

Analysis Of The Use Of Chatgpt And Deepseek In Solving HOTS Physics Problems In Kinematics Teaching Material

Nadya Aidilla Fitri Tanjung

Universitas Negeri Medan, Indonesia

Sahyar

Universitas Negeri Medan, Indonesia

Corresponding Author: nadyatanjung@mhs.unimed.ac.id

Abstract

This study analyzes the ability of ChatGPT 5.1 and DeepSeek V3 to solve Higher Order Thinking Skills (HOTS) problems in kinematics, in terms of answer accuracy and response consistency. A descriptive qualitative design with a desk-research approach was used. The sample consisted of 20 expert-validated HOTS items at cognitive levels C4 (analyzing) and C5 (evaluating), ten text-based and ten image-based. Accuracy was scored in binary form against three criteria: correctness of the final answer, conceptual understanding, and correct use of equations. Consistency was assessed by posing every item 30 times, each in a new conversation. On text-based items, both chatbots reached 80% accuracy (Good), with average consistency of 89% for ChatGPT and 91% for DeepSeek (both Very Good). On image-based items, ChatGPT achieved 100% accuracy and 100% consistency (Very Good), whereas DeepSeek achieved 60% accuracy and 73.7% consistency (Fairly Good). C5 items were the hardest for both chatbots, and DeepSeek's errors on image-based items came mainly from misreading graphs rather than from wrong formulas. Chatbot output should therefore be verified by teachers and students before it is used as a learning aid in physics.

Keywords: ChatGPT; DeepSeek; HOTS; Kinematics; Accuracy; Consistency.

INTRODUCTION

Higher Order Thinking Skills (HOTS) are among the competencies emphasized in the Merdeka Curriculum and in 21st-century education policy. They cover analyzing (C4), evaluating

3306

IJEVSS, Volume 05 Issue 04, 2026



Copyright at authors some right reserved this work is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/)

(C5), and creating (C6) in the revised Bloom's Taxonomy (Anderson & Krathwohl, 2001). In physics learning, these skills help students understand abstract concepts and apply them to real situations.

Studies nevertheless show that students ability to answer HOTS questions is still low. Magfira et al. (2021) reported that students tend to memorize rather than understand, and that misconceptions appear when they analyze images, distance and displacement in word problems, and graphs of linear motion. Andriyeni and Zakir (2023) likewise found that many students failed to connect mathematical representations with descriptions of motion, even when they were given similar example problems. School evaluations point the same way: in the midterm and final examinations of Physics at SMAN 1 Medan, the average score on items classified as HOTS was below 60.

Another problem is the shortage of practice material. Most high school physics textbooks emphasize questions in the lower cognitive domains (C1–C3), and varied questions that require analysis and evaluation are scarce (Giancoli, 2014). Students and teachers who solve HOTS questions by hand often work without a systematic reasoning process, and many students apply formulas mechanically without understanding the physical meaning of each quantity (Docktor & Mestre, 2014).

At the same time, generative artificial intelligence (GenAI) has developed rapidly. Chatbots built on large language models (LLMs), such as ChatGPT and DeepSeek, interpret natural-language input and generate responses that resemble human reasoning, including on concept-based problems (Brown et al., 2020). Earlier studies show both promise and limits. Mustofa et al. (2024) found that ChatGPT gives quick feedback and logical explanations in physics learning but still makes conceptual errors in applying physical laws. Nurhuda et al. (2023) reported that ChatGPT answered only about 60% of illustrated physics problems correctly, with errors concentrated in interpreting the images. DeepSeek, a newer model, is claimed to have stronger scientific and mathematical reasoning: DeepSeek-R1 scored 97.3% on the MATH-500 benchmark and 79.8% on AIME 2024 (DeepSeek-AI, 2025). Even so, little research has tested its performance on high school physics HOTS problems.

This study therefore analyzes how well ChatGPT 5.1 and DeepSeek V3 solve HOTS (C4–C5) problems in kinematics, using answers compiled from standard high school physics textbooks and validated by experts as the reference. Two research questions guide the work: (1) how well do the two chatbots solve text based HOTS kinematics questions in terms of answer accuracy and



response consistency, and (2) how well do they solve HOTS kinematics questions that involve images and graphs on the same two measures? The study does not involve students or classroom intervention. It is expected to support research on AI-based physics education and to help teachers and students use chatbots as a critical and responsible learning aid.

LITERATURE REVIEW

Generative AI and large language models

Russell and Norvig (2020) define artificial intelligence (AI) as a field of computer science that builds systems able to perform tasks that normally require human intelligence, such as pattern recognition, learning, reasoning, and problem solving. GenAI is a type of AI trained on data to create new content, such as text, images, audio, and video. Its best-known form in education is the chatbot, which interprets a user input and produces a contextually relevant response. LLMs are trained on very large text corpora and use the transformer architecture to process long contexts and relate words to each other (Brown et al., 2020). Wei et al. (2022) found that reasoning ability, including logical reasoning, mathematics, and complex problem solving, emerges as model and data size grow, which is relevant to physics at the high school level. In physics education, AI is expected to help students understand abstract concepts through visualizations, solve problems with systematic steps, receive instant feedback, and study independently (Hwang et al., 2020).

ChatGPT and DeepSeek

ChatGPT is an LLM developed by OpenAI, launched in November 2022 and built on the Generative Pre-trained Transformer (GPT) architecture (Brown et al., 2020). Successive versions improved reasoning and added multimodal input, so that text, graphs, and other visual representations can be processed. GPT-5.1, introduced in 2025, is described as having stronger reasoning, better long-context stability, and clearer explanations of solution steps in science problems, which is why it was chosen for this study. DeepSeek is an LLM developed by DeepSeek AI, a China-based company, with an emphasis on mathematical and scientific reasoning and low operating cost (DeepSeek AI, 2024). DeepSeek-V3 is its general-purpose model, based on a Mixture of Experts architecture with a context length of up to 128K tokens; DeepSeek-R1 is a separate model specialized for step-by-step reasoning, and DeepSeek-VL2 is a separate model for images and text.



Both chatbots depend on how the problem is posed. Effective prompts state all known information, request ordered solution steps, name the educational level and topic, and ask for justification of each step rather than only the final answer (Mollick & Mollick, 2023). Because chatbots can produce answers that sound reasonable but contain conceptual errors, output has to be verified by a teacher or student (Mustofa et al., 2024).

Answer accuracy and response consistency

Answer accuracy is the degree to which an answer is scientifically and mathematically correct relative to the answer key. Following Oblitas et al. (2025), it can be judged on three criteria: the accuracy of the final answer (numerical value and unit), conceptual understanding, and the correct use of equations. Good conceptual understanding means the AI not only gives a technically correct answer but also connects concepts and explains their relationships logically (Azizah & Mansyur, 2024). Response consistency is the ability to give the same, stable answer to the same question. Because chatbot output is probabilistic, consistency is tested by repeating the same prompt, as Funk et al. (2024) did by repeating medical examination questions. Polverini and Gregorcic (2024) recommend sending every repetition in a new conversation window instead of regenerating the answer, so that each repetition counts as a completely new task, and uploading screenshots of image-based items so that responses rest only on the image data.

HOTS and kinematics

The cognitive levels of the items follow the revised Bloom's Taxonomy (Anderson & Krathwohl, 2001), and C4–C6 are treated as HOTS (Sani, 2019). Brookhart (2010) describes HOTS questions as non-algorithmic, complex, open to more than one solution, dependent on nuanced judgment, and demanding self-regulation of thinking. This study uses only C4 and C5 because the items are multiple-choice with a single correct answer, which allows objective binary scoring; C6 (creating) requires students to produce their own design or product and is better measured with essays.

Kinematics is the branch of mechanics that describes the position, velocity, and acceleration of an object as functions of time without considering the force that causes the motion (Halliday et al., 2013). Mastery of linear motion, motion graphs, and the relations among these quantities is a common source of difficulty for students (Giancoli, 2014). The material in this study covers uniform straight motion (GLB), $x = x_0 + vt$, and uniformly accelerated straight



motion (GLBB), $x = x_0 + v_0t + \frac{1}{2}at^2$, $v = v_0 + at$, and $v^2 = v_0^2 + 2a\Delta x$, together with the interpretation of position time and velocity time graphs (Kemendikbudristek, 2022).

Related research

Andriyeni and Zakir (2023) found that students' ability to understand physics with AI tools was in the high category. Nurhuda et al. (2023) tested ChatGPT 3.5 on ten resistor circuit problems and obtained 60% accuracy when the image was described correctly in the prompt. Mustofa et al. (2024) concluded that ChatGPT explains concepts and gives systematic solution steps but sometimes lacks detail and needs human verification. Polverini and Gregorcic (2024) tested eight vision capable chatbots on a kinematics graph test and found that OpenAI chatbots performed best and that language based tasks were easier than tasks requiring visual interpretation. Oblitas et al. (2025) likewise found lower accuracy on image based than on text based questions. None of these studies compared ChatGPT 5.1 and DeepSeek V3 on HOTS kinematics items with both text and image formats, which is the gap addressed here.

RESEARCH METHOD

Research design

This was a descriptive qualitative study using literature review and observation, carried out as desk research because all data were collected through digital devices; it was planned for April to June 2026. Qualitative research yields descriptive data in the form of words from the observed object (Moleong, 2017), and descriptive research describes a phenomenon systematically, factually, and accurately without manipulating it (Sukmadinata, 2011). The literature review (Zed, 2008) was used to collect the HOTS kinematics questions from a HOTS physics book, and systematic non-participant observation (Sugiyono, 2017) was used to document the chatbots' solutions.

Sample and instrument

The population was all HOTS kinematics questions for senior high school in textbooks and standard question banks at levels C4 and C5. Purposive sampling produced 20 multiple choice items with five options: ten text-based items (Q1–Q10) and ten image-based items (Q11–Q20), the latter including four graph-based items (Q13, Q14, Q16, Q17). Twelve items were at level C4



and eight at level C5. The items covered GLB, GLBB, and the interpretation of position time and velocity time graphs. The researcher solved every item by hand to form the answer key.

Before use, two expert validators judged the instrument on 14 indicators in three aspects (material, construct, and language), scoring each indicator 1 (agree) or 0 (disagree). Validity was calculated as the percentage of the maximum score and classified using the criteria of Arikunto (2018): 76–100% is highly valid. Agreement between the two validators was examined with Cohen's Kappa (Sabharwal, 2021).

Data collection

The free versions of ChatGPT 5.1 and DeepSeek V3 were used. Text-based questions were entered manually and verbatim, and image based questions were uploaded as screenshots so that responses depended only on the image. Each item was posed 30 times to each chatbot, every time in a new conversation rather than by regenerating the answer, following Funk et al. (2024) and Polverini and Gregorcic (2024). Usage limits of the free versions (usually reset every three or four hours) did not affect the data. Responses were stored as screenshots or text copies, recorded on observation sheets, and supplemented with field notes.

Assessment and data analysis

Accuracy was assessed with the rubric in Table 1, modified from Oblitas et al. (2025), and scored in binary form: 1 if all evaluation points were met and 0 if any was not met. Consistency was assessed by scoring each of the 30 responses to an item as 1 if its final answer matched the answer key and 0 otherwise. The percentages were computed as follows:

Table 1. Rubric for evaluating the accuracy of answers

Criterion	Evaluation points
Final answer accuracy	The chosen option and the final numerical value, with a unit appropriate to the physical context, match the answer key; the answer is not based on guessing or on picking the closest value.
Conceptual understanding	The answer identifies the correct area of physics or concept; the explanation states or implies a correct physical principle; the explanation is consistent with the interpretation of the graph, image, or diagram.

3311



Criterion	Evaluation points
Accuracy of equation usage	The equations are consistent with the physics concepts and the answer key; there are no calculation, unit-conversion, or variable-substitution errors.

Source: modified from Oblitas et al. (2025).

$$\text{Accuracy (\%)} = (\Sigma x / n) \times 100\% \quad (1)$$

$$\text{Consistency (\%)} = (\Sigma x / N) \times 100\% \quad (2)$$

$$\text{Mean per chatbot} = \Sigma (\text{Consistency or Accuracy \%}) / \text{number of items} \quad (3)$$

In Equation (1), Σx is the number of answers scored 1 and n the number of questions; in Equation (2), Σx is the number of consistent answers and N the number of trials (30). Results were classified as Very Poor (0–20%), Poor (21–40%), Fairly Good (41–60%), Good (61–80%), or Very Good (81–100%) (Azizah & Mansyur, 2024). The analysis followed data reduction, data display, and conclusion drawing (Miles & Huberman, 1992): responses were scored against the rubric, presented in tables and narrative, and then compared to identify the strengths and weaknesses of each chatbot.

RESULTS

Instrument validity

The 20 item instrument reached a validity of 100% in each of the three aspects (material, construct, and language), an overall average of 100% in the Highly Valid category. The two validators gave identical scores on all 14 indicators for all items, so the proportion of agreement was $P = 1.00$ and Cohen's Kappa is mathematically undefined (0/0); this perfect agreement is interpreted as the Near Perfect category. One validator suggested aligning the item indicators with the revised Bloom's Taxonomy, and the suggestion was applied before data collection. The items were revised three times before both validators judged them valid, so the instrument was considered suitable for use.

Text based HOTS questions

Table 2 gives the accuracy and consistency results for all 20 items. On the ten text based items, ChatGPT 5.1 and DeepSeek V3 each scored 8 of 10, an accuracy of 80% (Good). ChatGPT did not meet the rubric on Q1 and Q5. On Q1 its final answer was correct, but it did not state the general equation or verify the result, so criterion 3 was not met. On Q5 it said that none of the



options was correct. DeepSeek did not meet the rubric on Q5, where it chose D instead of E, and on Q6, where it substituted values without naming the formula or principle. Q5, a C5 item, was the only item that neither chatbot solved. The two chatbots differed on Q1 (ChatGPT 0, DeepSeek 1) and Q6 (ChatGPT 1, DeepSeek 0).

For consistency, ChatGPT averaged 89% and DeepSeek 91%, both Very Good. DeepSeek was fully consistent (100%) on nine of ten items and reached only 10% on Q5 (correct in 3 of 30 trials). ChatGPT was fully consistent on seven items, scored 96,7% on Q9 and 93,3% on Q10, and scored 0% on Q5, where it never gave a valid answer option in 30 trials. Field notes recorded a “Too many requests” message for ChatGPT on Q2 (trial 17, 8 May 2026) and on Q5 (trial 24, 15 May 2026), which were repeated later; an error on Q9 (trial 23, 17 May 2026) that produced an invalid response scored 0; and one DeepSeek answer in English to Q1 (trial 4, 19 May 2026) although the question was in Indonesian, which could still be evaluated.

Table 2. Accuracy (score 0/1) and consistency (% of 30 trials) of ChatGPT 5.1 and DeepSeek V3 on 20 HOTS kinematics items

Item	Level	Key	ChatGPT accuracy	ChatGPT consistency (%)	DeepSeek accuracy	DeepSeek consistency (%)
1	C4	B	0	100	1	100
2	C4	B	1	100	1	100
3	C4	B	1	100	1	100
4	C5	C	1	100	1	100
5	C5	E	0	0	0	10
6	C4	C	1	100	0	100
7	C5	E	1	100	1	100
8	C4	D	1	100	1	100
9	C4	D	1	96.7	1	100
10	C5	E	1	93.3	1	100
Text			80%	89%	80%	91%
11	C4	D	1	100	1	100
12	C5	E	1	100	0	93.3
13	C4	D	1	100	0	43.3
14	C4	B	1	100	0	56.7



Item	Level	Key	ChatGPT accuracy	ChatGPT consistency (%)	DeepSeek accuracy	DeepSeek consistency (%)
15	C4	C	1	100	1	100
16	C5	E	1	100	0	6.7
17	C5	B	1	100	1	36.7
18	C4	C	1	100	1	100
19	C4	C	1	100	1	100
20	C5	C	1	100	1	100
Image			100%	100%	60%	73.7%

Note: Items 1–10 are text-based and items 11–20 are image-based. Bold rows give accuracy (%) and mean consistency (%) per question type.

Image based HOTS questions

On the ten image based items, ChatGPT 5.1 answered all correctly (100%, Very Good), whereas DeepSeek V3 answered six correctly (60%, Fairly Good). DeepSeek missed four items. On Q12 and Q13 it chose the correct option (E and D) but did not use the equations correctly, so criterion 3 was not met. On Q14 it chose A instead of B, and on Q16 it gave no valid answer option. On the four graph-based items (Q13, Q14, Q16, and Q17), DeepSeek was correct only on Q17 (25%), while ChatGPT was correct on all four.

ChatGPT's consistency was 100% on every image based item, a mean of 100% (Very Good). DeepSeek averaged 73,7% (Fairly Good). It was fully consistent on five items (Q11, Q15, Q18, Q19, and Q20), and scored 93,3% on Q12, 43,3% on Q13 (no valid option in 17 trials), 56,7% on Q14 (answers varied among A, B, and E, with some trials giving no answer), 6,7% on Q16, and 36,7% on Q17. In the 30th trial of Q16 (24 May 2026), DeepSeek listed two options, B and C, without selecting one, and the response was scored 0.

Summary and performance by cognitive level

Table 3. Summary of performance by question type

Question type	Metric	ChatGPT 5.1	DeepSeek V3
Text (Q1–Q10)	Accuracy	80% (Good)	80% (Good)
	Consistency	89% (Very Good)	91% (Very Good)
Image (Q11–Q20)	Accuracy	100% (Very Good)	60% (Fairly Good)

3314



Question type	Metric	ChatGPT 5.1	DeepSeek V3
	Consistency	100% (Very Good)	73.7% (Fairly Good)

Table 4 groups the results of all 20 items by cognitive level and question indicator. For C4 (12 items), ChatGPT reached 91,7% accuracy and DeepSeek 75%; for C5 (8 items), ChatGPT reached 87,5% and DeepSeek 62,5%. The lowest values for both chatbots were on the “Selecting” indicator (Q5), and DeepSeek was also weak on “Concluding” (Q13 and Q15).

Table 4. Performance by HOTS level and indicator (%)

Level / indicator	Items	ChatGPT accuracy	ChatGPT consistency	DeepSeek accuracy	DeepSeek consistency
C4 (Analyzing), overall	12 items	91.7	99.7	75.0	91.7
Analyzing	Q1, 3, 6, 8, 11, 14, 19	85.7	100	71.4	93.8
Measuring	Q2	100	100	100	100
Relating	Q9, 18	100	98.3	100	100
Concluding	Q13, 15	100	100	50.0	71.7
C5 (Evaluating), overall	8 items	87.5	86.7	62.5	68.3
Assessing	Q4, 10, 16	100	97.8	66.7	68.9
Comparing	Q7, 12, 17, 20	100	100	75.0	82.5
Selecting	Q5	0	0	0	10.0

DISCUSSION

Text based questions

Both chatbots reached the same accuracy (80%, Good) on text based items. This agrees with Oblitas et al. (2025), who found that chatbots perform well on text-based tasks, which LLMs are optimized to process. Their weaknesses, however, differed in kind: ChatGPT lost points on the precise use of equations (Q1), and DeepSeek on conceptual grounding (Q6). Q1 shows how the rubric separates these cases. The question asks for the height at which two balls meet when the first is thrown upward at 10m/s and the second, one second later, at 25m/s ($g = 10\text{m/s}^2$); the

3315



correct answer is B (4.8m). Both chatbots chose B, took the first throw as the time reference, equated the two heights, and obtained $t = 1.2\text{s}$ and $h = 10(1.2) - 5(1.2)^2 = 4.8\text{m}$. ChatGPT wrote the equations only in substituted form, $h_1 = 10t - 5t^2$ and $h_2 = 25(t - 1) - 5(t - 1)^2$, without the general equation, the motion time of the second ball at the meeting point (0.2s), or a check of the height with the second balls equation. DeepSeek organized its answer into six labeled steps, wrote the general form first, stated that the second ball had moved for 0.2 s, and verified $h = 25(0.2) - 5(0.2)^2 = 4.8\text{m}$. The same final answer therefore scored 0 for ChatGPT and 1 for DeepSeek, which is why accuracy was judged on three criteria and not on the final answer alone. This is consistent with Azizah and Mansyur (2024), who argue that a strong conceptual understanding includes explaining the relations among concepts logically.

The consistency results (89% and 91%) were both Very Good, in line with Funk et al. (2024), who treat stable answers to repeated input as an indicator of model quality. DeepSeek was slightly more stable, but the 2 point gap is small, and the main difference is on Q5, where neither chatbot produced the correct answer in most trials. The lapses in ChatGPT's consistency on Q9 and Q10, the error and "Too many requests" messages, and DeepSeek's one switch to English reflect the non deterministic nature of LLMs, whose outputs vary even for identical input (Polverini & Gregorcic, 2024).

Image based questions

The largest difference in this study appears on image-based items: ChatGPT achieved 100% accuracy and consistency, while DeepSeek achieved 60% and 73,7%. The 26,3 point gap in consistency is much larger than the 2 point gap on text based items. This supports Oblitas et al. (2025) and Polverini and Gregorcic (2024), who found that visual representations are harder for chatbots than text. ChatGPT, however, showed no decline from text to images, unlike the results of Oblitas et al. (2025) and Nurhuda et al. (2023). A likely explanation is that ChatGPT 5.1 is a newer multimodal model and that images were uploaded as screenshots instead of being described in the prompt, as they were in Nurhuda et al. (2023); this study did not test that explanation directly. For Q12 and Q13, DeepSeek chose the correct option without demonstrating the correct use of equations, which may indicate that it matched answer patterns without fully processing the visual information.

Q16 shows the source of the difference most clearly. It is a C5 item based on a velocity time graph of two objects over 4s: object A moves from 2m/s to 6m/s and object B from 1m/s to



8m/s, and students choose the correct statement about their accelerations and distances. The correct answer is E (object B travels 18m). ChatGPT read the four velocity values correctly and matched each to the right object, computed $a_A = 1\text{m/s}^2$ and $a_B = 1.75\text{m/s}^2$, found the distances from the area under the graph $((2 + 6)/2 \times 4 = 16\text{m}$ for A and $(1 + 8)/2 \times 4 = 18\text{m}$ for B), checked all five options, and chose E. DeepSeek assumed that both objects started from rest and swapped the final velocities, obtaining $a_A = 2\text{m/s}^2$ and $a_B = 1.5\text{m/s}^2$, concluded that no option was correct, and suspected an error in the question without doubting its own reading of the graph. The formulas it used ($a = \Delta v/\Delta t$ and $s = v_0t + \frac{1}{2}at^2$) were correct, but they were applied to wrong data. The difference between the two chatbots thus lies in reading the graph, not in the physics formulas, in line with Polverini and Gregoric (2024), who locate the difficulty in visual interpretation rather than in text-based problem solving.

The instability of DeepSeek (6,7% on Q16, 36,7% on Q17, 43,3% on Q13, 56,7% on Q14, and a response listing two options in one trial) means that the same graph can lead to different answers from one conversation to the next. Because physics problems rely heavily on images and graphs, a chatbot that misreads them can give students confident but wrong help, and its output has to be checked against the figure.

Cognitive level and indicator

C5 items were harder than C4 items for both chatbots. According to Anderson and Krathwohl (2001), C4 requires analyzing information and identifying relationships among parts, while C5 requires judgments based on criteria. In this instrument, C5 items required the chatbot to evaluate several statements or options against one criterion, which adds steps and points of error. DeepSeek's weakness on C5 is not due only to graph-based items: excluding Q13, Q14, Q16, and Q17, its accuracy was still lower on C5 (66,7%) than on C4 (90%). For ChatGPT, the gap between levels comes almost entirely from Q5. "Measuring" and "Relating" were the strongest indicators for both chatbots because they rest on direct calculation with kinematic equations, whereas "Selecting" and, for DeepSeek, "Concluding" involve choosing among statements and interpreting graphs.

Several limitations should be noted. The study used 20 items from a single topic, the free versions of the chatbots at one point in time, and binary scoring in which one unmet evaluation point yields a score of 0. The results describe these two models on this instrument and should not be generalized to other versions, topics, or prompting techniques.



CONCLUSION

On text-based HOTS kinematics questions, ChatGPT 5.1 and DeepSeek V3 achieved the same accuracy of 80% (Good), with average consistency of 89% and 91%, respectively (both Very Good). Their weaknesses differed: ChatGPT was less precise in applying equations, and DeepSeek was less secure conceptually. Questions at level C5 were the main weakness of both chatbots. On image-based questions, ChatGPT 5.1 achieved 100% accuracy and 100% consistency (Very Good), whereas DeepSeek V3 achieved 60% accuracy and 73,7% consistency (Fairly Good), mainly because it misread graphs. Overall, ChatGPT 5.1 outperformed DeepSeek V3 on kinematics HOTS questions that involve visual representations, and the two were on par on text-based questions.

SUGGESTIONS

1. Students and teachers who use chatbots as physics learning aids should verify the problem-solving process and the concepts used, and should not rely only on the final answer.
2. Future researchers should explore prompting techniques for image-based HOTS questions, especially for DeepSeek, such as chain of thought and self verification prompts, so that errors in image interpretation can be separated from computational errors.
3. AI developers, particularly those working on DeepSeek V3, should improve visual interpretation and train on multimodal physics data to raise the accuracy and consistency of answers to problems with images.

REFERENCES

- Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.
- Andriyeni, & Zakir. (2023). Analisis penggunaan artificial intelligence dalam pemahaman pembelajaran fisika di SMAN 1 Ampek Angkek. *Jurnal Ilmiah Research Student*, 2(1).
- Arikunto, S. (2018). *Dasar-dasar evaluasi pendidikan*. Bumi Aksara.
- Azizah, M. A., & Mansyur, J. (2024). Perbandingan kapabilitas respons ChatGPT dan Gemini terhadap pertanyaan konseptual tentang optik. *JPFT (Jurnal Pendidikan Fisika Tadulako Online)*, 12(1), 15–25.
- Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. ASCD.



- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- DeepSeek AI. (2024). *DeepSeek technical documentation*. <https://www.deepseek.com>
- DeepSeek-AI. (2025). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Docktor, J. L., & Mestre, J. P. (2014). Synthesis of discipline-based education research in physics. *Physical Review Special Topics - Physics Education Research*, 10(2), 020119.
- Funk, P. F., Hoch, C. C., Knoedler, S., Knoedler, L., Cotofana, S., Sofo, G., Dezfouli, A. B., Wollenberg, B., Guntinas-Lichius, O., & Alfertshofer, M. (2024). ChatGPT's response consistency: A study on repeated queries of medical examination questions. *European Journal of Investigation in Health, Psychology and Education*, 14(3), 657–668.
- Giancoli, D. C. (2014). *Physics: Principles with applications* (7th ed.). Pearson.
- Halliday, D., Resnick, R., & Walker, J. (2013). *Fundamentals of physics* (10th ed.). Wiley.
- Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 1, 100001.
- Kemendikbudristek [Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi]. (2022). *Buku siswa fisika untuk SMA/MA kelas X*. Pusat Kurikulum dan Perbukuan.
- Magfira, A., Kamaluddin, & Darsikin. (2021). Kesalahan siswa dalam memahami konsep kinematika gerak lurus. *Jurnal Pendidikan Fisika*.
- Miles, M. B., & Huberman, A. M. (1992). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). SAGE Publications.
- Moleong, L. J. (2017). *Metodologi penelitian kualitatif* (Edisi revisi). PT Remaja Rosdakarya.
- Mollick, E. R., & Mollick, L. (2023). Using AI to implement effective teaching strategies in classrooms: Five strategies, including prompts. *The Wharton School Research Paper*.
- Mustofa, H. A., Bilad, M. R., & Grendis, N. W. B. (2024). Utilizing AI for physics problem solving: A literature review and ChatGPT experience. *Lensa: Jurnal Kependidikan Fisika*, 12(1), 78–97.
- Nurhuda, Kumala, & Widiyatun. (2023). Analysis of ChatGPT artificial intelligence in solving illustrated physics problems on resistors.



- Oblitas, L., Adeeb, S., & Cruz-Noguez, C. (2025). Performance of AI chatbots on the fundamentals of engineering civil exam. *Discover Artificial Intelligence*, 5(1), 378.
- Polverini, G., & Gregorcic, B. (2024). Evaluating vision-capable chatbots in interpreting kinematics graphs: A comparative study of free and subscription-based models. *Frontiers in Education*, 9, 1452414.
- Russell, S. J., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Pearson.
- Sabharwal, C. L. (2021). Cohen's Kappa statistic and new Kappa statistic for measuring and interpreting inter-rater agreement. *International Journal of Research in Engineering and Science*, 9(2), 23–28.
- Sani, R. A. (2019). *Cara membuat soal HOTS (High Order Thinking Skills)*. Tira Smart.
- Sugiyono. (2017). *Metode penelitian kuantitatif, kualitatif, dan R&D*. Alfabeta.
- Sukmadinata, N. S. (2011). *Metode penelitian pendidikan*. PT Remaja Rosdakarya.
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*.
- Zed, M. (2008). *Metode penelitian kepustakaan* (2nd ed.). Yayasan Obor Indonesia.

