

Performance of ChatGPT in solving wave problems

Muhammad Aswin Rangkuti^{1,2,*} Wawan Bunawan^{1,2} Hendra Y. Agustian³

Tuti Hardianti^{1,2} Dewi Wulandari¹ and Nazwa Hudaiby¹

¹*Department of Physics, Universitas Negeri Medan, Jl Willem Iskandar Psr V, Medan 20211, Indonesia*

²*Center of Excellence for Future Innovative Science Education, Universitas Negeri Medan, Jl Willem Iskandar Psr V, Medan 20211, Indonesia*

³*Department of Science Education, University of Copenhagen, Universitetsparken 5, 2100, Copenhagen, Denmark*



(Received 10 November 2025; accepted 15 July 2026; published 6 August 2026)

The ability of ChatGPT, especially its latest versions, attracts significant attention from educators and researchers in fields such as physics. Several studies on this topic explore the performance of ChatGPT in answering physics questions across different versions of ChatGPT. In this study, we evaluate the development of ChatGPT in answering physics questions using the Mechanical Waves Conceptual Survey (MWCS) to examine whether its capabilities improve across different versions. A total of 75 responses per item are generated using ChatGPT versions 4, 4.o, and 5.2, resulting in 4950 responses for comprehensive analysis. Based on the comparison between the three versions of ChatGPT, we find that some difficult types of questions that are answered incorrectly by previous versions, as reported in earlier studies, can now be answered correctly, although there are still minor mistakes, especially when examining the reasoning. A generalized linear mixed model (GLMM) analysis, together with *post hoc* comparisons, indicates that ChatGPT 5.2 significantly outperforms both ChatGPT 4 and 4.o, suggesting that the capabilities of the latest version included in this study significantly improve. These findings highlight a future direction for this type of research, as our study shows that the latest version of ChatGPT can correctly answer more types of questions in MWCS.

DOI: [10.1103/4dgy-1dgl](https://doi.org/10.1103/4dgy-1dgl)

I. INTRODUCTION

The rapid development of large language models, including generative artificial intelligence (GAI), has impacted various sectors, especially education. Educators have voiced concerns about the potential marginalization of human teaching and the impact on academic integrity [1,2]. For instance, students might become overly reliant on GAI as a problem-solving tool, which could stifle creativity and critical thinking and reduce the learning experience [1,3,4]. Educators are also concerned about the misuse of GAI for academic dishonesty, such as cheating and plagiarism [5–7]. Despite these concerns, previous studies show that the benefits of GAI in education often outweigh its challenges. Systematic reviews on the use of GAI in education indicate that it generally improves both instructional quality and student learning experiences [8]. As GAI

continues to evolve, it becomes increasingly important to examine how its capabilities develop over time.

In the context of physics education, recent studies have begun to investigate the ability of ChatGPT to answer physics questions [9–12]. These studies show that ChatGPT still has difficulty answering certain types of questions. In particular, difficulties are found in questions that involve graph interpretation [9], questions that require complex mathematical representations [12], and questions that involve spatial analysis [13]. These findings suggest that the performance of ChatGPT in physics is still developing and may improve in newer versions. However, most previous studies focus on a single version of ChatGPT, and only limited research examines how its performance changes across different versions.

This paper aims to examine the development of ChatGPT capabilities across different versions in answering physics conceptual questions. In this study, ChatGPT versions 4, 4.o, and 5.2 are used to answer the Mechanical Waves Conceptual Survey (MWCS). The MWCS consists of several types of questions, including multiple-choice questions, multiple-choice questions with images, and two-tier multiple-choice questions that require reasoning. The results obtained from the different versions of ChatGPT are then compared with previous studies by

*Contact author: aswin_ray@unimed.ac.id

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/). Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

Tongchai *et al.* [14] and Barniol and Zavala [15] to evaluate how the performance of ChatGPT compares with human performance and to determine whether its capabilities improve across versions.

II. THEORETICAL FRAMEWORK

A. AI in science education

AI has rapidly advanced over recent decades, with significant growth expected in education [16–19]. An empirical study on Artificial Intelligence in Education (AIED) from 1993 to 2020 shows the USA and China lead in research output [16]. The number of publications from 2009 to 2019 shows growing interest in exploring the relationship between education and technology [17]. This growth highlights the need for collaboration between educators and AI developers to guide future research in AIED [20].

GAI integration can significantly enhance the science teaching and learning experience [1,21–23]. For instance, preservice science teachers have incorporated ChatGPT into their instructional strategies, recognizing its value in supporting learning processes such as cooperative learning models and learning cycles [24]. A key feature of GAI is its ability to generate personalized responses and provide immediate feedback, which can improve student engagement and learning outcomes [25]. In one study, prompt engineering enabled ChatGPT to deliver feedback on questions related to Newton’s second and third laws in multiobject systems, with 70% of the responses requiring only minor or no revisions from instructors [26].

Additionally, GAI can support the development of students’ inquiry skills by encouraging them to pose questions and explore their curiosity in a responsive learning environment [21]. For instance, intelligent personal assistants offered responses that challenged students’ initial understanding of Newton’s first law of motion [27]. However, despite the promising potential of GAI in education, more rigorous evidence is needed to evaluate its practical outcomes, especially its effectiveness in contributing to quality education for all [19].

A major concern in the literature is the ethical integration of GAI in science education [19,28,29]. Issues of academic integrity, including cheating, plagiarism, and fabrication, must be addressed, requiring teachers to verify students’ work [19]. Moreover, using GAI to generate a complete solution is considered an ethical issue [5]. A study showed that nearly half of physics students trusted ChatGPT’s answers without verification [30], highlighting the importance of teaching students to critically evaluate AI-generated responses.

Despite the challenges, AI offers significant potential in education, with teacher readiness being key to unlocking its advantages. AI readiness refers to a teacher’s ability to understand and effectively use AI tools in their practice.

Higher AI readiness is associated with lower perceived threats from AI, leading to more innovation and job satisfaction [31]. Therefore, teachers need to follow training programs to keep up with evolving AIED so that they can use it productively in teaching and learning [32].

B. GAI as a problem-solving tool

The challenges of using GAI in science education are a vital research area. As GAI use expands, science and mathematics education researchers are exploring its performance in problem solving. In math, ChatGPT’s performance is heavily dependent on the prompt, with GPT 4 outperforming GPT 3.5 in clarity and correctness [33]. Compared to human responses, GPT’s problem solving has been found to be similar to that of an average human. In chemistry, GPT’s performance depends on the question type, prompts, tools, and version [23,34–36].

Physics questions, which combine principles and mathematical tools, can be challenging for students [37–40]. Interestingly, GAI encounters these same challenges. For instance, ChatGPT performs well on simple definitions and basic calculations but struggles with multistep problems requiring complex reasoning [12,41]. Furthermore, it fails to interpret graph-based questions accurately [9], highlighting the need for further research and development in this area.

Despite its limited performance, ChatGPT shows promising potential in solving physics problems. Studies show that it performs well on conceptual problems [11,42,43]. For instance, it successfully tackled the Force Concept Inventory, achieving an average of 83% with few challenging items [10]. ChatGPT 4 even achieved a perfect score on multiple-choice questions and can solve conceptual and reasoning problems [11]. Furthermore, it has a 65.5% success rate in solving physics problems but struggles with calculation errors and assumptions [43].

C. Mechanical Waves Conceptual Survey

The MWCS is a 22-item multiple-choice survey developed by Tongchai *et al.* [14] to measure the conceptual understanding of mechanical waves. The questions are divided into four types: standard questions, those with images, those with graphs, and two-tier reasoning items. Topics include propagation, superposition, reflection, and standing waves. The MWCS is then used to identify students’ difficulties and guide instructional interventions [14,15,44,45].

Studies show that students struggle with the conceptual understanding of wave mechanics when administered the MWCS [14,15,44]. For instance, students achieved an average score of only 9.86 out of 22 questions, with the most difficult topics being propagation and superposition [15]. Another study found that students provided correct answers with an average score of 4.69 out of 22 questions. Students who had already learned about waves before the

MWCS was administered also struggled to achieve a high score, with an average of only 24.41% correct [45]. In this paper, we utilize ChatGPT to answer the MWCS and examine whether its performance mirrors previous student results [14,15,45]. We then reflect on our findings to underscore the importance of educators in the context of AI in education, particularly in the design of problems and solutions.

III. RESEARCH QUESTIONS

Grounded in the theoretical framework, this study seeks to address the following research questions:

- (1) How does the performance of different versions of ChatGPT change in answering the MWCS?
- (2) How does the performance of different versions of ChatGPT compare with that of students in answering the MWCS?

IV. METHODS

A. Data collection

The data collection procedure in this study was inspired by previous research that evaluates ChatGPT's performance in answering kinematics graph questions. In that study, the authors used the TUG-K assessment and analyzed the performance of ChatGPT 4 by prompting each question 40 times [42]. In the present study, the MWCS was used to examine the development of ChatGPT capabilities across different versions. The data were collected between May 2024 and February 2026.

To answer RQ1, screenshots of each MWCS question are input into ChatGPT without any additional prompts. Shared statements, images, or graphs are combined with each question screenshot. Each response and its reasoning are then collected for analysis. This procedure is repeated 75 times for each question, and a new chat is initiated for every repetition. This results in 1,650 responses for each ChatGPT version and 4,950 responses in total for versions 4, 4.o, and 5.2. If ChatGPT does not select any answer, the response is categorized as incorrect. The reasoning generated by ChatGPT is also organized and analyzed to examine the consistency of the problem-solving strategies and final answers. At the time of data collection, ChatGPT 5.2 was the most recent version included in this study. However, available versions may change over time.

To address RQ2, the results obtained in this study are compared with human responses reported in Ref. [14]. Tongchai *et al.* [14] administered the MWCS to several groups of students in two different countries who have received different levels of formal physics instruction. In addition, the reasoning generated by ChatGPT is compared with students' reasoning reported by Barniola and Zavala [15]. In that study, the MWCS was administered to undergraduate students who have completed a course in fluids, waves, and thermodynamics [15].

The comparison with student data should be interpreted cautiously. The student results are drawn from different populations and instructional contexts and therefore do not represent a direct comparison under equivalent conditions. In this study, the comparison, including the difficulty and discrimination indices, is used to provide context for interpreting performance patterns, particularly in relation to question types and conceptual difficulties, rather than to establish relative performance between ChatGPT and students.

B. Data analysis

1. Statistical analysis

We conducted three statistical analyses on ChatGPT's final MWCS answers: the difficulty index, discrimination index, and GLMM. The difficulty index measured each question's difficulty. An index value between 0.2 and 0.8 indicates adequate difficulty; a value below 0.2 is very difficult, while a value above 0.8 is too easy [46].

The discrimination index assesses how effectively a question differentiates between high and low performers. A value greater than 0.3 is highly discriminating, with a higher value indicating greater discrimination [46]. A GLMM was used to analyze performance differences across ChatGPT versions while accounting for fixed effects of version and repetition; model validity was checked by confirming convergence. Overall, GLMM significance was evaluated to test whether performance differs across versions in general. Fixed-effects results were examined to assess performance patterns and stability across repetitions, including the version $\times$ repetition interaction, and *post hoc* estimated marginal means (EMM) analyses were conducted to identify specific repetitions that showed significant differences between versions.

2. Reasoning analysis

The MWCS consists of four question types: those without images or graphs, image-based, graph-based, and two-tier questions. The five two-tier items, three of which included images, were all classified as two-tier for analysis. We analyzed the reasoning from all three ChatGPT versions for consistency with their final answers. The first and second authors reviewed a total of 4950 responses.

In this study, reasoning is classified as correct when it is consistent with expert physics understanding, meaning that the explanation applies the relevant concepts and relationships in a conceptually valid way.

Of the 4950 total responses, 167 disagreements were found. Most of these occurred in lengthy explanations, such as those for question 21, where complex reasoning increased the likelihood of misinterpretation. All disagreements were subsequently discussed between the two

authors to reach a consensus. We calculate the percentage of agreement using the following formula, where TRG is the “total reasoning generated” and TDA is the “total disagreement analyzed”:

$$\% \text{Agreement} = \frac{(\text{TRG} - \text{TDA})}{\text{TRG}}$$

$$\% \text{Agreement} = \frac{(4950 - 167)}{4950} = 96.62\%$$

V. RESULTS

A. Overall performance of ChatGPT

The performance of ChatGPT shows a clear improvement across versions, with ChatGPT 4.o outperforming ChatGPT 4, and further substantial gains were observed in ChatGPT 5.2. Overall, ChatGPT 4 and ChatGPT 4.o correctly answer 58.18% and 64.12% of the 1650 repetitions, respectively. In comparison, ChatGPT 5.2 achieves 1469 correct responses out of 1650, corresponding to 89.03%, indicating a significant advancement in performance.

Figure 1 compares the performance of the three versions and highlights substantial improvement in ChatGPT 5.2, particularly on questions that are most difficult for ChatGPT 4 and 4.o. While ChatGPT 4.o generally demonstrates greater consistency than ChatGPT 4, both earlier versions show significant limitations on specific items.

The most notable improvements are observed in questions that involve graphical representations, which are the most challenging for both ChatGPT 4 and 4.o, particularly questions 9, 11, 13–16, and 20. In these items, the performance of the earlier versions is substantially lower. For instance, in questions 13–16, ChatGPT 4 achieves only 19, 10, 11, and 8 correct responses, while ChatGPT 4.o shows only modest improvement with 26, 14, 17, and 12 correct responses. In contrast, ChatGPT 5.2 demonstrates marked improvement, achieving 60, 60, 22, and 35 correct responses for the same questions. A similar pattern is observed in question 20, where ChatGPT 4 and 4.o achieve only 4 and 2 correct responses, respectively, whereas ChatGPT 5.2 answers all 75 repetitions correctly.

B. Statistical analysis

1. Difficulty index

Table I presents the difficulty index of the MWCS for different groups of students reported by Tongchai *et al.* [14] and for ChatGPT versions 4, 4.o, and 5.2. The results show a clear progression in performance across versions. The mean difficulty index for ChatGPT 4 is 0.59, which is comparable to that of first-advanced students (0.56), suggesting similar overall performance. ChatGPT 4.o shows improvement, with a mean difficulty index of 0.64, approaching that of second-advanced students (0.70). In contrast, ChatGPT 5.2 achieves a substantially higher mean difficulty index of 0.89, exceeding all student groups reported in the study. This result indicates a significant

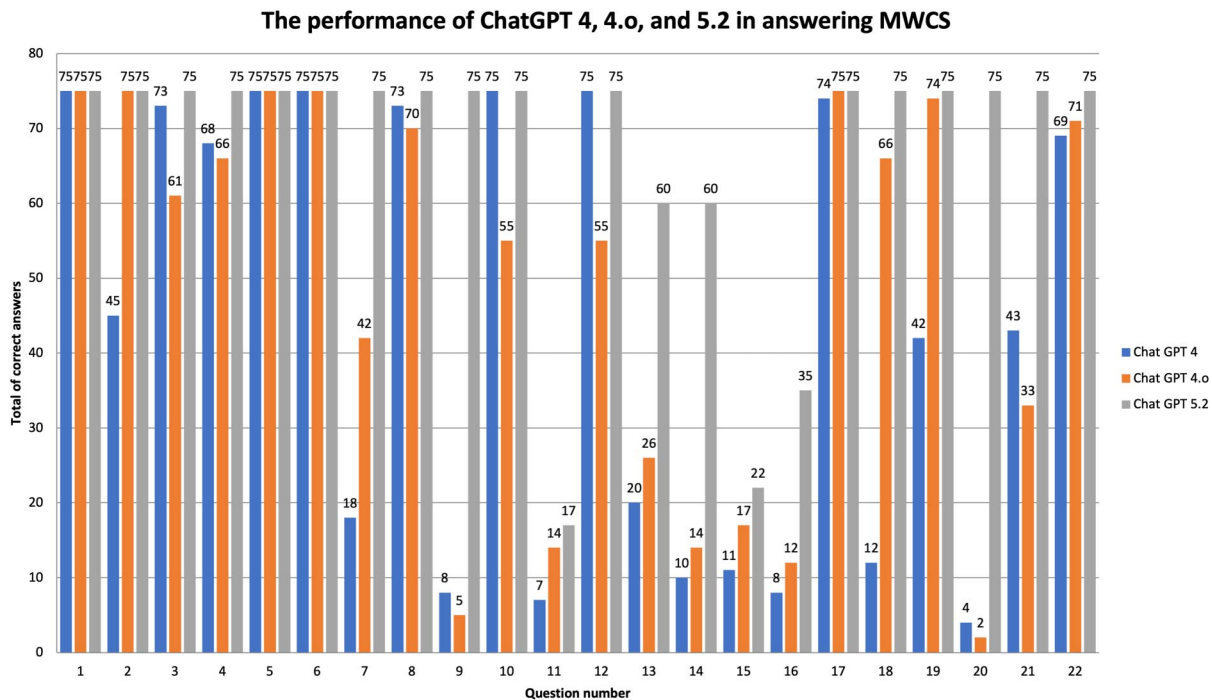


FIG. 1. The comparison of ChatGPT 4, 4.o, and 5.2 in answering MWCS.

TABLE I. The comparison of difficulty index between students from Tongchai *et al.* [14] and ChatGPT. Column headings denote the following student groups: 1st Fund (first-year Fundamentals), SydHigh (Sydney senior high school), ThaiHigh (Thai senior high school), 1st Reg (first-year Regulars), 2nd Reg (second-year Regulars), 1st Adv (first-year Advanced), and 2nd Adv (second-year Advanced).

Question	1st Fund <i>n</i> = 123	SydHigh <i>n</i> = 54	ThaiHigh <i>n</i> = 270	1st Reg <i>n</i> = 287	2nd Reg <i>n</i> = 48	1st Adv <i>n</i> = 69	2nd Adv <i>n</i> = 51	GPT 4 <i>n</i> = 75	GPT 4.o <i>n</i> = 75	GPT 5.2 <i>n</i> = 75
<i>Propagation</i>										
1	0.59	0.89	0.69	0.83	0.75	0.96	0.94	1.00	1.00	1.00
2	0.37	0.44	0.31	0.53	0.65	0.93	0.86	0.60	1.00	1.00
3	0.33	0.37	0.24	0.50	0.65	0.90	0.78	0.97	0.81	1.00
4	0.11	0.26	0.17	0.22	0.31	0.52	0.67	0.91	0.88	1.00
5	0.29	0.30	0.35	0.26	0.48	0.39	0.63	1.00	1.00	1.00
6	0.13	0.50	0.20	0.48	0.58	0.77	0.88	1.00	1.00	1.00
7	0.11	0.41	0.17	0.43	0.54	0.77	0.84	0.24	0.56	1.00
8	0.07	0.41	0.14	0.39	0.48	0.67	0.78	0.97	0.93	1.00
<i>Superposition</i>										
9	0.11	0.33	0.33	0.39	0.48	0.65	0.73	0.11	0.07	1.00
10	0.35	0.59	0.71	0.70	0.88	0.86	1.00	1.00	0.73	1.00
11	0.30	0.44	0.38	0.44	0.48	0.64	0.69	0.09	0.19	0.23
12	0.34	0.56	0.51	0.52	0.81	0.77	0.94	1.00	0.73	1.00
<i>Reflection</i>										
13	0.17	0.30	0.61	0.32	0.40	0.41	0.73	0.27	0.35	0.80
14	0.11	0.11	0.47	0.15	0.29	0.26	0.65	0.13	0.19	0.80
15	0.18	0.20	0.26	0.21	0.40	0.46	0.55	0.15	0.23	0.29
16	0.21	0.26	0.41	0.28	0.38	0.29	0.35	0.11	0.16	0.47
<i>Standing waves</i>										
17	0.32	0.20	0.41	0.57	0.42	0.65	0.59	0.99	1.00	1.00
18	0.14	0.09	0.13	0.16	0.21	0.29	0.53	0.16	0.88	1.00
19	0.25	0.46	0.26	0.26	0.21	0.29	0.43	0.56	0.99	1.00
20	0.09	0.02	0.31	0.13	0.33	0.22	0.67	0.05	0.03	1.00
21	0.14	0.04	0.07	0.17	0.19	0.19	0.37	0.57	0.44	1.00
22	0.24	0.26	0.14	0.30	0.44	0.52	0.71	0.92	0.95	1.00
Mean	0.22	0.34	0.34	0.37	0.47	0.56	0.70	0.58	0.64	0.89

development in capability, with the latest version outperforming even the most advanced student cohort.

The improvement is particularly evident in question types that are most challenging for earlier versions. For wave reflection questions, ChatGPT 4 and 4.o consistently show low difficulty index values, often below 0.30, indicating substantial difficulty. Their performance is comparable to that of 1st Fund and SydHigh student groups, which also exhibit low scores on these items. However, ChatGPT 5.2 demonstrates marked improvement in this topic, with difficulty index values increasing to 0.80 and 0.80 for questions 13 and 14, respectively, and moderate gains in questions 15 and 16.

A similar pattern is observed in standing wave questions, where the largest differences between versions are evident. For question 20, which is one of the most difficult items, ChatGPT 4 and 4.o achieve extremely low difficulty index values of 0.05 and 0.03, respectively, both far below the highest student performance of 0.67. In contrast, ChatGPT 5.2 achieves a perfect difficulty index of 1.00. Improvements are also observed in question 21, where performance increases from 0.57 and 0.44 in earlier versions to 1.00 in ChatGPT 5.2. ChatGPT 5.2 also

improves on question 22, reaching a difficulty index of 1.00 compared to 0.92 and 0.95 in ChatGPT 4 and 4.o so that its performance on the standing wave items is uniformly high.

According to Table II, ChatGPT 4 and 4.o struggled with graph-based questions on the MWCS. ChatGPT 4 had a mean difficulty index of 0.36 for graph-related questions, while 4.o scored slightly lower at 0.33. In contrast, ChatGPT 5.2 showed a substantial improvement, with a mean difficulty index of 0.70. Although a score above 0.3 is acceptable, these values for ChatGPT 4 and 4.o are significantly lower than their performance on other question types, which had difficulty indices above 0.5. This highlights that interpreting graphs remained a challenge for the earlier models, while ChatGPT 5.2 improved considerably in this area.

ChatGPT performed well on questions 10 and 12, with ChatGPT 4 achieving a difficulty index of 1.00. Both questions focused on wave profiles after superposition, and ChatGPT correctly identified that the wave shapes would remain unchanged following constructive (Q10) and destructive (Q12) superposition. In contrast, it struggled with questions 9 and 11, which required

TABLE II. Comparison of ChatGPT 4, ChatGPT 4.o, and ChatGPT 5.2 performance on different question topics.

Type of questions	GPT 4 (<i>n</i> = 75)	GPT 4.o (<i>n</i> = 75)	GPT 5.2 (<i>n</i> = 75)
<i>Questions without images/graphs</i>			
1	1.00	1.00	1.00
2	0.60	1.00	1.00
3	0.97	0.81	1.00
Mean	0.86	0.94	1.00
<i>Questions with images</i>			
4	0.91	0.88	1.00
5	1.00	1.00	1.00
6	1.00	1.00	1.00
7	0.24	0.56	1.00
8	0.97	0.93	1.00
20	0.05	0.03	1.00
Mean	0.70	0.73	1.00
<i>Questions with graphs</i>			
9	0.11	0.07	1.00
10	1.00	0.73	1.00
11	0.09	0.19	0.23
12	1.00	0.73	1.00
13	0.27	0.35	0.80
14	0.13	0.19	0.80
15	0.15	0.23	0.29
16	0.11	0.16	0.47
Mean	0.36	0.33	0.70
<i>Two-tier questions</i>			
17	0.99	1.00	1.00
18	0.16	0.88	1.00
19	0.56	0.99	1.00
21	0.57	0.44	1.00
22	0.92	0.95	1.00
Mean	0.64	0.85	1.00

identifying wave profiles just before peak overlap. Both questions had difficulty indices below 0.2, highlighting ChatGPT's difficulty in interpreting dynamic wave interactions at intermediate stages, particularly when graphs are involved. In comparison, ChatGPT 5.2 showed clear improvement on question 9, achieving a perfect difficulty index of 1.00, although question 11 remained difficult with a low difficulty index of 0.23, indicating that some inconsistencies persisted on these intermediate-stage graph items.

Other question types had difficulty index values above 0.6, suggesting they were relatively easy for ChatGPT. Questions without images or graphs were the easiest, with mean scores of 0.86 for ChatGPT 4 and 0.94 for 4.o, while ChatGPT 5.2 achieved a perfect mean score of 1.00. Both models also performed well on questions with images and two-tier formats. However, questions 7 and 20 remained notably more challenging for ChatGPT 4 and 4.o. In contrast, ChatGPT 5.2 showed improvement on these questions, answering both question 7 and question 20 correctly in all repetitions.

2. Discrimination index

Table III presents the discrimination index for ChatGPT compared to student groups. Tongchai *et al.* [14] reported that the mean discrimination index across all student samples was acceptable, with a value of 0.48. This indicates that the test was effective in distinguishing between high- and low-performing students. The discrimination index, however, is defined for a population of test-takers who differ in ability. In the present study, the ChatGPT observations originate from independent runs of a single model under identical conditions rather than from individuals, so the high- and low-scoring groups on which the index depends do not correspond to differences in ability. We therefore report the ChatGPT discrimination values in an exploratory capacity only. Computed on this basis, ChatGPT 4 and ChatGPT 4.o show low mean indices (0.175 and 0.177), and ChatGPT 5.2 a slightly lower overall mean (0.089), all well below the student mean. The individual subtopic values are also unstable and sensitive to the arbitrary assignment of near-identical runs to the high and low groups and should not be read as conventional item discrimination.

3. GLMM analysis

In the first GLMM, we analyzed whether an interaction effect existed between the ChatGPT versions and the questions. The results, presented in Tables IV and V, show that the model failed to converge. As shown in Table V, the fixed-effect estimates are accompanied by exceptionally large standard errors (on the order of 3.3×10^3 to 4.6×10^3), yielding *z* values near zero and *p* values close to 1.00 for almost every parameter. These inflated standard

TABLE III. Discrimination index of all subtopics between students from Tongchai *et al.* [14] and ChatGPT.

Subtopics	Mean (human groups)	Mean (ChatGPT)	1st Fund (<i>n</i> = 123)	SydHigh (<i>n</i> = 54)	ThaiHigh (<i>n</i> = 270)	1st Reg (<i>n</i> = 287)	2nd Reg (<i>n</i> = 48)	1st Adv (<i>n</i> = 69)	2nd Adv (<i>n</i> = 51)	GPT 4 (<i>n</i> = 75)	GPT 4o (<i>n</i> = 75)	GPT 5.2 (<i>n</i> = 75)
Propagation	0.55	0.080	0.33	0.60	0.23	0.58	0.70	0.36	0.38	0.125	0.114	0
Superposition	0.60	0.164	0.34	0.16	0.58	0.52	0.33	0.43	0.25	0.081	0.326	0.084
Reflection	0.46	0.296	0.24	0.11	0.69	0.30	0.52	0.43	0.29	0.246	0.238	0.403
Standing waves	0.34	0.126	0.24	0.18	0.28	0.21	0.31	0.51	0.51	0.256	0.123	0
Mean	0.48	0.147	0.29	0.31	0.39	0.42	0.49	0.42	0.38	0.175	0.177	0.089

TABLE IV. The first model of GLMM to analyze the effect between ChatGPT versions and MWCS questions.

Group	Name	Variance	Std. dev
Repetition	(Intercept)	0.03087	0.1757

TABLE V. Summary of fixed effects of the first model.

Fixed effects	Estimate	Std. error	z Value	$\Pr(> z)$
(Intercept)	20.490	3252.000	0.006	0.995
ChatGPT4.o	-0.003	4596.000	0.000	1.000
ChatGPT5.2	0.003	4603.000	0.000	1.000
Q2	-20.090	3252.000	-0.006	0.995
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
ChatGPT4.o:Q22	0.438	4596.000	0.000	1.000
ChatGPT5.2:Q22	18.170	5771.000	0.003	0.997

errors are driven by the near-perfect accuracy of ChatGPT 5.2, which answered most questions correctly in all repetitions and thereby caused data separation. Table IV further shows that the random effect for repetition is negligible (variance = 0.031, SD = 0.176), confirming that little variation is attributable to repetition in this model. Because the model did not converge, reliable estimates for the parameters, including the interaction term, could not be obtained, and no conclusions about the interaction effect between ChatGPT versions and their performance on the MWCS could be drawn from this analysis.

The second model was used to analyze the interaction effect between ChatGPT versions and repetitions. Before interpreting these results, we note that each repetition is an independent execution of the same prompt in a new chat session; the repetition variable therefore reflects stochastic variability in response generation rather than any change in model ability, and statistically significant repetition or version $\times$ repetition terms should not be read as substantively meaningful performance trends. As shown in Table VI, the estimation results yielded a random effect for the (1|Question) component with a variance of 9.97 and a standard deviation of 3.16. This suggests that unexplained variance exists between questions, which is captured by the random effect.

Overall, the baseline GLMM results indicate that there is no statistically significant difference in initial

TABLE VI. The second model of GLMM to analyze the interaction of ChatGPT versions and repetitions.

Group	Name	Variance	SD
Question	(Intercept)	9.97	3.16

TABLE VII. Summary of fixed effects of the second model.

Fixed effects	Estimate	Std. error	z Value	$\Pr(> z)$
(Intercept)	2.955	0.981	3.012	0.003
ChatGPT4.o	-0.953	0.983	-0.970	0.332
ChatGPT5.2	1.729	1.114	1.552	0.121
G2	-2.778	0.966	-2.877	0.004
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
ChatGPT4.o:G75	1.406	1.367	1.028	0.304
ChatGPT5.2:G75	2.346	1.560	1.504	0.133

performance between the ChatGPT versions. As shown in Table VII, the version main effects were not statistically significant (ChatGPT 4.o: $p = 0.332$; ChatGPT 5.2: $p = 0.121$). A number of repetition terms and version $\times$ repetition interaction terms reached statistical significance ($p < 0.05$). Consistent with the caution noted above, however, these effects appear only in a limited number of repetitions, do not follow a systematic pattern, and are attributable to run-to-run variability in generation rather than to any change in model capability. We therefore do not interpret them as evidence of a performance trend across repetitions.

The distribution of points in Fig. 2 reflects the relative consistency of each version across repetitions. ChatGPT 4 shows the most dispersed distribution, with values spread across a wide range, indicating greater run-to-run variability in the proportion of items answered correctly. ChatGPT 4.o is somewhat less dispersed, while ChatGPT

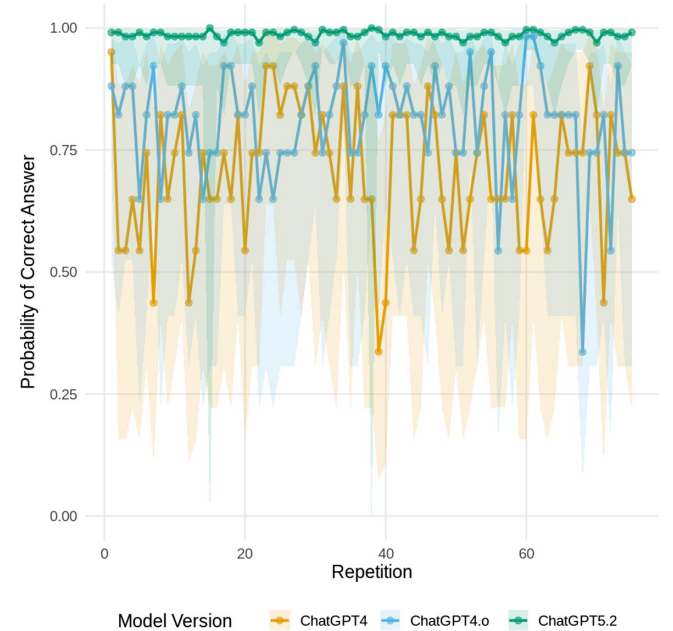


FIG. 2. Proportion of MWCS items answered correctly by ChatGPT 4, 4.o, and 5.2 across the 75 independent repetitions, with 95% confidence intervals.

TABLE VIII. Estimated marginal means analysis between versions of ChatGPT at repetition 75.

Repetition	Contrast	Odds ratio	<i>p</i> -Value
75	ChatGPT4 / ChatGPT4.o	0.636	0.882
75	ChatGPT4 / ChatGPT5.2	0.017	0.0006*
75	ChatGPT4.o / ChatGPT5.2	0.027	0.0029*

Signif. codes: 0.05 ‘’

5.2 is concentrated near the top of the scale with little variation across repetitions. This pattern illustrates that the newer version answers the survey with high and stable accuracy, whereas the variability seen in the earlier versions reflects stochastic differences between independent runs rather than any systematic change across the repetition sequence.

Table VIII shows the *post hoc* estimated marginal means (EMM) analysis used to assess whether there were any significant differences between the versions at the end of the test cycle (Repetition 75).

From Table VIII, it is noted that at the end of the experimental repetitions, the difference between ChatGPT 4 and ChatGPT 4.o was not statistically significant ($p = 0.882$). In contrast, statistically significant differences were detected when comparing the older models to the newest iteration: ChatGPT 5.2 significantly outperformed both ChatGPT 4 ($p = 0.0006$) and ChatGPT 4.o ($p = 0.0029$). This confirms that ChatGPT 5.2 maintains a markedly higher accuracy than the older models at the end of the extended testing cycle.

The repetition variable in the GLMM represents repeated independent runs of the same prompt under identical conditions. As such, it does not reflect changes in model ability but variation in response generation across runs. Although some repetition and interaction terms are statistically significant, these effects appear only in a limited number of repetitions and do not show a consistent pattern. We emphasize that statistical significance in this setting does not imply practical significance with 4950 observations, even small, purely random fluctuations in generation can reach conventional significance thresholds without carrying substantive meaning. In practical terms, the results indicate that the model responses are generally stable across repetitions, with only minor variation between runs.

C. Reasoning analysis

1. Overall reasoning comparison

Table IX shows the performance of ChatGPT 4, 4.o, and 5.2 based on answer accuracy and reasoning. ChatGPT 4 gave correct answers with correct reasoning in 57% of cases, while 4.o performed slightly better at 64%. ChatGPT 5.2 shows a substantial improvement, achieving correct answers with correct reasoning in 85% of cases.

TABLE IX. Reasoning and answers by ChatGPT for all questions.

Reasoning	Answers						Total
	GPT 4		GPT 4o		GPT 5.2		
	I	C	I	C	I	C	
Incorrect	185	22	109	7	38	70	431
Correct	505	938	483	1051	143	1399	4519
Total	1650		1650		1650		4950

ChatGPT 4 had 11% incorrect answers with incorrect reasoning, 31% incorrect answers with correct reasoning, and 1% correct answers with incorrect reasoning. In comparison, 4.o showed 7%, 29%, and 0% in those respective categories. ChatGPT 5.2 shows further improvement, with 2% incorrect answers with incorrect reasoning and 9% incorrect answers with correct reasoning, while 4% of the responses were correct answers with incorrect reasoning.

2. Questions without image/graph

There are three questions in MWCS that do not include images or graphs. As shown in Table X, these types of questions were relatively easy for ChatGPT, with both ChatGPT 4 and 4.o achieving 86% and 94% accuracy in providing correct answers with correct reasoning, with ChatGPT 4.o performing slightly better. ChatGPT 5.2 shows further improvement, achieving 100% accuracy, with all responses categorized as correct answers with correct reasoning.

While ChatGPT generally performed well on questions without images or graphs, it struggled with question 2. This question assesses whether students understand that the speed of a sound wave remains constant regardless of frequency. As Table I shows, ChatGPT 4 had a difficulty index of 0.6, which is acceptable but lower than similar questions that are considered very easy (indices > 0.8). Barniol and Zavala [15] found 40% of students answered question 2 correctly, with many incorrectly believing that higher frequency means higher speed. This is because students tended to approach $v = \lambda f$ mathematically

TABLE X. Reasoning and answers by ChatGPT for questions without images or graphs.

Reasoning	Answers						Total
	GPT 4		GPT 4o		GPT 5.2		
	I	C	I	C	I	C	
Incorrect	0	0	0	0	0	0	0
Correct	32	193	14	211	0	225	675
Total	225		225		225		675

without considering the physical relationships between them [15,47].

Similar to students' performance, ChatGPT 4 answered question 2 correctly in 60% of the 75 attempts. In the remaining 40%, it gave incorrect answers but still produced correct reasoning, recognizing that the speed of sound depends on the properties of air and not amplitude. Despite this understanding, it struggled to match the correct reasoning with the appropriate answer. This highlights a key difference in performance where students often approached physics equations mathematically while ChatGPT had the correct reasoning but failed to align it with the correct answer.

3. Questions with image

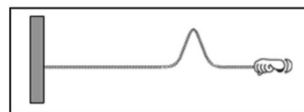
ChatGPT versions 4 and 4.o performed similarly on questions that included images, with both giving correct answers and reasoning in about 70% of the attempts. As shown in Table XI, ChatGPT often generated accurate reasoning for image-based questions but did not always align it with the correct answers.

Questions 4 and 5 were the easiest for ChatGPT to answer with correct reasoning and high consistency. These questions assess the understanding that wave speed depends on the string's tension and density, rather than on changes in hand movement when shaking the string. We argue that ChatGPT's high consistency in answering these questions may be due to the clear textual descriptions provided, as illustrated in Fig. 3. In this case, ChatGPT was only required to interpret the text, without the need to analyze any accompanying images. These results further support the observation that ChatGPT performs better on questions that rely on text-based descriptions. This pattern is consistent across all versions, with ChatGPT 5.2 maintaining perfect performance.

The inconsistency in earlier versions is shown in question 7, which asks about the motion of a dust particle in front of a speaker when the pitch is increased while keeping the amplitude the same. The correct answer is that the particle moves back and forth faster due to the increase in frequency. Earlier versions correctly recognized that a higher frequency would cause the particle to move faster; however, instead of describing the motion as back and

Consider the following description and answer questions 4 – 5.

One end of a long taut string is tied to a distant pole while the other end of the string is held by a girl (see figure below). This girl quickly flicks her hand up and down to create a pulse moving towards the pole.



4) She now wants to produce a pulse that takes a shorter time to reach the pole. How can she do this?

- A. Flick the string harder to push more force into the pulse.
- B. Flick the string faster to create a pulse with higher frequency.
- C. Flick the string further up and down to create a pulse with larger amplitude.
- D. Flick the string a shorter distance up and down to create a pulse with smaller amplitude.
- E. Wait until the first pulse is reflected back then flick again to add the pulses together.
- F. None of the above would produce a pulse that takes a shorter time to reach the pole.

FIG. 3. One example of a question in MWCS with an image and description of the image.

forth, they stated that the particle moves up and down, as if it were a transverse wave. This indicates a mismatch between conceptual understanding and the final answer. This inconsistency is improved in ChatGPT 5.2, where the reasoning is consistently aligned with the final answers.

Question 20 (Fig. 4) was the most challenging for both ChatGPT 4 and 4.o. Unlike questions 4 and 5, which rely on text-based descriptions to assess basic conceptual understanding, question 20 requires identifying the fundamental harmonic of air displacement in an open-end cylinder using image-based answer choices without supporting text. Although ChatGPT often generated accurate reasoning, such as stating that in an open-end cylinder, the fundamental frequency occurs when there is a node at the

20) A tuning fork is hit and held in front of a one open-ended cylinder. The first harmonic resonance (the fundamental) is generated. Which diagram best describes the pattern of the displacement of air molecules inside the tube.

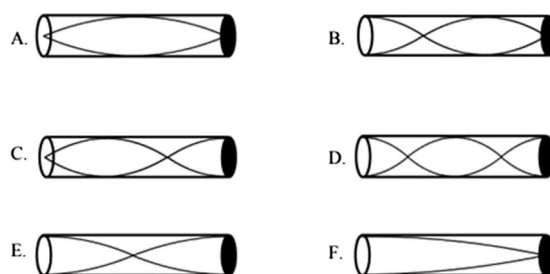


FIG. 4. Question 20, which is the most challenging question for ChatGPT 4 and 4.o.

TABLE XI. Reasoning and answers by ChatGPT for questions with image.

Answers							
GPT 4		GPT 4o		GPT 5.2			
Reasoning	I	C	I	C	I	C	Total
Incorrect	1	0	0	0	0	0	1
Correct	136	313	120	330	0	450	1349
Total	450		450		450		1350

closed end with no displacement of air molecules and an antinode at the open end with maximum displacement of air molecules, it failed to correctly identify the image representing the appropriate wavelength in the cylinder. In contrast, ChatGPT 5.2 showed improvement on this question, with a higher number of correct responses and better alignment between reasoning and answers. We observed that ChatGPT 5.2 includes improved image processing capabilities, which may allow it to interpret visual information more clearly and thus improve its analysis.

4. Questions with graph

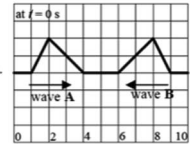
As discussed in the difficulty index section, questions involving graphs were the most challenging for ChatGPT to answer. According to Table XII, ChatGPT 4 and 4.o answered these questions correctly with correct reasoning in only about a third of the total repetitions. However, when considering the total number of correct reasoning instances, ChatGPT 4 achieved 71% and 4.o 82%. This indicates that while ChatGPT can generate logical reasoning to address MWCS questions, it struggles to select the correct answers due to its limitations in interpreting graphs. In contrast, ChatGPT 5.2 shows a substantial improvement, with a much higher proportion of correct answers with correct reasoning.

ChatGPT 4 and 4.o correctly answered questions 10 and 12, which evaluated the concept of wave profiles after superposition. However, it failed to identify the wave profile just before the two wave peaks overlapped, as seen in questions 9 and 11. To examine the reasoning and final answers in this case, we analyzed responses to questions 9 and 10, as shown in Fig. 5.

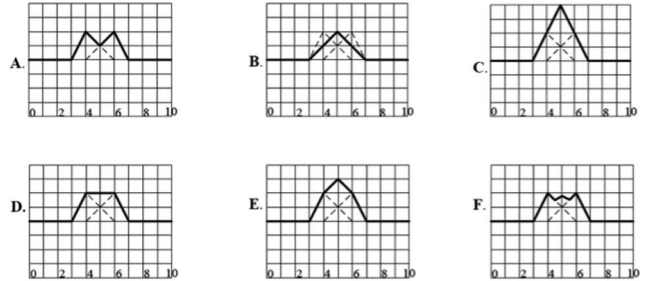
Figure 5 presents a question on constructive superposition, where question 9 assesses the understanding of the wave profile at the moment just before the two wave peaks overlap, and question 10 assesses the wave profile after superposition. Both ChatGPT 4 and 4.o struggled with question 9 due to misinterpreting the final wave position. Based on the reasoning, ChatGPT correctly calculated that each pulse would move 2 cm toward the center after 2 s at a speed of 1 cm/s. However, it mistakenly concluded that the pulses had already overlapped,

Consider the following description and answer questions 9-10.

Two pulses are moving towards each other. Each pulse has a speed of 1 cm/s. The figure on the right shows the pulses at time $t = 0$ s. Each square width corresponds to 1 cm x 1 cm. The dashed lines indicate the correct positions of the individual pulses after 2 s.



9) Select the drawing that corresponds to the shape of the resultant pulse after 2 s.



10) Select the drawing and explanation that corresponds to the shape of the resultant pulse after 5 s.

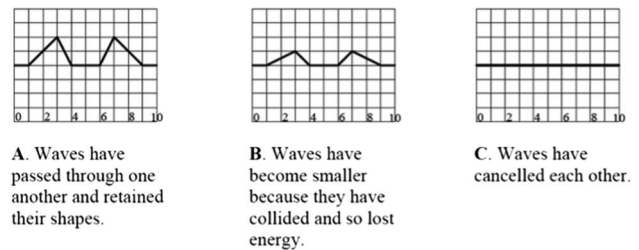


FIG. 5. Graphical questions of 9 and 10 which assess the understanding of constructive superposition.

interpreting this as the point of maximum constructive interference. As a result, it selected option C, which depicts the wave profile after the peaks have fully overlapped. This answer is incorrect, as the question asks for the wave profile just before the overlap occurs. The correct answer is option D.

In contrast, ChatGPT 5.2 shows substantial improvement in question 9. It is able to identify that the shape of the two pulses just before maximum superposition forms a trapezoidal profile, reflecting the partial overlap of the two waves before reaching the peak. This indicates that ChatGPT 5.2 not only improves in selecting the correct answer but also demonstrates a better understanding of the intermediate stage of wave superposition.

However, both ChatGPT 4 and 4.o answered question 10 correctly with high consistency. This likely occurred because the answer choices were described using text, removing the need to interpret graphical representations. Based on its reasoning, ChatGPT recognized that after 5 s the waves would pass through each other without losing energy or changing shape. This understanding matched the description in option A, leading ChatGPT to consistently select the correct answer. ChatGPT 5.2 also maintained high performance on this question.

TABLE XII. Reasoning and answers by ChatGPT for questions with graph.

Answers								
		GPT 4		GPT 4o		GPT 5.2		
Reasoning	I	C	I	C	I	C	Total	
Incorrect	162	13	101	5	38	20	339	
Correct	224	201	301	193	143	399	1461	
Total	600		600		600		1800	

Overall, the findings show that text descriptions accompanying graphical representations help ChatGPT answer questions more accurately. This is consistent with the findings of Polverini and Gregorcic [9], who also reported that ChatGPT struggles to interpret graphs but performs better when text descriptions are provided. The improvement observed in ChatGPT 5.2 suggests that its ability to interpret graphical information has increased, particularly in identifying intermediate wave shapes.

5. Two-tier questions

ChatGPT 4 and 4.o are both capable of answering two-tier questions in the MWCS, as shown in Table XIII. ChatGPT 4 provided correct answers with reasoning 62% of the time, while ChatGPT 4.o achieved 85%, indicating improved performance. ChatGPT 5.2 shows further improvement, answering all two-tier items correctly across every repetition, with 86.7% of responses (325 of 375) also supported by correct reasoning. In the remaining 50 correct responses, however, the reasoning was flawed, so the alignment between reasoning and answer is not perfect.

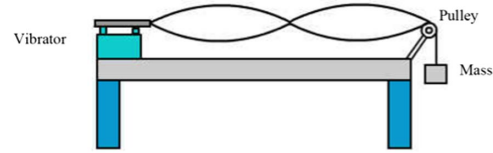
To explore these questions further, we examined questions 17–19, which assess students' understanding of the relationship between the tension in a string and the wavelength of standing waves (Fig. 6). On question 17, both ChatGPT 4 and 4.o answered with near-perfect consistency, correctly answering 74 and 75 of the 75 attempts, respectively, and ChatGPT 5.2 likewise answered all 75 correctly.

However, inconsistencies appeared in ChatGPT 4's responses to questions 18 and 19, which use the same apparatus but vary in the quantities involved. Question 18 asks how the wavelength changes when the hanging mass is increased by a factor of 4, thereby increasing the string tension, and question 19 asks about the effect of using a thicker rope, thereby increasing the linear mass density. Because the vibrator drives the string at a fixed frequency, the wavelength is governed by $\lambda = v/f$ together with $v = \sqrt{T/\mu}$. Increasing the tension raises the wave speed and therefore increases the wavelength so that a different harmonic is established, whereas a heavier rope lowers the wave speed and therefore decreases the wavelength.

TABLE XIII. Reasoning and answers by ChatGPT for two-tier questions.

Reasoning	Answers						Total
	GPT 4		GPT 4o		GPT 5.2		
	I	C	I	C	I	C	
Incorrect	22	9	8	2	0	50	91
Correct	113	231	48	317	0	325	1034
Total	375		375		375		1125

A standing wave is produced with a fixed length string, one end of which is attached to a vibrator, the other end of which is placed on a pulley and hung with a mass. Using the vibrator, the second harmonic standing wave is created (see figure below). The length between the vibrator and the pulley does not change!



17) If the frequency of the vibrator is doubled while everything else stays the same, a different harmonic standing wave is created. How would the wavelength of the new harmonic standing wave change?

Answer:

- A. Increase
- B. Decrease
- C. Stay the same

Reason:

1. The wavelength depends on the amount of energy flowing in the waves.
2. The wavelength is proportional to frequency since the velocity doesn't change.
3. The wavelength depends on the length of the rope the frequency doesn't affect the wavelength.
4. The wavelength is inversely proportional to the frequency since the velocity doesn't change.

FIG. 6. Question 17, which assesses the understanding of the relationship between the tension in a string and the wavelength of standing waves.

The difficulty ChatGPT 4 encountered on these items is illustrated by its responses to question 18. In some cases, ChatGPT 4 reasoned correctly, recognizing that increasing the mass raises the tension, that a higher tension increases the wave speed through $v = \sqrt{T/\mu}$, and that, because the vibrator holds the frequency fixed, the relationship $v = f\lambda$ requires the wavelength to increase. In these cases, it correctly selected the answer that the wavelength increases. However, it did not consistently pair this answer with the corresponding reason option, at times selecting a reason statement that did not match its own explanation. This pattern indicates that ChatGPT 4's difficulty on this item lay less in the physical reasoning itself than in reliably aligning that reasoning with the structured two-tier answer format. In contrast, ChatGPT 5.2 answered these items more consistently, though its alignment between reasoning and answer options was not uniformly perfect.

While the specific difficulties differ, the responses of ChatGPT 4 to questions 18 and 19 can be compared with the student difficulties reported by Barniol and Zavala [15]. Many students believed that increasing the tension would result in a decrease in wavelength, and although some offered seemingly logical explanations such as the frequency increasing with tension, their justifications were conceptually flawed. ChatGPT 4's failures on these items were of a different character. Rather than reflecting a consistent misconception about how tension affects the wavelength, they arose from run-to-run variability and from an inconsistent mapping between its reasoning and the two-tier answer options.

VI. DISCUSSION

Our study offers insights into the role of GAI in solving physics problems, particularly in terms of the development of capabilities across ChatGPT versions. This section highlights two main findings: ChatGPT's performance in selecting final answers on the MWCS and the quality of its reasoning when answering questions.

In general, ChatGPT performs well in selecting correct answers across different types of questions in the MWCS, including those without images or graphs, questions with images, and two-tier questions. The results show a clear progression across versions where ChatGPT 4's performance is comparable to that of first-year university students, ChatGPT 4.o approaches that of second-year students, and ChatGPT 5.2 surpasses all student groups in overall performance. These findings are consistent with previous studies showing that ChatGPT can match or exceed student performance on conceptual physics assessments [11,48].

More importantly, our results indicate that many of the previously reported limitations of ChatGPT have been substantially reduced in the latest version. Earlier studies and our results show that graph-based and visual representation questions were among the most challenging types [9,49,50]. However, ChatGPT 5.2 demonstrates significant improvement in these areas. For example, it is able to correctly identify intermediate wave shapes in superposition problems and shows improved alignment between reasoning and answers in image-based and two-tier questions. Although some inconsistencies remain in certain cases, the overall trend suggests that ChatGPT can now answer most types of questions within MWCS with high accuracy.

Another important finding relates to the alignment between reasoning and final answers. Earlier versions of ChatGPT often produced correct reasoning but failed to select the correct answer, particularly in questions involving graphs or visual interpretation. This mismatch is reduced in ChatGPT 5.2, where reasoning and answers are more consistently aligned.

The results of this study have important implications for classroom practice. Rather than viewing ChatGPT as a tool that should be restricted, educators may consider ways to use it to support learning [25,51]. For instance, the reasoning generated by ChatGPT often provides correct conceptual elaborations, which may be useful in helping students grasp underlying principles in physics problems. In addition, ChatGPT's errors can be used as a basis for discussion and critical evaluation. Such use may encourage students to engage more deeply with the material and develop higher-order thinking skills [52].

The strong ability of ChatGPT to answer physics problems also raises broader questions about the role of GAI in the classroom. Prior studies suggest that students may become overly reliant on ChatGPT to solve

problems, which can weaken critical thinking skills and reduce meaningful engagement with problem-solving tasks [12,41,53,54]. This also raises ethical concerns, as ChatGPT often produces responses that appear highly accurate, which may lead students to trust its answers without applying their own reasoning [11]. At the same time, previous studies report that some physics teachers view ChatGPT positively and explore ways to use it to support problem solving and navigate complex tasks more effectively [55]. In this context, studies that examine the performance of GAI in solving physics problems, such as the present work, can inform discussions about future instructional design and the possible integration of GAI into physics classrooms.

Another important aspect that needs to be addressed is the influence of language on the performance of GAI in answering physics questions. In this study, the MWCS was administered in English; however, previous research has shown that ChatGPT's performance varies significantly across languages [56–58]. For instance, one study that tested assessments from PhysPort, including the MWCS, found that GAI struggled when questions were presented in different languages. In the case of the MWCS, ChatGPT achieved a score of only 35 in Thai, compared to 79 in English [56]. Similarly, ChatGPT has been shown to struggle with physics terminology in other languages, such as German [58]. Therefore, existing reports on GAI performance should be interpreted carefully, particularly when applying findings across different linguistic contexts.

These findings suggest an important shift in the direction of research on GAI in physics education. Previous studies have focused on evaluating whether ChatGPT can correctly answer physics questions. However, based on the results of this study, particularly with ChatGPT 5.2, this type of evaluation may no longer be the most relevant research focus. The model demonstrates the ability to answer most conceptual questions correctly, including those that were previously considered challenging. Therefore, future research should move beyond performance evaluation and instead focus on how GAI can be effectively integrated into physics education.

The limitation of this study is the limited temporal validity of model-specific benchmarks. The performance reported here reflects the capabilities of ChatGPT at the time of data collection and may change as newer versions are released. Therefore, the findings should be interpreted as time-sensitive.

VII. CONCLUSION

This study demonstrates that ChatGPT shows significant development across versions in answering conceptual questions related to mechanical waves. While ChatGPT 4 and 4.o already demonstrate strong performance, ChatGPT 5.2 is able to answer most types of MWCS questions with

high accuracy, including those involving graphs, images, and two-tier reasoning, which were previously challenging. In addition, the alignment between reasoning and final answers improves substantially in the latest version.

These findings suggest that ChatGPT has reached a level where it can successfully handle standard conceptual physics questions commonly used in introductory assessments. As a result, research that focuses solely on evaluating whether ChatGPT can answer physics questions correctly may no longer be necessary. Instead, future research should focus on how GAI can be used to enhance teaching and learning in physics classrooms.

The presence of GAI in education requires a shift in instructional approaches. Rather than designing questions that attempt to avoid AI use, educators should consider how to incorporate GAI as a learning tool. For example, ChatGPT can be used to support inquiry-based learning,

provide explanations, and generate alternative perspectives on physics problems. At the same time, students should be guided to critically evaluate AI-generated responses and understand their limitations.

Overall, this study highlights the need to move from evaluating GAI performance to exploring its pedagogical implementations. By integrating GAI effectively into the classroom, educators can leverage its strengths to support deeper conceptual understanding while fostering critical thinking and problem-solving skills.

DATA AVAILABILITY

The data that support the findings of this article are available upon reasonable request. We thank the students who helped us with this research. We also thank the reviewers for their relevant and valuable suggestions.

-
- [1] E. A. Alasadi and C. R. Baiz, Generative AI in education and research: Opportunities, concerns, and solutions, *J. Chem. Educ.* **100**, 2965 (2023).
 - [2] Y.-J. Lan and N.-S. Chen, Teachers' agency in the era of LLM and generative AI, *Educ. Technol. Soc.* **27**, 1 (2024).
 - [3] H. Lin and Q. Chen, Artificial intelligence (AI)-integrated educational applications and college students' creativity and academic emotions: Students and teachers' perceptions and attitudes, *BMC Psychol.* **12**, 487 (2024).
 - [4] S. Habib, T. Vogel, X. Anli, and E. Thorne, How does generative artificial intelligence impact student creativity?, *J. Creativity* **34**, 100072 (2024).
 - [5] M. Giannakos, R. Azevedo, P. Brusilovsky, M. Cukurova, Y. Dimitriou, D. Hernandez-Leo, S. Järvelä, M. Mavrikis, and B. Rienties, The promise and challenges of generative AI in education, *Behaviour Inf. Technol.* **1** (2024).
 - [6] C. K. Y. Chan, A comprehensive AI policy education framework for university teaching and learning, *Int. J. Educ. Technol. Higher Educ.* **20**, 38 (2023).
 - [7] A. Guilherme, AI and education: The importance of teacher and student relations, *AI & Society* **34**, 47 (2019).
 - [8] L. Chen, P. Chen, and Z. Lin, Artificial intelligence in education: A review, *IEEE Access* **8**, 75264 (2020).
 - [9] G. Polverini and B. Gregorcic, Performance of ChatGPT on the test of understanding graphs in kinematics, *Phys. Rev. Phys. Educ. Res.* **20**, 010109 (2024).
 - [10] F. Kieser, P. Wulff, J. Kuhn, and S. Küchemann, Educational data augmentation in physics education research using ChatGPT, *Phys. Rev. Phys. Educ. Res.* **19**, 020150 (2023).
 - [11] D. Tong, Y. Tao, K. Zhang, X. Dong, Y. Hu, S. Pan, and Q. Liu, Investigating ChatGPT-4's performance in solving physics problems and its potential implications for education, *Asia Pac. Educ. Rev.* **25**, 1379 (2023).
 - [12] V. López-Simó and M. F. Rezende, Jr., Challenging ChatGPT with different types of physics education questions, *Phys. Teach.* **62**, 290 (2024).
 - [13] S. Aldazharova, G. Issayeva, S. Maxutov, and N. Balta, Assessing AI's problem solving in physics: Analyzing reasoning, false positives and negatives through the force concept inventory, *Contemp. Educ. Technol.* **16**, ep538 (2024).
 - [14] A. Tongchai, M. D. Sharma, I. D. Johnston, K. Arayathanitkul, and C. Soankwan, Developing, evaluating and demonstrating the use of a conceptual survey in mechanical waves, *Int. J. Sci. Educ.* **31**, 2437 (2009).
 - [15] P. Barniola and G. Zavala, The mechanical waves conceptual survey: An analysis of university students' performance, and recommendations for instruction, *Eurasia J. Math. Sci. Technol. Educ.* **13**, 929 (2016).
 - [16] K. Zhang and A. B. Aslan, AI technologies for education: Recent research and future directions, *Comput. Educ.* **2**, 100025 (2021).
 - [17] X. Chen, D. Zou, H. Xie, G. Cheng, and C. Liu, Two decades of artificial intelligence in education, *Educ. Technol. Soc.* **25**, 28 (2022).
 - [18] W. Holmes, K. Porayska-Pomsta, K. Holstein, E. Sutherland, T. Baker, S. B. Shum, O. C. Santos, M. T. Rodrigo, M. Cukurova, and I. I. Bittencourt, Ethics of AI in education: Towards a community-wide framework, *Int. J. Artif. Intell. Educ.* **32**, 504 (2022).
 - [19] S. Nemorin, A. Vlachidis, H. M. Ayerakwa, and P. Andriotis, AI hyped? A horizon scan of discourse on artificial intelligence in education (AIED) and development, *Learn. Media Technol.* **48**, 38 (2023).
 - [20] X. Zhai, X. Chu, C. S. Chai, M. S. Y. Jong, A. Istenic, M. Spector, J.-B. Liu, J. Yuan, Y. Li, and N. Cai, A review of artificial intelligence (AI) in education from 2010 to 2020, *Complexity* **2021**, 8812542 (2021).

- [21] U. Mittal, S. Sai, and V. Chamola, A comprehensive review on generative AI for education, *IEEE Access* **12**, 142733 (2024).
- [22] F. Almasri, Exploring the impact of artificial intelligence in teaching and learning of science: A systematic review of empirical research, *Res. Sci. Educ.* **54**, 977 (2024).
- [23] E. A. Alasadi and C. R. Baiz, Multimodal generative artificial intelligence tackles visual problems in chemistry, *J. Chem. Educ.* **101**, 2716 (2024).
- [24] G.-G. Lee and X. Zhai, Using ChatGPT for science learning: A study on pre-service teachers' lesson planning, *IEEE Trans. Learn. Technol.* **17**, 1643 (2024).
- [25] A. M. Bettayeb, M. Abu Talib, A. Z. Sobhe Altayasinah, and F. Dakalbab, Exploring the impact of ChatGPT: Conversational AI in education, *Front. Educ.* **9**, 1379796 (2024).
- [26] T. Wan and Z. Chen, Exploring generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning, *Phys. Rev. Phys. Educ. Res.* **20**, 010152 (2024).
- [27] N. M. Nasri, N. Nasri, N. F. Nasri, and M. A. Abd Talib, The impact of integrating an intelligent personal assistant (IPA) on secondary school physics students' scientific inquiry skills, *IEEE Trans. Learn. Technol.* **16**, 232 (2023).
- [28] W. Holmes and I. Tuomi, State of the art and practice in AI in education, *Eur. J. Educ.* **57**, 542 (2022).
- [29] B. Williamson and R. Eynon, *Historical Threads, Missing Links, and Future Directions in AI in Education* (Taylor & Francis, London, 2020).
- [30] L. Krupp, S. Steinert, M. Kiefer-Emmanouilidis, K. E. Avila, P. Lukowicz, J. Kuhn, S. Küchemann, and J. Karolus, Challenges and opportunities of moderating usage of large language models in education, *MUM'24: Proceedings of the International Conference on Mobile and Ubiquitous Multimedia* (2024), pp. 249–254, [10.1145/3701571.3701590](#).
- [31] X. Wang, L. Li, S. C. Tan, L. Yang, and J. Lei, Preparing for AI-enhanced education: Conceptualizing and empirically examining teachers' AI readiness, *Comput. Hum. Behav.* **146**, 107798 (2023).
- [32] J. Kim, H. Lee, and Y. H. Cho, Learning design to support student-AI collaboration: Perspectives of leading teachers for AI in education, *Educ. Inf. Technol.* **27**, 6069 (2022).
- [33] S. Schorcht, N. Buchholtz, and L. Baumanns, Prompt the problem—investigating the mathematics educational quality of AI-supported problem solving by comparing prompt techniques, *Front. Educ.* **9**, 1386075 (2024).
- [34] M.-L. Tsai, C. W. Ong, and C.-L. Chen, Exploring the use of large language models (LLMs) in chemical engineering education: Building core course problem models with Chat-GPT, *Educ. Chem. Eng.* **44**, 71 (2023).
- [35] T. M. Clark, Investigating the use of an artificial intelligence chatbot with general chemistry exam questions, *J. Chem. Educ.* **100**, 1905 (2023).
- [36] T. M. Clark, E. Anderson, N. M. Dickson-Karn, C. Soltanirad, and N. Tafini, Comparing the performance of college chemistry students with ChatGPT for calculations involving acids and bases, *J. Chem. Educ.* **100**, 3934 (2023).
- [37] M. A. Rangkuti and R. Karam, Conceptual challenges with the graphical representation of the propagation of a pulse in a string, *Phys. Rev. Phys. Educ. Res.* **18**, 020119 (2022).
- [38] M. A. Rangkuti and R. Karam, Encouraging students to understand the 1D wave equation, *Phys. Rev. Phys. Educ. Res.* **19**, 020161 (2023).
- [39] L. C. McDermott, M. L. Rosenquist, and E. H. van Zee, Student difficulties in connecting graphs and physics: Examples, *Am. J. Phys.* **55**, 503 (1987).
- [40] M. Planinic, Z. Milin-Sipus, H. Katic, A. Susac, and L. Ivanjek, Comparison of student understanding of line graph slope in physics and mathematics, *Int. J. Sci. Math. Educ.* **10**, 1393 (2012).
- [41] W. Yeadon and T. Hardy, The impact of AI in physics education: A comprehensive review from GCSE to university levels, *Phys. Educ.* **59**, 025010 (2024).
- [42] G. Polverini and B. Gregorcic, How understanding large language models can inform the use of ChatGPT in physics education, *Eur. J. Phys.* **45**, 025701 (2024).
- [43] K. D. Wang, E. Burkholder, C. Wieman, S. Salehi, and N. Haber, Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving, *Front. Educ.* **8**, 1330486 (2024).
- [44] P. Barniol and G. Zavala, Mechanical waves conceptual survey: Its modification and conversion to a standard multiple-choice test, *Phys. Rev. Phys. Educ. Res.* **12**, 010107 (2016).
- [45] S. Puttisanwimon and S. Wuttirom, Mechanical wave concepts of Thai high school students: Comparing learned and unlearned groups, *Int. J. Innovation Sci. Math. Educ.* **31** (2023).
- [46] T. Kubiszyn and G. D. Borich, *Educational Testing and Measurement* (John Wiley & Sons, New York, 2024).
- [47] M. Kryjevskaja, M. K. R. Stetzer, and P. R. L. Heron, Student understanding of wave behavior at a boundary: The relationships among wavelength, propagation speed, and frequency, *Am. J. Phys.* **80**, 339 (2012).
- [48] X. Zhai, M. Nyaaba, and W. Ma, Can generative AI and ChatGPT outperform humans on cognitive-demanding problem-solving tasks in science?, *Sci. Educ.* **34**, 649 (2025).
- [49] G. Polverini, J. Melin, E. Önerud, and B. Gregorcic, Performance of ChatGPT on tasks involving physics visual representations: The case of the brief electricity and magnetism assessment, *Phys. Rev. Phys. Educ. Res.* **21**, 010154 (2025).
- [50] G. Polverini and B. Gregorcic, Evaluating vision-capable chatbots in interpreting kinematics graphs: A comparative study of free and subscription-based models, *Front. Educ.* **9**, 1452414 (2024).
- [51] Y. Liang, D. Zou, H. Xie, and F. L. Wang, Exploring the potential of using ChatGPT in physics education, *Smart Learn. Environ.* **10**, 52 (2023).
- [52] D. M. Colaco and D. Antao, Exploring self-learning efficacy: A comparative analysis of ChatGPT and Gemini, *Interact. Technol. Smart Educ.* **22**, 688 (2025).
- [53] M. E. Frenkel and H. Emara, ChatGPT-3.5 and-4.0 and mechanical engineering: Examining performance on the

- FE mechanical engineering and undergraduate exams, *Comput. Appl. Eng. Educ.* **32**, e22781 (2024).
- [54] C. K. Y. Chan and L. H. Y. Tsi, Will generative AI replace teachers in higher education? A study of teacher and student perceptions, *Stud. Educ. Eval.* **83**, 101395 (2024).
- [55] H. Jang and H. Choi, A double-edged sword: Physics educators' perspectives on utilizing ChatGPT and its future in classrooms, *J. Sci. Educ. Technol.* **34**, 267 (2025).
- [56] G. Kortemeyer, M. Babayeva, G. Polverini, R. Widenhorn, and B. Gregorcic, Multilingual performance of a multimodal artificial intelligence system on multisubject physics concept inventories, *Phys. Rev. Phys. Educ. Res.* **21**, 020101 (2025).
- [57] G. Revalde, M. Zholdakhmet, A. Abola, and A. Murzagaliyeva, Can ChatGPT pass a physics test?, *Technol. Knowl. Learn.* **30**, 2459 (2025).
- [58] M. N. Dahlkemper, S. Z. Lahme, and P. Klein, How do physics students evaluate artificial intelligence responses on comprehension questions? A study on the perceived scientific accuracy and linguistic quality of ChatGPT, *Phys. Rev. Phys. Educ. Res.* **19**, 010142 (2023).