STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **16**, July 2026, pp 494–509.
Published online in International Academic Press (www.IAPress.org)

Addressing Question Repetition in Academic Assessments: An Empirical Study and Generative AI-Based Solution

Zohair Elmourabit*, Asmaâ Retbi

RIME Team-MASI Laboratory-Mohammadia School of Engineers (EMI), Mohammed V University in Rabat, Morocco

Abstract Frequent reuse of exam questions harms the integrity of assessments and pushes students towards learning by automatism rather than understanding. It is a fundamental problem, yet the scientific literature has paid little attention to it so far. Our research addresses precisely this lack. Initially, we measured the concrete impact of this phenomenon through a survey among 194 actors from the academic community. Faced with this observation, we have developed an automatic question generation system based on the fine-tuned Transformers, in order to ensure a continuous renewal of the proposed topics. The results of the survey confirm the relevance of this approach: 80.4% of respondents state that they regularly encounter questions already seen, a situation that 63.4% of them consider penalising. In addition, 72.7% of participants advocate for an accelerated renewal of evaluation content. To meet this identified need, we designed an automatic question generation (AQG) architecture based on transformers, by fine-tuning the T5-small, T5-base and BART models on the SQuAD dataset. The comparative evaluation, supported by the BLEU, ROUGE and METEOR metrics as well as a multi-domain qualitative semantic analysis, established the constant superiority of the T5-based model over the other approaches (BLEU = 0.1765; ROUGE-1 = 0.5317; METEOR = 0.4085). These findings empirically validate the urgency of renewing assessments and demonstrate the effectiveness of transformer-based systems in ensuring diversity of tests, while easing teachers' workload. This study thus establishes the pedagogical necessity, as well as the technical feasibility, of an AI-assisted generation of questions in the service of educational equity.

Keywords Education, AI, Question Generation, Generative AI, Educational Assessment, Transformer Models, T5, BART, Academic Evaluation, Natural Language Processing

AMS 2010 subject classifications 97C70, 68T50, 62P25

DOI: 10.19139/soic-2310-5070-3577

1. Introduction

The art of questioning has always structured human thought. From the dialogues of Socrates to the contemporary works of Michel Meyer, the question has never been reduced to a simple linguistic tool: it is considered by researchers as the primary driver of knowledge acquisition [1, 2, 3]. However, in the context of contemporary educational assessment, the effectiveness of questioning is tested by the static nature of exam content. The integrity of evaluation is based above all on the validity of inferences drawn from the results, an issue that is all the more crucial in a context marked by the rapid evolution of digital tools [4, 5]. However, reusing proofs from one session to another seriously compromises their validity. This bias manifests in two ways: on the one hand, through an underrepresentation of the construct, when repetitive and limited questions fail to cover all learning objectives [6]; on the other hand, through a variance not relevant to the construct (CIV), where the student's score ends up reflecting their access to the annals rather than their actual skill [7, 8]. This is not just a psychometric problem, but a matter of fairness. The repetition of questions systematically disadvantages students who have no access to

*Correspondence to: Zohair Elmourabit (Email: z.elmourabit@research.emi.ac.ma). RIME Team-MASI Laboratory-Mohammadia School of Engineers (EMI), Mohammed V University in Rabat, Morocco

ISSN 2310-5070 (online) ISSN 2311-004X (print)
Copyright © 2026 International Academic Press

the ‘archives’ of past exams, disproportionately penalising first-generation learners and those from disadvantaged backgrounds [9]. Despite these proven risks, the phenomenon of “academic overfitting” where students memorize response patterns instead of mastering concepts remains largely unmeasured [6, 10].

Although research in recent years has focused on assessment security in the age of AI [11, 12] or on the technical aspects of Automatic Question Generation (AQG) [13, 14], there remains a notable gap in the literature. Indeed, no research has ever attempted to quantify the phenomenon of question repetition across disciplines while also offering a technically valid solution. Moreover, research on Automatic Question Generation often overlooks the educational aspects of the phenomenon, favoring a more algorithmic approach [15]. This research aims to fill this gap. Indeed, by relying on the research results, we aim to provide the necessary information for renewal systems [16]. To address these research challenges, this work aims to pursue two interconnected research objectives. On one side, the research aims are: the empirical aim. On the other side, the research aims are: the technical aim.

To address these challenges, our research is guided by three interconnected Research Questions (RQs):

- RQ1 (Empirical): To what extent does question repetition occur across academic disciplines, and how does it influence student learning strategies (the ”Academic Overfitting” effect)?
- RQ2 (Technical): Which Transformer-based architecture (T5-Small, T5-Base, or BART) is most effective at generating questions that maintain semantic consistency while providing the lexical variation needed to mitigate repetition?
- RQ3 (Integration): How can these technical models be integrated into a pedagogically sound ”Human-in-the-loop” workflow to ensure assessment validity?

The focus of this research is to fine-tune transformer models to ensure that they meet pedagogical standards and promote higher cognitive skills. The structure of this article is as follows: In Section II, we discuss the literature review regarding assessment validity.. In Section III, we explain our mixed-methods methodology, including a discussion of our survey and the fine-tuning of our models. In Section IV, we integrate our results, including the calculation of ”academic overfitting” and a comparison of our AI models. In Section V, we present our implications for promoting equity in education and suggestions for future work.

2. LITERATURE REVIEW

2.1. Assessment Validity and the Threat of Item Repetition

Current validity theory is based on the idea that “the meaning of test scores ultimately hinges on the appropriateness of the inferences and actions based on those scores” [17]. The Standards for Educational and Psychological Testing [18] warn against item overexposure: “When test content becomes known in advance, the validity of inferences based on test performance may be compromised” (p. 84). Research has clearly shown that retrieval practice greatly enhances our ability to remember information over time[19].It is important to note, however, that the mnemonic benefit of retrieval practice is contingent on genuine cognitive engagement. When students are repeatedly exposed to identical items, the cognitive load associated with active retrieval diminishes progressively, giving way to recognition-based processing that does not support deep encoding or long-term retention [20]. This shift from generative retrieval to passive recognition is not merely a learning inefficiency it represents a fundamental corruption of the assessment purpose, since scores begin to reflect familiarity with specific items rather than command of the underlying knowledge domain.

2.2. Automated Question Generation: Technical Evolution

To reduce manual effort, Automated Question Generation (AQG) technology has advanced significantly. Although initially rule-based AQG systems used syntactic templates and semantic parsing, these generated questions were only pedagogically shallow, being grammatically correct but lacking substance [22]. However, with the advent of neural approaches, especially sequence-to-sequence models with attention, these systems have greatly improved in terms of fluency and relevance[23]. As shown in Figure 1, Transformer-based models have set new performance standards in NLG systems. The introduction of self-attention mechanisms by Vaswani et al [24]. enabled parallel

processing, helping overcome the major drawbacks of recurrent models. The basic idea behind applying the self-attention mechanism is Scaled Dot-Product Attention, which computes attention for each token by pairwise interactions among queries (Q), keys (K), and values (V). The attention scores are computed as per Equation 1:

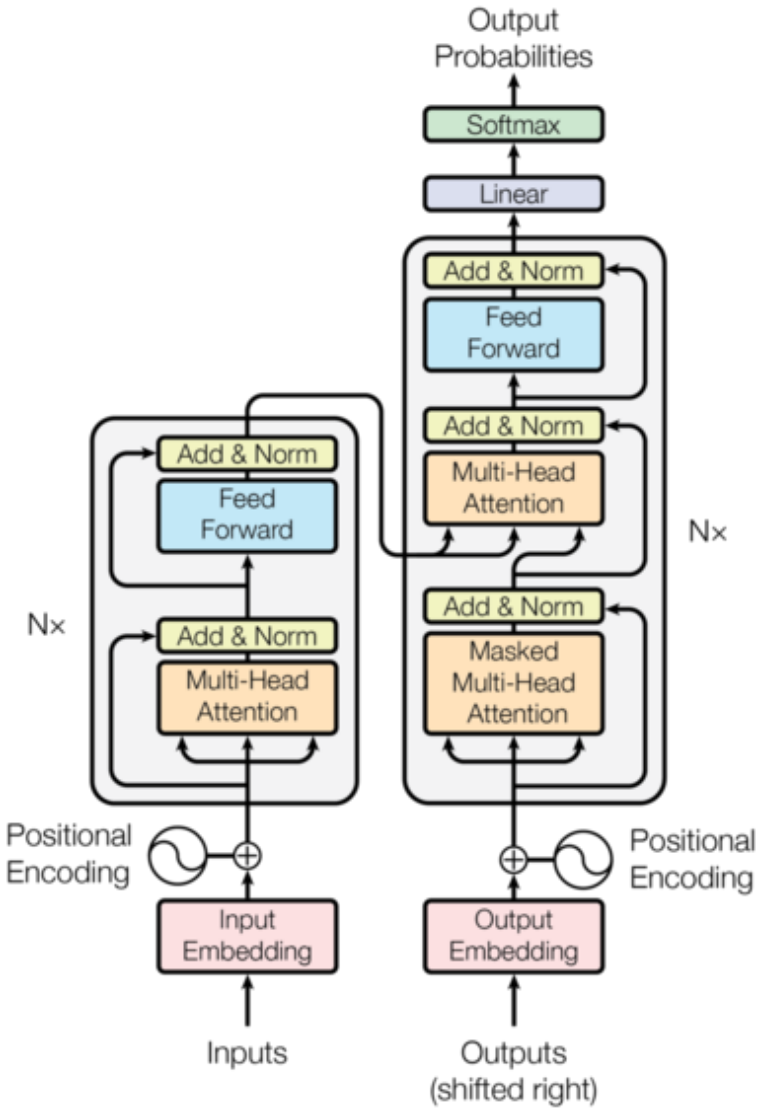

Figure 1. The Transformer architecture proposed by Vaswani et al. [24], utilizing self-attention mechanisms to process long-range dependencies.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Where d_k is the dimensionality of the keys, and the scaling factor $\frac{1}{\sqrt{d_k}}$ prevents gradient saturation in deep networks. To address the requirement of understanding the various contextual relationships, multiple attention heads are applied. Because transformer models lack an inherent understanding of sequence order, positional encoding is applied to the tokens. Following the introduction of the above concepts, there are currently three major styles of architectures used:

- Models like BERT [25], RoBERTa [26] are based on the concept of bidirectional context understanding through the masking of random tokens in the input sequence (MLM – Masked Language Modeling). These are incapable of generating output while performing understanding-based tasks like classification and sentiment analysis.
- Decoder-Only Models: (e.g., GPT series [27]): These architectures employ causal masking (left-to-right processing) to predict the next token in a sequence. While powerful for open-ended generation, they often struggle with tasks requiring bidirectional context understanding.
- Encoder-Decoder Models: Models like T5 [28], BART [29]) are based on the concept of unifying bidirectional context understanding into sequence-to-sequence tasks.

For question generation specifically, encoder-decoder architectures have demonstrated superior performance due to their explicit conditioning on source context [30].

2.3. Educational Applications and Pedagogical Constraints

Recent deployments of AQG in educational contexts highlight enduring challenges. Kurdi et al. in their systematic review [31] identified three fundamental aspects of quality: linguistic acceptability, semantic relevance, and pedagogical adequacy . According to Kumar et al. [32], fine-tuned T5 models were found to outperform standard models wwith contrastive learning objectives. However, their evaluation was based on technical indicators rather than pedagogical utility.

Crucially, existing works have been treating question generation as a purely technical optimization problem without empirical justification of the pedagogical need. While Muse et al. [33] demonstrated the empirical need for pre-training the question generation system with scientific texts for the improvement of question quality, Lamsiyah et al. [34] utilized reinforcement learning for the regulation of question difficulty. However, the question of whether the generated questions address the existing bottlenecks of evaluation remains unaddressed.

2.4. Research Gap Summary

Although the theory of assessment has evolved alongside the development of natural language generation techniques, three major gaps exist:

- **Empirical gap:** No prior study has quantified the phenomenon of question repetition with varying academic disciplines and stakeholder perspectives.
- **Integration gap:** While the technical advancement of the AQG problem has been conducted without empirical justification of the pedagogical need.
- **Comparative gap:** Although the comparison of the T5 and BART models for question generation has been conducted, the characteristics of the generated question have been under-explored.

Our study addresses all three gaps through an integrated empirical-technical research design that bridges the divide between psychometric validity concerns and generative AI capabilities.

3. METHODOLOGY

3.1. Study Design and Research Context

In our current research, we used a two-phase mixed-methods approach to address the complex issue of assessment repetition. In Phase 1 (Quantitative), we used a cross-sectional survey to determine the prevalence of “academic overfitting” and measure stakeholders’ perceptions. Cross-sectional design is particularly well-suited to our research problem, as it enables us to take a snapshot of the population and explore relationships between variables at a single specific point in time [35]. In Phase 2 (Technical), we adopted a comparative experimental design to evaluate three different architectures of transformers in automated question generation with standardized datasets and metrics in natural language processing. The two phases of this study are not independent but tightly connected.

The survey in Phase 1 defines the practical requirements that guide the technical design. In particular, the finding that 69% of repeated questions follow predictable surface patterns highlights the need for true semantic transformation rather than simple rewording. This directly motivated the use of encoder–decoder models such as T5 and BART, which generate questions based on both context and answers. The data collection took place within the 2025-2026 academic year at the Mohammed V University in Rabat, Morocco, which is a comprehensive public university with approximately 25,000 students in various academic programs. our research design conforms to the ethical standards of the research committee at the university and the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

3.2. Phase 1: Survey Methodology

3.2.1. Target Population and Sampling We identified the target population as people with direct and recent experience of academic assessment in higher education. This included enrolled students, instructors responsible for designing exams, and recent graduates now employed in engineering or technical fields.

To access the specialized population, we used the Moroccan Researcher Network (MRN) as the primary sampling frame. While the MRN effectively links diverse stakeholders, we acknowledge a potential sampling bias as the resulting participants skewed toward postgraduate respondents (67.1%). While this provides 'academically mature' insights from individuals with long-term exposure to assessment practices, it may not fully capture the specific experiences of early-stage undergraduate students.

In order to ensure that the sample is a true representation of the population, purposive sampling with quota allocation was used. The use of strict quotas was necessary to ensure that the sample is a true representation of the population. The population included 60% students, 15% faculty, and 25% professionals. The population was further divided by discipline: 35% engineering, 25% science, 15% computing, and 25% other disciplines. In order to determine the sample size, an a priori power analysis was done using G*Power 3.1.9.7 [36]. The required sample size for the study, considering a medium effect size (f2 = 0.15), is 92 to have 0.80 power. However, a sample size of 200 was used to account for survey non-completion. The required sample size of 194 was achieved, and we have 0.99 power, which exceeds the minimum required.

3.2.2. Instrument Development and Validation The survey instrument consists of 15 items that measure four distinct domains. The domains include "Demographics," "Repetition Prevalence," "Impact Perception," and "Renewal Preferences." The instrument development was very rigorous. The synthesis of the literature regarding the validity of the assessment was done first. The policies on assessment at the educational institution were also taken into consideration. After that, the instrument was refined, and inputs from three subject matter experts were incorporated. The experts included an educational measurement expert, a curriculum director, and an assessment coordinator. The survey was pilot-tested with 25 participants who were not included in the actual study. The pilot test also confirmed the instrument was "clear and concise." The average time it took to complete the instrument was a mere 8.3 minutes. The instrument was then evaluated by the experts for content validity. The experts evaluated the questionnaire and assessed the items' clarity and relevance. The content validity index was excellent at 0.93. The instrument's reliability was also evaluated. The instrument was highly reliable, with a reliability coefficient of 0.84. The item-total correlations ranged from 0.42 to 0.71.

3.2.3. Data Collection Procedures We used Google Form for the data collection because it works on all devices, and also for telephones, which made it easier for everyone to participate. The link to the form was shared through the groups and communities confident with the MRN. The data is collected over 6 weeks, from October 11, 2025, to November 15, 2025. To increase participation, the invitation to join this study has been relaunched weekly. These invitations were sent to 850 active members of the network. Among them, 194 responded completely (defined as ¡ 90% completion of the element), yielding a response rate of 22.8%. Although this rate is statistically powered at 0.99 to detect medium effects, we recognize it as a limitation that may restrict the broad generalizability of our findings across all diverse Moroccan higher education contexts. To eliminate potential non-response bias, we

compared early and late respondents (n = 67 and 42, respectively) on key demographic variables [38]. The chi-square tests did not reveal significant differences for academic role (x 2 = 2.14, p = 0.34) and discipline (x 2 = 4.67, p = 0.32), indicating that non-response did not bias our sample in any material way.

3.2.4. Ethical Considerations and Participant Protection In our study, we canceled the ethical review and approval because of the non-interventional nature of the anonymous survey, so all the procedures for the collection and use of data are in accordance with the institutional directives for the processing and confidentiality of participants' data. Before accessing the anonymous survey, each participant carefully read a digital informed consent page outlining the purpose of the study, the possibility to withdraw from the study at any time, and the confidentiality measures put in place. Participants were required to actively click "I agree to participate" to access the anonymous survey. We have strictly kept anonymity; no name, student ID, email address, or IP address has been collected. The data were stored in password-encrypted files that are only accessible to our research team. According to Moroccan law on data protection 09-08, the raw data will be retained and used during this study before being securely deleted.

3.2.5. Data Analysis To analyze the data quantitatively, descriptive statistics were used to characterize the sample and estimate prevalence. To study the links between variables, for example, whether engineering students report higher repetition rates than students in other disciplines, we used chi-square tests with effect sizes reported in the form of Cramer's V. For all statistical tests, the level of significance $\alpha = 0.05$ was set.

3.3. Phase 2: Technical Methodology (AI Model)

3.3.1. Dataset Preparation we used the Stanford Question Answer Dataset (SQuAD v1.1)[39] to train and evaluate our modules in the technical part of this study. SQuAD is a reading benchmark consisting of over 100,000 question-answer pairs derived from 536 high-quality Wikipedia articles.

We selected SQuAD v1.1 over the newer v2.0 because our primary objective is *generation* rather than *abstention*; v1.1 contains only answerable questions [39], which provides a cleaner signal for training generative models to produce valid interrogatives [23]. Our Dataset covers several themes, ranging from STEM fields to the humanities, ensuring that our model learns to generate applicable questions across various academic disciplines. As detailed in the table 1:

Table 1. Characteristics of the SQuAD Dataset Utilized in Training [39]

Aspect	**Details**
Source and Structure	Derived from Wikipedia articles across multiple domains. The dataset contains over 100,000 question-answer pairs (v1.1).
Thematic Coverage	**STEM:** Science, Technology, Medicine, Health Sciences, Geography. **Humanities:** History, Political Science, Religion, Philosophy, Arts. **Social Sciences:** Economics, Sociology, Psychology.

Before training our models, we performed data preprocessing. Our dataset is currently as follows: training sets (80%), validation (10%), and test (10%). To ensure the integrity of the technical evaluation and prevent data leakage, we implemented a context-level split. This ensures that no Wikipedia article (context) appearing in the training set was present in the test set. This rigorous separation forces the models to demonstrate true generalization capabilities on entirely unseen topics rather than merely memorizing patterns from familiar contexts.. We applied text normalization (lowercasing and tokenization) and filtered examples with more than 512 tokens to align with the input constraints of classical transformer architectures. The input was structured according to the conditional generation task:

$$Input : \text{answer: } [A] \text{ context: } [C] \rightarrow Target : [Q] \tag{2}$$

This framing forces the model to generate a question Q conditioned specifically on the answer A within the context C, preventing the generation of unanswerable or hallucinated questions.

3.3.2. Model Architectures A comparative analysis of three pre-trained transformer models was conducted, which have already demonstrated effectiveness in sequence-to-sequence problems. According to the specifications presented in Table 2, the models selected are those that balance efficiency with high semantic coherence.

Table 2. Specifications of Transformer Models Evaluated

Model	Params	Architecture	Pre-training Objective
T5-Small [28]	60M	Enc-Dec	**Span Corruption:** Treats all tasks as "text-to-text," predicting masked spans of text.
T5-Base [28]	220M	Enc-Dec	**Span Corruption:** Larger capacity for capturing complex syntactic structures.
BART-Base [29]	140M	Enc-Dec	**Denoising Autoencoder:** Reconstructs original text from corrupted input (shuffling/masking).

To evaluate the effect of model size (60M and 220M), the T5 (Text-to-Text Transfer Transformer) model family was used. To evaluate the effect of the denoising objective, which integrates BERT encoding and GPT decoding, the BART (Bidirectional Auto-Recurrent Transformers) model family was used.

3.3.3. Experimental Setup and Fine-Tuning All the models are fine-tuned using the Hugging Face library on an NVIDIA T4 GPU. The AdamW optimizer and a linear learning rate scheduler are utilized. To avoid overfitting, early stopping with a patience of 1 epoch is used. The specifications are presented in Table 3. Although the batch size and sequence length are kept constant, the learning rate is fine-tuned based on the model's pre-training objective.

Table 3. Hyperparameter Configuration by Model Architecture

Model	Learning Rate	Batch Size	Epochs	Input Length	Output Length
T5-Small	3×10^{-4}	8	3	512	64
T5-Base	3×10^{-4}	8	3	512	64
BART-Base	2×10^{-5}	8	3	512	64

3.3.4. Evaluation Metrics To thoroughly evaluate how well our generated questions hold up in terms of quality and linguistics, we use a three-part assessment that incorporates natural language processing metrics. This thorough method enables a detailed look at what the model produces, capturing both basic word similarities and deeper meanings and structures. Our goal is to synthesize diverse viewpoints so we can provide a fair and trustworthy evaluation of how well the generation performs.

1) BLEU (Bilingual Evaluation Understudy) BLEU serves as a measure of how accurately words line up, assessing how much the questions created by a machine match up with those written by humans. To calculate the metric, we take the geometric mean of the adjusted n-gram precisions and apply a brevity penalty BP, which helps avoid rewarding responses that are too short and could misleadingly boost the precision numbers. So, in technical terms, we can say that BLEU is defined as:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \quad BP = \min\left(1, e^{1-r/c}\right) \tag{3}$$

In this formulation, r and c denote the lengths of the reference and generated questions, respectively; w_n represents the weight assigned to each n-gram order; and p_n corresponds to the modified precision of matching n-grams [40]. The brevity penalty ensures that the metric balances precision with adequate output length.

2) ROUGE (Recall-Oriented Understudy for Gisting Evaluation) To complement the precision-oriented nature of BLEU, we incorporate ROUGE, which emphasizes recall and therefore evaluates how comprehensively the text captures the references. Specifically, we consider ROUGE-N, which measures n-gram overlap, and ROUGE-L, which relies on the longest common subsequence (LCS) to assess sentence-level structural similarity. The corresponding formulations are given by:

$$\text{ROUGE-N} = \frac{\sum_{S\in\{Ref\}}\sum_{gram_n\in S}\text{Countmatch}(gram_n)}{\sum S \in \{Ref\}\sum_{gram_n\in S}\text{Count}(gram_n)}, \quad \text{ROUGE-L} = \frac{LCS(Ref, Gen)}{|Ref|} \tag{4}$$

Through the use of the LCS, ROUGE-L captures structural alignment between sentences even when matching terms are not contiguous, thereby reflecting deeper syntactic correspondence beyond strict word adjacency [41].

3) METEOR To better understand the differences in meaning and avoid the pitfalls of just comparing surface-level features, we also use METEOR. Rather than just focusing on precise word matches, METEOR, which stands for Metric for Evaluation of Translation with Explicit Ordering, also considers factors like stemming and synonyms. This way, it can identify valid linguistic variations. We calculate the score using a harmonic mean of unigram precision P and recall R, putting more emphasis on recall. Additionally, it includes a fragmentation penalty that helps maintain the coherence of word order. The metric is defined as follows:

$$\text{METEOR} = F_{mean} \cdot (1 - \text{Penalty}), \quad F_{mean} = \frac{10PR}{R + 9P}, \quad \text{Penalty} = 0.5 \cdot \left(\frac{c}{u_m}\right)^3 \tag{5}$$

Here, c denotes the number of contiguous matched chunks, and u_m represents the total number of matched unigrams. The penalty term reduces the score when matches are excessively fragmented, thereby rewarding outputs that preserve logical flow and structural coherence [42]. While BLEU, ROUGE, and METEOR provide useful indicators of linguistic similarity and fluency, they do not directly capture pedagogical quality. In particular, these metrics do not assess whether a generated question aligns with specific learning objectives, cognitive difficulty levels, or assessment validity criteria. Therefore, their use in this study is complemented by qualitative analysis grounded in Bloom's Taxonomy[43] and pedagogical adequacy frameworks.

4. Experimental results and Discussion

4.1. Phase 1: Survey Findings (The Empirical Justification)

To establish the empirical motivation for our study, we analyzed the responses of 194 academic stakeholders. The results reveal a structural dependence on repeated questions, which compromises the validity and integrity of the evaluation.

4.1.1. Sample Characteristics Our final sample represents a highly qualified cross-sectional slice of the university community, with N = 194. Most of the participants in our survey were students (59.8%, n = 116), followed by engineering employees (26.8%, n = 52), and then faculty members (10.3%, n = 20). What is particularly notable is the high level of expertise represented by our sample. Table 4 below indicates that the overwhelming majority of our survey respondents, 67.1%, hold or are pursuing a postgraduate degree, including a PhD, an engineering degree,

or a master's degree. This predominance of postgraduate respondents strengthens the credibility of the reported perceptions, as these individuals bring extensive and sustained experience of academic assessment practices across multiple institutional contexts.

Table 4. Demographic Distribution of Survey Respondents ($N = 194$)

Category	**Count** (n)	**Percentage** (%)
Role		
Students	116	59.8%
Engineering Employees	52	26.8%
Faculty	20	10.3%
Job Seekers	6	3.1%
Education Level		
PhD (Student/Graduate)	64	33.0%
Bachelor's Degree	49	25.3%
Master's Level	43	22.2%
Engineering Degree	32	16.5%

4.1.2. Prevalence of Question Repetition Our results show that question repetition is not an isolated occurrence but the norm. Indeed, an astonishing 80.4% of the total sample reported the presence of question repetition with considerable regularity: 51.0% of the total sample answered that question repetition occurred "sometimes," and 29.4% responded with "very often." On the other hand, only 3.1% of the total sample reported never seeing repeated questions. Additional analysis revealed statistically significant variation across disciplines (χ^2 test, $p < 0.05$), with engineering and exact sciences reporting higher repetition rates compared to other fields. This pattern is consistent with the structured nature of problem-based assessments in technical domains. As demonstrated in Figure 2, the field of Engineering Sciences had the largest incidence of repetition, with n = 67. This is likely due to the static problem sets commonly employed in technical fields. This was followed closely by the field of Exact Sciences (n = 52), then by Computing (n = 19). As expected, the fields of study that rely on algorithm- or formula-based problem sets are the most stagnant.

4.1.3. The Academic Overfitting Phenomenon The most significant result of our survey is arguably the adaptation of the student body's behavior to the predictable repetition of questions. Indeed, 71.1% of the total sample revealed their regular review of previous examination materials as part of their academic preparation. This has given rise to what we call the Academic Overfitting Phenomenon. This has resulted in an alarming gap between the academic environment. Indeed, 63.4% of the total sample revealed their feeling of being at a disadvantage due to the academic environment. This is largely due to the "asymmetry of equity" of the academic environment. Moreover, 69% of question repetition instances were revealed to be part of a pattern, with 34.5% of these instances being "slight modifications" and 34.5% being "same structure, different data." There is therefore a strong need to resolve this issue technologically, with 72.7% of stakeholders calling for the renewal of questions. This calls for our proposed AQG solution, since renewal is evidently not sufficient. The survey findings serve as the foundational design requirements for our technical solution. Specifically, the fact that 69% of reported repetitions involve "slight modifications" or "same structure, different data" indicates that simple rule-based shuffling is insufficient. Mitigating "Academic Overfitting" requires a generative model capable of deep semantic paraphrasing and conceptual re-contextualization. Consequently, we selected Transformer models, specifically T5 and BART,

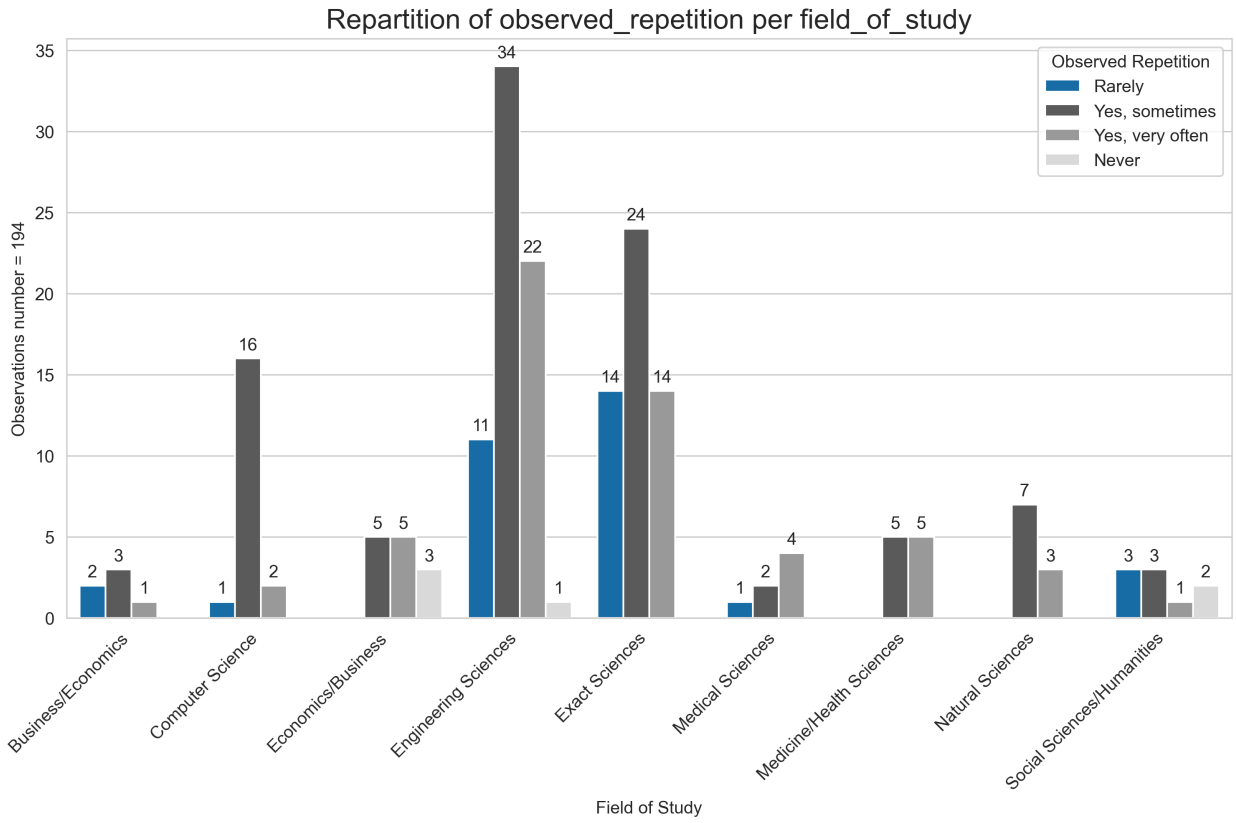

Figure 2. Frequency of reported question repetition disaggregated by academic discipline. Engineering ($n = 67$) and Exact Sciences ($n = 52$) exhibit the highest saturation of repeated assessment items.

due to their proven ability to handle complex sequence-to-sequence tasks that require more than surface-level lexical changes.

4.2. Phase 2: AQG Model Evaluation (The Technical Solution)

To meet the Phase 1 mandate in the assessment renewal process, we compared three transformer models: T5-Small, T5-Base, and BART-Base. The aim was to select one that could produce high-quality, legally accurate, and semantically robust questions.

4.2.1. Quantitative Performance (RQ3) A comparative evaluation was carried out using standard NLP evaluation metrics, as presented in Table 5. The evaluation revealed a distinct hierarchy in model performance. The model's size and pre-training objective significantly impact the quality of generated text.

Table 5. Comparative Performance of AQG Models on SQuAD Dataset

Model	**BLEU**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**METEOR**
T5-Small	0.1488	0.4905	0.2694	0.4514	0.3791
T5-Base	**0.1765**	**0.5317**	**0.3070**	**0.4905**	0.4085
BART-Base	0.1444	0.5020	0.2852	0.4671	**0.4094**

T5-Base performed better than other models in almost all structural evaluation metrics. It had the highest BLEU score (0.1765), indicating that T5-Base generates n-grams that best match human-written reference texts. Additionally, T5-Base scored highest in ROUGE-L (0.4905), which implies that its sentence structure is better than that of BART (0.4671) and T5-Small (0.4514).

BART-Base achieved lower precision but had the highest METEOR score (0.4094), consistent with its architecture, which is designed to be a denoising autoencoder.

4.2.2. Qualitative Semantic Analysis: A Framework-Grounded Evaluation To complement the automated metric evaluation, we conducted a structured qualitative analysis of model outputs across four academic domains: Programming, Literature, Economics, and Education. Rather than relying on the authors' subjective judgment, this analysis evaluates generated questions against three dimensions drawn from Kurdi et al.'s [31]validated framework for AQG quality assessment: (1) *linguistic acceptability* — grammatical correctness and natural phrasing; (2) *semantic relevance* — fidelity to the source passage and answer; and (3) *pedagogical adequacy* — alignment with recognizable cognitive demands, mapped to Bloom's Taxonomy levels [43, 13].

Programming Domain (Python List Comprehension) Given the passage: *"List comprehension in Python provides a concise way to create lists using a single line of expression."* Answer: *"a single line of expression."*

Table 6. Qualitative evaluation of generated questions — Programming domain

Model	Generated Question	Bloom's Level	Linguistic	Semantic
T5-Small	"What is the syntax structure of list comprehension?"	Remember	Acceptable	Partial
BART-Base	"What is the syntactical pattern employed in the construction of list comprehension expressions?"	Remember	Over-formal	Adequate
T5-Base	"What is the standard format used to write a list comprehension in Python?"	Understand	Natural	High

Table 7. Pedagogical adequacy assessment — Programming domain

Model	Pedagogical Assessment
T5-Small	Low — targets terminology recall only, without requiring the student to demonstrate operational understanding.
BART-Base	Low — unnecessarily complex phrasing risks confusing test-takers and does not improve cognitive demand.
T5-Base	Moderate — operationalizes the concept without reducing it to mere terminology; more consistent with how syntax comprehension is assessed in programming courses.

T5-Base's question targets the *application* of the format, mapping more naturally to the Understand/Apply levels of Bloom's Taxonomy and aligning more closely with standard practice in programming course assessment.

Education Domain (Bloom's Taxonomy) Drawing from a paragraph that describes the six progressive cognitive dimensions of Bloom's Taxonomy, each model generated a distinct question as follows:

Across both domains, T5-Base consistently generated questions that required more than simple recall, moving toward the Understand and Apply levels of Bloom's Taxonomy. T5-Small tended to produce terminological recall

Table 8. Qualitative evaluation of generated questions — Education domain

Model	Generated Question	Bloom's Level	Assessment
T5-Small	"What are the cognitive levels described in this framework?"	Remember	Grammatically correct but pedagogically shallow — listing does not assess understanding.
BART-Base	"What hierarchical cognitive levels comprise the taxonomic framework described?"	Remember	Jargon-heavy phrasing reduces accessibility and replicates the surface language of the passage rather than probing comprehension.
T5-Base	"How many cognitive levels does Bloom's Taxonomy include, and what is the function of each?"	Understand–Analyze	Pedagogically strongest — requires the student to enumerate *and* functionally distinguish levels, engaging higher-order thinking.

questions with limited cognitive depth, while BART-Base, despite its semantic richness, generated questions whose phrasing was excessively formal and occasionally misaligned with the vocabulary level appropriate for the target audience.

Limitations of the Qualitative Analysis We recognize that this analysis, while structured by an established framework, remains limited by the number of examples evaluated and the absence of external raters. A rigorous human evaluation study, in which a panel of domain educators blindly rates a larger random sample of generated questions on clarity, cognitive level, and answerability, would significantly strengthen these conclusions and is explicitly identified as a priority for future work (see Section 5).

4.2.3. Conclusion of Evaluation Through a comprehensive evaluation, **T5-Base** emerged as the superior model for educational applications. While T5-Small is efficient, it lacks depth, and while BART is semantically rich, it occasionally drifts into unnecessary formality. T5-Base provides the necessary balance of lexical precision (high BLEU) and pedagogical clarity required to automate the renewal of exam questions effectively. Despite strong overall performance, the models exhibit several limitations. In some cases, generated questions were overly generic, lacked sufficient specificity, or introduced minor semantic inconsistencies. Additionally, occasional hallucinations were observed when contextual cues were insufficient. These findings highlight the importance of incorporating validation mechanisms in practical deployment.

4.3. Discussion

The empirical findings of Phase 1 directly inform the design requirements of the proposed AQG system. In particular, the high prevalence of repetition patterns characterized by "slight modifications" (69%) indicates that simple surface-level variation is insufficient to ensure assessment validity. Instead, this pattern highlights the need for models capable of generating semantically diverse yet contextually consistent questions. Consequently, the selection of encoder–decoder architectures, and specifically T5-Base, is motivated by its ability to capture deep semantic relationships and produce meaningful variations beyond syntactic paraphrasing.

4.3.1. Interpretation of Survey Findings The empirical findings confirm that academic overfitting is not an isolated or anecdotal phenomenon but a structurally embedded feature of the current assessment landscape, reported by more than 80% of respondents. The finding that 63.4% of participants feel penalized by question repetition is

particularly striking because it inverts the conventional assumption that familiarity with assessment content benefits all students equally. As the data make clear, the advantage accrues selectively to those with access to archived exam materials, while students who lack such access, disproportionately first-generation learners and those from under-resourced backgrounds, are systematically disadvantaged. This asymmetry constitutes a direct threat to construct validity: when a student's score reflects their access to prior exam papers rather than their mastery of the assessed content, the inferences drawn from that score are no longer defensible. The gap between observed behavior and expressed preferences is equally telling. While 80.4% of respondents reported encountering repeated questions, 72.7% actively advocated for accelerated renewal of assessment content. This divergence suggests that the persistence of question repetition is not a reflection of stakeholder preference but rather of systemic inertia, an absence of practical tools to support renewal at scale. The cognitive science literature reinforces the urgency of this finding: Roediger and Butler[19] demonstrated that retrieval practice enhances retention only when it requires active cognitive engagement, and Cepeda et al. [21] showed that repeated exposure to identical items produces diminishing learning returns relative to varied but conceptually related material. A static question bank does not merely fail to challenge students, it actively undermines the long-term learning that assessment is designed to promote.

4.3.2. Technical Implications: Why T5-Base? The superiority of T5-Base is not only reflected in its quantitative performance but also in its alignment with the pedagogical requirements identified in Phase 1. Unlike models that primarily reproduce surface-level variations, T5-Base demonstrates a stronger capacity for semantic generalization, which is essential for generating novel questions that mitigate repetition effects. This capability directly addresses the observed need for deeper variation in assessment items, rather than simple reformulations of existing questions. It bears emphasizing that strong performance on BLEU, ROUGE, and METEOR does not constitute evidence of pedagogical validity. These metrics quantify surface-level linguistic alignment with human-written references; they do not assess whether a generated question targets a specific learning objective, operationalizes an appropriate cognitive level, or avoids construct-irrelevant variance [7, 8]. A question may achieve a high BLEU score by closely mirroring the syntactic structure of the reference while still being answerable by recognition rather than understanding precisely the failure mode that item repetition produces [6, 20]. T5-Base should therefore be understood as a viable generation engine whose output meets the threshold of surface-level linguistic quality, a necessary but not sufficient condition for deployment in formal assessment [17, 31]. A pedagogical review layer, as formalized in the Human-in-the-Loop workflow described in Section 4.3.5, remains an indispensable component of any responsible implementation [43, 15].

4.3.3. Integrated Contribution: Closing the Gap Between Pedagogical Need and Technical Capability The value of this study's dual-phase design lies not in the sum of its parts but in the direct correspondence between what the empirical findings reveal and what the technical solution is designed to address. Our empirical findings show that question repetition is largely systematic rather than random, with 69% following predictable surface patterns. This insight directly justifies the selection of T5-Base, as simpler approaches based on paraphrasing or template substitution would likely reproduce the very patterns that stakeholders already recognize and exploit [9, 6]. The model's performance (BLEU = 0.1765; ROUGE-L = 0.4905) confirms that it generates structurally distinct questions, supporting genuine novelty rather than superficial variation . On the technical side, our approach directly addresses the fairness concerns identified in the survey. Since 63.4% of respondents report feeling disadvantaged by repeated questions — often due to unequal access to past exam materials [9] — our system mitigates this imbalance by generating novel, semantically coherent questions independently of any archived materials. Moreover, T5-Base's conditioning on both context and answer helps limit hallucination risks [28], which is essential for maintaining assessment validity [17, 18]. We acknowledge, however, that this integration remains partial. While repetition is particularly prevalent in complex, multi-step problems in technical disciplines [6], the current system is primarily effective for reading-comprehension-style questions [39]. Extending this approach to domain-specific problem types remains a key direction for future work, as discussed in Sections 4.3.4 and 5.

4.3.4. Scope of the Current Solution and Path Toward Domain-Specific AQG The survey shows that question repetition is particularly pronounced in Engineering and Exact Sciences, where structured and formula-based problems are easier to reproduce. However, the AQG system evaluated here focuses on reading-comprehension-style questions derived from text, which are more representative of humanities and introductory courses than advanced STEM assessments. This represents a clear limitation. The current system cannot generate domain-specific problems such as thermodynamics or circuit analysis tasks. Rather, it demonstrates the feasibility of the core generation approach: producing linguistically sound and semantically varied questions using encoder–decoder models.

To extend this approach to STEM domains, future work should focus on (1) domain-specific training data, (2) support for mathematical and multi-step reasoning formats, and (3) alignment with pedagogical objectives and difficulty levels.

4.3.5. Proposed Implementation: A Human-in-the-Loop Workflow To ensure that AQG does not introduce "Construct-Irrelevant Variance," we propose a collaborative workflow rather than an autonomous one:

- AI Generation: The fine-tuned T5-Base model generates a pool of question variants based on a specific syllabus context.
- Pedagogical Review: An instructor reviews the pool, filtering for scientific accuracy and alignment with learning outcomes.
- Difficulty Calibration: The instructor selects and manually adjusts the complexity (e.g., adding distractors for MCQs).
- Deployment: The verified questions are randomized within the Learning Management System (LMS) to ensure no two students receive the identical repetition pattern.

5. Conclusion AND Future Work

This paper addresses the gap between theory and practice in education by presenting the first empirical investigation of question repetition in higher education. Our survey of 194 stakeholders showed that 80.4% of the academic community encounters repeated assessment items at regular intervals, creating the phenomenon of "Academic Overfitting," where the key to success lies in the availability of old papers rather than mastery of the subject. Considering that 72.7% of stakeholders wanted a quicker renewal of the assessment process, we developed a transformer-based framework for Automated Question Generation. We found that the T5-Base model is the best fit, striking a solid balance between lexical and semantic accuracy. Even though the survey results highlight some limitations related to people's perceptions and the differing national contexts, the overall findings support the idea that our proposed framework can indeed create valid, high-quality assessment items and help bring back the essential equity in the education system. In future work, we will extend the proposed framework by incorporating Automated Short Answer Grading models, adversarial AI detection mechanisms, and direct integration with Learning Management Systems. Despite its contributions, this study has several limitations. The survey relies on self-reported perceptions, which may not fully reflect actual repetition rates in practice. Additionally, the sampling frame, based on the Moroccan Researcher Network, may introduce bias toward academically advanced participants, as evidenced by the high proportion of postgraduate respondents (67.1%). These factors may limit the generalizability of the findings. Another limitation concerns the use of the SQuAD dataset, which is derived from Wikipedia and may not fully represent the complexity of real-world academic assessments, particularly in technical or domain-specific disciplines. Future work should focus on fine-tuning models on domain-specific educational datasets to improve contextual relevance and applicability.. This will utilize Item Response Theory to adapt the difficulty of assessment items based on student proficiency.

Acknowledgement

The authors express their sincere gratitude to the 194 students, faculty members, and engineering professionals who participated in the survey, providing the empirical foundation for this study. The survey instrument is available online.[†]

REFERENCES

1. Aristotle, *Topiques*, in Le Traité du dialogue, 2006.
2. M. Meyer, *Qu'est-ce que le questionnement?*, Paris: Vrin, 2017.
3. É. Benveniste, *Problèmes de linguistique générale*, Paris: Gallimard, 1966.
4. I. Levy-Feldman, *The Role of Assessment in Improving Education and Promoting Educational Equity*, Education Sciences, vol. 15, no. 2, p. 224, 2025.
5. M. N. A. Besar, K. H. Abd Aziz, and H. A. Halim, *The validity of Multiple-True-False and One-Best-Answer in the final professional undergraduate medical examination*, Education in Medicine Journal, vol. 17, no. 2, pp. 5–21, 2025.
6. M. Panczyk, A. Zarzeka, M. Malczyk, and J. Gotlib, *Does repetition of the same test questions in consecutive years affect their psychometric indicators?–Five-year analysis of in-house exams at Medical University of Warsaw*, EURASIA Journal of Mathematics, Science and Technology Education, vol. 14, no. 7, pp. 3301–3309, 2018.
7. X. Zhai, K. C. Haudek, C. Wilson, and M. Stuhlsatz, *A framework of construct-irrelevant variance for contextualized constructed response assessment*, Frontiers in Education, vol. 6, p. 751283, 2021.
8. T. Wongvorachan and O. Bulut, *Detecting Construct-Irrelevant Variance: A Comparison of Network Psychometrics and Traditional Psychometric Methods Using the HEXACO-PI Dataset*, Psychology International, vol. 7, no. 4, p. 88, 2025.
9. T. Fawns, *Identifying what our students have learned: a framework for assessment validity*, Assessment & Evaluation in Higher Education, 2026.
10. J. T. Davis, K. Adams, and A. Morgan, *The effect of exam retakes on future exam performance in a large classroom setting*, Advances in Physiology Education, vol. 48, no. 4, pp. 685–689, 2024.
11. K. Bittle and O. El-Gayar, *Generative AI and academic integrity in higher education: A systematic review and research agenda*, Information, vol. 16, no. 4, p. 296, 2025.
12. E. D. L. Evangelista, *Ensuring academic integrity in the age of ChatGPT: Rethinking exam design, assessment strategies, and ethical AI policies in higher education*, Contemporary Educational Technology, vol. 17, no. 1, p. ep559, 2025.
13. N. Scaria, S. Dharani Chenna, and D. Subramani, *Automated educational question generation at different Bloom's skill levels using large language models: Strategies and evaluation*, in International Conference on Artificial Intelligence in Education, Springer Nature Switzerland, Cham, pp. 165–179, 2024.
14. K. Boonkasem, T. Soonklang, and T. Supnithi, *Automatic question generation system for learning to create linear programming models*, Science, Engineering and Health Studies, 2025, Art. no. 25020004.
15. G. Ilieva, T. Yankova, M. Ruseva, and S. Kabaivanov, *A framework for generative AI-driven assessment in higher education*, Information, vol. 16, no. 6, p. 472, 2025.
16. P. Lorber, *Generative AI, law schools and assessment: where next?*, The Law Teacher, vol. 59, no. 3, pp. 372–390, 2025.
17. M. T. Kane, *Validating the Interpretations and Uses of Test Scores*, Journal of Educational Measurement, vol. 50, no. 1, pp. 1–73, 2013.
18. D. R. Eignor, *The standards for educational and psychological testing*, American Educational Research Association, 2013.
19. H. L. Roediger and A. C. Butler, *The critical role of retrieval practice in long-term retention*, Trends in Cognitive Sciences, vol. 15, no. 1, pp. 20–27, 2011.
20. E. L. Bjork and R. A. Bjork, *Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning*, in Psychology and the Real World, 2011.
21. N. J. Cepeda et al., *Spacing effects in learning: A temporal ridgeline of optimal retention*, Psychological Science, vol. 19, no. 11, pp. 1095–1102, 2008.
22. M. Heilman and N. A. Smith, *Good question! statistical ranking for question generation*, in Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 609–617, 2010.
23. X. Du, J. Shao, and C. Cardie, *Learning to ask: Neural question generation for reading comprehension*, in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1342–1352, 2017.
24. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and I. Polosukhin, *Attention is all you need*, Advances in Neural Information Processing Systems, vol. 30, 2017.
25. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, *BERT: Pre-training of deep bidirectional transformers for language understanding*, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long and Short Papers), pp. 4171–4186, 2019.
26. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, and V. Stoyanov, *RoBERTa: A robustly optimized BERT pretraining approach*, arXiv preprint arXiv:1907.11692, 2019.

[†]Survey data collection form: https://docs.google.com/forms/d/e/1FAIpQLSdO1nLcf-QUVRKmw_RGRkSrpUbTcPTayiNdUVmPhp-ExPTBZA/viewform?usp=dialog

27. T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, and D. Amodei, *Language models are few-shot learners*, Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901, 2020.
28. C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, and P. J. Liu, *Exploring the limits of transfer learning with a unified text-to-text transformer*, Journal of Machine Learning Research, vol. 21, no. 140, pp. 1–67, 2020.
29. M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, and L. Zettlemoyer, *BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7871–7880, 2020.
30. A. Ushio, F. Alva-Manchego, and J. Camacho-Collados, *Generative language models for paragraph-level question generation*, in Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 670–688, 2022.
31. G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, *A systematic review of automatic question generation for educational purposes*, International Journal of Artificial Intelligence in Education, vol. 30, no. 1, pp. 121–204, 2020.
32. S. Kumar, A. Chauhan, and P. Kumar C, *Learning enhancement using question-answer generation for e-book using contrastive fine-tuned T5*, in International Conference on Big Data Analytics, Springer Nature Switzerland, Cham, pp. 68–87, 2022.
33. H. Muse, S. Bulathwela, and E. Yilmaz, *Pre-training with scientific text improves educational question generation (student abstract)*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 13, pp. 16288–16289, 2023.
34. S. Lamsiyah, A. El Mahdaouy, A. Nourbakhsh, and C. Schommer, *Fine-tuning a large language model with reinforcement learning for educational question generation*, in International Conference on Artificial Intelligence in Education, Springer Nature Switzerland, Cham, pp. 424–438, 2024.
35. J. P. Takona, *Research design: qualitative, quantitative, and mixed methods approaches*, Quality & Quantity, vol. 58, no. 1, pp. 1011–1013, 2024.
36. F. Faul, E. Erdfelder, A. Buchner, and A. G. Lang, *Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses*, Behavior Research Methods, vol. 41, no. 4, pp. 1149–1160, 2009.
37. D. D. Nulty, *The adequacy of response rates to online and paper surveys: what can be done?*, Assessment & Evaluation in Higher Education, vol. 33, no. 3, pp. 301–314, 2008.
38. J. S. Armstrong and T. S. Overton, *Estimating nonresponse bias in mail surveys*, Journal of Marketing Research, vol. 14, no. 3, pp. 396–402, 1977.
39. 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, pp. 2383–2392, 2016.
40. 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, pp. 311–318, 2002.
41. C. Y. Lin, *ROUGE: A package for automatic evaluation of summaries*, in Text Summarization Branches Out, pp. 74–81, 2004.
42. 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, pp. 65–72, 2005.
43. B. S. Bloom, M. D. Engelhart, E. J. Furst, W. H. Hill, and D. R. Krathwohl, *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain*, David McKay Company, New York, 1956.