Id: 1906667

Participatory Educational Research (PER)
Vol.13(6), pp. 77-100, November 2026
Available online at http://www.perjournal.com
ISSN: 2148-6123
http://dx.doi.org/10.17275/per.26.79.13.6

Artificial Intelligence's Geometry Exam: Competencies in Generating Geometry Questions According to Bloom's Taxonomy

Gizem Ünsal*
Department of Mathematics and Science Education, Faculty of Education, Tokat Gaziosmanpaşa University, Tokat, Türkiye
ORCID: 0000-0002-4379-423X

Behice Nur Kalkan
Ministry of National Education, Ankara, Türkiye
ORCID: 0009-0003-5017-6346

Nazan Sezen Yüksel
Department of Mathematics and Science Education, Faculty of Education, Hacettepe University, Ankara, Türkiye
ORCID: 0000-0002-0539-3785

Article history

Received:
10.03.2026

Received in revised form:
01.06.2026

Accepted:
08.07.2026

Key words:
generative AI; bloom taxonomy; middle school mathematics; geometry; question generation

This study examines the extent to which generative artificial intelligence tools (ChatGPT, Gemini, and DeepSeek) can generate middle school geometry questions aligned with the cognitive levels of Bloom's Revised Taxonomy. The study also evaluates the quality and taxonomic appropriateness of these questions based on expert judgments. A convergent parallel mixed-methods design was employed. A total of 30 geometry questions, including open-ended and multiple-choice items, were generated by the AI tools using structured prompts. Ten mathematics education experts evaluated each question using a rubric that assessed alignment with the intended Bloom level. Quantitative data were analyzed using frequencies, percentages, and Gwet's AC1 agreement coefficient, while qualitative data were analyzed through thematic content analysis to identify patterns in expert justifications. The findings indicate that all three AI tools perform relatively well at lower and middle cognitive levels (understanding, applying, and analyzing), whereas their performance declines at higher levels (evaluating and creating). DeepSeek and Gemini showed higher overall alignment with targeted cognitive levels compared to ChatGPT. However, expert agreement decreased significantly at higher-order levels, indicating inconsistencies in how these levels are interpreted and represented. Qualitative findings revealed issues related to insufficient cognitive depth, weak distractors in multiple-choice items, and mismatches between intended and actual cognitive levels. Overall, the results highlight both the potential and limitations of generative AI in educational assessment and emphasize the necessity of expert supervision in ensuring pedagogical and taxonomic accuracy.

* Correspondency: gizemunsal06@gmail.com

Introduction

In education, writing high-quality questions plays a vital role in revealing students' cognitive processes and supporting deep understanding and critical thinking (Adams, 2015; Kogce & Baki, 2009). Exam and activity questions assess students' cognitive processes, and Bloom's taxonomy is widely used to classify questions according to these levels (Adams, 2015).

Preparing exam questions that reflect different cognitive levels requires considerable time, subject knowledge, and pedagogical attention from teachers (Koman et al., 2023). In this context, artificial intelligence-based tools have significant potential to reduce the time spent developing materials such as assignments, exams, and tests (Elkins et al., 2024; Wankhede & Kiwelekar, 2016). However, the need to verify the accuracy, ethical appropriateness, and pedagogical alignment of AI-generated content requires teachers to develop AI literacy (Eren et al., 2025; Zhang et al., 2024).

Generative Artificial Intelligence and Automatic Question Generation

The integration of generative artificial intelligence technologies into the education ecosystem has the potential to fundamentally transform traditional measurement and evaluation processes (Yosumaz, 2025). In particular, the speed and scalability of large language models in question generation offer significant opportunities to reduce educators' workload (Owan et al., 2023; Scaria et al., 2024). In the context of mathematics education, these tools' ability to construct and solve complex problems is noteworthy for their capacity to support students' analytical thinking skills (McGalliard & Otten, 2024; Nayıroglu & Tutak, 2023). However, the effectiveness of artificial intelligence in producing educational materials has also sparked academic debates about the pedagogical quality of the content and its compatibility with teaching programs (Gopal et al., 2024; Liu et al., 2026).

In automated question-generation processes, it is critical that questions not only be technically accurate but also exhibit a certain cognitive depth (Owan et al., 2023). At this point, Bloom's Revised Taxonomy provides a hierarchical framework for classifying questions (Maity & Deroy, 2024; Yaacoub et al., 2025a). Previous studies show that AI tools perform well at lower cognitive levels but are less consistent at higher levels such as analyzing, evaluating, and creating (Akdeniz et al., 2025; Maity & Deroy, 2024). This necessitates an in-depth examination of the taxonomic consistency of AI-generated items and the extent to which they serve the intended learning outcomes (Maity & Deroy, 2024).

In the measurement and evaluation literature, it is emphasized that the psychometric properties of items generated by artificial intelligence tools, such as item difficulty, discriminative power, and validity, must be rigorously evaluated (Bhandari et al., 2024; Li et al., 2025). The fact that artificial intelligence can sometimes produce erroneous content, referred to as "hallucinations," or fail to solve complex geometric problems that require mathematical reasoning, highlights the importance of "human-in-the-loop" approaches (Chen, 2025; Li et al., 2025). The oversight of AI outputs by subject matter experts and educators regarding pedagogical value and measurement quality is an indispensable step for these tools to become reliable teaching assistants (Grévisse et al., 2024; Li et al., 2025).

The literature indicates that while generative artificial intelligence tools are efficient in generating question stems, they face important limitations. These include difficulties in adhering to question-writing principles, accurately distinguishing between cognitive levels,

and producing high-quality distractors, particularly in multiple-choice items (Grevisse et al., 2024; Maity & Deroy, 2024). For example, teacher-expert-LLM collaborative approaches report that two-stage designs are needed, where correct-answer generation is separated from distractor and feedback generation (Chen, 2025; Li et al., 2025); human expertise remains decisive in aligning distractors with common student errors. A significant proportion of current studies has been conducted in the context of health sciences and at the higher education level; comparative studies examining question generation aligned with national curriculum outcomes and targeted at Bloom's cognitive levels at the middle school level remain more limited.

Therefore, it is important to examine whether current AI tools such as ChatGPT, Gemini, and DeepSeek can generate geometry questions aligned with Bloom's Revised Taxonomy and to evaluate their quality through expert judgment (Akdeniz et al., 2025; Gündeger Kılcı, 2025; Kotsis, 2025; Maity & Deroy, 2024; Omopekunola & Kardanova, 2024; Rahman et al., 2025; Yaacoub et al., 2025b). The primary purpose of this study is to evaluate the capacity of these generative AI models to construct middle school geometry questions aligned with Bloom's Revised Taxonomy and to assess their pedagogical quality through expert review. Accordingly, the research questions are: (RQ1) To what extent do ChatGPT, Gemini, and DeepSeek generate middle school geometry questions that are aligned with the cognitive levels of Bloom's Revised Taxonomy across different question types (open-ended and multiple-choice)? (RQ2) How do mathematics education experts evaluate the taxonomic alignment, content quality, and pedagogical appropriateness of AI-generated geometry questions?

Methodology

Research Design

In this study, a mixed-methods design was used to thoroughly examine the taxonomic alignment and pedagogical appropriateness of 5th-grade middle school mathematics questions generated by different artificial intelligence tools in the context of expert opinions. In order to comprehensively address the research questions, this study employed a convergent parallel mixed-methods design (Creswell & Plano Clark, 2017) within an evaluation research framework. This specific design was selected because it enables the simultaneous collection and independent analysis of both quantitative and qualitative data, which are then merged for a holistic interpretation. To answer the first research question (RQ1), quantitative data were gathered by measuring the alignment of AI-generated questions with Bloom's Taxonomy, utilizing binary expert ratings and statistical agreement coefficients (Gwet's AC1). Concurrently, to address the second research question (RQ2), qualitative data were collected through the experts' written justifications on the same evaluation rubric. In accordance with the convergent parallel approach, the numerical scores and in-depth qualitative feedback were merged during the interpretation phase. This allowed the qualitative content analysis to provide a deeper pedagogical context for the statistical performance metrics of the AI tools, ensuring a comprehensive evaluation. During this data collection process, 10 mathematics education experts evaluated the alignment of the generated questions with the curriculum outcomes and the cognitive process levels of Bloom's Taxonomy.

Data Collection Tools

Two primary data sources were used in the research:

(1) ChatGPT, Gemini, and Deepseek Outputs: A total of 30 questions, including open-ended and multiple-choice questions tailored to each level of Bloom's Revised Taxonomy through customized prompts based on learning outcomes, formed the main object of analysis.

(2) Assessment Rubrics: A rubric was developed to help mathematics teachers assess each question systematically and consistently. The rubric is two-tiered, labeled "Appropriate" or "Inappropriate," and is designed to determine whether each question aligned with the targeted level of Bloom's Revised Taxonomy. If it is inappropriate, an explanation is requested. Finally, a section for suggestions was included. For the overall assessment, the evaluation criteria for the AI-generated questions were not determined arbitrarily; rather, they were established based on well-documented item-writing guidelines and mathematical assessment frameworks in the literature. Specifically, the evaluation categories were constructed as follows: i) suitability to learning outcomes, to ensure content validity and strict alignment with the intended curriculum (Haladyna et al., 2002; Webb, 1997); ii) suitability to middle school level, addressing the developmental and cognitive appropriateness of the vocabulary and mathematical complexity (Anderson & Krathwohl, 2001); iii) realistic scenario creation, drawing upon the principles of authentic assessment and contextualized mathematical literacy frameworks commonly used in international assessments (OECD, 2023); and iv) consistency, verifying the structural integrity, absence of ambiguity, and internal logical coherence of the generated items (Haladyna et al., 2002). To ensure content validity, the evaluation rubrics underwent a preliminary assessment by three mathematics education experts. These experts reviewed the criteria for clarity and taxonomic alignment. Following their qualitative feedback, the rubrics were revised until full consensus was reached regarding their validity, prior to the main expert review phase.

Participants

The participants in the study consisted of ten experienced mathematics teachers (mathematics education specialists/educators) selected through criterion purposive sampling. The selection of experts was based on the following key criteria: (i) holding bachelor's and master's degrees in mathematics education, (ii) actively pursuing a doctoral degree, and (iii) possessing theoretical and practical expertise in the middle school geometry curriculum and Bloom's Revised Taxonomy, (iv) voluntary participation in the study. The involvement of subject matter experts was critical for ensuring the pedagogical and technical accuracy of the materials generated by artificial intelligence (human-in-the-loop).

The participants' demographic characteristics, professional experience, and areas of expertise are presented in Table 1.

Table 1. Demographic information of teachers

Expert Teacher (ET) Code	Educational Level	Professional Experience
ET1	PhD	16
ET2	PhD	5
ET3	PhD	10
ET4	PhD	5
ET5	PhD	15
ET6	PhD	7
ET7	PhD	6
ET8	PhD	13
ET9	PhD	12
ET10	PhD	5

Experts can evaluate both content (geometry gains) and process, as required by the nature of the work. Participants have an average of 9 years of professional experience, and all have received academic-level training in measurement and evaluation principles. This contributed to the reliability of the qualitative evaluation of the AI-generated questions.

Pilot Application and Data Collection Process

In this study, all AI-generated questions were systematically structured around a specific middle school geometry learning outcome to ensure methodological consistency. The targeted learning outcome was defined as follows: “Be able to interpret polygons as closed shapes formed by lines intersecting sequentially on a plane. a) Examines situations where at least three lines intersect on a plane in a way that the last line intersects the first line. b) Creates various polygons with the sequential intersections of at least three lines on a plane, with the last line intersecting the first line. c) States that polygons are formed by the sequential intersections of at least three lines on a plane, with the last line intersecting the first line.” This comprehensive outcome was kept constant across all prompts to accurately and reliably measure the models' capacity to shift between different cognitive levels of Bloom's Taxonomy without introducing the confounding variable of varying topic complexity.

Special prompts were prepared for ChatGPT, DeepSeek, and Gemini AI tools to generate geometry questions suitable for each of Bloom's cognitive levels at the middle school level. The prompts used in this study were designed by researchers with demonstrated academic expertise in artificial intelligence applications in mathematics education. The researchers have conducted doctoral-level academic work at the intersection of artificial intelligence and mathematics education, delivered seminars on AI's capacity to generate assessment items in educational contexts, and contributed to book chapters and articles in this field. This academic background supported the development of the prompts, which were aligned with the mathematical assessment context and grounded in theoretical principles. This specialized background ensured that the prompts were structured using evidence-based prompt engineering principles specifically optimized for mathematical assessment. The goal was for the model to generate both open-ended (requiring interpretation or solution) and multiple-choice questions. A comprehensive expert consensus process was conducted to ensure the structural and pedagogical validity of the 36 geometry items (open-ended and multiple-choice) generated by the AI models. This process aimed to assess not only the technical accuracy of the items but also their alignment with the targeted cognitive levels in Bloom's Revised Taxonomy.

Following the evaluations, it was agreed that the gg level should be excluded from the analysis because, by its very nature, it focuses solely on short-answer questions that require

recalling the concept or appropriate word corresponding to the definition from memory. This decision was based on the similarity of the questions generated by the three tools. For each of the five levels of Bloom's Revised Taxonomy, two questions were generated by each AI tool from three different artificial intelligence tools, in the form of open-ended and multiple-choice questions, this resulted in a total of 30 items included in the Expert Evaluation Rubric (Figure 1).

Bloom's Taxonomy – AI Tool Evaluation Form

Bloom's Level	Question Type	AI Tool Used	Appropriate	Not Appropriate	Reason (If Not Appropriate)
III. Application	Open-Ended	Gen-1	□	□	
		Gen-2	□	□	
		Gen-3	□	□	
	Multiple Choice	Gen-1	□	□	
		Gen-2	□	□	
		Gen-3	□	□	
Suggestion Area:					

GENERAL EVALUATION

Criterion	Evaluation	Notes & Suggestions
Alignment with Learning Outcome (MAT.5.3.5)	Fully Aligned / Partially Aligned / Not Aligned	
Suitability for Middle School Level Creating realistic scenarios	Yes / No	
Consistency of Question Type (Open-Ended / Multiple Choice)	Yes / No	
How successful are the distractors in multiple-choice questions at targeting students' common conceptual misconceptions rather than just presenting wrong information?		
Have the open-ended questions been directed sufficiently to guide the expected student response? (That is, does the answering process actually measure the intended Bloom's level, or does it remain at a lower level?)		
Which AI tool consistently stays true to Bloom's Taxonomy levels?		
Based on the suggestions made in the feedback section, which AI tool can most easily improve the question format and level?		

Figure 1.Sample screenshot of the expert evaluation rubric

The names of the artificial intelligence tools used at this stage were anonymized. To ensure the experts' impartiality and prevent bias stemming from their personal experience with the artificial intelligence tools, it was not specified which question was generated by which tool. The three tools were anonymized as GAI-1, GAI-2, and GAI-3. The question forms and evaluation rubrics were emailed to 10 experts. The experts conducted their evaluations independently of one another, ensuring that they assessed the generated question items without knowing which AI tool had generated them. The process flowchart is shown in Figure 2.

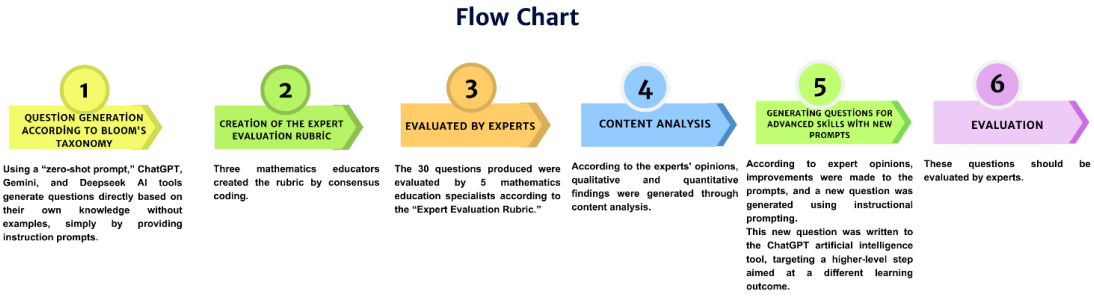


Figure 2. Flow Chart

ChatGPT, Gemini, Deepseek Prompts

Zero-shot prompting is a method in which large language models are expected to generate answers directly from task instructions, without any input-output examples (Debnath et al., 2025; Wan et al., 2023). This approach has recently been used in education to evaluate students' responses with zero-shot chain-of-thought prompting and to analyze educational survey feedback with zero-shot frameworks that achieve human-level performance (Cohn et al., 2024). This study aimed to evaluate the ability of artificial intelligence tools such as ChatGPT, Gemini, and DeepSeek to generate geometry questions solely from instructions, and to assess these models' ability to understand complex geometry concepts and generate questions aligned with Bloom's Taxonomy using only the knowledge embedded in the models without additional examples.

Figure 3 presents an example prompt template.

ChatGPT, Deepseek, and Gemini Prompt Entered:
Generate a question that fully meets the following criteria:
• Level: Middle School (Grades 5–8)
• Theme: Geometry
• Topic: Polygons
Learning Outcome: Be able to interpret polygons as closed shapes formed by lines intersecting sequentially on a plane. a) Examines situations where at least three lines intersect on a plane in a way that the last line intersects the first line. b) Creates various polygons with the sequential intersections of at least three lines on a plane, with the last line intersecting the first line. c) States that polygons are formed by the sequential intersections of at least three lines on a plane, with the last line intersecting the first line.
*• Cognitive level: **Bloom's Taxonomy – Application***
*• Question type: **Multiple choice***
• Target behavior: The question should require the student to apply their knowledge of polygon formation to a given situation.
• The language should be appropriate for middle school students.
According to these criteria, create a multiple-choice application-level question with four options, including the correct answer and three distractors, on the topic of polygons.

Figure 3. Prompts entered into artificial intelligence tools

To provide illustrative examples of the outputs obtained from the three artificial intelligence tools via zero-shot prompting, Figure 4 displays the multiple-choice geometry questions developed by the models corresponding to the applying level of Bloom's Revised Taxonomy.

ChatGPT:

Three lines connecting points A, B, and C are drawn on a plane. Each line intersects the previous line, and the last line intersects the first line, closing the shape. Considering this situation, which of the following is true?

A) The three lines form a circle.
B) The three lines form a triangle.
C) Three lines form a quadrilateral.
D) Three straight lines form a pentagon.

Deepseek:
A student draws shapes by connecting the points given below in order:
Shape 1: A → B → C → A
Shape 2: K → L → M → N
Shape 3: X → Y → Z → X → Y
Which of these shapes, if any, fits the definition of a polygon?
OPTIONS:
A) Only Shape 1
B) Only Figure 2
C) Figure 1 and Figure 3
D) Figure 2 and Figure 3
Correct Answer: A
Distracting Explanations:
Figure 2 (K-L-M-N): Since the last point (N) does not connect to the first point (K), it does not form a closed shape and is not a polygon.
Figure 3 (X-Y-Z-X-Y): Although the line segments intersect sequentially, after the shape is closed (X-Y-Z-X), an extra line segment (X-Y) is drawn. This creates a situation that does not fully comply with the basic polygon definition (closed and simple) and is not considered a pure polygon. Therefore, it acts as a distractor.

GEMİNİ:
Mark places the 4 straight sticks he has on the table. He begins to create a polygon model by connecting these sticks end to end. After joining the first three sticks, he adds the fourth stick to the end of the third stick, but notices that the starting point and ending point of the shape do not meet.
Which of the following actions should Mark take at this stage to transform the shape he has created into a polygon?
A) He should connect the starting and ending points with a curved line using a flexible string.
B) He should close the shape by adding another straight stick starting from the end of the last stick and going to the starting point of the first stick.
C) Remove one of the rods and arrange the remaining rods parallel to each other so that they do not touch.
D) Add another rod to the open part of the shape and leave its end in the void.
Correct Answer: B

Each question received a dual evaluation from 10 experts: Appropriate (aligned with the targeted Bloom level) and Not Appropriate (misaligned). For each question, n(Appropriate) and n(Not Appropriate) were reported separately. The decision rule is as follows: if n(Appropriate) ≥ 6, then 'Appropriate'; if n(Appropriate) ≤ 4, then 'Inappropriate'; if n(Appropriate) = 5, then 'Undetermined' (tied vote). This reporting is mandatory for the transparency and re-exam inability of quantitative findings.

Figure 4. Examples of multiple-choice questions generated by generative AI tools

Data Analysis

For the quantitative analysis, the experts' assessments regarding the alignment of the AI-generated questions with the targeted Bloom's Taxonomy levels were quantified using a binary coding system. Specifically, an assessment of 'Appropriate' (consistent with the targeted Bloom level) was coded as 1, while 'Inappropriate' (inconsistent) was coded as 0. Based on the responses from the 10 expert evaluators, a majority decision rule was established for the overall evaluation of each question: if n(Appropriate) ≥ 6, the item was deemed 'Appropriate'; if n(Appropriate) ≤ 4, it was deemed 'Inappropriate'; and if n(Appropriate) = 5, it was classified as 'Undetermined' (a tie). This transparent reporting of frequencies supports the reproducibility of the quantitative findings. Finally, descriptive statistics (frequencies and percentages) were used to reveal the overall agreement levels, and inter-rater agreement was assessed using Gwet's AC1 coefficient.

To assess agreement between expert evaluations in the study, Gwet's AC1 (First-Order Agreement Coefficient), a chance-adjusted metric, was used in addition to the raw agreement percentage (Gwet, 2014). Reporting only the percentage agreement value can yield misleadingly high results because it does not account for agreement that occurs by chance (McHugh, 2012). Measures commonly preferred in the literature, such as Cohen's or Fleiss's Kappa, can produce mathematically low coefficients despite high raw agreement in unbalanced marginal distributions where expert votes are concentrated in a specific category, leading to the "Kappa paradox" (Feinstein & Cicchetti, 1990). Gwet's AC1 overcomes this paradox by handling probabilities more evenly when calculating chance agreement and provides a more stable indicator of evaluator agreement in the literature (Gwet, 2008; Wongpakaran et al., 2013). In this study, Gwet's AC1 coefficients were calculated using a custom data analysis script developed in the Python programming language, precisely implementing the mathematical formulas proposed by Gwet (2014).

The qualitative data, consisting of the experts' written responses in the "Justification" and "Recommendation Area" sections of the evaluation rubric, were analyzed using thematic content analysis. All opinions were compiled into a single dataset, systematically coded, and grouped into categories. These categories were then organized into three main themes based on their conceptual relationships. To ensure the reliability of the analysis, the coding of the expert opinions was performed by two different researchers, who subsequently reached a full consensus. The main objective of this qualitative phase was to gain an in-depth understanding of the reasons behind the experts' agreements or disagreements and to provide qualitative support for the quantitative findings.

Results

This section assesses the suitability of middle school geometry questions generated by three generative artificial intelligence tools (GAT-1 ChatGPT, GAT-2 DeepSeek, GAT-3 Gemini) for the targeted Bloom cognitive levels and their overall question quality, based on evaluations by 10 mathematics education experts. The relevant learning outcome involves the student "forming a polygon with the sequential intersection of at least three lines" and expressing the condition for a closed shape.

Alignment of AI-Generated Geometry Questions with Bloom's Taxonomy Across Question Types (RQ1)

This section examines the extent to which ChatGPT, Gemini, and DeepSeek generate middle school geometry questions aligned with Bloom's cognitive levels, with particular attention to differences across open-ended and multiple-choice question types.

When comparing overall performance, DeepSeek and Gemini's question sets received higher expert approval than ChatGPT (average n_suitable per item: ChatGPT = 5.7/10; Deepseek = 7.2/10; Gemini = 7.1/10). This difference is particularly evident in multiple-choice questions when examining Table 2 (ChatGPT multiple-choice average n_appropriate = 4.4/10, while Deepseek = 7.0/10 and Gemini = 6.4/10). As the Bloom level increases (especially during the evaluate and create level), expert approval rates decrease. The most significant decrease is in the Creation–Multiple Choice cell: the create-level questions produced by all three tools in this format were mostly rated "inappropriate" (average n_appropriate = 3.67/10). In contrast, although most open-ended questions at the create level were rated "appropriate," the approval rate was moderate (mean n_appropriate = 6.33/10). Table 2 presents a comparative analysis of the performance of ChatGPT, Gemini, and DeepSeek across Bloom's Revised Taxonomy levels and question types.

Table 2. Comparison of AI Tools by Bloom Level and Question Type

Bloom Level	Question Type	ChatGPT (Decision / Vote)	Deepseek (Decision / Vote)	Gemini (Decision / Vote)
II. Understand	Open-ended	Appropriate (10 - 0)	Appropriate (7 - 3)	Appropriate (8 - 2)
	Multiple-choice	Appropriate (6 - 4)	Appropriate (9 - 1)	Appropriate (8 - 2)
III. Apply	Open-ended	Appropriate (7 - 3)	Appropriate (7 - 3)	Appropriate (8 - 2)
	Multiple-choice	Inappropriate (4 - 6)	Appropriate (7 - 3)	Appropriate (6 - 4)
IV. Analyze	Open-ended	Appropriate (8 - 2)	Appropriate (8 - 2)	Appropriate (9 - 1)
	Multiple-choice	Uncertain (5 - 5)	Appropriate (9 - 1)	Appropriate (7 - 3)
V. Evaluate	Open-ended	Inappropriate (4 - 6)	Appropriate (8 - 2)	Appropriate (8 - 2)
	Multiple-choice	Inappropriate (4 - 6)	Appropriate (6 - 4)	Appropriate (7 - 3)
VI. Create	Open-ended	Appropriate (6 - 4)	Appropriate (7 - 3)	Appropriate (6 - 4)
	Multiple-choice	Inappropriate (3 - 7)	Inappropriate (4 - 6)	Inappropriate (4 - 6)
Summary	Average (Open-	7.0 / 10	7.4 / 10	7.8 / 10

Statistics	ended)			
	Average (Multiple-choice)	4.4 / 10	7.0 / 10	6.4 / 10
	Overall Average	5.7 / 10	7.2 / 10	7.1 / 10
Cross-Tool Averages	Create Level (Open-ended) 6.33 / 10 *(Mean of 3 tools)*			
	Create Level (Multiple-choice) 3.67 / 10 *(Mean of 3 tools)*			

Gwet's AC1 (First-Order Agreement Coefficient) statistic was used to determine the agreement between expert evaluations in the study (Gwet, 2014). The raw agreement percentages for each item in Table 3 are reported alongside Gwet's AC1 values.

Table 3. Assessment of Questions Produced by AI tools for Compliance with Bloom's Taxonomy Levels - Inter-Expert Consensus

Bloom Level	Question Type	Tool	Majority Decision	Vote Distribution (Appropriate - Inappropriate)	Agreement Percentage (%)	Gwet AC1
II. Understand	Open-ended	ChatGPT	Appropriate	10 - 0	100.0	1.000
		Deepseek	Appropriate	7 - 3	70.0	195
		Gemini	Appropriate	8 - 2	80.0	477
	Multiple-choice	ChatGPT	Appropriate	6 - 4	60.0	-26
		Deepseek	Appropriate	9 - 1	90.0	756
		Gemini	Appropriate	8 - 2	80.0	477
III. Apply	Open-ended	ChatGPT	Appropriate	7 - 3	70.0	195
		Deepseek	Appropriate	7 - 3	70.0	195
		Gemini	Appropriate	8 - 2	80.0	477
	Multiple-choice	ChatGPT	inappropriate	4 - 6	60.0	-26
		Deepseek	Appropriate	7 - 3	70.0	195
		Gemini	Appropriate	6 - 4	60.0	-26
IV. Analyze	Open-ended	ChatGPT	Appropriate	8 - 2	80.0	477
		Deepseek	Appropriate	8 - 2	80.0	477
		Gemini	Appropriate	9 - 1	90.0	756
	Multiple-choice	ChatGPT	Uncertain	5 - 5	50.0	-111
		Deepseek	Appropriate	9 - 1	90.0	756
		Gemini	Appropriate	7 - 3	70.0	195
V. Evaluate	Open-ended	ChatGPT	inappropriate	4 - 6	60.0	-26

		Deepseek	Appropriate	8 - 2	80.0	477
	Multiple-choice	Gemini	Appropriate	8 - 2	80.0	477
		ChatGPT	inappropriate	4 - 6	60.0	-26
		Deepseek	Appropriate	6 - 4	60.0	-26
		Gemini	Appropriate	7 - 3	70.0	195
VI. Create	Open-ended	ChatGPT	Appropriate	6 - 4	60.0	-26
		Deepseek	Appropriate	7 - 3	70.0	195
		Gemini	Appropriate	6 - 4	60.0	-26
	Multiple-choice	ChatGPT	inappropriate	3 - 7	70.0	195
		Deepseek	inappropriate	4 - 6	60.0	-26
		Gemini	inappropriate	4 - 6	60.0	-26

When examining Gwet's AC1 coefficients, moderate to very high levels of agreement among experts are observed at lower cognitive levels (comprehension, application, and analysis) (AC1 = .477–1.000). At these levels, agreement is acceptable or good, and therefore stable and independent of chance. In contrast, at the higher cognitive levels (evaluate and create), the agreement coefficients were low and even negative in some cases (AC1 = −.026). The confidence intervals are very close to 0. This indicates that experts struggled to reach a consensus on the taxonomic appropriateness of questions that represent higher cognitive processes in particular.

When examining Table 3 separately for open-ended and multiple-choice questions at the comprehension level of Bloom's Revised Taxonomy, the experts' evaluations of open-ended questions show high agreement. The open-ended questions generated by the ChatGPT tool were evaluated by all experts as appropriate for the targeted cognitive level, achieving 100% agreement and a Gwet's AC1 coefficient of 1.000. Similarly, for open-ended questions generated by the DeepSeek and Gemini tools, agreement rates were 70% and 80%, respectively, and the AC1 coefficients were .195 and .477, respectively. This indicates a low to moderate level of agreement among experts.

When examined through multiple-choice questions, differences emerge among generative AI tools. Although the agreement rate for ChatGPT-generated questions was 60%, the negative AC1 coefficient (-.026) indicates no meaningful agreement among experts. On the other hand, the agreement rate for questions generated by DeepSeek was 90%, and the AC1 coefficient was 0.756. This indicates a high level of expert agreement. The Gemini tool achieved an 80% agreement rate and an AC1 value of .477. This result reflects a moderate level of agreement. The item with the greatest divergence among experts is the multiple-choice question created by ChatGPT in the analysis step. The ambiguity and distracting quality at the root of this question have divided the experts. When examining the questions in the evaluation stage of Bloom's taxonomy, a greater range of agreement among experts was observed. For open-ended questions, the DeepSeek and Gemini-generated questions show moderate agreement, with 80% consensus and an AC1 value of 0.477. In contrast, the consensus rate for questions generated by ChatGPT is 60%, and the AC1 coefficient is -0.026.

At the highest level of Bloom's Revised Taxonomy (creating), there was generally a lower level of agreement among experts. Experts' opinions differ on whether the questions at the create level meet the requirements of this stage. Although the majority of questions generated by ChatGPT and Gemini tools for open-ended questions were rated as appropriate by experts, the agreement rate remained at 60%, and the AC1 coefficient was -.026. For questions generated by the Deepseek tool, the agreement rate was 70% and the AC1 value was 0.195, indicating limited agreement. In the question items prepared at the create level, artificial intelligence uses verbs such as "Design" and "Create" in the question stem, which are thought to reflect the create level. In the qualitative explanations, some experts considered these verbs technically sufficient and accepted the question as "Create." In contrast, others characterized this situation as "pseudo-creativity", arguing that the student did not actually create anything new but only applied the given rules. Therefore, the items remained at the "Application" level.

Overall, it was observed that questions generated by artificial intelligence tools were evaluated more consistently by experts at the lower and middle cognitive levels (comprehension, application, and analysis). In contrast, both agreement rates and Gwet's AC1 coefficients are lower at higher cognitive levels (evaluate and create). This situation indicates that artificial intelligence tools have limitations in producing questions that measure higher cognitive processes, and that differences of opinion among experts regarding the taxonomic appropriateness of these questions are more pronounced. The inter-expert chance-adjusted agreement coefficient (Gwet AC1) and 95% Confidence Intervals were calculated based on the 30-question dataset obtained in the study. The analyses are grouped by Bloom's levels and the artificial intelligence tools used, and are presented in Table 4.

Table 4. Expert Alignment (Gwet AC1) and Confidence Intervals According to Bloom's Levels and Generative AI Tools

Variable / Category	Gwet AC1	95% Confidence Interval (CI)
Overall Average (30 Questions)	0.237	[0.152, 0.323]
According to Bloom's Taxonomy		
II. Understand	0.532	[0.304, 0.759]
III. Apply	0.137	[0.041, 0.232]
IV. Analyze	0.446	[0.270, 0.623]
V. Evaluate	0.122	[-0.010, 0.254]
VI. Create	-0.022	[-0.077, 0.033]
According to GenerativeAI Tools		
ChatGPT	0.115	[-0.091, 0.321]
DeepSeek	0.322	[0.194, 0.450]
Gemini	0.294	[0.178, 0.409]

When the results are examined, the average agreement coefficient across all questions is AC1 = .237. This value indicates a low level of agreement among experts. In other words, experts show limited consensus regarding the suitability of the questions generated by artificial

intelligence tools for the targeted Bloom levels. When examining the consistency coefficients calculated according to Bloom's cognitive levels, it is evident that consistency among experts varies across levels. The AC1 = .532 value at the comprehension level indicates moderate agreement among experts. At the application level, the agreement coefficient is AC1 = .137. This value indicates a weak level of agreement among experts. This finding reveals that experts differ in their assessments of the extent to which the generated questions reflect cognitive processes at the application level. Again, when higher-level cognitive stages are examined, it is seen that the agreement coefficients decrease significantly. The agreement coefficient at the evaluation stage was calculated as AC1 = .122. The fact that the confidence interval includes zero indicates that there is no statistically significant agreement between the experts' evaluations. The lowest agreement is observed at the create level. The calculated AC1 = -.022 value for this level indicates that the experts' evaluations are indistinguishable from random agreement. In other words, there is no meaningful consensus among experts on whether the questions at the create level truly represent this higher-order level of Bloom's taxonomy.

Expert Evaluation of AI-Generated Geometry Questions (RQ2)

This section examines how mathematics education experts evaluate AI-generated geometry questions in terms of taxonomic alignment, content quality, and pedagogical appropriateness. This section presents the qualitative evaluations by mathematics education experts regarding the overall pedagogical quality, depth, and appropriateness of the AI-generated questions for middle school students.

Within the scope of the second research question (RQ2), the underlying justifications for the experts' evaluations of the suitability and quality of the AI-generated questions were analyzed through thematic analysis. Qualitative findings were grouped under three main themes: taxonomic alignment, content quality and mathematical validity, and pedagogical appropriateness and tool comparison. The content analysis was conducted in three stages: preparation, organization, and reporting. Codes were derived from similar expressions and grouped into themes.

Taxonomic Alignment

When expert opinions on the theme of taxonomic alignment are examined, three fundamental problems are identified, as shown in Table 5. These were insufficient representation of higher-order cognitive skills, particularly at the evaluate and create levels; limitations of the multiple-choice question format in revealing thinking processes at the analyzing and creating levels; and a discrepancy between the targeted Bloom level and the cognitive level actually measured by the question.

Expert opinions such as "This is not a question at the create level" and "It does not reflect a higher-order skill" were evaluated under the subheading "Failure in higher-order skills." Expert evaluations show that AI tools are more successful at lower cognitive levels (Understanding and Application) but struggle to reach the targeted level at higher cognitive levels (analyzing, evaluating, and creating). While some experts noted that the tools could generate questions consistent with the learning outcome (MAT.5.3.5) and covering the subject matter, others noted that even questions at the Comprehension level sometimes remained at the Recall level or contained linguistic problems.

Table 5. Codes and frequency distribution obtained from expert opinions on the taxonomic compatibility theme

Code (Subtitle)	Citing Experts	Frequency (f)
Failure in Higher-Level Skills	Expert 1, Expert 2, Expert 3, Expert 4, Expert 8, Expert 10	6
Limitations of Question Format	Expert 1, Expert 3, Expert 4, Expert 8, Expert 10	5
Level Shift	Expert 2, Expert 4, Expert 5, Expert 8, Expert 9	5

Experts' evaluations indicated a level shift in AI-generated questions, with many items intended to assess higher-order thinking remaining at no higher than the applying level in Bloom's Revised Taxonomy. For example, Expert-2 stated the following about the questions produced for the Create level: "It is not a question at the Create level. There is no new polygon design involved; it is at the applying level."

Expert-3 emphasized that the Analysis questions were at the "maximum application level." Expert 5 stated that "Gemini's question at the create level could be an appropriate question for the Application stage," and that, for the questions at the Application stage, "it does not seem to be different from the above level of understanding," emphasizing that the distinction between the stages was unclear. Experts identified a mismatch between the targeted and measured Bloom levels, indicating that as the intended level increases, AI-generated questions tend to shift one or two levels lower, often remaining at the Application level.

The experts' emphasis on "multiple-choice is insufficient/inappropriate," "multiple choice questions are not useful," and "inadequacy in producing distractors" was evaluated under the heading of question format limitations. At this point, it was observed that the experts emphasized that the multiple-choice format, in particular, was inadequate for measuring higher-level skills. Expert 7 criticized the use of multiple-choice question stems in the assessment stage, stating, "Why should a student prefer one shape? A context or situation could be given for them to make an appropriate choice. This way, their choice becomes arbitrary rather than having a logical justification." Expert 3 again refers to the limitations of the question format, stating, "I think multiple-choice questions cannot fully meet levels such as analysis and create in the taxonomy." Expert 10 stated, "Multiple-choice questions are not useful for measuring higher-level steps such as analyzing and evaluating," indicating that they did not consider multiple-choice items appropriate for assessing higher-order cognitive skills, not specifically in relation to questions generated by artificial intelligence tools, but rather as a general assessment issue rather than one specific to AI-generated questions. Expert 1, with the statement, "The distractors are quite weak, and it can be considered a recall level," refers to the strength of distractors in multiple-choice items and states that because the distractors are weak, the question is below the targeted cognitive level.

Content Quality and Mathematical Validity

As shown in Table 6, expert evaluations indicate weaknesses in mathematical and conceptual accuracy, linguistic ambiguities, and non-functional distractors in some multiple-choice questions. The measurement literature on distractor writing principles emphasizes that options should reflect students' typical errors and be consistent with the question stem; therefore, distractor weakness should be considered a factor that directly reduces item quality.

Table 6. Codes and frequency distribution obtained from expert opinions on the theme of Content Quality and Mathematical Validity

Code (Subtitle)	Citing Experts	Frequency (f)
Mathematical and Conceptual Errors Lack of definition, incorrect use of terms, geometric error	Expert 1, Expert 2, Expert 3, Expert 5, Expert 6, Expert 8	6
Distractor weakness Irrational distractors, options that suggest the answer	Expert 1, Expert 3, Expert 4, Expert 5, Expert 8, Expert 9, Expert 10	7
Language and Clarity Spelling errors, awkward phrasing, confusing instructions	Expert 2, Expert 3, Expert 4, Expert 7, Expert 8, Expert 9	6

Expert 4 stated, "There are linguistic problems again; it does not start from the right point, it has no beginning or end. Placing concepts in the wrong place creates different problems. Expert 1 similarly noted "The definition sentence is weaker; the emphasis on closure is not made in the definition," emphasized that AI-generated questions used mathematical definitions (line, ray, polygon closure) incorrectly or incompletely. Similarly, Expert 7's warning regarding the drawings in the Application question, "The arrow signs between the points may be perceived as rays by students," supports the finding that artificial intelligence deviates from mathematical precision in its visual descriptions.

On the other hand, seventy percent of the experts stated that the AI tools did not generate effective distractors that identify conceptual errors and that it fails in distractor writing. It was noted that distractors are far from targeting conceptual errors and consist of options that make the answer very obvious or are irrelevant. Expert 2 stated, "They are not successful. The distractors only focus on the wrong answer," Expert 4 stated, "The way the question is asked again suggests that the answer is B." Expert 1 stated, "The distractors have not been assigned as geometric concepts," supporting this finding. Similarly, Expert 9 emphasized the poor quality of distractors in multiple-choice questions generated by artificial intelligence tools, stating, "Distractors should target conceptual misunderstandings rather than merely providing incorrect information."

Another issue that experts frequently emphasize is that the language of the questions is sometimes ambiguous or gives away the answer. Experts have particularly criticized ChatGPT and DeepSeek tools for the clarity and linguistic accuracy of the questions. Expert 8 stated, "The question is a bit confusing; the language used should be simpler for the fifth-grade level," while Expert 4 stated, "The question is not clear and concise... The question is not clear. There is a linguistic problem," emphasizing that linguistic problems reduce the quality of measurement. Expert 3 noted that "it may be difficult to understand the question while reading it," pointing out that some questions are difficult for students to understand and emphasizing that the distractors are insufficient in revealing the student's thought process.

Pedagogical Suitability and Tool Comparison

Table 7 suggests that Gemini produces more realistic contexts, while other tools tend to remain definition-focused when targeting the same learning outcomes. This theme also explains the "multiple-choice create" failure seen in RQ1: even if a realistic context is

established, if the options fail to elicit higher-order reasoning, the item may remain at the applying level rather than reaching the creating level.

Table 7. Codes and frequency distribution obtained from expert opinions on the theme of Pedagogical Suitability and Tool Comparison

Code (Subtitle)	Experts Cited	Frequency (f)
Context and Scenario Success	Expert 1, Expert 2, Expert 5, Expert 7, Expert 8, Expert 9, Expert 10	*7*
Appropriateness for students' level	Expert 1, Expert 2, Expert 4, Expert 6, Expert 7, Expert 8, Expert 9, Expert 10	*8*
Dissenting Opinion Criticisms and Pedagogical Risks	Expert 1, Expert 3, Expert 4, Expert 5, Expert 6, Expert 8	*6*

Most experts agree that Gemini is the most successful tool for storytelling and context creation. Expert 2 stated, "The third tool created realistic scenarios. The others failed to do so," Expert 1 stated, "It appears that the third tool, unlike the others, placed context within the question," and Expert 3 stated, "The third tool seemed better." As seen, Expert 1 and Expert 2 praised Gemini's storytelling and context-building, noting that the other tools' questions were more "dry" or "only definition-focused." Similarly, Expert 5's statement, "The third tool always provided questions within context and was generally consistent," and Expert 7's statement about Gemini, "This question is very well done; it is more understandable when there is context," show that they praised this tool's ability to create scenarios.

Expert assessments regarding situations considered pedagogically risky, potentially leading to misconceptions, or having weak measurement power have been evaluated under the heading of Dissenting Opinion. While experts found artificial intelligence quite successful at establishing context, they adopted a more cautious, critical approach to pedagogical accuracy (the correct use of concepts and the measurement of higher-level skills). In particular, the opinions of Experts 3 and 6 contain strong warnings about the misuse of mathematical terminology by artificial intelligence. Expert 3 emphasizes the low pedagogical measurement value of questions generated by artificial intelligence, stating, "The questions are not asked correctly, and the concepts are not expressed correctly. The way some questions are asked directly suggests the answer." While some experts highlighted differences among the tools in context creation and the accurate use of definitions and concepts, Expert 4 expressed a dissenting view. This expert argued that the three tools were similar in overall quality and that no meaningful differences were evident.

Discussion

Previous studies show that generative AI tools can produce questions with efficiency and psychometric properties comparable to human-written items, particularly at lower cognitive levels (Bhandari et al., 2024; Kıyak & Emekli, 2024). Consistent with this literature, the present study found that AI tools performed relatively well in generating questions aligned with the lower and middle levels of Bloom's Revised Taxonomy (remembering, understanding, applying), while demonstrating limitations in higher-order skills such as analysis, evaluation, and creation (Maity & Deroy, 2024; Akdeniz et al., 2025).

AI models still struggle to generate high-quality distractors targeting student misconceptions

in multiple-choice questions, particularly in visually and conceptually dense domains such as geometry. Consistent with previous research, experts frequently reported that distractors did not effectively target common student misconceptions and often made the correct answer obvious. Similarly, Lee et al. (2024) found that only 37% of distractors produced by large language models were considered valid by teachers, while Gündeğer Kılcı (2025) emphasized that distractor design remains one of the weakest aspects of AI-generated assessment items. We argue that these limitations likely stem from the underlying architecture of generative AI systems, which are primarily optimized to generate logically and mathematically correct outputs rather than to model and predict typical human cognitive errors or specific geometric misconceptions. The study's findings also show that, particularly during the analysis stage, the ambiguity in the stems of the multiple-choice questions generated by ChatGPT and the poor quality of the distractors sharply divided experts, with an equal split among experts. This disagreement among experts indicates uncertainty regarding the validity and reliability of AI-generated items.

Another important finding relates to the variance in contextualization and scenario creation among different AI models. In the current study, Gemini was perceived by experts as significantly more successful in constructing realistic problem contexts for geometry, suggesting emerging differences in the narrative capabilities of models like Gemini, DeepSeek, and ChatGPT. Similarly, Karabıyık (2024) reported that ChatGPT is generally effective in solving math problems, particularly in topics such as limits and derivatives, providing conceptually and operationally correct, understandable explanations. However, the same study emphasizes that although artificial intelligence is successful with basic geometry concepts such as right triangles and the Pythagorean theorem, it can be inadequate for more complex or in-depth geometry questions, and that expert support is always needed in such cases. Furthermore, ChatGPT, the most studied tool in the literature, showed the lowest and statistically uncertain performance in terms of expert alignment in this study (AC1 = .115, CI [-.091, .321]), while DeepSeek (AC1 = .322) and Gemini (AC1 = .294) achieved more stable, chance-adjusted agreement, indicating that architectural and pedagogical alignment differences among new-generation large language models are beginning to emerge.

The finding that artificial intelligence tools are inadequate in the analyzing, evaluating, and creating levels and that questions were evaluated at the Application level by domain experts is consistent with the literature. Maity and Deroy (2024) noted that GPT-4 struggled to distinguish between closely related high-level categories such as Analysis and Evaluate. Similarly, Akdeniz et al. (2025) emphasized that AI questions can compete with human-generated questions at lower skill levels but that humans still have an advantage at higher skill levels. These results suggest that AI experiences a "cognitive level shift," often producing pseudo-creativity by mechanically using verbs such as "design" or "create" without generating genuinely higher-order questions.

When examining the Gwet AC1 statistics, the negative values at the "Evaluation" (-0.122) and "Creation" (-0.022) levels reveal that the disagreement among experts goes beyond mere chance, indicating substantial disagreement among experts regarding the taxonomic alignment of AI-generated questions at higher cognitive levels. As the targeted cognitive processes become more complex such as synthesis and evaluation, it can be interpreted that the structural predictability of artificial intelligence outputs is lost, leading to different pedagogical interpretations among subject matter experts. The literature also emphasizes that although artificial intelligence can effectively mimic teaching tasks, it struggles to produce the originality required at the "creation" stage, thus highlighting the need for teacher

supervision (Shaunessy, 2000; VanTassel-Baska, 2014; Akdeniz et al., 2025). Although this performance varies depending on the prompt engineering used—where prior research indicates that more structured prompts can significantly improve output quality (Omopekunola & Kardanova, 2024; Karakaya et al., 2024)— the literature generally suggests that AI-generated content should not be used directly in assessments without careful expert review without careful expert review (Kıyak & Emekli, 2024; May et al., 2025; Siegle, 2024). Overall, the results suggest that generative AI tools can serve as valuable assistants in generating geometry items; however, they appear more suitable for supporting the item development and revision process rather than replacing human item writers.

Conclusions and Recommendations

Specific Conclusions

In conclusion, this study demonstrates that while ChatGPT, Gemini, and DeepSeek can function as effective draft item generators for middle school geometry, their taxonomic alignment is inconsistent. The models perform adequately at the "Understanding" level, but show substantial limitations in generating valid multiple-choice items at the "Analyzing" and "Evaluating" levels without explicit human intervention. Furthermore, a critical deficiency across all tested models is the generation of non-functional distractors that fail to capture typical student misconceptions in geometry. Conversely, Gemini appears to perform better in constructing realistic contextual scenarios for geometry problems than the other models.

Recommendations for Educational Practice and Future Research

For Mathematics Educators: We recommend that educators avoid using AI-generated geometry items directly for formal assessment. Instead, they should implement a mandatory "expert-review protocol" using a validated pedagogical rubric, specifically targeting the revision of distractors. Educators should revise AI-generated distractors manually to reflect empirically known student misconceptions in geometry, rather than relying on the AI's mathematically logical but pedagogically weak alternatives.

For Prompt Engineering: Future applications should move away from zero-shot prompting in item generation. We recommend designing and testing structured, multi-step prompts in which common geometric misconceptions are explicitly provided to the AI models beforehand.

For Future Research: Subsequent studies should quantitatively evaluate whether providing targeted instructional prompts (e.g., "create a distractor based on the misconception of confusing perimeter with area") significantly increases the validity index of AI-generated distractors compared with the baseline results observed in this study.

Limitations

This study is limited to the topic of polygons in the fifth-grade mathematics curriculum. Future research should compare the context-generation capabilities of artificial intelligence across different mathematical topics and disciplines. The difficulty in producing higher-level questions in this study may be related to the inherently visual and conceptual nature of geometry, which poses challenges for large language models. As noted by Koman et al. (2023), generating questions in visually intensive domains such as geometry is already

demanding for teachers, and the primarily text-based architecture of large language models makes tasks involving spatial reasoning and visual construction particularly challenging.

Consistent with this interpretation, experts in the present study reported that AI tools were generally successful in generating questions involving definitions or simple applications but were less effective in tasks requiring shape analysis or the construction of new geometric structures. Similar findings were reported by Karabıyık (2024), who showed that while AI models can explain algebraic solutions step by step, they struggle to represent geometric relationships or visualize complex shapes.

The integration of artificial intelligence into education also raises important concerns regarding algorithmic bias, ethics, and pedagogical reliability (Roll & Wylie, 2016; Xu & Ouyang, 2022). Although taxonomic frameworks such as Bloom's taxonomy provide useful guidance for structuring assessment tasks, relying solely on these frameworks in AI-based question generation is insufficient. Expert evaluation remains essential to ensure pedagogical and psychometric quality. Previous studies similarly emphasize that although AI systems can produce linguistically fluent text, expert oversight is necessary to evaluate the instructional and measurement validity of the generated items (Tan et al., 2025; Karakaya et al., 2024; Maity & Deroy, 2024; Lee et al., 2024; Omopekunola & Kardanova, 2024).

The lack of full consensus among the ten experts in this study further demonstrates the variability in evaluating AI-generated questions. Some experts focused primarily on mathematical accuracy, while others emphasized overall pedagogical coherence, indicating that shared evaluation standards among experts remain underdeveloped. Future research should involve larger expert groups and explore how human–AI collaboration can improve item quality, particularly by integrating real student error patterns and misconception libraries into AI-supported question generation systems. Additionally, systematically comparing prompt design strategies (e.g., chain-of-thought prompting, persona-based prompts, and example-based prompting) across different AI models may help improve performance across the levels of Bloom's Revised Taxonomy.

Declarations
Acknowledgments: There are no acknowledgements to be made within the scope of this study.
Funding: This research has not received financial support from any public institution, commercial organization, or non-profit organization.
Ethics Statements: Ethical principles were followed throughout the study. Participation was voluntary, participants' identities were kept confidential, and all data were anonymized and used solely for research purposes.
Conflict of Interest: The author declares that there is no conflict of interest related to this research, the authorship process, or the publication of the article.
Informed Consent: Informed consent was obtained from the teachers participating in the research.
Data availability: The data obtained in this research is not publicly available in order to protect the privacy of the participants. However, it can be obtained from the corresponding author upon reasonable request.

References

Adams, N. E. (2015). Bloom's Taxonomy of Cognitive Learning Objectives. *Journal of the Medical Library Association : JMLA*, *103*(3), 152–153. doi:10.3163/1536-5050.103.3.010

Akdeniz, H., Clark, T., & Roberts, J. L. (2025). Can AI generate questions aligned with Bloom's taxonomy? A framework for gifted education to support teachers. *Journal of Advanced Academics*, 36(4), 671-694. https://doi.org/10.1177/1932202X25134991

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: Complete edition*. Addison Wesley Longman, Inc.

Bhandari, S., Liu, Y., Kwak, Y., & Pardos, Z. A. (2024). Evaluating the psychometric properties of chatgpt-generated questions. *Computers and Education: Artificial Intelligence*, *7*, 100284. https://doi.org/10.1016/j.caeai.2024.100284

Chen, Z. (2025). *Reliable generation of isomorphic physics problems using generative AI with prompt-chaining and tool use*. https://doi.org/10.48550/arXiv.2508.14755

Cohn, C., Hutchins, N., Le, T., & Biswas, G. (2024). A chain-of-thought prompting approach with LLMs for evaluating students' formative assessment responses in science. In Proceedings of the AAAI Conference on Artificial Intelligence, 38(21), 23182–23190. doi: 10.48550/arXiv.2403.14565

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and Conducting Mixed Methods Research*. Thousand Oaks, CA: Sage Publications.

Debnath, T., Siddiky, M. N., Rahman, M. E., Das, P., Guha, A. K., Rahman, M. R., & Kabir, H. M. (2025). *A Comprehensive Survey of Prompt Engineering Techniques in Large Language Models*. doi:10.36227/techrxiv.174140719.96375390/v2

Elkins, S., Kochmar, E., Cheung, J. C. K., & Serban, I. (2024). How teachers can use large language models and Bloom's taxonomy to create educational quizzes. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(21), 23084-23091. doi: 10.1609/aaai.v38i21.30353

Eren, E., Ayata, N., Çetinkaya, N. B., Yardımcı, T., Koç, A., & Akpınar, E. N. (2025). Eğitimde yapay zekâ kullanımı: Öğretmenlerin perspektifinden bir değerlendirme [The use of artificial intelligence in education: An evaluation from teachers' perspectives]. *Toplum, Eğitim ve Kültür Araştırmaları Dergisi [Journal of Society,*

Education and Cultural Research], 4(1), 23–38. https://doi.org/10.5281/zenodo.15460358

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 543–549. doi: 10.1016/0895-4356(90)90158-L

Gopal, A., Chen, P., Oh, W., Xu, S. X., & Sarker, S. (2024). On crafting effective theoretical contributions for empirical papers in economics of Information Systems: Some editorial reflections. *Information Systems Research, 35*(3), 917–935. doi: 10.1287/isre.2024.editorial.v35.n3

Grévisse, C., Pavlou, M. A. S. , & Schneider, J.G. (2024). Docimological quality analysis of LLM-generated multiple choice questions in computer science and medicine. *Sn Comput. Sci*. 5, 636. https://doi.org/10.1007/s42979-024-02963-6

Gündeğer Kılcı, C. (2025). ChatGPT vs. Deepseek: A comparative psychometric evaluation of AI tools in generating multiple-choice questions. *International Journal of Assessment Tools in Education*, 12(4), 1055-1079. doi: 10.21449/ijate.1674995

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 29–48. https://doi.org/10.21449/ijate.1674995

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters* (4th edition). Advanced Analytics, LLC.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education*, *15*(3), 309–333. https://doi.org/10.1207/S15324818AME1503_5

Karabıyık, Ü. (2024). Matematik eğitiminde yenilikçi bir yaklaşım: ChatGPT'nin rolü [An innovative approach in mathematics education: The role of ChatGPT]. *Uşak Üniversitesi Eğitim Araştırmaları Dergisi [Usak University Journal of Educational Research], 10*(1), 26–46. https://doi.org/10.29065/usakead.1393487

Karakaya, İ., Öç, U., & Gümüş, G. (2024). Üst düzey zihinsel becerilerin ölçülmesinde açık uçlu maddelerin yapay zekâ destekli uygulamalarda hazırlanması [Developing open-ended items supported by artificial intelligence for assessing higher-order thinking skills]. *Journal of Applied Measurement and Assessment, 1*(2), 59–70. https://izlik.org/JA72EM45TW

Kıyak, Y. S., & Emekli, E. (2024). ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: A literature review. *Postgraduate Medical Journal, 100*(1189), 858–865. https://doi.org/10.1093/postmj/qgae065

Köğce, D., & Baki, A. (2009). Matematik öğretmenlerinin yazılı sınav soruları ile ÖSS sınavlarında sorulan matematik sorularının Bloom taksonomisine göre karşılaştırılması [A comparison of high-school mathematics teachers' examination questions and mathematics questions in university entrance examinations according to Bloom's taxonomy]. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi [Pamukkale University Journal of Education], 26*(26), 70–80. https://izlik.org/JA53FX47SB

Koman, I., Bayrakceken, S., Oktay, O., & Canpolat, N. (2023). Evaluation of science teachers' exam and high school entrance exam science questions based on the revised Bloom taxonomy. *Science Insights Education Frontiers*, *16*(2), 2505-2536. doi: 10.15354/sief.23.or335

Kotsis, K. T. (2025). Artificial intelligence as a catalyst for changes in university-level science education. *EIKI Journal of Effective Teaching Methods*, 3(3). https://doi.org/10.59652/jetm.v3i3.618

Lee, J., Smith, D., Woodhead, S., & Lan, A. (2024). Math multiple choice question generation via human-large language model collaboration. *ArXiv Preprint Arxiv:2405.00864*.

Li, H., Fang, Y., Zhang, S., Lee, S. M., Wang, Y., Trexler, M., & Botelho, A. F. (2025). *ARCHED: A human-centered framework for transparent, responsible, and collaborative AI-assisted instructional design* (ArXiv:2503.08931). https://doi.org/10.48550/arXiv.2503.08931

Liu, Y., Zheng, Z., Deng, N., & Li, T. (2026). How ai-driven technological transformation advances the modernization integrated transition of Energy. *Economic Analysis and Policy*, *91*, 206–223. https://doi.org/10.1016/j.eap.2026.03.003

Maity, S., & Deroy, A. (2024). Human-centric explainable AI in education. *Arxiv Preprint Arxiv:2410.19822*. https://doi.org/10.48550/arXiv.2410.19822

May, T. A., Fan, Y. K., Stone, G. E., Koskey, K. L. K., Sondergeld, C. J., Folger, T. D., Archer, J. N., Provinzano, K., & Johnson, C. C. (2025). An effectiveness study of generative artificial intelligence tools used to develop multiple-choice test items. *Education Sciences,* 15(2), 144. doi: 10.3390/educsci15020144

McGalliard, W., & Otten, S. (2025). AI responses to challenging problems and educator responses to AI availability. *Digital Experiences in Mathematics Education*, 11, 319–332. https://doi.org/10.1007/s40751-024-00167-4

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282. https://pmc.ncbi.nlm.nih.gov/articles/PMC3900052/

Nayıroğlu, B., & Tutak T. (2023). The effect of using Web 2.0 tools in algebra teaching on student success and attitude. *International e-Journal of Educational Studies*, 7(14), 416–425. https://doi.org/10.31458/iejes.1270732

Omopekunola, M. O., & Kardanova, E. Y. (2024). Automatic generation of physics items with large language models (LLMs). *REID (Research and Evaluation İn Education)*, *10*(2), 4. https://doi.org/10.21831/reid.v10i2.76864

Organisation for Economic Co-operation and Development (OECD). (2023). *Education at a glance 2023: OECD indicators.* OECD Publishing. https://doi.org/10.1787/e13bef63-en

Owan, V. J., Abang, K. B., Idika, D. O., Etta, E. O., & Bassey, B. A. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. *Eurasia Journal of Mathematics, Science and Technology Education, 19*(8), https://doi.org/10.29333/ejmste/13428

Rahman, A. B. M., Yang, J., Zhan, Z., Hossain, M. K., Xu, Y., & Feng, W. (2025). Pedagogical transformation through artificial intelligence: Implementation barriers and adoption strategies in business education. *ICEB 2025 Proceedings*, *50*, 314-322. Retrieved from https://aisel.aisnet.org/iceb2025/50

Roll, I., & Wylie, R. (2016). Evolution and revolution in artificial intelligence in education. *International Journal of Artificial Intelligence in Education*, 26, 582-599. doi: 10.1007/S40593-016-0110-3

Scaria, N., Dharani Chenna, S., & Subramani, D. (2024). Automated educational question generation at different Bloom's skill levels using large language models: Strategies and evaluation. In *Artificial intelligence in education* (Lecture Notes in Computer Science, Vol. 14830, pp. 165–179). Springer. https://doi.org/10.1007/978-3-031-64299-9_12

Shaunessy, E. (2000). Techniques questioning in the gifted classroom? *Gifted Child Today*, 23(5), 14–21. doi: 10.4219/Gct-2000-752

Siegle, D. (2024). Using artificial intelligence (AI) technology to support the three legs of talent development. *Gifted Child Today*, 47(3), 221–227. doi: 10.1177/10762175241242495

Tan, X., Cheng, G., & Ling, M. H. (2025). Artificial Intelligence in teaching and teacher professional development: A systematic review. *Computers and Education: Artificial Intelligence*, *8*, 100355. doi:10.1016/j.caeai.2024.100355

Vantassel-Baska, J. (2014). Curriculum issues: Artful inquiry: the use of questions in working with the gifted. *Gifted Child Today*, 37(1), 48–50. doi: 10.1177/1076217513509621

Wan, X., Sun, R., Nakhost, H., Dai, H., Eisenschlos, J. M., Arık, S. Ö., & Pfister, T. (2023). Universal self-adaptive prompting. *In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 7437–7462). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.461

Wankhede, H. S., & Kiwelekar, A. W. (2016). Qualitative assessment of software engineering examination questions with Bloom's taxonomy. *Indian Journal of Science and Technology, 9*(6). doi: 10.17485/ijst/2016/v9i6/85012

Webb, N. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. National Institute for Science Education Madison, WI. Retrieved from http://facstaff.wceruw.org/ normw/WEBBMonograph6criteria.pdf.

Wongpakaran, N., Ward, P. R., Wongpakaran, T., & Sirisengtaksin, V. (2013). A comparison of reliability methods for psychiatric diagnosis: Cohen's kappa, Gwet's AC1, and Fleiss' kappa. *Annals of General Psychiatry*, *12*(1), Article 17. doi: 10.1186/1744-859X-12-17

Xu, W., & Ouyang, F. (2022). The application of AI technologies in STEM education: A systematic review from 2011 to 2021. *International Journal of STEM Education, 9*, Article 59. https://doi.org/10.1186/s40594-022-00377-5

Yaacoub, A., Assaghir, Z., Prevost, L., & Da-Rugna, J. (2025a). Analyzing feedback mechanisms in AI-generated MCQS: Insights into readability, lexical properties, and levels of challenge. *Arxiv Preprint Arxiv:2504.21013*. https://doi.org/10.48550/arXiv.2504.21013

Yaacoub, A., Da-Rugna, J., & Assaghir, Z. (2025b). Assessing AI-generated questions' alignment with cognitive frameworks in educational assessment. *International Journal of Computer Theory and Engineering*, 17 (3), 114-125. https://doi.org/10.7763/IJCTE.2025.V17.1374

Yoşumaz, İ. (2025). Generative artificial intelligence and usage in academia. *Fırat Üniversitesi Sosyal Bilimler Dergisi [Firat University Journal of Social Sciences], 35*(1), 1–24. https://doi.org/10.18069/firatsbed.1423208

Zhang, X., Zhang, P., Shen, Y., Wang, Y., Li, M., & Chen, H. (2024). A systematic literature review of empirical research on applying generative artificial intelligence in education. *Frontiers of Digital Education*, *1(3)*, 223–245. doi:10.1007/s44366-024-0028-5