognitively aligned and structurally valid assessment items, putting models such as Llama and Qwen in a strong position for integration into various educational software solutions.

VIII. CONCLUSION AND FUTURE WORK

This study demonstrates the feasibility of generating cognitively aligned assessment questions efficiently using open-source, instruction-tuned language models, based on structured data augmentation. By incorporating evidence-based prompts and Bloom's taxonomy labels, the proposed framework improves both the structural validity and semantic alignment of the generated questions.

These results highlight the importance of selecting the appropriate architecture and designing effective prompts in

educational natural language processing systems, showing that decoder-only models are particularly well-suited for constrained generation tasks.

Future work should expand the scope of human evaluation of cognitive alignment by incorporating more detailed annotations and broader expert validation, developing bias-aware explanation strategies, and testing robustness under random decoding to further enhance educational reliability.

AUTHORS' CONTRIBUTIONS

Conceptualization, Mahmoud Badry and Shereen A. Taie; methodology, Mahmoud Badry; software, Mahmoud Badry; validation, Mahmoud Badry, Asmaa Sweidan, and Walaa Medhat; formal analysis, Mahmoud Badry; investigation, Mahmoud Badry; data curation, Mahmoud Badry; writing—original draft preparation, Mahmoud Badry and Asmaa Sweidan; writing—review and editing, Shereen A. Taie, Asmaa Sweidan, and Walaa Medhat; supervision, Shereen A. Taie and Asmaa Sweidan.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, TX, USA, 2016, pp. 2383–2392.

[2] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, "RACE: Large-Scale Reading Comprehension Dataset from Examinations," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark, 2017, pp. 785–794.

[3] G. Chen, J. Yang, C. Hauff, and G.-J. Houben, "LearningQ: A Large-Scale Dataset for Educational Question Generation," in *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, vol. 12, no. 1, Stanford, CA, USA, Jun. 2018.

[4] Qwen Team, "Qwen2.5: A Party of Foundation Models," https://qwenlm.github.io/blog/qwen2.5/, 2024, qwen Blog.

[5] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, Virtual Event, 2022.

[6] A. Dubey, "The Llama 3 Herd of Models," *arXiv preprint arXiv:2407.21783*, 2024.

[7] H. W. Chung, "Scaling Instruction-Finetuned Language Models," *arXiv preprint arXiv:2210.11416*, 2022.

[8] A. Hadifar, "EduQG: A Multi-Format Multiple-Choice Dataset for the Educational Domain," *IEEE Access*, vol. 11, pp. 20 885–20 896, 2023.

[9] O. Holl, "EduQuest: Lecture Texts and Questions for Higher Education," in *Proceedings of the 17th International Conference on Educational Data Mining (EDM)*, Atlanta, GA, USA, 2024.

[10] S. Lamsiyah, "Fine-Tuning a Large Language Model with Reinforcement Learning for Educational Question Generation," in *Proceedings of the International Conference on Artificial Intelligence in Education (AIED)*, 2024.

[11] M. A. Ehsan, A. S. Hasan, K. B. Shahnoor, and S. S. Tasneem, "Automatic Question and Answer Generation Using Generative Large Language Models (LLMs)," Ph.D. dissertation, Department of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh, 2024.

[12] Q. Zhuge, H. Wang, and X. Chen, "TwinStar: A Novel Design for Enhanced Test Question Generation Using Dual-LLM Engine," *Applied Sciences*, vol. 15, no. 6, p. 3055, 2025.

[13] P. Meyers, "Focal: A Proposed Method of Leveraging LLMs for Automating Assessments," in *Proceedings of the International Conference on Computers in Education (ICCE)*, 2023.

[14] S. García-Méndez, F. de Arriba-Pérez, and M. del C. Somoza-López, "A Review on the Use of Large Language Models as Virtual Tutors," *Science & Education*, vol. 34, no. 2, pp. 877–892, 2025.

[15] M. Stamatakis, "Enhancing the Learning Experience: Using Vision-Language Models to Generate Questions for Educational Videos," in *Proceedings of the International Conference on Artificial Intelligence in Education (AIED)*, 2025.

[16] N. Scaria, S. D. Chenna, and D. Subramani, "Automated Educational Question Generation at Different Bloom's Skill Levels Using Large Language Models: Strategies and Evaluation," in *Proceedings of the International Conference on Artificial Intelligence in Education (AIED)*, 2024.

[17] T. Huber and C. Niklaus, "LLMs Meet Bloom's Taxonomy: A Cognitive View on Large Language Model Evaluations," in *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, 2025.

[18] A. Q. Jiang, "Mistral 7B," *arXiv preprint arXiv:2310.06825*, 2023.

[19] J.-P. Corbeil, "A Modular Approach for Clinical SLMs Driven by Synthetic Data with Pre-Instruction Tuning, Model Merging, and Clinical-Tasks Alignment," *arXiv preprint arXiv:2505.10717*, 2025.

[20] Gemma Team, "Gemma 3 Technical Report," *arXiv preprint arXiv:2503.19786*, 2025.

[21] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020, https://arxiv.org/abs/1904.09675.

[22] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 3982–3992, https://arxiv.org/abs/1908.10084.

[23] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, USA, 2002, pp. 311–318, https://aclanthology.org/P02-1040/.

[24] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Proceedings of the Workshop on Text Summarization Branches Out*, Barcelona, Spain, Jul. 2004, pp. 74–81.

[25] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, MI, USA, 2005, https://aclanthology.org/W05-0909/.