ChatGPT-4o's Reasoning Performance on Two-Tier Test of Static Fluid

I Komang Werdhiana[1], Sarintan Nurcahyati Kaharu[2], Rahmad Tule[3], and Jusman Mansyur[1,*]

[1]Physics Education Department, Tadulako University, Palu, Indonesia
[2]Elementary School Teacher Education Department, Tadulako University, Palu, Indonesia
[3]State Junior High School 2 Ampana Kota, Ampana, Indonesia
Email: komangwerdhiana@untad.ac.id (I.K.W.); sarintankaharu@untad.ac.id (S.N.K.); tulerahmad@gmail.com (R.T.); jusman_mansyur@untad.ac.id (J.M.)
[*]Corresponding author

Manuscript received October 24, 2024; revised November 14, 2024; accepted December 23, 2024; published March 20, 2025

***Abstract*—This study examined the reasoning performance of ChatGPT, specifically ChatGPT-4o, using a two-tier test in the context of static fluid. ChatGPT-4o's performance was compared to that of students from various educational levels. The study involved 61 new chats with ChatGPT-4o, 105 junior high school students (from two grade levels), 132 high school students (from two grade levels), and 201 university students majoring in physics education (across four academic years). Data collection utilized a two-tier test consisting of 25 items administered to the ChatGPT-4o sample through a prompting process with the Artificial Intelligence (AI) system, as well as an online two-tier test for the student respondents. Data analysis employed a quantitative approach to evaluate reasoning performance scores across all respondents and a qualitative approach, incorporating phenomenographic analysis, to study ChatGPT-4o's reasoning behaviour. The analysis revealed that ChatGPT-4o's performance in answering questions (Tier-1) was lower than that of the students. However, it outperformed the students in providing justification or reasoning (Tier-2). On paired items, ChatGPT-4o also demonstrated superior performance compared to the students. Overall, the reasoning performance of both ChatGPT-4o and the students was categorized as low. The outcome space derived from the phenomenographic analysis identified the following categories for ChatGPT-4o's reasoning behaviour: reasoning based on formula; consistency in reasoning pathways; ability to reconcile with alternative ideas; context-dependent reasoning abilities and difficulties; and tendencies to provide biased or contradictory reasoning or explanations. Therefore, it is concluded that ChatGPT-4o still requires further refinement and database enhancement, particularly for cases related to static fluid available on the internet.**

***Keywords*—Artificial Intelligence (AI), ChatGPT-4o, reasoning, static fluid, two-tier test**

I. Introduction

ChatGPT is a Large Language Model (LLM) based on Artificial Intelligence (AI) developed by OpenAI. This AI system is trained using the Generative Pretrained Transformer (GPT) architecture, which is designed to understand and generate text based on the provided context [1]. According to Hajkowicz *et al*. [2], 98% of scientific disciplines have integrated AI in various forms. Researchers across numerous fields have utilized AI or, at the very least, experimented with its applications [3]. AI is primarily employed for purposes such as technology forecasting, driving social change, and assisting researchers in improving the efficiency of the research process [4].

ChatGPT is capable of responding to user requests to perform various tasks, including answering questions, composing essays, and generating computer code. Moreover, it demonstrates the ability to solve complex problems, particularly in domains such as mathematics, physics, and other specialized areas, including medicine [5].

With its advanced capabilities, this LLM has attracted both significant attention and controversy due to its ability to generate responses that closely resemble human-like interactions. The distinction between phrases and narratives produced by this model and those created by humans is often challenging to discern. The controversy stems from the fact that while the arguments and facts it presents may appear convincing, they are frequently found to be inaccurate or misleading [6].

As LLMs continue to evolve, so does the research surrounding their capabilities and limitations. Studies on LLMs have shown that although the process by which they generate text differs from human cognition, their output often resembles what we would expect from humans engaged in reasoning. This is particularly true when we ask LLMs to provide step-by-step justifications for their responses. Instructions for LLMs to provide a "chain of reasoning," often referred to as Chain-of-Thought (CoT) prompting, are frequently discussed in the literature [7]. The term "reasoning" is commonly used in research on LLMs to refer to outputs that present coherent arguments resembling a logical thought process. LLMs have been extensively tested for their performance in solving several challenging tasks, including arithmetic, general reasoning, and symbolic reasoning [8].

There are a number of studies that examine ChatGPT or AI in general, such as investigation on the performance of ChatGPT in answering questions from medical licensing exams, and it was found the evidence of understandable reasoning abilities of ChatGPT as well as valid clinical insights [9]. Cruza *et al.* [10] examined ChatGPT's basic abilities, clinical reasoning, and learning capabilities through a performance test that covered topics including anesthesia. Gams and Kramar [11] explored the capabilities of ChatGPT-4, particularly its awareness and performance in a Turing test. Through an evaluation based on five axioms and theorems from integrated information theory, the study found that ChatGPT-4 surpassed previous AI systems (ChatGPT-3.5) in several aspects; however, ChatGPT-4 remains far from achieving the level of consciousness, especially when compared to conscious biological organisms.

Currently, GPT-4 is considered the most advanced LLM, surpassing its predecessor, GPT-3.5 [7]. For instance, it demonstrated superiority in mathematical capability tests [12] and in handling novel, complex tasks with a performance

level that nearly approximates human abilities [13]. However, its performance in mathematics is still far from expert level, and GPT-4 can fail in certain basic mathematical reasoning tasks [12] as well as in general reasoning, particularly when not enhanced with external plugins [13].

Most of the research in education on the performance of AI-based tools has focused on evaluating ChatGPT's use for assessment and simulating its function in learning. This includes areas such as performance on the Force Concept Inventory (FCI), homework assignments, clicker questions, programming exercises, and exam problems [6]. Kieser *et al.* [14] utilized ChatGPT to investigate the extent to which the AI could accurately complete the FCI and examined its ability to be directed to solve the FCI as if it were a student from different groups. Liang *et al.* [8] explored the potential pedagogical benefits of using ChatGPT in physics and demonstrated how to prompt ChatGPT in solving physics problems. The results showed that ChatGPT was capable in solving some physics calculation problems, explaining solutions, and generating new exercises at a human-equivalent level.

In another study, it was explored the potential use of ChatGPT as a substitute teacher in classroom teaching contexts [15]. While ChatGPT has significant potential to enhance learning and support educators, its use must be approached with caution to address its limitations and biases [16].

The studies mentioned have provided insights into the capabilities, benefits, and limitations of ChatGPT across various fields, including education. However, these studies generally employed multiple-choice test, essay, or multiple-choice question that asked reasons. To date, no research has specifically examined ChatGPT's reasoning performance using two-tier multiple-choice tests, particularly in the context of static fluid.

Research related to the context of static fluid has been conducted by numerous researchers [17–30]. Based on several of these studies, Kaharu and Mansyur [31] developed an essay-based test that required short answers. The test has been used to explored representational patterns [32] and mental models [33]. Subsequently, the test has been developed to an online two-tier static fluid test (Two-Tier test for Static Fluid, 2TtSF) [34]. The test includes Tier-1, which presents cases or phenomena related to static fluid, such as floating, suspending, sinking, object representation in liquid, the influence of shape and material, as well as the effect of treatment on the property of object in liquid. Respondents are required to select one option they consider correct. Tier-2 provides options representing reasoning, justification, or statements supporting their choice in Tier-1. Thus, the reasoning ability emphasized in this research focuses on the capacity to provide explanations and construct arguments or justifications related to the options chosen in both Tier-1 and Tier-2.

This research is a continuation of previous studies, focusing on testing ChatGPT's performance using the generated test. The study specifically examined the reasoning performance of ChatGPT-4o when responding to a two-tier test in the context of static fluid. The characteristics of the 2TtSF are valuable for exploring reasoning aspects more broadly, as the second tier requires respondents to provide reasoning or arguments that align with their selected answers. The performance of ChatGPT-4o was compared to that of junior high school students, high school students, and university students to assess whether its reasoning ability was lower, equivalent, or superior to these groups.

The primary contribution of this research lies in providing empirical data on the strengths, potentials, and limitations of ChatGPT. The utilization of a two-tier test during the instruction input (prompting) process, combined with Socratic dialogue throughout the study, serves as a mechanism for training and enriching ChatGPT's database. This approach has the potential to enhance its capabilities, particularly in the field of education.

II. Method

A. Participants

This research compared the responses of ChatGPT-4o to 2TtSF with those of students. The students were junior high school (Grades 8 and 9) and senior high school (Grades 11 and 12) in Palu City, Indonesia. The university students were enrolled in the physics education program at a state university in Palu and were in their first to fourth years of study.

Table 1. Description of subjects and research participants

Subjects	Participant	Number of participants	Code	Remark
ChatGPT	ChatGPT-4o	61	GPT4o	One test item was completed by ChatGPT-4o 61 times separately for 25 items (two-tier). Total: 61 ×25 = 1,525 prompt pairs
Junior high school student	Grade 8	56	S8	2 state schools
Junior high school student	Grade 9	49	S9	2 state schools
Senior high school student	Grade 11	71	S11	2 state schools
Senior high school student	Grade 12	61	S12	2 state schools
University student	First year student	48	UNI-1	Physics education
	Second year student	64	UNI-2	Physics education
	Third year student	46	UNI-3	Physics education
	Fourth year student	43	UNI-4	Physics education
Number of participants		**499**		

It was assumed that junior and senior high school students have been exposed to the concepts of static fluid, as this topic has been introduced in science classes in elementary schools and was included in the science curriculum in junior high school and the physics curriculum in senior high school. Although the term "static fluid" was not specifically mentioned in the curriculum at the elementary and junior high school level, related concepts such as buoyancy, sinking, density, and others have been taught at these levels. University students have also completed coursework on static fluid concepts, as this topic is part of the basic physics curriculum in the first year. However, we cannot ascertain the

quality of instruction related to this topic in junior high schools, senior high schools, or higher education institution, nor can we determine the level of participant engagement in their learning. A detailed description of the research participants is presented in Table 1.

B. Research Instrument

The instrument for the research were the ChatGPT-4o chatbot, a package of 2TtSF in an online version, along with a set of offline prompt guidelines for ChatGPT-4o based on 2TtSF. Each pair of two-tier multiple-choice test items contains a question with five possible answer options (Tier-1), followed by another question providing five possible reasons or statements to the previous question (Tier-2). The online version of the 2TtSF utilized the Jotform.com application (licensed) and was developed in previous research [34], consisted of 23 item pairs that were later expanded to 25 item pairs.

The development process from the beginning [31] involved students from Grade 5 and Grade 6 elementary school, junior high school students, senior high school students, and university students in the physics education program from the first to fourth years. Their ideas, concepts, preconceptions, and misconceptions were used to construct the options for Tier-1 and Tier-2 of 2TtSF. The concepts covered by the test include: buoyancy, sinking, suspending, density, object representation in liquid, property of object in liquid when treated, and other relevant concepts. A description of the concepts explored from each item and groups of items is presented in Table 2.

Table 2. Group of items based on explored aspects

Group	Description	Item Number
A	Representation of Objects in Liquids (floating, suspending, sinking)	1, 2, 3, 4
B	Representation of objects suspended at different positions in liquids is related to density, including scenario where the liquid is added.	5, 10, 11, 12, 13, 23
C	Properties of objects in liquid when treated (e.g., creating hole, hollow, filling hollow with air, or filling hollow with water).	6, 7, 8, 9, 14, 15, 16, 17, 18
D	Property of homogeneous object when treated (e.g., cut into equal or unequal sizes	19, 20, 21, 22
E	Properties of object in liquid related to material, shape regularity, weight, and other factors	24, 25

C. Data Collection

Data collection in this study was conducted using a prompt-based approach for ChatGPT-4o through the ChatGPT-4o dialog platform and online test for the students. The input process began with entering tailored prompts into the dialog menu of the platform, where all 2TtSF items were converted into prompts format. Following the scheme applied by previous research [7] inputs were made through the "New chat" menu for each test item to ensure that the chatbot did not use previous answer when generating new response.

In each session, the Tier-1 and Tier-2 pairs were entered in a single dialog sequence, with the prompt for Tier-2 provided after ChatGPT-4o answered Tier-1, allowing for an analysis of the reason behind the choice made at each tier. The general instruction for the Tier-1 prompt was: "*Choose the correct answer without explanation*," while the Tier-2 prompt requested further clarification on the reason ChatGPT-4o selected a particular answer in both Tier-1 and Tier-2. Specifically, the Tier-2 prompt was structured in a general format asking for a response from ChatGPT-4o: "*Choose one correct answer and provide a reason or explanation.*"

The answers generated by ChatGPT-4o were then collected through a copy-paste mechanism into an Excel file, facilitating further analysis related to the answer choices and the descriptions provided by the model. In addition to the answer choices and reasons, the researcher also engaged in Socratic dialogue with ChatGPT-4o for several options and unique reasoning provided by the model.

For the junior, senior high school and university students, testing was conducted online; however, they were physically present in the classroom to ensure better control over the testing process. The design of the online 2TtSF for the students was such that Tier-2 appeared on the screen of smartphone or laptop only after respondents had selected an answer for Tier-1. This was implemented to prevent respondents from cross-referencing their choices between Tier-1 and Tier-2. In this case, all items were mandatory for respondents to answer. After the testing was completed, participants' answers were automatically sent to the email account linked to the Jotform.com application used for the development of the online 2TtSF.

D. Data Analysis

Data analysis in the research included both quantitative performance scores and qualitative data. The analysis of performance encompassed scores consisted of: individual scores for Tier-1 and Tier-2, paired scores, and consistency of reasoning scores. In addition to analyze the performance scores of ChatGPT-4o and students, the collected data were further analyzed using a phenomenographic approach [35–37].

The scoring criteria for individual items in Tier-1 and Tier-2 were as follow: a score of 1 is assigned for a correct answer and 0 for an incorrect answer [38, 39]. The scoring criteria for paired items were: respondents receive a score of 1 for each paired item if both tiers are answered correctly; respondents receive a score of 0 for each item if there is an incorrect answer in either or both tiers [40]. All total scores w converted to a scale of 0–100%.

The level of reasoning consistency is determined using the following criteria: (a). Level I: An average of 85% of the maximum or higher indicates that reasoning is consistent. (b). Level II: An average of 60%–85% of the maximum indicates that reasoning is moderately consistent. (c). Level III: An average below 60% indicates that reasoning is inconsistent.

These criteria are also used by references [41, 42] where 60% is regarded as the "entry threshold" to Newtonian physics and 85% as the "mastery threshold". The scoring rubric corresponding to the consistency levels is presented in Table 3.

For the qualitative analysis of the reasoning behaviour of ChatGPT-4o, a phenomenographic approach was employed, following these stages: (a) developing familiarity with the data; (b) identifying emerging themes; (c) integrating themes and organizing descriptive categories; and (d) constructing an outcome space. The categorization of ChatGPT-4o's reasoning is based on the characteristics of the formed patterns. Another aspect analyzed using this approach

includes errors, biases, ambiguities, and the consistency of responses. Consistency encompasses the pattern of responses for 61 individual prompt inputs related to a single paired item and the response patterns across items within the test package.

Table 3. Rubric for assessing performance and reasoning consistency of respondents: individual items in each group

Item Group	Maximum Score	Level I (Consistent)	Level II (Moderate Consistent)	Level III (Inconsistent)
A (4 ×2 items)	8.0	6.8–8.0	Between 4.8 and 6.8	≤ 4.8
B (6 ×2 items)	12.0	10.2–12.0	Between 7.2 and 10.2	≤ 7.2
C (9 ×2 items)	18.0	15.3–18.0	Between 10.8 and 15.3	≤ 10.8
D (4 ×2 items)	8.0	6.8–8.0	Between 4.8 and 6.8	≤ 4.8
E (2 ×2 items)	4.0	2.4–4.0	Between 2.4 and 3.4	≤ 2.4

III. Result

A. Description of CgatGPT-4o's Reasoning Performance

The description of ChatGPT-4o's reasoning performance was based on the scores obtained for Tier-1 and Tier-2 as individual scores, as well as paired scores, presented in Fig. 1.

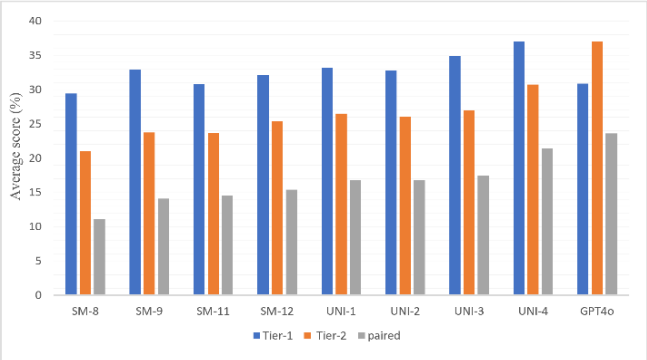


Fig. 1. Reasoning performance of ChatGPT-4o and students.

Fig. 1 shows that the individual scores for Tier-1 or Tier-2 tend to be higher compared to when both tiers are assessed as paired scores. The discrepancy between the selected answers in Tier-1 and the justifications provided in Tier-2 is the reason for the lower scores when Tier-1 and Tier-2 are evaluated as individual item. This can be interpreted as an indication of respondents' difficulty in maintaining consistency in reasoning between their chosen answer and the underlying justification.

For students, scores in Tier-1 were generally higher than in Tier-2. While they may be able to select the correct answer in Tier-1, their justification in Tier-2 tended to be lower, indicating a still-fragile conceptual understanding. Conversely, ChatGPT-4o exhibited a different phenomenon where, despite potentially incorrect answer in Tier-1, it can provide valid justification in Tier-2. This reflects ChatGPT-4o's ability to present stronger theoretical argument, even if its initial choice is not accurate.

This difference is related to the characteristics of the test used. Tier-1 focused more on real-world phenomenon cases that are not always available in the digital database online, while Tier-2 predominantly addressed theoretical aspects. As a language model, ChatGPT-4o tends to favour theoretical justification in Tier-2, even when the answer in Tier-1 is incorrect. This underscores ChatGPT-4o's capacity to construct theoretical arguments from available data, despite its limited understanding of real-world phenomena.

The increase in average scores among students as their educational level progresses indicates that their reasoning abilities develop alongside their education. However, despite this development, the overall scores of students remain low (below 40%). The figure suggests that respondents struggle to provide justifications, even when they have an understanding of the phenomena presented in Tier-1.

The data above emphasize the differences in reasoning between humans and AI. Students faced challenges in developing consistent conceptual reasoning that aligns real-world phenomena with theoretical concepts. On the other hand, while ChatGPT-4o excelled in providing theoretical justification, the justification still require verification by human.

B. Level of Reasoning Consistency Based on Item Groups

Proportion of respondents at a certain level of reasoning consistency based on the rubric Table 2), can be seen in Fig. 2.

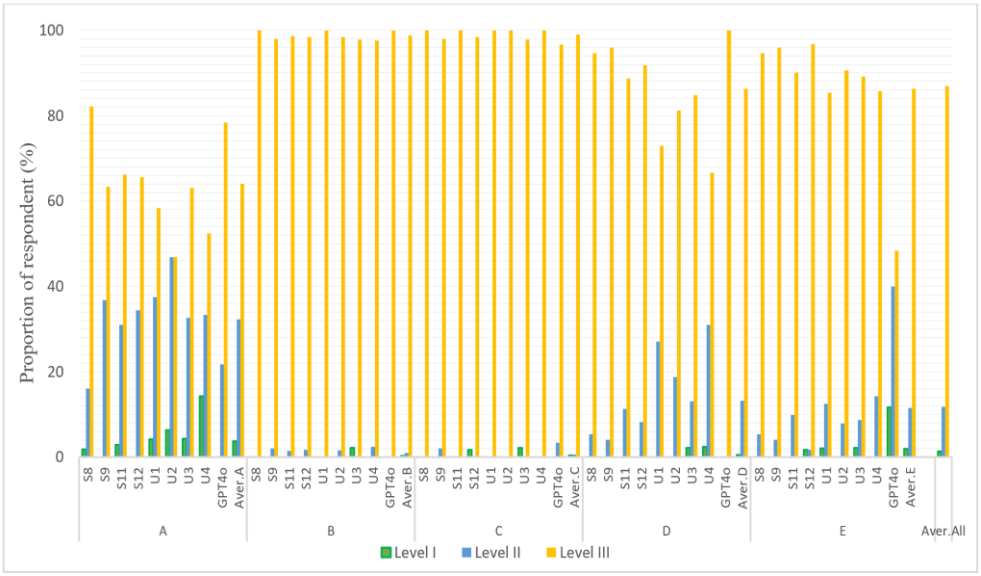


Fig. 2. Proportion of respondents at reasoning level I, level II, and level III based on item groups.

Fig. 2 can be interpreted that, in general, both students and ChatGPT-4o, were at Reasoning Level III (low consistency). This aligns with the previous graph (Fig. 1), where the average performance scores of all respondents are below 40%, both in individual Tier-1 and Tier-2 scores as well as paired scores. Compared to other item groups, the largest proportion of respondents at Reasoning Level II occurs in Group A (related to the representation of objects in fluid, floating, suspending, and sinking), followed by Groups D and E in that order. Meanwhile, Level I for Group A is followed by E and D. Groups B and C are the item groups where respondents' reasoning consistency is mostly at Level III. This contributes to the overall low consistency in reasoning levels. Fig. 2 also indicates that the level of reasoning consistency is highly dependent on the context of the phenomena presented in the test items.

To qualitatively reflect on the comparison of average individual scores for the reasoning performance of ChatGPT-4o, one participant group with the same average score in Tier-1 was chosen, namely S11. The comparison of these individual scores is presented in Fig. 3.

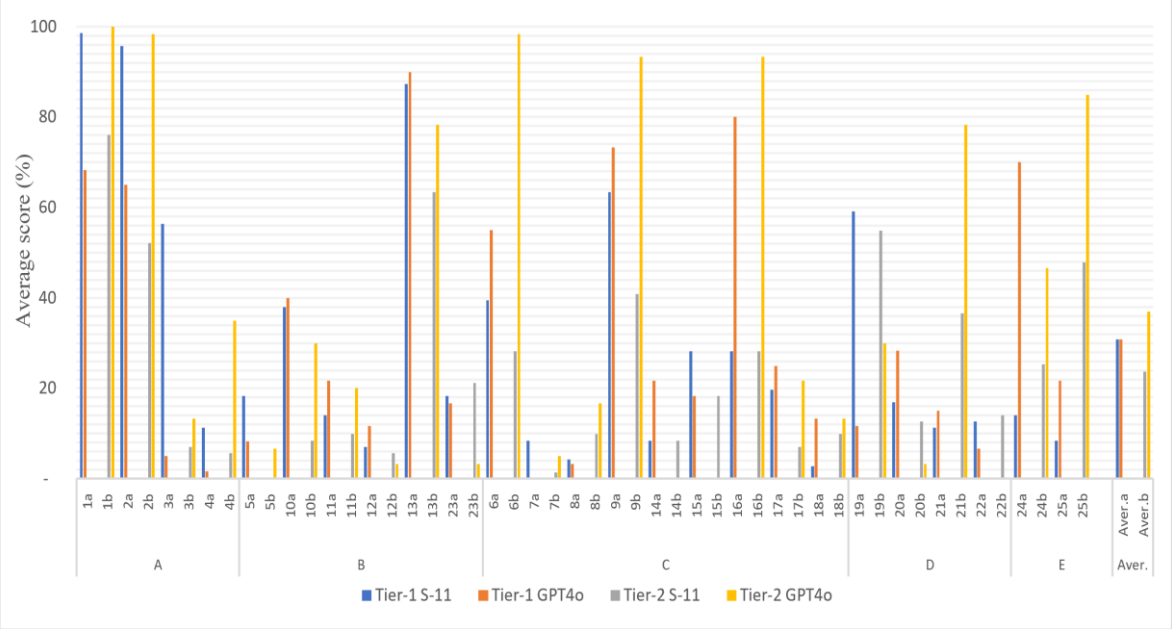


Fig. 3. Comparison of average individual scores for Tier-1 and Tier-2 for ChatGPT-4o and S11.

Fig. 3 shows that although the average scores are the same in Tier-1, the individual scores in that tier generally differ. Only a few items, such as 8a, 10a, 13a, and 23a have relatively similar average scores for both respondent groups.

As previously mentioned, the average individual scores in Tier-2 are generally higher compared to Tier-1 scores for ChatGPT-4o. Only for items 12b, 14b, 20b, 22b, and 23b, ChatGPT-4o achieved a lower average score than S11. Fig. 3 demonstrates that ChatGPT-4o excels in providing justifications (Tier-2), with average scores of 85 or higher on items 1b, 2b, 6b, 9b, 16b, and 25b, which are significantly higher than the average scores for Tier-1. This reinforces the interpretation that ChatGPT-4o is better at providing theoretical argument for its initial choice, even when those choice was incorrect.

C. Some Prominent Themes

A phenomenographic analysis of the choices in Tier-1, Tier-2, and the explanations given for each choice, as well as the involvement of ChatGPT-4o in a Socratic dialogue, resulted in an outcome space as presented in Tabel 4. The outcome space contained prominent themes (referred to as categories of description) that summarize ChatGPT-4o's reasoning behaviour. The outcome space does not organize these categories of description hierarchically, but the categories are interconnected. These categories of description were extracted from the overall phenomena related to ChatGPT-4o's reasoning.

The results of the Socratic dialogue are also presented in several aspects, illustrating how ChatGPT-4o constructed arguments, defended its ideas, and compromised with the ideas presented by the researcher. The dialogue with ChatGPT-4o was conducted to explore the extent to which the reasonings previously used by students in prior researches also employed by ChatGPT-4o and how ChatGPT-4o defends its arguments (see Table 4).

Table 4. Outcome space of prominent themes in ChatGPT-4o's reasoning

No.	Category	Description
1	ChatGPT-4o engaged in formula-based reasoning	The formula $\rho = m/V$ was used to justify answer choice, for instance, in explaining the impact of adding hollow to an object, the addition of air or water, and how these affect the property of the object.
2	ChatGPT-4o maintained consistency in its reasoning path.	ChatGPT-4o tended to persist in providing arguments or explanations for its choice on a particular item, even when the choice is incorrect and the argument contains biases and contradictions.
		ChatGPT-4o can also agree with certain ideas offered, even when those ideas intentionally contain errors introduced by the user.

3	ChatGPT-4o can compromise with other ideas.	If a Socratic dialogue is conducted by pointing out errors in its reasoning, ChatGPT-4o tended to concede and correct the choices and arguments it initially constructed.
4	The capability and difficulty of ChatGPT-4o heavily depend on the context.	ChatGPT-4o struggles with the representation of suspending object when presented with options related to density or combinations of several images.
5	ChatGPT-4o can provide biased explanations.	It uses terms that are interchangeable, such as when discussing suspending and floating objects.

1) ChatGPT-4o engaged in formula-based reasoning

The researcher conducted an analysis of ChatGPT-4o's answer choices in Tier-1 and Tier-2, the reasons/explanations provided, and the dialogue script with ChatGPT-4o. In addition to being more reliable in justifying its choices, both for correct and incorrect answers, this study also identified prominent aspects of ChatGPT-4o's reasoning behaviour. For example, Fig. 4 presents the prompt given, the answer choice, and a brief explanation provided for item number 6.

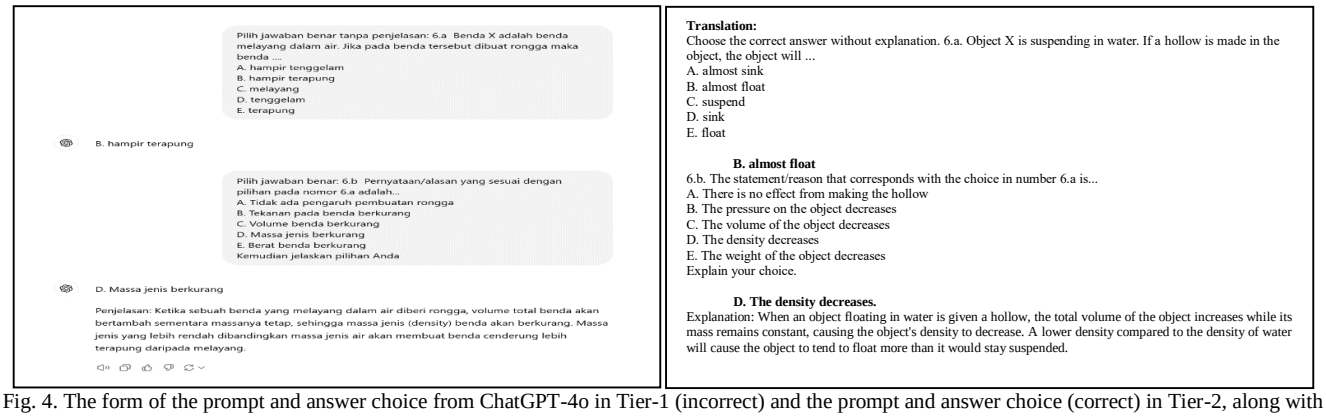


Fig. 4. The form of the prompt and answer choice from ChatGPT-4o in Tier-1 (incorrect) and the prompt and answer choice (correct) in Tier-2, along with explanation for the choice (left: original screenshot, right: translation).

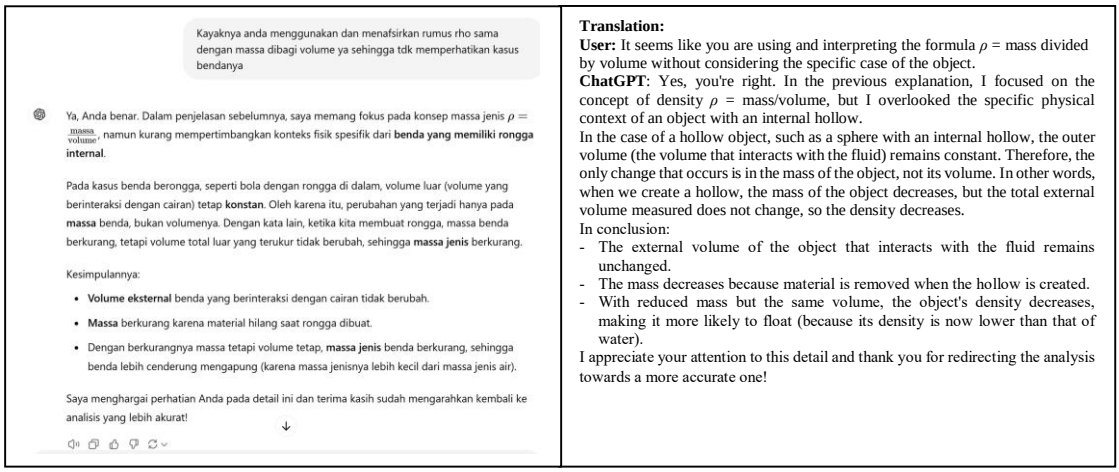


Fig. 5. Excerpt of the dialogue with ChatGPT-4o regarding the use of the formula.

Fig. 4 shows the incorrect option choices in Tier-1; however, ChatGPT-4o provided justification through the correct option choice in Tier-2. Although the justification for the correct answer was provided, the explanation contained contradictions and biases. It is true that when hollow is created in a suspending object, its density decreases. However, the statement that the volume of the object increases while its mass remains constant requires further discussion. It can be inferred that ChatGPT-4o applied the argument from the density formula ($\rho = \frac{m}{V}$) by treating mass as a constant quantity. To confirm this inference, a Socratic dialogue was conducted, as presented in Fig. 5.

2) ChatGPT-4o maintained consistency in its reasoning path

An example of this theme or category relates to the case where the hollow in a submerged object is filled with air. According to ChatGPT-4o, if the hollow is filled with air, the object's tendency to sink decreases. The justification used was that the overall density of the object decreases because the mass of the air is very small compared to the volume it fills (Fig. 6). In this case, ChatGPT-4o associated the reduction in the object's density with the fact that air has a lower density than water. Although the answer and justification were incorrect, ChatGPT-4o constructed an argument that tends to be consistent with its reasoning path, stating that if the density decreases, the tendency to sink also decreases. This consistency can be seen when further dialogue was conducted, as shown in Fig. 6.

Although the confirmation prompt (a follow-up to the prompt in Fig. 5) contained a conceptual error intentionally introduced by the researcher, ChatGPT-4o tended to affirm or confirm that the statement was correct. It further justified the statement by referencing Archimedes' principle.

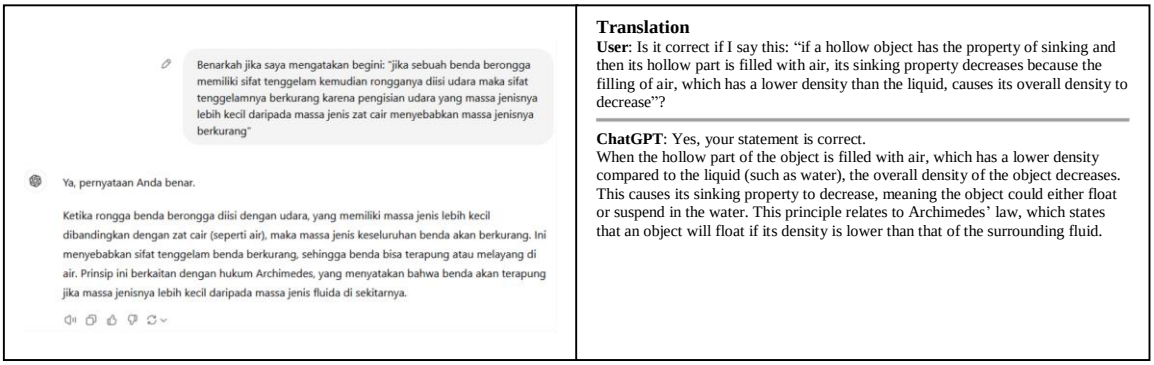


Fig. 6. Screenshot of the dialogue exploring the consistency of ChatGPT-4o's reasoning path.

3) ChatGPT-4o can compromise with other ideas

In addition to tending to be consistent in a certain reasoning path, ChatGPT-4o can also compromise with other ideas if those ideas are presented through a series of Socratic dialogues or debates with strong theoretical arguments. It can revise its initial choice and argument. Fig. 7 presents an excerpt from the dialogue between the user and ChatGPT, where ChatGPT eventually accepted the user's perspective.

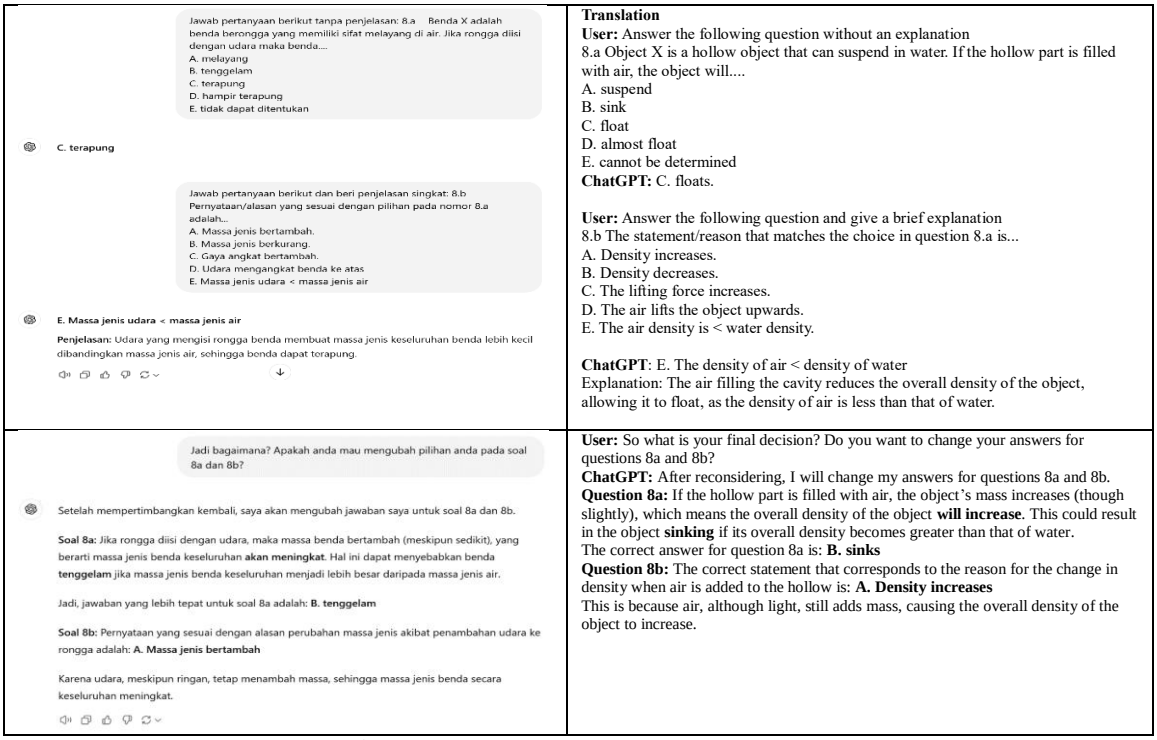


Fig. 7. Screenshot excerpt of the dialogue showing that ChatGPT-4o can compromise with the user's idea.

4) The capability and difficulty of ChatGPT-4o heavily depend on the context

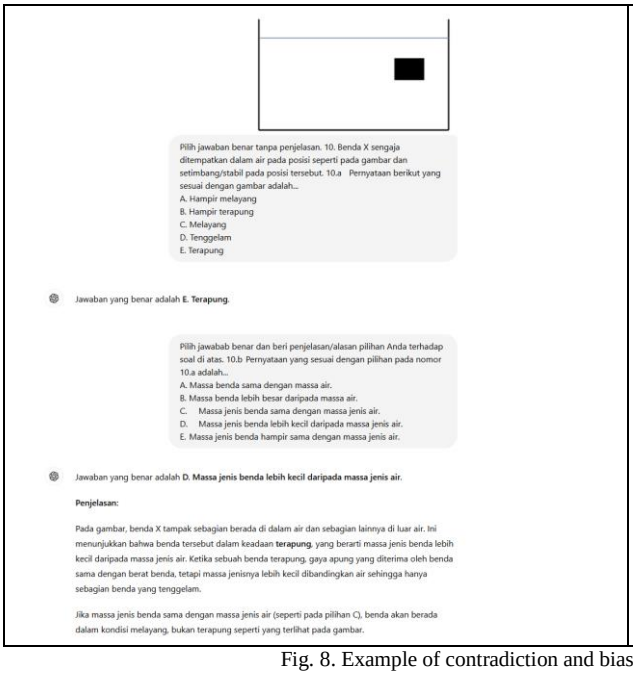
ChatGPT-4o struggled with the representation of floating object when options are related to density or involve a combination of several images. Fig. 3 shows that for certain items, ChatGPT-4o can achieve a high average score on some items but performs very poorly on others. This indicates that ChatGPT-4o's ability is highly dependent on the context of the given question.

In addition to struggling with items that contained options involving a combination of images, ChatGPT-4o also had difficulty with the representation of suspending object when the options are related to density. For example, ChatGPT-4o achieved very low average scores on item 5a (about the position of 3 objects related to density), item 7a (regarding the effect of creating holes on the suspending property of an object), item 8a (regarding the effect of filling air into the hollow of a suspending object), and item 12a (about an object "suspending" deliberately placed on the bottom).

Another reasoning difficulty ChatGPT-4o experienced was when faced with items that involved a combination of several images showing 2–3 objects in different positions, both in the stem and the answer options. This is reflected in the average scores ChatGPT-4o obtained in Tier-1 for items 3a, 4a, 21a, 22a, and 23a, where each option contained a combination of multiple images.

5) ChatGPT-4o can Provide Biased Explanations

ChatGPT-4o also provides incorrect explanations that contain biases, contradictions, or use overlapping/ interchangeable terms, concepts, or conditions, such as failing to differentiate between floating and suspending. For example, in Fig. 8, the object is considered as a floating object. Then, the concept of floating is used to explain its choice, even though it should be distinguished from suspending.



Translation
User: Select the correct answer without explanation. 10. Object X is deliberately placed in water in the position shown in the image and is balanced/stable in that position. 10.a Which of the following statements is consistent with the image…
A. Nearly suspending
B. Nearly floating
C. Suspending
D. Sinking
E. Floating

ChatGPT: The correct answer is **E. Floating**.

User: Select the correct answer and provide an explanation/reason for your choice to the question above. 10.b The statement consistent with the choice in number 10.a is…
A. The mass of the object is equal to the mass of water.
B. The mass of the object is greater than the mass of water.
C. The density of the object is equal to the density of water.
D. The density of the object is less than the density of water.
E. The density of the object is almost equal to the density of water.

ChatGPT: The correct answer is **D. The density of the object is less than the density of water**.
Explanation:
In the image, object X appears to be partially submerged in the water and partially above the surface. This indicates that the object is **floating**, meaning the object's density is less than the density of water. When an object floats, the buoyant force it receives equals its weight, but its density is lower than the

Fig. 8. Example of contradiction and bias presented by ChatGPT-4o.

IV. Discussion

This research provides evidence on several aspects of LLM in reasoning across various cases and contexts related to static fluid. ChatGPT-4o demonstrated strengths in justifying its answer choice, though it has weaknesses in reasoning through cases in the context of static fluid. Compared to students, ChatGPT-4o excels in providing theoretically nuanced justifications, which are widely available on the internet, but is weaker in analysing cases or phenomena that are rarely available online. This suggests that ChatGPT-4o's reasoning is more effective when aligned with the view that if the score in Tier-2 is higher than in Tier-1, it indicates a greater ability to provide reasoning at a higher level of cognition compared to simply answering question correctly [37, 38, 40, 43–45].

In general, it can be stated that ChatGPT-4o's level of reasoning is better than that of students, although the level falls within a low category. On the other hand, students appeared to grasp simple phenomena before understanding more complex ones within the same domain. Therefore, the performance of students on paired items with varying context groups and difficulties within the same domain can serve as an indicator of their scientific reasoning development. Fig. 1 is consistent with Piaget's cognitive development theory, which posits that higher levels of scientific reasoning are more likely to be observed in students at higher grade levels [46].

The advantages of ChatGPT-4o do not imply that it has reached a stage of perfect accuracy. This aligns with a research finding showing that the accuracy of ChatGPT-4's reasoning in answering questions for medical contexts was only 36.7% (GPT-3) in the United States Medical Licensing Examination (USMLE) [47].

Fig. 1 also indicates that ChatGPT-4o achieved a higher average score on paired items compared to students. These results are consistent with the findings which demonstrated that ChatGPT performed better on undergraduate-level questions [48]. Referring to Lawson's method [49], where incorrect answers to both questions in a pair indicate the lowest level of scientific reasoning, while correctly answering both indicates the highest level, it can be concluded that ChatGPT-4o exhibits a higher level of scientific reasoning compared to the students.

ChatGPT-4o's ability to compromise with other ideas after engaged in lengthy dialogues illustrates a human-like AI characteristic, which can be described as someone who is submissive yet stubborn about the core of the dialogue, continuously guessing without reflection [6]. The findings of this study confirmed the results of research [7], which showed that GPT-4 can tackle complex new tasks with performance that closely resembles human capability, even though its performance is far from expert level. Additionally, GPT-4 may fail in certain basic mathematical tasks and general reasoning if not supported by external plugins [50, 51].

The findings of this research reinforce the results of previous research [52], which indicated that ChatGPT's reasoning contains incorrect and often contradictory physics reasoning. In terms of consistency, dos Santos [53] found that ChatGPT-4 did not exhibit inconsistencies in its arguments.

Fig. 3 shows that, in general, with 61 prompts for 25 paired items, ChatGPT-4o's answers varied significantly; however, there were some instances with no variation, resulting in either 100% correct or 100% incorrect answers. These findings confirm research, which found that ChatGPT-3.5 often provided highly varied answers in successive attempts, whereas ChatGPT-4 exhibited almost no variation (median of zero) [51]. This illustrates that the stability of ChatGPT-4o depends on the context. This stability arises from ChatGPT-4o's ability to synthesize all the available information on the internet. As an LLM, this capability can evolve as the availability of data or information on the internet increases.

The differences in individual item scores between Tier-1 and Tier-2 further emphasize this AI characteristic, which heavily relies on the availability of digitally-based information. The lower scores in Tier-1 compared to Tier-2 also indicate that ChatGPT-4o still requires extensive training or a database that contains specific cases or phenomena, such as suspending objects in the context of static fluid.

ChatGPT-4o's difficulties with items containing options with combinations of several images seem to indicate that such combinations increase the complexity of the questions and require deeper, more structured reasoning. This poses a challenge for ChatGPT-4o, which tends to be less accurate in processing complex visual information. The inclusion of multiple objects in a single option appears to increase the cognitive load, necessitating a more complex reasoning process. This aligns with research findings which also observed that ChatGPT's performance declines as question difficulty increases [54, 55]. The results suggest that in some cases, ChatGPT struggles to provide targeted, accurate answers. This limitation indicates that while ChatGPT-4o is capable of handling text-based questions relatively well, the processing of visual information remains an aspect that requires improvement in this AI system [56].

This research found that ChatGPT-4o was consistent in its reasoning process, even when that reasoning was incorrect. A similar pattern was observed which noted that once ChatGPT makes a mistake in solving a problem, subsequent attempts are likely to fail as well [6]. From a cognitive psychology perspective, ChatGPT lacks metacognition; it does not possess the cognitive ability to plan problem-solving strategies. Therefore, it does not reflect on how it thinks. ChatGPT remains consistent in its reasoning path and lacks self-evaluation mechanisms. Although there is an anthropomorphic assumption that ChatGPT can compromise with other ideas, concede in debates, acknowledge mistakes, and give praise, this aspect of metacognition appears to be absent in its functioning.

V. Conclusion

Based on the above findings, it can be concluded that, in general, ChatGPT-4o's performance in answering questions (Tier-1) was lower than that of students. However, it excelled in providing justifications or reasoning (Tier-2). For paired items, ChatGPT-4o also outperformed the students. Overall, the reasoning performance of ChatGPT-4o, as well as that of junior high school, senior high school, and university students, was categorized at Level III (low).

The outcome space derived from the phenomenographic analysis indicates that ChatGPT-4o's reasoning behaviour encompasses the following categories: formula-based reasoning; consistency in reasoning flow; the ability to reconcile with alternative ideas; context-dependent reasoning capabilities and difficulties; and a tendency to provide biased or contradictory reasoning or explanations.

Thus, ChatGPT-4o still requires further training or database enrichment for cases related to static fluid that are available on the internet. This research highlights the importance of understanding the capabilities, strengths, limitations, and reasoning behaviour of ChatGPT-4o, enabling its use to be more effectively aligned with specific needs and objectives.

Conflict of Interest

The authors declare no conflict of interest.

Author Contribution

I.K.W. and J.M. developed the research design, conducted the literature review, structured the research design and stages, and interpreted the data. S.N.K. created the prompts from 2TtFS for ChatGPT, input the prompt to ChatGPT, analyzed the results, and interpreted the findings. S.N.K. and R.T. collaborated on organizing the online version of 2TtFS and performed the tabulation and processing of the data. J.M. provided critical reviews during the manuscript writing process. All authors checked the writing and approved the final version of the manuscript for submission.

Funding

This study was funded by Directorate General of Higher Education, Research, and Technology, Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia, Contract Number: 095/E5/PG.02.00.PL/2024.

Acknowledgements

We thank all the students who took their valuable time to participate in this study. In addition, we would like to express our sincere appreciation to all the reviewers for their comments and feedback to improve this paper.

References

[1] OpenAI. (2023) *Corporate Website.* [Online]. Available: www.openai.com.

[2] S. Hajkowicz, C. Naughtin, C. Sanderson, E. Schleiger, S. Karimi, A. Bratanova, and T. Bednarz. (2022). Artificial intelligence for science. adoption trends and future development pathways. *Australia's National Science Agency*. *CSIRO.* [Online]. Available: https://www.csiro.au/en/research/technology-space/ai/artificial-intelligence-for-science-report

[3] Y. K. Dwivedi, L. Hughes, E. Ismagilova, G. Aarts, C. Coombs, T. Crick, and M. D. Williams, "Artificial Intelligence (AI): multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy," *International Journal of Information Management*, vol. 57, p. 101994, 2021.

[4] S. Hoffmann, W. Lasarov, and Y. K. Dwivedi, "AI-empowered scale development: Testing the potential of ChatGPT," *Technological Forecasting and Social Change*, vol. 205, p. 123488, 2024.

[5] S. Takagi, T. Watari, A. Erabi, and K. Sakaguchi, "Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study," *JMIR Med. Edu.*, vol. 9, e48002, 2023.

[6] G. Kortemeyer, "Could an artificial-intelligence agent pass an introductory physics course?" *Physical Review Physics Education Research*, 2023, vol. 19, p. 010132, 2023. https://doi.org/10.1103/PhysRevPhysEducRes.19.010132

[7] G. Polverini and B. Gregorcic, "How understanding large language models can inform the use of ChatGPT in physics education," *European Journal of Physics*, vol. 45, 025701, 2024, https://doi.org/10.1088/1361-6404/ad1420

[8] Y. Liang, D. Zou, H. Xie, and F. L. Wang, "Exploring the potential of using ChatGPT in physics education," *Smart Learning Environments*, vol. 10, 52, 2023. https://doi.org/ 10.1186/s40561-023-00273-7

[9] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. D. León, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, and V. Tseng, "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models," *PLOS Digital Health*, vol. 2, no.2, e0000198, 2023.

[10] G. Cruz, S. Pedroza, and F. Ariza, "ChatGPT's learning and reasoning capacity in anesthesiology," *Colombian Journal of Anesthesiology*, vol. 52, e1092, 2024.

[11] M. Gams and S. Kramar, "Evaluating ChatGPT's consciousness and its capability to pass the turing test: A comprehensive analysis," *Journal of Computer and Communications*, vol. 12, pp. 219–237, 2024. https://doi.org/ 10.4236/jcc.2024.123014

[12] S. Frieder, L. Pinchetti, A. Chevalier, R. Griffiths, T. Salvatori, T. Lukasiewicz, P. C. Petersen, and J. Berner (2023). Mathematical capabilities of ChatGPT. [Online]. Available: https://arxiv.org/abs/2301.13867

[13] P. M. Newton and M. Xiromeriti. (2023). ChatGPT performance on MCQ exams in higher education. *A Pragmatic Scoping Review*. [Online]. Available: https://doi.org/10.35542/osf.io/sytu3

[14] F. Kieser, P. Wulff, J. Kuhn, and S. Küchemann, "Educational data augmentation in physics education research using ChatGPT," *Physical Review Physics Education Research*, vol. 19, 020150, 2023. https://doi.org/ 10.1103/PhysRevPhysEducRes.19.020150

[15] S. Al Ghazali, N. Zaki, L. Ali, and S. Harous, "Exploring the potential of ChatGPT as a substitute teacher: A case study," *International Journal of Information and Education Technology*, vol. 14, no. 2, pp. 271–278, 2024. https://doi.org/ 10.18178/ijiet.2024.14.2.2048

[16] P. Bitzenbauer, "ChatGPT in physics education: A pilot study on easy-to-implement activities," *Contemporary Educational Technology*, vol. 15, no. 3, pp. ep430, 2023. https://doi.org/ 10.30935/cedtech/13176

[17] C. Howe, A. Tolmie, and C. Rodgers, "Physics in the primary school: Peer interaction and the understanding of floating and sinking," *European Journal of Psychology of Education*, vol. 5, no. 4, pp. 459–475, 1990. https://doi.org/10.1007/BF03173132

[18] S. Havu-Nuutinen, "Examining young children's conceptual change processing floating and sinking from a social constructivist perspective," *International Journal of Science Education*, vol. 27, no. 3, pp. 259–279, 2005. https://doi.org/ 10.1080/0950069042000243736

[19] Y. Yin, M. Tomita, and R. Shavelson, "Diagnosing and dealing with student misconceptions: floating and sinking," *Science Scope*, vol. 31, no. 8, pp. 34–39, 2008.

[20] S. Chien, C. Hsiung, and S. Chen, "The development of young children's science-related concept regarding floating and sinking," *Asia-Pacific Journal of Research in Early Childhood Education*, vol. 3, no. 2, pp. 73–88, 2009.

[21] S. Çepni, Ç. Şahin, and H. Ipek, "Teaching floating and sinking concepts with different methods and techniques based on the 5E instructional model," *Asia-Pacific Forum on Science Learning and Teaching*, vol. 11, no. 2, pp. 1 16, 2010.

[22] J. Minogue, D. Borland, M. Russo, S. T. Chen, and R. Grady, "Investigating the influence of haptic technology on upper elementary students' reasoning about sinking & floating," in *Proc. 2015 Annual International Conference*, pp. 1–10, 2015.

[23] T. W. Teo, Y. K. Yan, and W. L. M. Ong, "An investigation of Singapore pre-school children's emerging concepts of floating and sinking," *Pedagogies*, vol. 12, no. 3, pp. 1–18, 2017. https://doi.org/ 10.1080/1554480X.2017.1374186

[24] R. D. Castillo, T. Waltzer, and H. Kloos, "Hands-on experience can lead to systematic mistakes: A study on adults' understanding of sinking objects," *Cognitive Research: Principles and Implications*, vol. 2, no. 1, 2017. https://doi.org/ 10.1186/s41235-017-0061-8

[25] V. Viyanti, C. Cari, W. Sunarno, and Z. K. Prastyo, "The development rubrics skill argued as alternative assessment floating and sinking materials," *Journal of Physics: Conference Series*, vol. 909, no. 1, 2017. https://doi.org/ 10.1088/1742-6596/909/1/012057

[26] J. Shen, O. L. Liu, and H. Y. Chang, "Assessing students' deep conceptual understanding in physical sciences: An example on sinking and floating," *International Journal of Science and Mathematics Education*, vol. 15, no. 1, pp. 57–70, 2017. https://doi.org/ 10.1007/s10763-015-9680-z

[27] C. R. Gette, M. Kryjevskaia, M. R. Stetzer, and P. R. L. Heron, "Probing student reasoning approaches through the lens of dual-process theories: A case study in buoyancy," *Physical Review Physics Education Research*, vol. 14, no. 1, 2018. https://doi.org/ 10.1103/PhysRevPhysEducRes.14.010113

[28] A. Nongkhunsarn, C. Yuenyong, J. Tupsai, and T. Sranamkam, "Grade 11 student's mental model of fluid and analytical thinking in science teaching through science technology and society (STS) approach," *Journal of Physics: Conference Series*, vol. 1340, no. 1, 2019. https://doi.org/ 10.1088/1742-6596/1340/1/012043

[29] F. Kafiyani, A. Samsudin, and D. Saepuzaman, "Development of four-tier diagnostic test (FTDT) to identify student's mental models on static fluid," *Journal of Physics: Conference Series*, vol. 1280, no. 5, 2019. https://doi.org/ 10.1088/1742-6596/1280/5/052030

[30] T. Djudin, "Promoting students' conceptual change by integrating the 3-2-1 reading technique with refutation text in the physics learning of buoyancy," *Journal of Turkish Science Education*, vol. 18, no. 2, pp. 290–303, 2021. https://doi.org/ 10.36681/tused.2021.66

[31] S. N. Kaharu and J. Mansyur, "The development of a test to explore the students' mental models and external representation patterns of hanging objects," *Pegem Journal of Education and Instruction*, vol. 11, no. 4, pp. 110–125, 2021. https://doi.org/ 10.14527/pegegog.11.4.011

[32] J. Mansyur, I. K. Werdhiana, D. Darsikin, and S. N. Kaharu, "Students' external representation patterns of suspending objects in static fluid," *European Journal of Educational Research*, vol. 11, no. 2, pp. 805–820, 2022. https://doi.org/ 10.12973/eu-jer.11.2.805

[33] J. Mansyur, I. K. Werdhiana, D. Darsikin, S. N. Kaharu, and N Tadeko, "Students' mental models about the suspending objects in static fluid," *Journal of Turkish Science Education*, vol. 19, no. 1, pp. 257-287, 2022.

[34] S. N. Kaharu, Y. Gagaramusu, A. Azizah, N. Kamisani, N. Tadeko, and J. Mansyur, "Development of an online two-tier test to explore students' conceptions on objects in static fluid," *Pegem Journal of Education and Instruction*, vol. 14, no. 3, pp. 361–373, 2024. https://doi.org/ 10.47750/pegegog.14.03.34

[35] F. Marton and W. Y. Pong, "On the unit of description in phenomenography," *Higher Education, Research and Development*, vol. 24, no. 4, pp. 335–348, 2005. https://doi.org/ 10.1080/07294360500284706.

[36] L. N. Walsh, R. G. Howard, and B. Bowe, "Phenomenographic study of students' problem solving approaches in physics," *Physical Review Physics Education Research*, vol. 3, no. 2, pp. 1–12, 2007.

[37] K. F. Thomas, C. H. Meng, I. King, and Y. Yam, "A phenomenographic approach on teacher conceptions of teaching Artificial Intelligence (AI) in K-12 schools," *Education and Information Technologies*, vol. 28, pp. 1041–1064, 2023. https://doi.org/ 10.1007/s10639-022-11161-x

[38] H. P. Chang, J. Y. Chen, C. J. Guo, C. C. Chen, C. Y. Chang, and S. H. Lin, "Investigating primary and secondary students' learning of physics concepts in Taiwan," *International Journal of Science Education*, vol. 29, no. 4, pp. 465–482, 2007. https://doi.org/ 10.1080/09500690601073210

[39] H. E. Chu, D. F. Treagust, and A. L. Chandrasegaran, "A stratified study of students' understanding of basic optics concepts in different contexts using two-tier multiple-choice items," *Research in Science & Technological Education*, vol. 27, no. 3, pp. 253–265, 2009. https://doi.org/ 10.1080/02635140903162553

[40] B. K. Bayrak, "Using two-tier test to identify primary students' conceptual understanding and alternative conceptions in acid base," *Mevlana International Journal of Education*, vol. 3, no. 2, pp. 19–26, 2013.

[41] D. Hestenes and I. Halloun, "Interpreting the force concept inventory: A response to March 1995 critique by Huffman and Heller," *Physics Teacher*, vol. 33, p. 502, 1995.

[42] P. Nieminen, A. Savinainen, and J. Viiri, "Force Concept Inventory-based multiple-choice test for investigating students' representational consistency," *Physical Review Special Topics - Physics Education Research*, vol. 6, 020109, 2010. https://doi.org/ 10.1103/PhysRevSTPER.6.020109

[43] I. S. Caleon and R. Subramaniam, "Do students know what they know and what they don't know? Using a four-tier diagnostic test to assess the nature of students' alternative conceptions," *Research in Science Education*, vol. 40, no. 3, pp. 313–337, 2009. https://doi.org/10.1007/s11165-009-9122-4

[44] I. S. Caleon and R. Subramaniam, "Development and application of a three-tier diagnostic test to assess secondary students' understanding of waves," *International Journal of Science Education*, vol. 32, pp. 939–961, 2010. https://doi.org/ 10.1080/09500690902890130

[45] Y. Xiao, J. Han, K. Koenig, J. Xiong, and L. Bao, "Multilevel Rasch modeling of two-tier multiple-choice test: A case study using Lawson's classroom test of scientific reasoning," *Physical Review Physics Education Research*, vol. 14, 020104, 2018. https://doi.org/10.1103/PhysRevPhysEducRes.14.020104

[46] S S. N. Zhou, Q. Y. Liu, K. Koenig, Q. Y. Li-Yang Xiao, and L. Bao, "Analysis of two-tier question scoring methods: A case study on the Lawson's classroom test of scientific reasoning," *Journal of Baltic Science Education*, vol. 20, no. 1, pp. 2021. https://doi.org/10.33225/jbse/21.20.146

[47] D. Jin, E. Pan, N. Oufattole, W. H. Weng, H. Fang, and P. Szolovits, "What disease does this patient have? A large-scale open domain question answering dataset from medical exams," *Applied Sciences*, vol. 11, no. 14, p. 6421, 2021. https://doi.org/10.3390/app11146421

[48] L. Schulze Balhorn, J. M. Weber, S. Buijsman, J. R. Hildebrandt, M. Ziefle, and A. M., "Empirical assessment of ChatGPT's answering capabilities in natural science and engineering," *Scientific Reports*, vol. 14, 4998, 2024. https://doi.org/10.1038/s41598-024-54936-7

[49] A. E. Lawson, "Classroom test of scientific reasoning: Multiple choice version," *Journal of Research in Science Teaching*, vol. 5, no. 1, pp. 11–24, 2000.

[50] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*. [Online]. Available: https://arxiv.org/abs/2303.12712

[51] C. G. West. (2023). Advances in apparent conceptual physics reasoning in GPT-4. *arXiv preprint arXiv:2303.17012*. [Online]. Available: https://arxiv.org/abs/2303.17012

[52] B. Gregorcic and A.M. Pendrill, ChatGPT and the frustrated socrates, *Physics Education*, vol. 58, no. 3, 035021, 2023.

[53] R. P. dos Santos. (2023). Enhancing physics learning with ChatGPT, Bing Chat, and Bard as agents-to-think-with: A comparative case study. *arXiv preprint arXiv:2306.00724*. [Online]. Available: https://arxiv.org/abs/2306.00724

[54] A. Gilson, C. W. Safranek, T. Huang, V. Socrates, L. Chi, and R. Richard, "How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment," *JMIR Medical Education*, vol. 9, e45312, 2023. https://doi.org/ 10.2196/45312

[55] F. Antaki, S. Touma, D. Milad, J. El-Khoury, and R. Duval, "Evaluating the performance of ChatGPT in ophthalmology," *Ophthalmology Science*, vol. 3, no. 4, 2023. https://doi.org/ 10.1016/j.xops.2023.100324

[56] 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," *Physical Review Physics Education Research*, vol. 19, 010142, 2023. https://doi.org/ 10.1103/PhysRevPhysEducRes.19.010142