REID (Research and Evaluation in Education), 10(2), 2024, 168-185
Available online at: http://journal.uny.ac.id/index.php/reid

Automatic generation of physics items with Large Language Models (LLMs)

Moses Oluoke Omopekunola*; Elena Yu Kardanova
National Research University Higher School of Economics (HSE University), Russian Federation
*Corresponding Author. E-mail: omopekunolamoses@gmail.com

ARTICLE INFO

Article History
Submitted:
08 August 2024
Revised:
23 September 2024
Accepted:
23 September 2024

Keywords
AIG; Bloom's taxonomy; ChatGPT; Gemini; LLM; physics items

Scan Me:

ABSTRACT

High-quality items are essential for producing reliable and valid assessments, offering valuable insights for decision-making processes. As the demand for items with strong psychometric properties increases for both summative and formative assessments, automatic item generation (AIG) has gained prominence. Research highlights the potential of large language models (LLMs) in the AIG process, noting the positive impact of generative AI tools like ChatGPT on educational assessments, recognized for their ability to generate various item types across different languages and subjects. This study fills a research gap by exploring how AI-generated items in secondary/high school physics aligned with educational taxonomy. It utilizes Bloom's taxonomy, a well-known framework for designing and categorizing assessment items across various cognitive levels, from low to high. It focuses on a preliminary assessment of LLMs ability to generate physics items that match the Bloom's taxonomy application level. Two leading LLMs, ChatGPT (GPT-4) and Gemini, were chosen for their strong performance in creating high-quality educational content. The research utilized various prompts to generate items at different cognitive levels based on Bloom's taxonomy. These items were assessed using multiple criteria: clarity, accuracy, absence of misleading content, appropriate complexity, correct language use, alignment with the intended level of Bloom's taxonomy, solvability, and assurance of a single correct answer. The findings indicated that both ChatGPT and Gemini were skilled at generating physics assessment items, though their effectiveness varied based on the prompting methods used. Instructional prompts, particularly, resulted in excellent outputs from both models, producing items that were clear, precise, and consistently aligned with the Application level of Bloom's taxonomy.

To cite this article (in APA style):
Omopekunola, M., & Kardanova, E. (2024). Automatic generation of physics items with Large Language Models (LLMs). *REID (Research and Evaluation in Education), 10*(2), 168-185. doi:https://doi.org/10.21831/reid.v10i2.76864

INTRODUCTION

High-quality items used for assessment play a crucial role in informing instructional decisions and providing valuable data for teachers, enabling them to effectively diagnose students' comprehension and challenges, track their progress, and assess the effectiveness of pedagogical methods (Darling-Hammond, 2015; Mystakidis et al., 2021). They serve as a powerful tool for teachers to implement timely interventions and provide support for student development, thereby directly impacting student achievement and grades (Archibald et al., 2011; Burns et al., 2017). They contribute to creating an inclusive learning environment by addressing different learning preferences and levels of prior knowledge, ultimately enhancing the overall quality of education (Collins et al., 1989; Tomlinson, 2017, 2023). Also, these items not only support immediate instructional goals but also contribute to long-term educational outcomes by fostering critical thinking, problem-solving, and a deeper understanding of the subject matter (Kong, 2014; Miri et al., 2007).

To effectively develop or design such items that cater to a wide range of students' needs, educational taxonomies have become essential as a gauge (Irvine, 2021; Veal & MaKinster, 1999).

ISSN: 2460-6995 (Online)

Bloom's taxonomy has gained more popularity across disciplines among the existing educational taxonomies (Agarwal, 2019; Tabrizi & Rideout, 2017). Bloom's taxonomy and its revised version by Krathwohl and Anderson (2010) is an important tool for categorizing assessment items based on cognitive skills (Bloom et al., 1956; Krathwohl & Anderson, 2010). This taxonomy provides a framework for designing and classifying assessment items into different levels, ranging from low to high. Bloom's taxonomy consists of six levels: knowledge, comprehension, application, analysis, evaluation, and synthesis, with the latter three requiring higher-order thinking and critical thought (Bloom et al., 1956). The revised taxonomy by Krathwohl and Anderson (2010) refines these categories and introduces new verbs for higher-order thinking, emphasizing deeper thinking and critical thinking skills, aiming to better reflect human thought and learning complexity. Higher-order cognitive skills, such as problem-solving and higher understanding, are essential for student learning (Alsubait, 2015). Therefore, it is necessary for educational assessments to address both lower and higher levels of cognitive skills. This comprehensive approach ensures that assessments measure a wide range of abilities, providing a more comprehensive understanding of student learning (Adams, 2015; Crowe et al., 2008; Krathwohl, 2002).

The cognitive level of an item can be determined by assessing the cognitive processes, sometimes called "thinking skills" required to engage with the item (Li, 2021; Scully, 2017). In the field of education, Bloom's taxonomy is widely used to categorize items into different levels according to their thinking skills. The original (knowledge, comprehension, application, analysis, evaluation, and synthesis) and its revised version consist of six cognitive levels, where the verb forms replace the noun forms of the original category labels. They are remembering, understanding, applying, analyzing, evaluating, and creating (Krathwohl & Anderson, 2010). Remembering involves retrieving relevant knowledge from long-term memory. Its associated verbs are *state, define, choose, name*, and *write*, among others. Understanding is about constructing meaning from messages through various cognitive processes. Its verbs include *draw, explain, give a reason,* and *show*. The next level is applying, which refers to using procedures or implementing knowledge. Its verbs are *calculate, solve, sketch, determine*, and *describe*. Analyzing involves breaking down material and understanding relationships. The verbs are *analyze, illustrate*, and *classify*. Evaluating requires making judgments based on criteria. The associated verbs are *evaluate, formulate,* and *derive*. Creating involves putting elements together to form a coherent form. Its verbs are *create, discuss, modify, estimate,* and *design*. While the verbs listed under each level are not exhausted, they are only common verbs related to physics vocabulary. These levels of learning are essential for effective education and cognitive development (Krathwohl & Anderson, 2010).

There are various approaches to such categorization, such as statistical methods by utilizing natural language processing (NLP) techniques or machine learning (ML) (Abduljabbar & Omar, 2015; Chang & Chung, 2009; Yahya et al., 2012). Another approach is manual classification using subject matter expert (SME), which involves matching the keywords in the item with the cognitive verbs associated with each level of the taxonomy and determining the cognitive level (Ingwersen, 1996; Karabenick et al., 2007; Mohammed & Omar, 2020). For this study, the latter was used by researchers.

Due to the increasing demand for more items with good psychometric properties for large-scale summative and formative assessments, automatic item generation (AIG) has gained popularity in more good items for large-scale assessments (Attali, 2018). An earlier attempt at AIG uses computer algorithms to generate items based on cognitive and item models developed by human experts, followed by expert evaluation (Arendasy & Sommer, 2007; Bejar, 2002; Embretson, 2005; Embretson & Yang, 2007; Gierl & Haladyna, 2012; Glas & van der Linden, 2003; Gorin, 2006). Although effective, AIG heavily relies on human input and may not produce diverse items in terms of content and structure. The method has proven effective in generating high-quality items but still requires human input. This approach addresses challenges like the cost and difficulty of maintaining a large inventory, as well as the need for inventive item categories for higher abilities (Attali, 2018).

AIG also allows for item replacement at any time, mitigating item exposure (Attali, 2018; Kurdi et al., 2019). However, both cognitive and item approaches are not without limitations, including limited distinct item generation and potential deletion of items due to insufficient psychometric characteristics (Arendasy & Sommer, 2012; Embretson, 2005). Hence, test developers are now turning to exploring large language models (LLMs) as alternative methods of creating items automatically (Borji, 2023; Dao et al., 2023; Gregorcic & Pendrill, 2023; Küchemann et al., 2023; Offerijns et al., 2020) due to their capacity to create coherent and information-rich sequences of text (Wang et al., 2021; Yu et al., 2022). This approach can reduce costs in item generation, provided the tools are trained to perform at a level comparable to human item writers.

Large language models (LLMs) have significantly transformed the field of natural language processing (NLP) through the application of deep learning methodologies, particularly leveraging transformer architectures to facilitate the understanding and generation of human-like text, as evidenced by the foundational work of Vaswani et al. (2017). These models undergo extensive pre-training on large-scale datasets, followed by fine-tuning for specific tasks, enabling them to execute a variety of complex language-related functions, including text summarization, translation, and conversational AI, as highlighted by Feng et al. (2023). The emergence of advanced models such as OpenAI's GPT-4 and Google's Gemini has established new benchmarks in language comprehension; GPT-4 demonstrates superior capabilities in natural language understanding and generation, while Gemini showcases enhanced multimodal functionalities that integrate both textual and visual data, according to Roumeliotis et al. (2023). LLMs exhibit remarkable proficiency in capturing contextual relationships, managing long-distance dependencies, and generalizing across diverse language tasks, thereby surpassing traditional approaches like n-grams and Hidden Markov Models in efficiency and accuracy, as discussed by Perikos et al. (2021). Their fine-tuning aspect allows specialization in applications ranging from creative writing and customer service to legal research, code generation, and medical diagnostics, as documented by Hou et al. (2023) and Liu et al. (2023). Furthermore, advancements in neural transformer architectures have enabled LLMs to generate long, coherent, and information-rich sequences of text (Devlin et al., 2019; Radford et al., 2019; Vaswani et al., 2017). Since the launch of the first GPT series and subsequent iterations (Devlin et al., 2019; Radford et al., 2019; Zhang & Li, 2021), ongoing studies have evaluated their performance in generating high-quality items across various contexts (see Bulut et al., 2024; Tan et al., 2024). These capabilities make LLMs especially useful in Automated Item Generation (AIG), where they can efficiently produce diverse, high-quality test items by leveraging their deep understanding of language patterns and content generation (Shen et al., 2023).

Küchemann et al. (2023) provided empirical evidence on the integration of ChatGPT 3.5 in the creation of physics tasks by prospective physics teachers for 10th-grade students. The study involved evaluating the quality of items developed using ChatGPT 3.5 compared to traditional textbook methods. The evaluation criteria included specificity, clarity, correctness, context, relevance, and overall quality. The findings revealed disparities in the quality and effectiveness of tasks created with ChatGPT 3.5. While some tasks demonstrated high levels of specificity, clarity, relevance, and correctness, others faced challenges in these areas. A comparison with tasks generated through traditional textbooks highlighted the potential advantages of incorporating large language models like ChatGPT 3.5 in physics task development.

Santos (2023) conducted a study to assess the effectiveness of GenAIbots, specifically ChatGPT-4, as thinking agents in physics education. The research involved collecting data from GenAIbot interactions and reflective journals and analyzing the data to identify recurring themes and patterns. The analysis focused on understanding how GenAIbots impact physics learning, considering tutor characteristics like subject knowledge, empathy, and understanding of the learning process. The study found that GenAIbots, particularly ChatGPT-4, could effectively serve as thinking agents in physics education, promoting critical thinking, problem-solving, and personalized learning. However, although the GenAIbots demonstrated subject-matter knowledge and empathy, they displayed inconsistencies, underscoring the necessity of human intervention in AI supported learning environments.

In mathematics, Bhandari et al. (2024) conducted a study comparing human-designed textbook items with ChatGPT-generated items in the context of college algebra assessments. The study aimed to evaluate the ability of ChatGPT-generated questions to differentiate between various ability levels and assessed their discriminating power compared to textbook questions. The research utilized a psychometric linking/equating strategy to ensure calibration results comparability across different test phases. The study focused on assessing the quality of items generated by ChatGPT in comparison to gold standard questions from a published College Algebra textbook. The researchers employed item response theory (IRT) as a psychometric measurement approach to analyze data from 207 test respondents answering items from an online standard mathematics textbook, OpenStax College Algebra. The results indicated that ChatGPT-generated items were comparable to human-authored textbook questions in assessing students' abilities in college algebra.

Laverghetta Jr and Licato (2023) introduced a new prompting strategy for item generation using GPT-3, aiming to enhance the diversity and quality of items by selecting those with both the best and worst properties. The study focused on natural language inference tasks and utilized GPT-3 to generate new items, emphasizing the importance of strategic item selection in improving item quality iteratively. The research highlighted the significance of careful prompting in leveraging large language models (LLMs) to enhance cognitive assessments. The study demonstrated that items generated by GPT-3 showed improved psychometric properties in many cases, as evidenced by analyses of item difficulty. These findings suggest that LLMs can positively contribute to the item development process in cognitive assessments by producing items with enhanced quality, which is essential for establishing the validity and reliability of assessments.

Bezirhan and von Davier (2023) shifted the focus to the field of reading comprehension, where they rigorously assessed the effectiveness of OpenAI's GPT-3 for generating passages suitable for educational assessments. They utilized specific prompts and parameters to generate informational and literary passages using GPT-3, conducting 10 replications for each setting. The text difficulty scores of the generated passages were computed, and passages with similar scores to the original ones were chosen. Human editors then checked these selected passages for errors, coherence, clarity, and consistency. In an online survey, the AI-generated passages were compared to the original PIRLS passages using a Likert scale to evaluate factors like reading level suitability, coherence, identifying the main topic, and engaging children. The results indicated that AI-generated passages had the potential to align with educational standards, with minor variations in coherence ratings.

In a more recent study, Doughty et al. (2024) presented a pioneering empirical on the use of LLMs for creating multiple-choice questions (MCQs) in programming education. The study assessed the quality and effectiveness of AI-generated 1,100 MCQs compared to traditionally crafted questions, focusing on factors such as clarity, single correct answer, distractor quality, and syntactic/logical correctness of code. The study also assessed how well these MCQs aligned with pre-defined module-level learning objectives by generating MCQs directly from learning objectives (LOs), a departure from traditional methods that use short course materials. This approach ensures a direct link between assessment and intended learning outcomes. The evaluation shows a stronger correlation between LLM questions with LOs than human-crafted questions, highlighting the potential of LLMs to generate assessments that directly target intended learning outcomes. The studies reviewed highlight the positive impact of GenAI tools like ChatGPT on educational assessment practices, emphasizing the importance of combining artificial intelligence with human expertise for efficient item generation and content quality, paving the way for further research into the collaborative potential of human and artificial intelligence.

Prompt engineering is a crucial technique in the field of artificial intelligence and machine learning, particularly for large language models (LLMs) and generative systems. It involves constructing input text, known as prompts, to effectively convey the user's intention and influence the model's output (Marvin et al., 2023). The quality and relevance of the model's output greatly depend on prompt engineering (Brown et al., 2020). As such, in generating test items using LLMs, prompt engineering plays a vital role in enhancing the effectiveness and quality of the items produced. Several strategies for designing precise prompts have been proposed, including setting clear objectives, using appropriate language and tone, providing context, examples, and references, specifying the expected output format, and incorporating essential details (Bozkurt & Sharma, 2023; Liu et al., 2023). By implementing these strategies, the generation of irrelevant or erroneous outputs can be mitigated, leading to enhanced overall performance in both educational and general applications of LLMs. By employing various prompting techniques, test developers and educators can tailor items to align with specific educational frameworks, such as Bloom's taxonomy, ensuring that the generated items address specific cognitive levels and enhance students' learning of complex concepts.

Some of the prompting strategies are zero-shot prompting, one-shot prompting, few-shot prompting, and instructional prompting. Zero-shot prompting is a method where a model is given a prompt without any prior examples or context, relying solely on its pre-trained knowledge to generate a response (Zhong et al., 2023). This approach has been proven effective in various standard natural language understanding tasks but may not always provide the necessary specificity and accuracy for specialized problems (Miao et al., 2024). One-shot prompting involves providing a model with a single example to guide its response, serving as a pattern or template (Polat et al., 2024). This approach is especially advantageous when the model needs to understand the desired format and depth of questions, such as distinguishing between basic knowledge questions and those requiring higher-order thinking skills (Song et al., 2023). Few-shot prompting involves providing a model with a small number of examples to guide its understanding of a task before presenting the main query. This approach offers more efficient in-context learning, resulting in more generalized and task-specific outcomes (Agarwal et al., 2024; Gao et al., 2020; Miao et al., 2024). However, the order in which examples are given can also impact the model's performance. Research has shown that providing the instruction before presenting the example yields better results (Li et al., 2024; Minaee et al., 2024). Instructional prompting is another strategy for providing LLMs with clear instructions in plain text. This method enhances performance and streamlines task completion by making complex instructions easily understandable. It is also effective for creating educational content and aligning with curriculum standards and testing objectives (Mishra et al., 2021).

However, it is important to note that the way prompts are framed plays a crucial role in the accuracy of task execution. Clear and straightforward prompts tend to yield higher accuracy compared to vague or overly detailed prompts. Therefore, it is important to apply different prompting strategies to evaluate the quality of items generated and provide documentation that outlines the purpose and provides guidance using prompt engineering strategies and the example of prompts with the items generated for physics educators.

Previous studies highlight the potential of LLMs in the AIG process in assessing students' abilities, and in physics in particular (Bhandari et al., 2024). They reveal the characteristics of generated items and the flexibility and efficiency of this approach in generating various types of items across languages and subject domains (Minaee et al., 2024; Rangapur & Rangapur, 2024; Tan et al., 2024). However, these AIG studies lack solid educational foundations in aligning items generated with assessment purposes and integrating measurement and learning theories into the AIG process. Hence, there is a need to provide a further study on the work of Küchemann et al. (2023) where evidence was gathered that ChatGPT generates items comparable to items generated from physics textbooks and expands the scope by generating items that align with educational taxonomies using different LLMs prompting methods.

The current study addresses the research gap by examining the alignment between AI-generated items in physics and educational taxonomies. Specifically, the research provides a preliminary evaluation of LLMs' abilities to generate physics items that align with Bloom's taxonomy. The research answers the following questions:

1. What is the quality of items generated by ChatGPT and Gemini under each prompting strategy?
2. To what extent can ChatGPT and Gemini generate items of different levels according to Bloom's taxonomy?
3. What is the best prompting strategy for generating high-quality physics items for both LLMs?

The evaluation focuses on predefined criteria such as clarity, correctness, non-misleading content, adequate difficulty, use of appropriate language, alignment with the intended Bloom's taxonomy, and that the item is solvable and leads to a single solution. This research lays the groundwork for future investigations that will include expert evaluation to refine and validate the items generated by these as well as other LLMs.

METHOD

The research was executed by adopting usability testing as a research method, involving a comparative analysis of items generated by LLMs. As an evaluative research method, usability testing assisted test developers in measuring the effectiveness of a concept or skill through specific items that resonate with the cognitive, affective, and psychomotor domains of learners (Barnum, 2020). Different prompts – instructional, zero-shot, one-shot, and few-shot – were utilized to generate items at different cognitive levels according to Bloom's taxonomy. All prompts were carefully crafted by the researchers following prompt engineering techniques. For each prompt strategy, several prompts were tested and the one that gave the best results according to the item evaluation criteria was chosen under each prompt strategy for item generation. This allowed for generating items of different levels of difficulty and specificity, enabling a tailored assessment of 50 items generated for each prompting strategy.

These items were evaluated based on criteria such as clarity, correctness, non-misleading content, adequate difficulty, use of appropriate language, and alignment with the intended Bloom's taxonomy, and the item is solvable and leads to a single solution. Clarity refers to how easily the test item can be understood. A clear item avoids ambiguity or confusing wording, ensuring that test-takers comprehend exactly what is being asked without misinterpretation. Correctness checks for the accuracy of the concepts and facts, ensuring the item is free from errors in content and structure. "Non-misleading content" means that the item must present its information and choices in a way that does not unintentionally confuse or mislead the test-taker. The phrasing should reflect the intended challenge without introducing unnecessary traps. Adequate difficulty means the item should be both easy and challenging, providing a suitable level of difficulty for the intended test-takers. Use of appropriate language is not only limited to the use of appropriate language but most importantly the use of physics vocabulary and familiar illustration. Alignment with the intended Bloom's taxonomy ensures that the item targets the intended cognitive level, such as remembering, understanding, or applying. The item should match the desired Bloom's taxonomy level as determined by the test's goals. This criterion "solvable and the item leads to a single solution" tests the solvability of an item and checks if it contains a single solution (in the case of multiple-choice items) without considering approximate options as solutions unless otherwise stated. Each parameter was evaluated on a binary scale with a score of 1 indicating a *yes* and 0 indicating a *no*. One out of the three major domains of the high physics curriculum (Mechanics, Optics, and Electricity), mechanics was purposefully chosen. Within the selected domain, the Projectile motion (or simply projectiles) topic was chosen for item generation.

The study used ChatGPT (GPT-4) and Gemini due to their advanced language models in generating high-quality educational content (Brown et al., 2020; Radford et al., 2019). Gemini, Google's next-generation large language model, has been praised for its capability to handle mathematical problems (Islam & Ahmed, 2024; Lorenzo, 2024). The current study employed the premium version of these LLMs to assess the effectiveness of applying the level of Bloom's taxonomy in generating physics assessment items for students. This is essential to understand how students can have more than mere knowledge from the subject and the need to have many items of this level as it occupies the largest percentage of physics assessment items among the six levels of Bloom's taxonomy (Motlhabane, 2017). Specifically, the selected cognitive level and item prompting strategies are expected to lead to items that contribute to the practical problem-solving abilities of the students, which usually involve calculations (Buick, 2011).

FINDINGS AND DISCUSSION

Comparison of the Quality of Items Generated by ChatGPT and Gemini under each Prompting Strategy

This section presents the results obtained from physics item generation by ChatGPT (GPT-4) and Gemini under different prompts and evaluation of items generated. The results obtained from this analysis provide a comparative analysis of the performance of ChatGPT and Gemini in generating high-quality items for the applying cognitive level under different prompting strategies. The answers are presented based on the research questions formulated earlier. In comparing the items generated by ChatGPT and Gemini, it is evident that both large language models (LLMs) have their strengths and weaknesses across different prompting strategies.

Prompting Strategy/Method:
Zero-Shot
Prompt Sample:
Generate a multiple-choice question on projectile motion that tests application skills as per Bloom's taxonomy. Include four options and identify the correct answer
Cognitive Level:
Apply
Item Generated by ChatGPT:
A basketball player throws a ball at a 60-degree angle to the horizontal with an initial speed of 15 meters per second. If the player releases the ball from a height of 2 meters, which of the following is the approximate time it will take for the basketball to hit the ground? Assume there is no air resistance.
A) 1.5 seconds
B) 2.1 seconds
C) 2.8 seconds
D) 3.5 seconds
Correct Answer:
C) 2.8 seconds
Item Generated by Gemini:
A soccer player needs to kick a ball from a distance of 30 meters to reach the goal directly. If the maximum speed the player can give the ball is 20 m/s, at approximately what angle should the ball be launched to reach the goal (neglect air resistance)? Use $g = 10$ m/s²
Options:
A. 30°
B. 45°
C. 60°
D. 75°
Correct Answer:
C. 60°

Figure 1. Items Generated Using Zero-Shot Prompting Strategy

Zero-Shot Prompting

Figure 1 shows examples of the items generated using zero-shot prompting, while the result of their evaluation is presented in Table 1. The items generated using this prompting strategy are multiple-choice type with four options (one key option and three distractors). ChatGPT performs well in clarity, alignment with the application level of Bloom's taxonomy, non-misleading content, adequate difficulty, appropriate language use, and solvability. However, it struggles with providing some details like indicating the numerical value of the gravitational constant *g*, which may confuse test takers. Gemini shows consistent performance, excelling in clarity, correctness, alignment, and non-misleading content.

Table 1. Rating of Zero-Shot Prompting

Evaluation Criteria	ChatGPT	Gemini
Clarity	1	1
Correctness	0	1
Alignment	1	1
Non-Misleading	1	1
Adequate Difficulty	1	0
Use of Appropriate Language	1	1
Solvable and Single Solution	1	0
Total	6	5

One-Shot Prompting

Figure 2 shows the examples generated using one-shot prompting. From the rating table (Table 2), both ChatGPT and Gemini perform well in most evaluation criteria for both items, scoring highly in clarity, correctness, alignment, non-misleading nature, and solvability. ChatGPT items show a significant deficiency in using appropriate language for the item. While Gemini demonstrates exemplary performance, achieving a perfect score, and consistently meeting all evaluation criteria.

Prompting Strategy/Method:
One-Shot
Prompt Sample:
Generate a question about projectile motion that involves calculating the height, time, and horizontal displacement (range) of the motion. Here is an example; During a fireworks display, a shell is shot into the air with an initial speed of 70.0 m/s at an angle of 75.0° above the horizontal, as illustrated in Figure 3.36. The fuse is timed to ignite the shell just as it reaches its highest point above the ground. (a) Calculate the height at which the shell explodes. (b) How much time passed between the launch of the shell and the explosion? (c) What is the horizontal displacement of the shell when it explodes?
Cognitive Level:
Apply
Item Generated by ChatGPT:
In a sports science experiment, a football is kicked with an initial speed of 50.0 m/s at an angle of 60.0° above the horizontal, as illustrated in the accompanying figure (not shown). The experiment is designed to measure the ball's behavior at its peak trajectory.
Questions:
(a) Calculate the height at which the football reaches its maximum point.
(b) How much time passed between the kick and the football reaching its maximum height?
(c) What is the horizontal displacement of the football when it reaches its maximum height?
Item Generated by Gemini:
A toy rocket is launched at an angle of 40 degrees above the horizontal with an initial speed of 25 m/s. Neglecting air resistance, calculate:
a) The maximum height reached by the rocket.
b) The time it takes for the rocket to reach its maximum height.
c) The horizontal distance the rocket travels before landing.

Figure 2. Items Generated Using One-Shot Prompting Strategy

Table 2. Rating of One-Shot Prompting

Evaluation Criteria	ChatGPT	Gemini
Clarity	1	1
Correctness	1	1
Alignment	1	1
Non-Misleading	1	1
Adequate Difficulty	1	1
Use of Appropriate Language	0	1
Solvable and Single Solution	1	1
Total	6	7

Prompting Strategy/Method:
Few-Shot
Prompt Sample:
Generate multiple-choice questions about projectile motion with moderate levels of difficulty. I will provide some examples of questions low, medium and high their difficulty levels
Example 1 (Low Difficulty)
A ball is thrown horizontally from a cliff. Which of the following is true about its initial vertical velocity?
A. Zero
B. Equal to its horizontal velocity
C. Equal to the acceleration due to gravity
D. It depends on the height of the cliff
Example 2 (Medium Difficulty)
A projectile is launched at an angle of 30 degrees above the horizontal with an initial speed of 30 m/s. What is the approximate time it takes for the projectile to reach its maximum height? (Use g = 10 m/s²)
A. 1.5 s
B. 3 s
C. 6 s
D. 9 s
Example 3 (High Difficulty)
Two projectiles, A and B, are launched from the same point with the same initial speed. Projectile A is launched at a steeper angle than projectile B. Neglecting air resistance, which of the following statements is true about the projectiles?
A. Projectile A will have a greater maximum height and a longer time of flight than Projectile B.
B. Projectile A will have a greater maximum height and a shorter time of flight than Projectile B.
C. Projectile A and B will have the same maximum height and the same time of flight.
D. The relationship between their maximum heights and times of flight cannot be determined with the given information.
Cognitive Level:
Apply
Item Generated by ChatGPT:
A football is kicked with an initial velocity of 20 m/s at an angle of 45 degrees to the horizontal. Assuming no air resistance and using g = 10 m/s² for gravity, calculate the total time the football spends in the air before hitting the ground.
A. 2 s
B. 3 s
C. 4 s
D. 5 s
Item Generated by Gemini:
A toy car is launched down a ramp inclined at 30 degrees with respect to the horizontal. The car leaves the ramp at a point 1 meter above the ground with an initial speed of 5 m/s. Neglecting air resistance, what is the horizontal distance the car travels before hitting the ground?
Options:
A. 1 meter
B. 2 meters
C. 3 meters
D. 4 meters

Figure 3. Items Generated Using Few-Shot Prompting Strategy

Few-Shot Prompting

Figure 3 presents the examples generated using few-shot prompting. The results (see Table 3) indicate that ChatGPT performed well in several metrics but scored 0 for adequate difficulty, suggesting a failure to produce items with adequate challenge. Gemini performed slightly better, scoring consistently high across almost all metrics, indicating clear and accurate items.

Table 3. Rating of Few-Shot Prompting

Evaluation Criteria	ChatGPT	Gemini
Clarity	1	1
Correctness	0	1
Alignment	1	1
Non-Misleading	1	1
Adequate Difficulty	0	0
Use of Appropriate Language	1	1
Solvable and Single Solution	0	0
Total	5	6

Prompting Strategy/Method:
Instructional Prompting
Prompt Sample:
As a good test developer, develop a multiple-choice question aimed at grade 10 students that assesses higher-order cognitive skills according to Bloom's taxonomy, specifically focusing on the 'applying' level. Ensure the question is formulated clearly and is appropriate for the educational level, using terminology and concepts suitable for grade 10 physics students. Indicate the correct answer and incorporating the relevant physics principles and the application of equations.
Chain:
Topic: Projectiles
Bloom's taxonomy Level: Applying
Question Format: Multiple choice with a focus on application of formulas and concepts
Difficulty: High
Cognitive Level:
Apply
Item Generated by ChatGPT:
A soccer player kicks a ball from the ground with an initial velocity of 20 m/s at an angle of 30° above the horizontal. Assume that air resistance is negligible and the acceleration due to gravity is 9.8 m/s².
Which of the following is closest to the horizontal distance (range) the ball will travel before hitting the ground?
A) 17.7 meters
B) 35.4 meters
C) 44.9 meters
D) 59.0 meters
Answer: B
Item Generated by Gemini:
A catapult launches a projectile that needs to clear a castle wall 15 m high and located 60 m away. What approximate minimum launch speed is required? (Assume the projectile is launched from ground level.) $g = 9.8$ m/s².
A) 90 m/s
B) 45 m/s
C) 30 m/s
D) 35.9 m/s
Answer: C

Figure 4. Items Generated Using the Instructional Prompting Strategy

Instructional Prompting

The examples of items generated using the instructional prompting strategy are presented in Figure 4. The results of their evaluation (Table 4) show that both ChatGPT and Gemini performed exceptionally well in generating items with perfect scores across all evaluated metrics. Gemini items engage requires a more complex calculation involving finding the optimal launch angle and then determining the minimum initial velocity needed to clear the wall. Gemini generated a harder item as indicated in the prompt. It is important to note that using the same prompt may lead to items that receive the same or close ratings; however, further study is needed to confirm these findings.

Table 4. Rating of Instructional-Shot Prompting

Evaluation Criteria	ChatGPT	Gemini
Clarity	1	1
Correctness	1	1
Alignment	1	1
Non-Misleading	1	1
Adequate Difficulty	1	1
Use of Appropriate Language	1	1
Solvable and Single Solution	1	1
Total	7	7

Alignment of LLMs Items with Bloom's Taxonomy Level

According to the analysis of items generated, both ChatGPT and Gemini can generate items that align with the application level of Bloom's taxonomy. This alignment is evaluated through the use of appropriate action verbs e.g., "calculate," "solve" and the cognitive requirements of the tasks. However, it is important to note that some words or phrases in physics vocabulary do not have the same meaning in their natural context, and such might be wrongly placed in different levels. For example, question words like "What", "How far", and "How fast" are words that can be used in place of calculating, calculating the distance, and calculating the time respectively.

Performance by Prompting Strategy

Under zero-shot prompting, the items generated by both models were clear and correct but lacked alignment with the targeted Bloom's taxonomy levels. The generated items were of low difficulty, making explicit categorization under the intended levels challenging. The use of one-shot and few-shot prompting significantly improved the models' performance. Both models achieved high scores in aligning with the application level of Bloom's taxonomy, particularly through the use of the action verb calculate, which was included in the examples provided to the models. Instructional prompting yielded the best results for both models regarding alignment with the targeted application level of Bloom's taxonomy and appropriate difficulty. This strategy excelled because the intended level was explicitly stated in the prompt, reducing ambiguity. The generated items effectively tested the application level of Bloom's taxonomy and were suitable for students familiar with basic projectile motion concepts. The items generated under instructional prompting were of adequate difficulty, requiring test takers to apply kinematic equations similar to calculating basic circuit variables. Gemini's item presented a higher cognitive demand compared to ChatGPT, as it involved deeper circuit analysis, including scenarios with an inductor and a capacitor in a steady-state condition. This question targeted fundamental concepts of LC circuit behavior and included a clear circuit diagram, guiding students to apply steady-state behaviors and phase relationships in the circuit.

Prompting Strategy for Generating High-Quality Physics Items

The quality of physics items produced by LLMs is significantly affected by the prompting strategy employed. As detailed in Figure 1, zero-shot prompting involves generating items without providing examples. This approach often results in basic items that may be incomplete or less challenging. For instance, while ChatGPT can produce clear and aligned questions, it sometimes omits critical details, such as the gravitational constant in physics problems, which negatively impacts correctness and solvability. Presented in Figure 2, one-shot prompting involves providing a

ISSN: 2460-6995 (Online)

single example. This strategy generally enhances the clarity, alignment, and relevance of the generated items, leading to improved performance compared to zero-shot prompting. Detailed in Figure 3, few-shot prompting includes several examples to guide the model. This technique tends to produce even higher-quality items by further enhancing clarity, correctness, and alignment. However, both ChatGPT and Gemini may still struggle with ensuring adequate difficulty and solvability, as the models often find it challenging to adjust the difficulty of items based on the limited examples provided. This limitation is a notable drawback of few-shot prompting.

As shown in Figure 4, instructional prompting involves detailed instructions and multiple examples, yielding the highest quality items. Under this strategy, both LLMs perform exceptionally well, generating items that are clear, accurate, appropriately challenging, and aligned with Bloom's taxonomy at the application level. Instructional prompting ensures that the generated items are comprehensive and capable of testing deeper understanding and complex concepts. The analysis indicates that instructional prompting is the most effective strategy for generating high-quality physics items. It allows for the creation of items that not only meet clarity and correctness standards but also challenge students' ability appropriately and align with educational frameworks.

CONCLUSION

The findings of this study contribute to the growing body of research on the application of Large Language Models (LLMs) for Automatic Item Generation (AIG). By employing Bloom's taxonomy as a framework, the study focuses on aligning the generated items with Bloom's taxonomy, with a particular emphasis on the application level of Bloom's taxonomy, thereby addressing a significant gap in the existing literature on AI-assisted educational assessments. The study employed a variety of prompting strategies, including zero-shot, one-shot, few-shot, and instructional prompts to assess the capabilities of ChatGPT (GPT-4) and Gemini in producing high-quality assessment items. These items were evaluated based on criteria including clarity, correctness, non-misleading content, adequate difficulty, use of appropriate language, alignment with the intended Bloom's taxonomy, and the item is solvable and leads to a single solution. Each parameter was evaluated on a binary scale with a score of 1 indicating a *yes* and 0 indicating a *no*.

The study extends previous work by comparing the performance of two prominent LLMs, ChatGPT and Gemini, in generating physics assessment items. The results indicate that both ChatGPT and Gemini can generate high-quality physics assessment items, although their effectiveness varies depending on the prompting method used. Instructional prompts, in particular, yielded exceptional results from both models, producing items that were clear, precise, and consistently aligned with the application level of Bloom's taxonomy. This finding underscores the importance of prompt engineering in guiding LLMs to generate desired outputs. The consistent results obtained from repeated trials using the same prompts under different prompting strategies highlight the ability of both models to generate clear, accurate, and cognitively appropriate assessment items demonstrating the potential of LLMs as valuable tools in Automatic Item Generation (AIG). This consistency implies that the models can reliably produce high-quality items, reinforcing their potential as tools for educators.

The potential of LLMs in AIG, as demonstrated in this study, aligns with the growing body of research on AI-assisted educational assessment. This study contributes to the existing literature on LLMs in educational assessment and provides additional evidence supporting their application in physics education. The study's focus on physics as a fundamental STEM subject highlights the importance of developing effective assessment tools in this area. By effectively meeting the increasing demand for high-quality assessment items, LLMs can alleviate the burden on educators and assessment developers. Additionally, the comparison between ChatGPT and Gemini offers new insights into the relative strengths of different LLMs in this context, which has not been extensively explored in previous literature. The findings support the viability of using LLMs for generating assessment items across STEM subjects, as evidenced by previous studies in physics and mathematics (Bhandari et al., 2024; Küchemann et al., 2023) which found that ChatGPT generated questions were comparable to human-authored textbook items by experts in assessing students' abilities. The ability of LLMs to generate diverse, contextually appropriate, and cognitively aligned items addresses a significant challenge in educational assessment, particularly in STEM subjects like physics. Physics education plays a vital role in developing critical thinking, problem-solving skills, and scientific literacy, which are essential for students' future academic and professional success. The ability to generate high-quality physics items that target specific cognitive levels can enhance the effectiveness of physics education by providing more tailored and diverse assessment opportunities.

The current study focused on generating and evaluating physics items only for the applying cognitive level of Bloom's taxonomy. While the evaluation criteria used were comprehensive, it was self-evaluation without empirical data. Future research could benefit from expert evaluation to further validate the quality and educational value of AI-generated items. Additionally, expanding the research to cover all levels of Bloom's taxonomy and gathering empirical data for piloting would provide a more complete understanding of LLMs' capabilities in educational assessment, hence, paving the way for LLMs' items for large-scale assessments.

DISCLOSURE STATEMENT

The authors declare that they have no conflict of interest to disclose.

REFERENCES

Abduljabbar, D. A., & Omar, N. (2015). Exam questions classification based on Bloom's taxonomy cognitive level using classifiers combination. *Journal of Theoretical and Applied Information Technology*, *78*(3), 447–455.

Adams, N. E. (2015). Bloom's taxonomy of cognitive learning objectives. *Journal of the Medical Library Association: JMLA*, *103*(3), 152–153. https://doi.org/10.3163/1536-5050.103.3.010

Agarwal, P. K. (2019). Retrieval practice & Bloom's taxonomy: Do students need fact knowledge before higher order learning? *Journal of Educational Psychology*, *111*(2), 189–209. https://doi.org/10.1037/edu0000282

Agarwal, R., Singh, A., Zhang, L. M., Bohnet, B., Chan, S., Zhang, B., Anand, A., Abbas, Z., Nova, A., Co-Reyes, J. D., Chu, E., Behbahani, F., Faust, A., & Larochelle, H. (2024). *Many-shot in-context learning. arXiv*. https://doi.org/10.48550/arXiv.2404.11018

Alsubait, T., Parsia, B., & Sattler, U. (2015). Generating multiple choice questions from ontologies: How far can we go? In P. Lambrix, E. Hyvönen, E. Blomqvist, V. Presutti, G. Qi, U. Sattler, Y. Ding, & C. Ghidini (Eds.), *Knowledge engineering and knowledge management (EKAW 2014): Lecture notes in computer science* (vol. 8982, pp. 66–79). Springer. https://doi.org/10.1007/978-3-319-17966-7_7

Attali, Y. (2018). Automatic item generation unleashed: An evaluation of a large-scale deployment of item models. In In C. P. Rosé, R. Martínez-Maldonado, H. U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren, & B. du Boulay (Eds.), *Artificial Intelligence in Education: The 19th International Conference, AIED 2018* (pp. 17–29). Springer. https://doi.org/10.1007/978-3-319-93843-1_2

Archibald, S., Coggshall, J. G., Croft, A., & Goe, L. (2011). *High-Quality professional development for all teachers: Effectively allocating resources* [Research & policy brief]. National Comprehensive Center for Teacher Quality. https://files.eric.ed.gov/fulltext/ED520732.pdf

Arendasy, M., & Sommer, M. (2007). Using psychometric technology in educational assessment: The case of a schema-based isomorphic approach to the automatic generation of quantitative reasoning items. *Learning and Individual Differences*, *17*(4), 366–383. https://doi.org/10.1016/j.lindif.2007.03.005

Barnum, C. M. (2020). *Usability testing essentials: Ready, set... test!* (2nd ed.). Morgan Kaufmann.

Bejar, I. I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 199–217). Lawrence Erlbaum Associates Publishers.

Bezirhan, U., & von Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. *Computers and Education: Artificial Intelligence*, *5*, 1–13. https://doi.org/10.1016/j.caeai.2023.100161

Bhandari, S., Liu, Y., Kwak, Y., & Pardos, Z. A. (2024). Evaluating the psychometric properties of ChatGPT-generated questions. *Computers and Education: Artificial Intelligence*, *7*, 1–9. https://doi.org/10.1016/j.caeai.2024.100284

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals (Handbook I: Cognitive domain)*. Longmans.

Borji, A. (2023). Stochastic Parrots or Intelligent Systems? A perspective on true depth of understanding in LLMs. *SSRN Electronic Journal*, 1–10. https://doi.org/10.2139/ssrn.4507038

Bozkurt, A., & Sharma, R. C. (2023). Challenging the status quo and exploring the new boundaries in the age of algorithms: Reimagining the role of generative AI in distance education and online learning. *Asian Journal of Distance Education*, *18*(1), 1–8. https://doi.org/10.5281/zenodo.7755273

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). *Language models are few-shot learners. arXiv*. https://doi.org/10.48550/arXiv.2005.14165

Buick, J. M. (2011). Physics assessment and the development of a taxonomy. *European Journal of Physics Education*, *2*(1), 7–15. https://files.eric.ed.gov/fulltext/EJ1053836.pdf

Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., Ormerod, C. M., Fabiyi, D. G., Ivan, R., Walsh, C., Rios, O., Wilson, J., Yildirim-Erbasli, S. N., Wongvorachan, T., Liu, J. X., Tan, B., & Morilova, P. (2024). *The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges. arXiv*. https://doi.org/10.48550/arXiv.2406.18900

Burns, M. K., Riley-Tillman, T. C., & Rathvon, N. (2017). *Effective school interventions: Evidence-based strategies for improving student outcomes* (3rd ed.). Guilford Press.

Chang, W. C., & Chung, M. S. (2009). Automatic applying Bloom's taxonomy to classify and analysis the cognition level of English question items. *Proceedings of the 2009 Joint Conferences on Pervasive Computing (JCPC)*, 727–734. https://doi.org/10.1109/JCPC.2009.5420087

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453–494). Routledge.

Crowe, A., Dirks, C., & Wenderoth, M. P. (2008). Biology in Bloom: Implementing Bloom's taxonomy to enhance student learning in biology. *CBE E–Life Sciences Education*, *7*(4), 368–381. https://doi.org/10.1187/cbe.08-05-0024

Dao, X. Q., & Le, N. B. (2023). LLMs performance on Vietnamese high school biology examination. *International Journal of Modern Education and Computer Science*, *15*(6), 14–30. https://doi.org/10.5815/ijmecs.2023.06.02

Darling-Hammond, L. (2015). *Getting teacher evaluation right: What really matters for effectiveness and improvement*. Teachers College Press.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv*. https://doi.org/10.48550/arXiv.1810.04805

Doughty, J., Wan, Z., Bompelli, A., Qayum, J., Wang, T., Zhang, J., Zheng, Y., Doyle, A., Sridhar, P., Agarwal, A., Bogart, C., Keylor, E., Kultur, C., Savelka, J., & Sakr, M. (2024). A comparative study of AI-generated (GPT-4) and human-crafted MCQs in programming education. *Proceedings of the 26th Australasian Computing Education Conference*, 114–123. https://doi.org/10.1145/3636243.3636256

Embretson, S. E. (2005). Measuring human intelligence with artificial intelligence: Adaptive item generation. In R. J. Sternberg & J. E. Pretz (Eds.), *Cognition and intelligence: Identifying the mechanisms of the mind* (pp. 251–267). Cambridge University Press.

Embretson, S., & Yang, X. (2007). Automatic item generation and cognitive psychology. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (vol. 26, pp. 747–768). Elsevier. https://doi.org/10.1016/S0169-7161(06)26023-1

Feng, S., Park, C. Y., Liu, Y., & Tsvetkov, Y. (2023). *From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. arXiv*. https://doi.org/10.48550/arXiv.2305.08283

Gao, T., Fisch, A., & Chen, D. (2020). *Making pre-trained language models better few-shot learners. arXiv*. https://doi.org/10.48550/arXiv.2012.15723

Gierl, M. J., & Haladyna, T. M. (2012). Automatic item generation: An introduction. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 3-12). Routledge.

Glas, C. A., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement*, *27*(4), 247–261. https://doi.org/10.1177/0146621603027004001

Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice, 25*(4), 21–35. https://doi.org/10.1111/j.1745-3992.2006.00076.x

Gregorcic, B., & Pendrill, A. (2023). ChatGPT and the frustrated Socrates. *Physics Education*, *58*(3), 1–9. https://doi.org/10.1088/1361-6552/acc299

Hou, X., Zhao, Y., Liu, Y., Yang, Z., Wang, K., Li, L., Luo, X., Lo, D., Grundy, J., & Wang, H. (2023). *Large language models for software engineering: A systematic literature review. arXiv*. https://doi.org/10.48550/arXiv.2308.10620

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation*, *52*(1), 3–50. https://doi.org/10.1108/eb026960

Irvine, J. (2021). Taxonomies in education: Overview, comparison, and future directions. *Journal of Education and Development*, *5*(2), 1–25. https://doi.org/10.20849/jed.v5i2.898

Islam, R., & Ahmed, I. (2024). Gemini-the most powerful LLM: Myth or truth. *Proceedings of the 2024 5th Information Communication Technologies Conference (ICTC 2024)*, 303–308. https://doi.org/10.1109/ICTC61510.2024.10602253

Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Bonney, C. R., de Groot, E., Gilbert, M. C., Musu, L., Kempler, T. M., & Kelly, K. L. (2007). Cognitive processing of self-report items in educational research: Do they think what we mean? *Educational Psychologist*, *42*(3), 139–151. https://doi.org/10.1080/00461520701416231

Kong, S. C. (2014). Developing information literacy and critical thinking skills through domain knowledge learning in digital classrooms: An experience of practicing flipped classroom strategy. *Computers & Education*, *8*, 160–173. https://doi.org/10.1016/j.compedu.2014.05.009

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice*, *41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Krathwohl, D. R., & Anderson, L. W. (2010). Merlin C. Wittrock and the revision of Bloom's taxonomy. *Educational Psychologist*, *45*(1), 64–65. https://doi.org/10.1080/00461520903433562

Küchemann, S., Steinert, S., Revenga, N., Schweinberger, M., Dinc, Y., Avila, K. E., & Kuhn, J. (2023). Can ChatGPT support prospective teachers in physics task development? *Physical Review Physics Education Research*, *19*(2), 1–14. https://doi.org/10.1103/PhysRevPhysEducRes.19.020128

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, *30*, 121–204. https://doi.org/10.1007/s40593-019-00186-y

Laverghetta Jr, A., & Licato, J. (2023). Generating better items for cognitive assessments using large language models. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, 414–428. https://doi.org/10.18653/v1/2023.bea-1.34

Li, J., Tang, T., Zhao, W. X., Nie, J. Y., & Wen, J. R. (2024). Pre-trained language models for text generation: A survey. *ACM Computing Surveys*, *56*(9), 1–39. https://doi.org/10.1145/3649449

Li, S. (2021). Measuring cognitive engagement: An overview of measurement instruments and techniques. *IJPS: International Journal of Psychology and Educational Studies*, *8*(3), 63–76. https://doi.org/10.52380/ijpes.2021.8.3.239

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhao, L., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., & Ge, B. (2023). Summary of ChatGPT-Related research and perspective towards the future of large language models. *Meta-Radiology*, *1*(2), 1–14. https://doi.org/10.1016/j.metrad.2023.100017

Lorenzo, C. M. (2024). Integrating large language models for real-world problem modelling: A comparative study. *Proceedings of the 18th International Technology, Education and Development (INTED 2024) Conference*, 3262–3272. https://doi.org/10.21125/inted.2024.0871

Marvin, G., Hellen, N., Jjingo, D., & Nakatumba-Nabende, J. (2024). Prompt engineering in large language models. In I. J. Jacob, S. Piramuthu, & P. Falkowski-Gilski (Eds.), Proceedings of the International Conference on Data Intelligence and Cognitive Informatics (ICDICI 2023) (pp. 387-402). Springer. https://doi.org/10.1007/978-981-99-7962-2_30

Miao, J., Thongprayoon, C., Suppadungsuk, S., Krisanapan, P., Radhakrishnan, Y., & Cheungpasitporn, W. (2024). Chain of thought utilization in large language models and application in nephrology. *Medicina*, *60*(1), 1–19. https://doi.org/10.3390/medicina60010148

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language models: A survey. *arXiv*. https://doi.org/10.48550/arXiv.2402.06196

Miri, B., David, B. C., & Uri, Z. (2007). Purposely teaching for the promotion of higher-order thinking skills: A case of critical thinking. *Research in Science Education*, *37*, 353–369. https://doi.org/10.1007/s11165-006-9029-2

Mishra, S., Khashabi, D., Baral, C., Choi, Y., & Hajishirzi, H. (2021). *Reframing instructional prompts to GPTk's language*. *arXiv*. https://doi.org/10.48550/arXiv.2109.07830

Mohammed, M., & Omar, N. (2020). Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec. *PLoS ONE*, *15*(3), 1–21. https://doi.org/10.1371/journal.pone.0230442

Motlhabane, A. (2017). Unpacking the South African physics-examination questions according to Blooms' revised taxonomy. *Journal of Baltic Science Education*, *16*(6), 919–931.

Mystakidis, S., Fragkaki, M., & Filippousis, G. (2021). Ready teacher one: Virtual and augmented reality online professional development for K-12 school teachers. *Computers*, *10*(10), 134. https://doi.org/10.3390/computers10100134

Offerijns, J., Verberne, S., & Verhoef, T. (2020). *Better distractions: Transformer-based distractor generation and multiple choice question filtering. arXiv*. https://doi.org/10.48550/arXiv.2010.09598

Perikos, I., Kardakis, S., & Hatzilygeroudis, I. (2021). Sentiment analysis using novel and interpretable architectures of Hidden Markov models. *Knowledge-Based Systems*, *229*, 1–18. https://doi.org/10.1016/j.knosys.2021.107332

Polat, F., Tiddi, I., & Groth, P. (2024). *Testing prompt engineering methods for knowledge extraction from text*. Semantic Web. https://www.semantic-web-journal.net/system/files/swj3719.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. OpenAI.

Rangapur, A., & Rangapur, A. (2024). *The battle of LLMs: A comparative study in conversational QA tasks. arXiv*. https://doi.org/10.48550/arXiv.2405.18344

Roumeliotis, K. I., & Tselikas, N. D. (2023). ChatGPT and Open-AI models: A preliminary review. *Future Internet*, *15*(6), 1–24. https://doi.org/10.3390/fi15060192

Santos, R. P. D. (2023). *Enhancing physics learning with ChatGPT, Bing Chat, and Bard as agents-to-think-with: A comparative case study. arXiv*. https://doi.org/10.48550/arXiv.2306.00724

Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research, and Evaluation*, *22*(1), 1–13. https://doi.org/10.7275/swgt-rj52

Shen, Y., Heacock, L., Elias, J., Hentel, K. D., Reig, B., Shih, G., & Moy, L. (2023). Reviews and commentary: ChatGPT and other large language models are double-edged swords. *Radiology*, *307*(2), 1–4. https://doi.org/10.1148/radiol.230163

Song, Y., Dhariwal, P., Chen, M., & Sutskever, I. (2023). *Consistency models. arXiv*. https://doi.org/10.48550/arXiv.2303.01469

Tabrizi, S., & Rideout, G. (2017). Active learning: Using Bloom's taxonomy to support critical pedagogy. *International Journal for Cross-Disciplinary Subjects in Education*, *8*(3), 3202–3209.

Tan, B., Armoush, N., Mazzullo, E., Bulut, O., & Gierl, M. (2024). *A review of automatic item generation techniques leveraging large language models. EdArXiv Preprints*. https://doi.org/10.35542/osf.io/6d8tj

Tomlinson, C. A. (2017). *How to differentiate instruction in academically diverse classrooms*. ASCD.

Tomlinson, C. A. (2023). The parallel curriculum model: A design to develop potential & challenge high-ability learners. In J. S. Renzulli, E. J. Gubbins, K. S. McMillen, R. D. Eckert, & C. A. Little (Eds.), *Systems and models for developing programs for the gifted and talented* (pp. 571–598). Routledge. https://doi.org/10.4324/9781003419426

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need. arXiv*. https://doi.org/10.48550/arXiv.1706.03762

Veal, W. R., & MaKinster, J. G. (1999). Pedagogical content knowledge taxonomies. *The Electronic Journal for Research in Science & Mathematics Education*, *3*(4). https://ejrsme.icrsme.com/article/view/7615

Wang, H., Guo, B., Wu, W., Liu, S., & Yu, Z. (2021). Towards information-rich, logical dialogue systems with knowledge-enhanced neural models. *Neurocomputing*, *465*, 248–264. https://doi.org/10.1016/j.neucom.2021.08.131

Yahya, A. A., Toukal, Z., & Osman, A. (2012). Bloom's taxonomy–based classification for item bank questions using support vector machines. In W. Ding, H. Jiang, M. Ali, & M. Li (Eds.), *Modern advances in intelligent systems and tools: Studies in computational intelligence* (vol. 431, pp. 135–140). Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-642-30732-4_17

Yu, W., Zhu, C., Li, Z., Hu, Z., Wang, Q., Ji, H., & Jiang, M. (2022). A survey of knowledge-enhanced text generation. *ACM Computing Surveys*, *54*(11s), 1–38. https://doi.org/10.1145/3512467

Zhang, M., & Li, J. (2021). A commentary of GPT-3 in MIT Technology Review 2021. *Fundamental Research*, *1*(6), 831–833. https://doi.org/10.1016/j.fmre.2021.11.011

Zhong, Q., Ding, L., Liu, J., Du, B., & Tao, D. (2023). *Can ChatGPT understand too? A comparative study on ChatGPT and fine-tuned Bert. arXiv*. https://doi.org/10.48550/arXiv.2302.10198