Edu Komputika 12 (2) (2025)

Edu Komputika Journal

https://journal.unnes.ac.id/journals/edukom/

Assessing Generative AI with Context-Augmented Zero-Shot Prompting for HOTS Question Generation Aligned with Bloom's Taxonomy

Saiful Ridlo[1)]✉, Ahmad Sehabuddin[2)], Syahroni Hidayat[3)], Taofan Ali Achmadi[4)], Uswatun Hasanah[5)], Indah Indi Afifah[6)], and Haikal Abror[7)]

[1]Natural Sciences Education Study Program, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Semarang, 50229, Indonesia

[2]Economic Education Study Program, Faculty of Economics and Business, Universitas Negeri Semarang, Semarang, 50229, Indonesia

[3] Electrical Engineering Study Program, Faculty of Engineering, Universitas Negeri Semarang, Semarang, 50229, Indonesia

[4]Family Welfare Education Study Program, Faculty of Engineering, Universitas Negeri Semarang, Semarang, 50229, Indonesia

[5]Computer Engineering Study Program, Faculty of Engineering, Universitas Negeri Semarang, Semarang, 50229, Indonesia

[6]Beauty Education Study Program, Faculty of Engineering, Universitas Negeri Semarang, Semarang, 50229, Indonesia

[7]Informatic and Computer Engineering Education Study Program, Faculty of Engineering, Universitas Negeri Semarang, Semarang, 50229, ndonesia

Article Info

Article History:
Received: 29 September 2025
Revised: 23 February 2026
Accepted: 26 February 2026

Keywords:
Automatic Question Generation, Bloom's Taxonomy, Context-Augmented Zero-Shot Prompting, Generative AI Higher Order Thinking Skills (HOTS).

Abstract

This study investigates the use of HOTS-GenAI, a generative AI system employing context-augmented zero-shot prompting, to automatically generate multiple-choice questions aligned with higher-order thinking skills (HOTS) in Bloom's taxonomy. A dataset of 200 items for vocational high schools was validated by three experts. The ground truth data demonstrated good quality with an inter-rater reliability of 0.75 (Gregory's Index). System performance across analysis (C4), evaluation (C5), and creation (C6) levels was evaluated using Content Validity Index (CVI), gap analysis, and confusion-matrix-based metrics. The findings revealed that HOTS-GenAI performed relatively well at the analytical level, where 70% of items met the HOTS threshold, supported by higher expert consensus. However, only 10% of items achieved the threshold for evaluation, and none for creation. CVI results indicated moderate validity overall, with stronger agreement for C4 than for C5 or C6. Confusion matrix analysis further confirmed this imbalance: accuracy and F1-scores were highest for analysis items but dropped sharply for evaluation and creation, where recall and precision were near zero. These results suggest that while HOTS-GenAI has potential in generating analytical questions, its capacity to model evaluative and creative tasks remains underdeveloped. Future research should involve larger datasets, refined prompt design, and more operational rubrics to enhance both validity and reliability in AI-generated HOTS assessments.

✉ Correspondence Address:
Faculty of Mathematics and Natural Sciences
UNNES, Sekaran, Gunung Pati, Semarang City, Central Java, 50229
E-mail: saiful_ridlo@mail.unnes.ac.id

ISSN 2252-6811
E-ISSN 2599-297X

INTRODUCTION

The development of Higher Order Thinking Skills (HOTS) has become a central priority in vocational education, especially in Indonesia's vocational high schools (SMK) (Sudana et al., 2020). Teachers are expected not only to prepare students with technical competencies but also to equip them with analytical, evaluative, and creative thinking skills aligned with Bloom's Taxonomy (Himawan & Suyata, 2023). In practice, however, many teachers still face difficulties in constructing questions that accurately represent HOTS indicators (Arfandi et al., 2021; Mulyono et al., 2019). This gap often leads to assessments dominated by lower-order thinking skills, which are easier to design but fail to stimulate students' deeper cognitive processes (Isnaeni et al., 2021).

Recent advances in generative artificial intelligence (GenAI) offer promising opportunities to address this challenge. Large language models, when combined with prompt engineering, have shown the ability to automatically generate educational content, including assessment questions (Lam et al., 2024; Pawar et al., 2024). Among various prompting strategies, zero-shot prompting has attracted attention because it allows AI to perform tasks without requiring explicit examples, making it potentially scalable for classroom assessment design (Lee et al., 2024). Despite this promise, the extent to which zero-shot prompting can produce valid HOTS-based questions suitable for vocational education remains underexplored.

Previous studies have examined the use of AI in question generation, particularly in higher education and language learning contexts (Bitew et al., 2024; Chan & Fan, 2019). However, these works often focused on surface-level quality measures such as fluency, grammatical correctness, or topical relevance, rather than on cognitive depth aligned with Bloom's Taxonomy (Atayolu & Kutlu, 2024). Moreover, most prior research has paid little attention to the distinction between HOTS and non-HOTS items, especially at the more demanding levels of Bloom's taxonomy, namely analysis (C4), evaluation (C5), and creation (C6). This indicates a critical gap in the literature, as the validity and reliability of AI-generated HOTS items are essential for meaningful assessment (Huber & Niklaus, 2025; Koto, 2025; Pawar et al., 2024).

Building on this gap, the present study evaluates the performance of Generative AI with context-augmented zero-shot prompting in generating HOTS-oriented questions for vocational high school contexts. The primary objective of this research is threefold. First, it seeks to evaluate the content validity of AI-generated questions across Bloom's higher levels analysis (C4), evaluation (C5), and creation (C6) using the Content Validity Index (CVI). Second, it aims to measure the system's success rate in achieving the HOTS threshold through Gap Analysis. Finally, this study intends to analyze the classification performance of the GenAI model in terms of accuracy, precision, recall, and F1-score by employing a confusion-matrix-based evaluation. By addressing these specific objectives, the study provides a comprehensive assessment of how GenAI can be effectively leveraged for designing cognitively demanding assessment items. Thus, this study not only tests a system, but also provides an empirical map of the 'cognitive limits' of generative AI and its implications for human-AI collaboration models in assessment design.

RESEARCH METHODS

The present study was conducted following the research procedure as illustrated in Figure 1, which outlines the systematic steps undertaken throughout the research process.

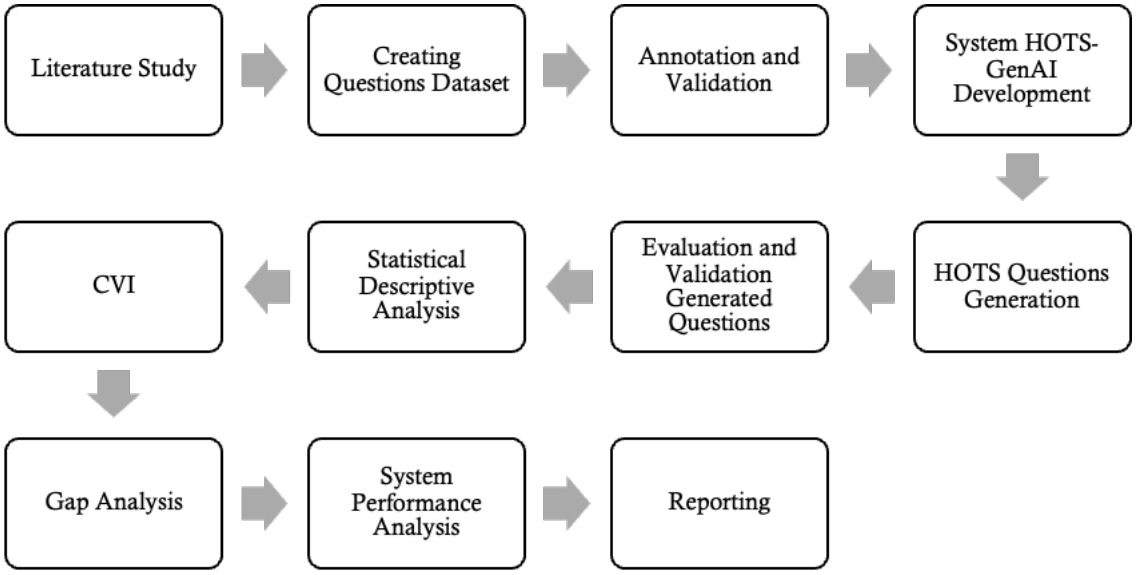

Figure 1. Research steps

A. Dataset

We compiled a corpus of approximately 1,000 multiple-choice questions drawn from past national and school examinations (Ujian Nasional and Ujian Sekolah) at vocational high schools (SMK). The dataset covered both general subjects (e.g., Indonesian Language, English) and vocational subjects (e.g., Office Administration, Accounting, Physics), each originally labeled according to HOTS/non-HOTS and Bloom's taxonomy levels (C1–C6). Only the question stems were extracted, excluding the answer options, to minimize bias related to distractor patterns. This approach aligns with the demand for HOTS-oriented assessment in vocational education and the practice of mapping Bloom's operational verbs in SMK test items ([Journal UNJ][1]). All examination items were obtained from publicly available online sources.

From this initial pool of more than 1,000 items, 200 questions were systematically selected to ensure feasibility and consistency in expert evaluation. These selected items underwent annotation by two evaluation experts, followed by validation procedures, and ultimately served as the learning documents for the development of the HOTS-GenAI system. This sampling strategy allowed the study to balance the richness of a large dataset with the practical need for rigorous expert review and analysis.

B. Annotation and Validation

The annotation process involved labeling the compiled corpus of questions based on Bloom's taxonomy levels (C1–C6) and categorization into HOTS and non-HOTS. Each question was annotated by two experts in educational evaluation, who employed operational verbs from the revised Bloom's taxonomy within the cognitive domain of vocational education to assign the appropriate C1–C6 labels. In addition, the experts also categorized each question as either HOTS or non-HOTS.

To verify the consistency of the assigned labels across Bloom's taxonomy levels and HOTS versus non-HOTS classification, Gregory's Index was applied. Gregory's Index is a measure of inter-rater reliability used to assess the degree of agreement between two raters when evaluating the same set of items (Retnawati, 2016). In educational research, it is frequently employed to validate labeling decisions for test items in relation to Bloom's taxonomy as well as HOTS classification (Arslan Mancar & Gülleroğlu, 2022). The principle of Gregory's Index is to compare the number of agreements and disagreements between raters, adjusting for the likelihood of agreement occurring by chance.

The mathematical formula for Gregory's Index is expressed as equation (1):

$$I_g = \frac{D}{D+B+C} \quad (1)$$

where D = the number of items labeled identically by both raters (agreement), B = the number of items labeled differently by the first rater but similarly by the second, and C = the number of items labeled similarly by the first rater but differently by the second. The value of Gregory's Index ranges from 0 to 1, with values

closer to 1 indicating a higher level of inter-rater agreement.

C. System HOTS Gen-AI

In this study, we developed and implemented a system referred to as HOTS-GenAI, which utilizes the Gemini 2.5 Pro reasoning model as its core engine (Pawar et al., 2024). To provide relevant domain-specific context for vocational education (SMK), we employed a knowledge injection strategy in the form of a single curated document summarizing the corpus. This approach was chosen to reflect the typical use case of teachers, who generally focus on a single subject domain.

The system was deployed in a controlled local computing environment to ensure reproducibility and to maintain a consistent evaluation setting (C. Shu et al., 2025). HOTS-GenAI was configured in a text-in/text-out mode, with low-to-moderate decoding temperatures (0.2–0.5) and top-p set to 0.9 to balance response consistency and diversity. The emphasis of this implementation was not on the technical specifications of Gemini 2.5 Pro itself, but on its practical capacity to generate assessment questions aligned with Bloom's taxonomy through carefully designed prompts.

The HOTS-GenAI system developed in this study consists of three main components: API integration, back-end processing, and front-end interface, as illustrated in Figure 2. This modular architecture was designed to ensure scalability, flexibility, and seamless communication between the generative AI model and the user-facing application.

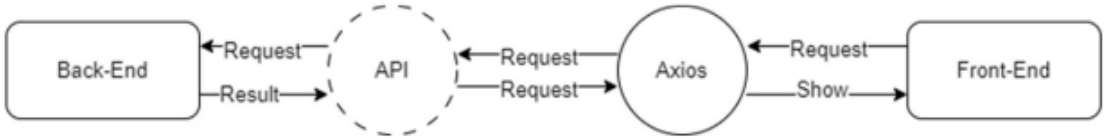

Figure 2. Illustrates the overall system architecture, showing the interaction between the API, back-end, and front-end components.

The research was conducted following this workflow, ensuring that each stage — from data input and model processing to output delivery — operates in a structured, integrated, and reproducible manner.

1. API Integration

At the core of the system, the question generation model utilizes Gemini 2.5 Pro, accessed through an API as the communication bridge between the developed interface and Google's Gemini service (Sallam et al., 2025). An API key is required to establish this connection, enabling the system to call the model directly and process user requests. Once integrated, the system can automatically generate HOTS-oriented questions by processing user-defined prompts. To narrow the domain and ensure subject-specific question generation, the system allows users to upload supporting documents such as learning modules or instructional materials. These documents are analyzed to identify relevant patterns, which guide the generation process to produce context-specific assessment items.

The API integration process begins with model configuration in Google AI Studio, including the addition of stop-words, prompt customization, token limit settings, and creativity adjustment to tailor the model's output to educational needs. Furthermore, the Cross-Lingual-Thought Prompting (XLT) technique was employed to enhance the quality of question generation. Upon successful configuration, the system generates an API key that can be utilized by the back-end, enabling real-time access to Gemini 2.5 Pro and facilitating adaptive question generation.

2. Back-End Processing

The back-end layer is responsible for transforming uploaded learning materials into automatically generated HOTS questions aligned with the targeted Bloom's taxonomy levels. This component was developed using Python, integrated with the configured Gemini 2.5 Pro API. Once the integration was established, the back-end was further developed using the Flask framework to manage incoming requests and outgoing responses efficiently. Beyond serving as a communication bridge between the model and the system, the back-end also manages data processing pipelines, ensuring structured, scalable, and efficient handling of input modules and generated output.

3. Front-End Interface

The front-end of HOTS-GenAI was designed to deliver a dynamic and interactive

web-based interface. It was developed using the Vue.js framework for the user interface and TailwindCSS for responsive and visually appealing design across devices. Data communication between the front-end and back-end was facilitated by Axios, which allows efficient data transmission via HTTP requests. Through this mechanism, the front-end sends instructional materials to the back-end, which processes them and returns generated questions in the form of HTTP responses. These are then rendered for users through the web interface, ensuring an intuitive and seamless user experience.

D. Prompt Design

In this study, the question generation process was guided by context-augmented zero-shot prompting (Caufield et al., 2024; Li, 2023). In this approach, the model was provided with concise textual instructions (zero-shot prompting) while being grounded in an external knowledge source a curated corpus of SMK exam questions stored in CSV format (Li, 2023). This strategy was chosen to reflect realistic classroom practice: teachers typically rely on simple prompts but benefit from a system that is contextually informed by domain-specific exam items.

The prompts instructed the system to generate multiple-choice question stems without answer options, aligned with Bloom's taxonomy levels and classified as HOTS or non-HOTS. Importantly, the instructions required the model to reference only those items that had been validated by both experts with the same HOTS and Bloom classification. This ensured that the generated questions were consistent with expert-validated exemplars while still being novel in formulation.

For experimental consistency, the system was configured to generate question sets of varying sizes (N = 6, 12, 18, 24, 30). These quantities were selected to allow balanced comparisons between HOTS and non-HOTS questions and to mirror the typical structure of multiple-choice examinations in Indonesian schools, which often consist of approximately 30 items. The following template illustrates the context-augmented zero-shot prompt design employed in this study:

"Now try to produce output in this format:
Generate multiple-choice questions (question stems only) directly from the given module, without introductory or closing sentences.

- *Create new questions based on the items approved by both experts, strictly referring to the CSV dataset provided.*
- *Each new item must indicate its reference, for example: "Question 1 is based on items 2, 6, 8 from the CSV data."*
- *Ensure that the referenced items are those validated by both experts with the same HOTS classification and Bloom's taxonomy level.*
- *Generate 10 multiple-choice questions."*

Despite containing detailed instructions, this approach is still categorized as zero-shot because the model is not given example pairs of questions and answers to imitate.

E. Evaluation

The evaluation was carried out on the questions generated by HOTS-GenAI in response to prompts designed to create HOTS-oriented items. To restrict the scope to specific subjects and learning sessions, a single lesson module was uploaded, and prompts were entered according to the predefined template. The generated questions were subsequently reviewed by three experts in educational evaluation. The experts were asked to determine whether each item qualified as a HOTS question and, if so, to which Bloom's taxonomy level analysis (C4), evaluation (C5), or creation (C6) it corresponded (Ahmed et al., 2025).

Each item was rated on a 5-point Likert scale, where 1 indicated Very Invalid (the item did not meet the expected criteria at all) and 5 indicated Very Valid (the item fully met the criteria and could serve as a model). The indicators for each taxonomy level are summarized in Table 1.

Table 1. Indicators for HOTS Question Evaluation

No	Indicator
1	HOTS – Analysis (C4): The item requires students to analyze information, break down ideas into smaller parts, and identify relationships (e.g., comparing, categorizing, organizing).
2	HOTS – Evaluation (C5): The item requires students to make judgments or decisions based on criteria and standards, and to provide justification (e.g., criticizing, defending an argument, validating).
3	HOTS – Creation (C6): The item requires students to formulate alternative solutions, design hypotheses, or predict outcomes of a given scenario.

The expert ratings were first subjected to descriptive statistical analysis to capture

measures of central tendency (mean, median) and variability (standard deviation, variance). This preliminary step provided an overview of rating distributions and potential discrepancies among raters before applying more specific validation measures. The 5-point Likert scores were then directly employed in three evaluation frameworks namely Content Validity Index (CVI) (Gupta et al., 2025), Gap Analysis (Aithal & Aithal, 2020), and Confusion Matrix to get the Accuracy (Acc), Precision (Pr), Recall (Re), and f1-score (F1) (P. Y. Reddy et al., 2025).

1. Content Validity Index (CVI)

CVI was used to quantify expert consensus on item relevance to HOTS indicators. Item-level CVI (I-CVI) for each question was calculated as equation (2) (Gupta et al., 2025).

$$I - CVI = \frac{n_e}{N} \quad (2)$$

where n_e is the number of experts assigning a score ≥ 4, and N is the total number of experts. Scale-level CVI (S-CVI/Ave) was then calculated as the mean of all I-CVI values using equation (3).

$$S - CVI = \frac{\sum_{i=1}^{k} I-CVI_i}{k} \quad (3)$$

with k is the total number of items. CVI values closer to 1 indicate stronger content validity.

2. Gap Analysis

Gap Analysis assessed the proportion of items that achieved the HOTS threshold. An item was considered HOTS if its mean expert score was ≥ 4 (Aithal & Aithal, 2020). The ratio was calculated as equation (4):

$$Gap\ Ratio = \frac{Number\ of\ items\ with\ mean \geq 4}{Total\ questions} \quad (4)$$

This represented the distribution of items that truly met HOTS criteria.

3. Confusion Matrix

Operationally, system outputs are mapped based on the intent of the generation process: any item generated by HOTS-GenAI using the HOTS-oriented prompt is a priori predicted by the system as HOTS. To measure the system's performance, a confusion matrix was constructed by comparing ground truth labels (based on majority expert judgment ≥4 = HOTS, <4 = non-HOTS) with system predictions (based on mean scores). The confusion matrix is illustrated in Table 2.

Tabel 2. Confusion Matrix

	Prediction HOTS	**Prediction Non-HOTS**
GT-HOTS	True Positive (TP)	False Negative (FN)
GT-Non-HOTS	False Positive (FP)	True Negative (TN)

From the matrix, standard performance metrics were derived, namely the Accuracy (Acc), Precision (Pr), Recall (Re), and f1-score (F1) using equations (5) – (8) (P. Y. Reddy et al., 2025).

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$Pr = \frac{TP}{TP+FP} \quad (6)$$

$$Re = \frac{TP}{TP+FN} \quad (7)$$

$$F1 = 2\ x\ \frac{Pr\ X\ Re}{Pr + Re} \quad (8)$$

This multi-step evaluation process—descriptive statistics, CVI, Gap Analysis, and confusion-matrix-based performance metrics—ensured a comprehensive assessment of HOTS-GenAI's ability to generate valid and reliable HOTS questions.

RESULT AND DISCUSSION

A. Dataset and Annotation

From the initial corpus of more than 1,000 exam questions, 200 items were systematically selected for expert annotation and validation. The inter-rater agreement between the two evaluators was measured using Gregory's Index, as summarized in Table 3.

Table 3. Inter-Rater Agreement using Gregory's Index

Type of Agreement	**Items Agreed**	**Total Items**	**Index**	**Interpretation**
HOTS & Bloom	150	200	0.750	Good
HOTS/Non-HOTS Only	173	200	0.865	Very Good
Bloom Taxonomy Only	153	200	0.765	Good

As shown in Table 3, the highest consistency was found in the general categorization of HOTS versus non-HOTS (86.5%), indicating that the experts were highly aligned in distinguishing between higher- and lower-order items. Agreement was also good when both HOTS classification and Bloom's taxonomy level were considered together (75%). Slightly lower agreement appeared in the specific assignment of Bloom's levels (76.5%), reflecting

natural variation in expert judgments when differentiating among the six cognitive stages.

Overall, these values fall within the good to very good range, confirming that the annotated corpus is sufficiently reliable to serve as training and validation input for the subsequent evaluation of the HOTS-GenAI system. The stronger consensus at the HOTS/non-HOTS level also justifies why the later stages of analysis primarily emphasize higher-order categories (C4–C6), where the intended cognitive depth is most relevant.

B. Descriptive Analysis

The descriptive statistical analysis revealed striking differences in expert scoring patterns across the three HOTS indicators shown in Table 4. For C4 (Analysis), the first rater (P1) consistently assigned lower scores, with a mean of 2.43, indicating skepticism regarding the analytical quality of the generated items. In contrast, raters two and three (P2 and P3) gave much higher average scores of 4.70, suggesting that they found most items sufficiently aligned with analytical thinking. This divergence implies that while HOTS-GenAI shows potential in generating analytical questions, expert consensus remains partial: acceptable for lenient evaluators but less convincing for a stricter standard.

The divergence became even more pronounced at C5 (Evaluation). P1 assigned very low scores (mean = 1.27), reflecting the perception that the items lacked explicit evaluative elements such as justification or judgment criteria. Meanwhile, P2 and P3 consistently scored items above 4.30, indicating strong confidence that the questions met evaluative standards. This polarization highlights the ambiguity in interpreting "evaluation" within Bloom's taxonomy, where the absence of a highly operational rubric may lead to starkly different judgments (see also Anderson & Krathwohl, 2001).

Table 4. The Descriptive Statistical Analysis of Expert Scoring Patterns Across The Three HOTS Indicators

Indicator	Evaluator (P)	Mean	Std Dev	Min	Median	Max	Varians
C4 Analysis	P1	2.43	0.73	2	2	5	0.53
	P2	4.70	0.47	4	5	5	0.22
	P3	4.70	0.47	4	5	5	0.22
C5 Evaluation	P1	1.27	0.52	1	1	3	0.27
	P2	4.37	0.49	4	4	5	0.24
	P3	4.37	0.49	4	4	5	0.24
C6 Creation	P1	1.23	0.43	1	1	2	0.19
	P2	4.00	0.00	4	4	4	0.00
	P3	4.00	0.00	4	4	4	0.00

The most extreme contrast appeared in C6 (Creation). P1 again assigned minimal scores (mean = 1.23), suggesting no item met the criteria for creative thinking, whereas P2 and P3 uniformly rated all items at 4.00, indicating that they considered every item to fulfill the creation dimension. Such full polarization raises a critical issue: whether the prompts used truly captured creative task features, or whether some raters interpreted any novel output as creative by default. This resonates with prior studies showing that AI-generated items often struggle to embody the authentic complexity of creation tasks (Stasaski et al., 2021).

Overall, the descriptive results indicate low inter-rater reliability. P1 consistently functioned as a strict evaluator with broader score variability, while P2 and P3 provided uniformly high ratings with low variance. This discrepancy demonstrates that the perceived quality of HOTS-GenAI outputs depends heavily on the evaluative stance of the rater. From an optimistic perspective (P2 and P3), the system appears already capable of producing acceptable HOTS questions, whereas from a conservative perspective (P1), the items fall short of higher-order requirements. These findings underscore the necessity of developing more operational evaluation rubrics, particularly for C5 and C6, to minimize subjectivity and achieve a more reliable assessment of GenAI's capacity to generate HOTS-aligned questions.

C. Content Validity Index (CVI)

The Content Validity Index (CVI) analysis revealed that the content validity of items generated by HOTS-GenAI remained in the moderate category, as shown in Table 5. Most

questions achieved an item-level CVI (I-CVI) of 0.67, which indicates that only two out of three experts judged them as relevant to HOTS indicators. Very few items reached full consensus (I-CVI = 1.00), and such cases were found exclusively in C4 (Analysis). For C5 (Evaluation) and C6 (Creation), almost no item obtained unanimous agreement. At the scale level, the mean CVI (S-CVI) was 0.69 for C4 and 0.67 for both C5 and C6, all below the commonly accepted threshold of 0.80 for adequate content validity (Polit & Beck, 2006).

This pattern underscores the sharp divergence among experts. The first rater (P1) consistently assigned low scores, preventing many items from reaching full agreement, particularly for evaluation and creation tasks. By contrast, the second and third raters (P2 and P3) provided more consistently high scores, enabling most items to exceed the majority threshold. This discrepancy reflects differing standards in interpreting what qualifies as a HOTS-aligned item, especially at higher cognitive levels where operational definitions are less straightforward.

Table 5. The Content Validity Index (CVI) analysis of items generated by HOTS-GenAI

Items	C4_I-CVI	C5_I-CVI	C6_I-CVI
1	0.67	0.67	0.67
2	0.67	0.67	0.67
3	0.67	0.67	0.67
4	0.67	0.67	0.67
…	…	…	…
12	**1.00**	0.67	0.67
14	**1.00**	0.67	0.67
…	…	…	…
30	0.67	0.67	0.67
S-CVI	**0.69**	**0.67**	**0.67**

The implications of these findings are twofold. First, HOTS-GenAI shows stronger performance in producing analytical items (C4), but its ability to generate evaluative (C5) and creative (C6) questions remains limited. This aligns with prior research showing that AI-based question generation tends to excel in analytical and recall-oriented tasks while struggling to represent evaluative judgment and creative problem-solving (Stasaski et al., 2021; Susanti et al., 2023). Second, the moderate CVI values highlight the need for more explicit design of stimuli to elicit evaluative reasoning or creative synthesis. In addition, operationalized scoring rubrics are essential to reduce subjectivity and improve inter-rater agreement, ensuring that future assessments of GenAI-generated items are more reliable.

Taken together, the CVI results suggest that while HOTS-GenAI demonstrates promise in generating analysis-level items, substantial refinement is required before it can reliably produce evaluation- and creation-level questions for educational assessment.

D. Gap Analysis

The gap analysis further highlights the uneven performance of HOTS-GenAI across different cognitive levels show in Table 6. At C4 (Analysis), 70% of the generated items exceeded the HOTS threshold (mean ≥4), suggesting that most questions successfully required students to analyze, compare, or organize information. This reinforces the earlier CVI findings, which also indicated stronger validity for analytical tasks.

In contrast, performance dropped sharply at C5 (Evaluation), where only 10% of the items met the HOTS criteria. This result indicates that most questions lacked explicit evaluative components such as justification, critique, or argument validation. The challenge of representing evaluation has also been documented in prior studies, where AI-generated items tended to prioritize surface-level reasoning over more nuanced evaluative judgment (Zhang et al., 2022; Susanti et al., 2023).

The most concerning outcome appeared at C6 (Creation), where no item reached the HOTS threshold. This means that HOTS-GenAI was entirely unable to produce questions requiring learners to hypothesize, predict outcomes, or design novel solutions. The complete absence of valid creation-level items echoes the polarization noted in the descriptive statistics, where expert judgments were split between consistently high and consistently low scores. It also underscores a fundamental limitation: while generative AI can imitate analytical structures, it struggles to instantiate tasks that embody creativity in an educationally rigorous sense (Stasaski et al., 2021).

Overall, the gap analysis demonstrates that HOTS-GenAI's strength lies in producing analysis-level items, but its capacity to generate valid evaluation- and creation-level questions remains extremely limited. This gap suggests the need for refined prompt design strategies, particularly with context augmentation explicitly geared toward evaluative reasoning and creative problem-solving. Without such enhancements, the system risks reinforcing lower- to mid-level HOTS (C4) while neglecting the higher-order dimensions (C5–C6) that are essential for fostering advanced cognitive skills.

Table 6. The Gap Analysis of 30 Questions Generated by HOTS-GenAI for item score >=4

Item	C4_Mean	C4_HOTS	C5_Mean	C5_HOTS	C6_Mean	C6_HOTS
1	4.00	1	3.33	0	3.33	0
2	4.33	1	4.00	1	3.33	0
3	4.33	1	4.00	1	3.33	0
4	4.33	1	4.00	1	3.33	0
5	3.33	0	3.33	0	3.33	0
6	4.33	1	3.00	0	3.00	0
7	4.00	1	3.00	0	3.00	0
8	3.33	0	3.00	0	3.00	0
9	3.33	0	3.67	0	3.00	0
10	3.33	0	3.67	0	3.00	0
11	3.33	0	3.67	0	3.00	0
12	4.67	1	3.33	0	3.33	0
13	4.00	1	3.00	0	3.00	0
14	5.00	1	3.67	0	3.33	0
15	4.00	1	3.00	0	3.00	0
16	4.33	1	3.00	0	3.00	0
17	4.33	1	3.00	0	3.00	0
18	4.00	1	3.67	0	3.00	0
19	4.00	1	3.67	0	3.00	0
20	4.00	1	3.00	0	3.00	0
21	4.33	1	3.67	0	3.00	0
22	4.33	1	3.67	0	3.00	0
23	4.00	1	3.67	0	3.00	0
24	4.00	1	3.00	0	3.00	0
25	3.33	0	3.00	0	3.00	0
26	3.33	0	3.00	0	3.00	0
27	4.00	1	3.00	0	3.00	0
28	3.33	0	3.00	0	3.00	0
29	4.00	1	3.00	0	3.00	0
30	3.33	0	3.00	0	3.00	0
HOTS Proportions	0.7	21	0.1	3	0	0

E. System Performance

The confusion-matrix analysis shown in Table 7 revealed significant variation in HOTS-GenAI's performance across cognitive levels. For C4 (Analysis), the system achieved an accuracy of 0.70 with perfect precision (1.00) and recall of 0.70, yielding an F1-score of 0.824. This indicates that every item classified as analytical HOTS by the system was correctly identified, yet around 30% of valid analytical items were missing. In other words, the system operated conservatively: it avoided false positives but at the cost of overlooking a considerable number of true HOTS items.

Table 7. The Confusions Matrix of 30 Questions Generated by HOTS-GenAI

Indicator	Acc	Pr	Re	F1
C4 – Analysis	0.70	1.00	0.70	0.824
C5 – Evaluation	0.10	1.00	0.10	0.182
C6 – Creation	0.00	0.00	0.00	0.000

Performance dropped sharply at C5 (Evaluation), where accuracy fell to 0.10. While precision remained high at 1.00, recall was only 0.10, producing a low F1-score of 0.182. This outcome suggests that the system recognized only a very small subset of evaluative items, failing to capture the majority. Such underperformance reflects the difficulty of modeling evaluative thinking, which requires explicit judgment criteria and justification—elements that are often underspecified in AI-generated questions (Zhang et al., 2022).

The most critical weakness emerged at C6 (Creation), where all performance metrics were zero. The system was entirely unable to identify creation-level items, indicating a complete breakdown in representing creative tasks such as hypothesis generation, prediction, or solution design. This mirrors earlier findings in the CVI and gap analyses, which likewise revealed a lack of valid creation-level items. Prior studies have similarly noted that generative AI tends to reproduce analytical structures but struggles to

embody authentic creative reasoning (Stasaski et al., 2021).

Taken together, these results suggest that HOTS-GenAI demonstrates clear potential in producing analysis-level questions but faces severe limitations at the evaluation and creation levels. Improving performance will require both refined prompt engineering to ensure evaluative and creative dimensions are explicitly embedded in the generated tasks and more operational rubrics for expert validation, which can help align evaluators' judgments and provide more reliable feedback to guide system refinement (Sallam et al., 2025).

CONCLUSION

This study evaluated the capacity of HOTS-GenAI, a generative AI system applying context-augmented zero-shot prompting, to produce multiple-choice questions aligned with Bloom's higher-order thinking skills (C4–C6). The results demonstrated that while the system showed promise in generating analysis-level items (C4), its performance at evaluation (C5) and creation (C6) levels remained limited. The Content Validity Index indicated only moderate agreement among experts, with consensus strongest for analytical tasks and weakest for evaluative and creative dimensions. A gap analysis revealed that 70% of items met the HOTS threshold at C4, compared to only 10% at C5 and none at C6. Confusion-matrix-based metrics reinforced these findings, highlighting relatively high accuracy for analytical items but severe deficiencies at the evaluative and creative levels.

These results underscore both the potential and the limitations of current AI-based question generation. Future work should expand the dataset beyond the 200 items used in this study to improve generalizability and robustness. Additionally, prompt engineering strategies must be refined to embed more explicit evaluative and creative elements, and expert rubrics should be operationalized to reduce subjectivity in scoring. Ultimately, our findings advocate for a fundamental paradigm shift from viewing AI merely as an independent "generator" to positioning it as a "collaborative assistant" in the assessment design process. By embracing this collaborative model, HOTS-GenAI can be further developed into a more reliable tool that enhances, rather than replaces, the teacher's expertise in crafting cognitively demanding assessments for vocational education.

ACKNOWLEDGEMENTS

This research was supported by the Directorate of Research and Community Service (DPPM) under the Ministry of Higher Education, Science, and Technology (Kemdiktisaintek) through the 2025 Research Grant, Contract Number. 104.2.6/UN37/PPK.11/2025.

REFERENCES

Ahmed, A., Kerr, E., & O'Malley, A. (2025). Quality assurance and validity of AI-generated single best answer questions. *BMC Medical Education*, *25*(1), 300. https://doi.org/10.1186/s12909-025-06881-w

Aithal, A., & Aithal, P. (2020). Development and validation of survey questionnaire & experimental data–a systematical review-based statistical approach. *International Journal of Management, Technology, and Social Sciences (IJMTS)*, *5*(2), 233–251.

Arfandi, A., SURYANI, H., Panennungi, P., & Sampebua, O. (2021). The Ability of Vocational High School Teachers to Developing HOTS Question. *International Journal of Environment, Engineering, and Education*, *3*(3), 1–6.

Arslan Mancar, S., & Gülleroğlu, H. D. (2022). Comparison of Inter-Rater Reliability Techniques in Performance-Based Assessment. *International Journal of Assessment Tools in Education*, *9*(2), 515–533. https://doi.org/10.21449/ijate.993805

Atayolu, Y., & Kutlu, Y. (2024). Similarity and Classification Analysis in Turkish Question Generation of AI tools According to Bloom's Taxonomy. *Tethys Env. Sci*, *1*(2), 87–98.

Bitew, S. K., Hadifar, A., Sterckx, L., Deleu, J., Develder, C., & Demeester, T. (2024). Learning to Reuse Distractors to Support Multiple-Choice Question Generation in Education. *IEEE Transactions on Learning Technologies*, *17*, 375–390. https://doi.org/10.1109/TLT.2022.3226523

C. Shu, N. Yao, Y. Chen, V. Wijeratne, L. Ma, J. Loo, K. K. Chai, A. Alam, & A. Abuelmaatti. (2025). Ai-Assisted

Multiple-Choice Questions Generation with Multimodal Large Language Models in Engineering Higher Education. *2025 IEEE Global Engineering Education Conference (EDUCON)*, 1–9. https://doi.org/10.1109/EDUCON62633.2025.11016449

Caufield, J. H., Hegde, H., Emonet, V., Harris, N. L., Joachimiak, M. P., Matentzoglu, N., Kim, H., Moxon, S., Reese, J. T., Haendel, M. A., Robinson, P. N., & Mungall, C. J. (2024). Structured Prompt Interrogation and Recursive Extraction of Semantics (SPIRES): A method for populating knowledge bases using zero-shot learning. *Bioinformatics*, *40*(3), btae104. https://doi.org/10.1093/bioinformatics/btae104

Chan, Y.-H., & Fan, Y.-C. (2019). A Recurrent BERT-based Model for Question Generation. *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, 154–162. https://doi.org/10.18653/v1/D19-5821

Gupta, S., Grover, S., Menon, V., Indu, P., Vidhukumar, K., & Chacko, D. (2025). Item generation and establishing face and content validity of a rating scale: A primer. *Indian Journal of Psychiatry*, *67*(8). https://journals.lww.com/indianjpsychiatry/fulltext/2025/08000/item_generation_and_establishing_face_and_content.12.aspx

Himawan, R., & Suyata, P. (2023). Analisis Sebaran Level Kognitif HOTS Berdasarkan Taksonomi Bloom pada Soal Penilaian Harian Materi Teks Pidato Persuasif di SMPN 1 Bambanglipuro Bantul. *Stilistika: Jurnal Pendidikan Bahasa Dan Sastra*, *16*(1), 89–100.

Huber, T., & Niklaus, C. (2025). *LLMs meet Bloom's Taxonomy: A Cognitive View on Large Language Model Evaluations*.

Isnaeni, W., Rudyatmi, E., Ridlo, S., Ingesti, S., & Adiani, L. R. (2021). Improving students' communication skills and critical thinking ability with ICT-oriented problem-based learning and the assessment instruments with HOTS criteria on the immune system material. *Journal of Physics: Conference Series*, *1918*(5), 052048. https://doi.org/10.1088/1742-6596/1918/5/052048

Koto, F. (2025). *Cracking the Code: Multi-domain LLM Evaluation on Real-World Professional Exams in Indonesia* (No. arXiv:2409.08564). arXiv. https://doi.org/10.48550/arXiv.2409.08564

Lam, Y. Y., Chu, S. K. W., Ong, E. L. C., Suen, W. W. L., Xu, L., Lam, L. C. L., & Wong, S. M. Y. (2024). Comparative Study of GenAI (ChatGPT) vs. Human in Generating Multiple Choice Questions Based on the PIRLS Reading Assessment Framework. *Proceedings of the Association for Information Science and Technology*, *61*(1), 537–540.

Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., & Kim, H. (2024). Few-shot is enough: Exploring ChatGPT prompt engineering method for automatic question generation in english education. *Education and Information Technologies*, *29*(9), 11483–11515.

Li, Y. (2023). A Practical Survey on Zero-shot Prompt Design for In-context Learning. *Proceedings of the Conference Recent Advances in Natural Language Processing - Large Language Models for Natural Language Processings*, 641–647. https://doi.org/10.26615/978-954-452-092-2_069

Mulyono, H., Istiyati, S., Atmojo, I. R. W., & Ardiansyah, R. (2019). *Kompetensi Guru Dalam Penyusunan Soal Higher Order Thinking Skills (HOTS) Berbasis Critical Thinking Sesuai Kurikulum 2013 Guna Mengakselerasi Education 4.0*. *7*(2), 108–1111.

P. Y. Reddy, N. S. Satya Aarthi, S. Pooja, B. Veerendra, & S. D. (2025). A Deep Learning Approach for Identifying LLM-Generated Text. *2025 International Conference on Artificial Intelligence and Data Engineering (AIDE)*, 359–364. https://doi.org/10.1109/AIDE64228.2025.10987319

Pawar, P., Dube, R., Joshi, A., Gulhane, Z., & Patil, R. (2024). Automated Generation and Evaluation of MultipleChoice Quizzes using Langchain and Gemini LLM. *2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT)*, *1*, 1–7. https://doi.org/10.1109/ICEECT61758.2024.10739326

Retnawati, H. (2016). Proving content validity of self-regulated learning scale (The comparison of Aiken index and

expanded Gregory index). *REID (Research and Evaluation in Education)*, *2*(2), 155–164. https://doi.org/10.21831/reid.v2i2.11029

Sallam, M., Snygg, J., Hamdan, A., Allam, D., Kassem, R., & Damani, M. (2025). Evaluating the Performance of Seven Large Language Models (GPT4.5, Gemini, Copilot, Claude, Perplexity, DeepSeek, and Manus) in Answering Healthcare Quality Management Inquiries. *Research and Advances in Education*, *4*(4), 39–50. https://doi.org/10.63593/RAE.2788-7057.2025.05.005

Sudana, I. M., Oktarina, N., Apriyani, D., & Achmadi, T. A. (2020). An Implementation of HOTS Based Learning Strategy in Vocational High Schools. *International Journal of Innovation*, *13*(12).