European Journal of Physics



# Performance and failure modes of AI chatbots on a novel concept inventory on relativity in classical mechanics

European Journal of Physics



PAPER

OPEN ACCESS

RECEIVED
7 May 2026

REVISED
29 June 2026

ACCEPTED FOR PUBLICATION
17 July 2026

PUBLISHED
3 August 2026



# Performance and failure modes of AI chatbots on a novel concept inventory on relativity in classical mechanics

**Eugenio Tufino**[1], **Caterina Giovanzana**[2,*], **Andrea Zamboni**[2], **Pasquale Onorato**[2] and **Stefano Oss**[2]

[1] Department of Physics, Informatics and Mathematics, University of Modena and Reggio Emilia, Via G. Campi, 213/A, Modena, Italy
[2] Department of Physics, University of Trento, Via Sommarive 14, 38123 Povo (Trento), Italy
[*] Author to whom any correspondence should be addressed.

**E-mail:** caterina.giovanzana@unitn.it

**Keywords:** Artificial Intelligence, classical relativity, physics education, LLM

Supplementary material for this article is available online

## Abstract

AI chatbots are increasingly used by students as study tools in physics, raising practical questions about their reliability on conceptual tasks. Existing evaluations of large language models (LLMs) on physics concept inventories rely almost exclusively on instruments that have been publicly available for years and likely appear in model training data, making it difficult to disentangle physics competence from familiarity with the test items themselves. We address this by evaluating three LLMs (GPT-5.2, Gemini 3 Pro, Gemini 3 Flash) on the Classical Relativity Concept Inventory, a recently developed and validated 21-item instrument on Galilean relativity that was not publicly available at the time of testing. Each item was administered 30 times per model, and all 1890 responses were qualitatively coded along three dimensions: visual interpretation, physics reasoning, and coordination. Because GPT-5.2 was tested in a no-reasoning configuration whereas the Gemini models were tested with their provider-default reasoning settings, cross-model accuracy comparisons are treated descriptively rather than as direct evidence of differential physics competence. Mean accuracy was 97% for Gemini 3 Flash, 89% for Gemini 3 Pro, and 73% for GPT-5.2 in the no-reasoning configuration (rising to 85%–86% when reasoning was enabled), compared to 62% for the student sample ($N = 267$). However, on a small number of items GPT-5.2 and Gemini 3 Pro drop to near-zero accuracy, while Gemini 3 Flash, although more robust, still misreads specific diagrams. The qualitative analysis shows that these failures stem predominantly from misinterpretations of visual content rather than from deficits in physics knowledge, and that LLM errors differ structurally from those of students: when models err, they converge on a single distractor with high consistency, whereas student errors are more broadly distributed. These findings indicate that chatbot reliability on conceptual physics is item-dependent and unpredictable, with direct implications for how concept inventories are administered.

## 1. Introduction

AI-powered chatbots, such as OpenAI ChatGPT and Google Gemini, are increasingly used by university students as study tools in physics courses. This raises a practical question for instructors: to what extent can these tools reliably handle the physics we teach?

This question is especially relevant for conceptual reasoning, where students might be tempted to use a chatbot as a study companion or even to answer assessment questions.

A growing body of research has begun to address this issue by testing large language models (LLMs) on physics concept inventories—the standardized multiple-choice tests widely used in physics education research to probe conceptual understanding [1, 2]. Early studies showed rapidly improving performance

on the Force Concept Inventory (FCI) [1], with GPT-4 achieving over 80% correctness [3]. At the same time, Wheeler and Scherr [4] observed that the chatbot's incorrect answers often mirrored common student misconceptions, raising the question of whether these systems genuinely reason about physics or simply reproduce patterns from their training data. More recent work by Polverini *et al* [5–7] introduced a systematic methodology for evaluating chatbot performance. Their approach involved repeatedly submitting each item of the Brief Electricity and Magnetism Assessment (BEMA) [8] and classifying errors as visual, reasoning, or coordination difficulties. Their findings showed that ChatGPT-4o outperforms average university students, while still struggling with items that require images interpretation. Kortemeyer *et al* [9] confirmed this pattern on a large scale, testing GPT-4o on 54 concept inventories across 35 languages.

However, all of these studies share a limitation: the instruments they used—the FCI [1], the Test of Understanding Graphs in Kinematics [10], the BEMA [8], and others—have been publicly available for years, sometimes decades. Their items and answer keys can be found online, making it plausible that they were included in the training data of the models being tested [9]. Therefore, we cannot be sure whether high performance reflects genuine physics competence or a sort of memorization.

In this study, we address this issue by testing three multimodal chatbots on a concept inventory they could not have seen during training. In particular, we use the Classical Relativity Concept Inventory (CRCI), a 21-item questionnaire on Galilean relativity and non-inertial reference frames that was recently developed and validated at the University of Trento [11]. Its validation paper was published in early 2026—after the knowledge cutoff of all models we tested—and the instrument has not been made freely available online or indexed by repositories such as PhysPort. The physics covered by the survey—reference frames, the principle of relativity, Galilean velocity composition, the weak equivalence principle—is standard textbook material, but the specific questions, scenarios, and distractors are new. This allows us to observe how chatbots handle familiar physics presented in an unfamiliar format.

We tested GPT-5.2, Gemini 3 Pro, and Gemini 3 Flash—three of the most widely accessible chatbots available to university students at the time of this study [12]. Each item was submitted 30 times per model, following the repeated-trial methodology of Polverini *et al* [6], for a total of 1890 responses. Many survey items feature hand-drawn illustrations of physical scenarios that require spatial interpretation, making the test particularly suited for probing the interplay between visual processing and physics reasoning—an area identified as a persistent weakness of current AI systems [5, 6, 9].

In physics, diagrams are not simple decorations: they carry information essential to the task. As Larkin and Simon [13] noted, a diagram and a text can carry the same information and yet not be equally easy to use: a diagram lets the reader read off spatial relations by perception, replacing effortful step-by-step reasoning. In physics items, the diagram is therefore not decoration but part of the information needed to solve the task, and a model that misreads it loses exactly that information.

This misreading is not specific to physics: in the broader multimodal literature the visual pathway has been identified as the weak component of current vision-language models, which rely on contrastive image encoders that can map visually distinct images to nearly identical representations and therefore misread fine detail even in strong systems [14].

In this paper, we present the results of this evaluation. We report overall and item-level performance for all three models, classify the types of errors that emerge from the chain-of-thought outputs, and compare AI performance and distractor choices to those of the 267 university students who took the same test [11]. Our findings are informative both for instructors who want to understand the capabilities and limitations of the chatbots their students use, and for the physics education research community investigating AI performance on conceptual physics tasks.

## 2. Setting and research questions

### 2.1. The instrument: the CRCI

The survey consists in a 21-item multiple-choice questionnaire designed to assess university students' understanding of core concepts in relative motion [11].

The inventory covers three main conceptual areas: (i) reference Frames, including the general case of transformations between frames with different accelerations; (ii) the Principle of Relativity and Galilean transformations, including trajectory transformations between inertial frames and Galilean velocity composition; and (iii) the weak Equivalence Principle, covering horizontal motion in free-fall scenarios, simple harmonic motion in free fall, and orbital motion.

See figure 1 for an example of a CRCI item corresponding to the conceptual area (i) listed above.

Q7. Alice is on a train moving at a constant speed. Roberto is at the station and sees the train rush by from left to right. Alice, leaning out the window, drops a ball. Neglecting air resistance, what is the trajectory observed from Alice's reference frame?

(A)

(B)

(C)

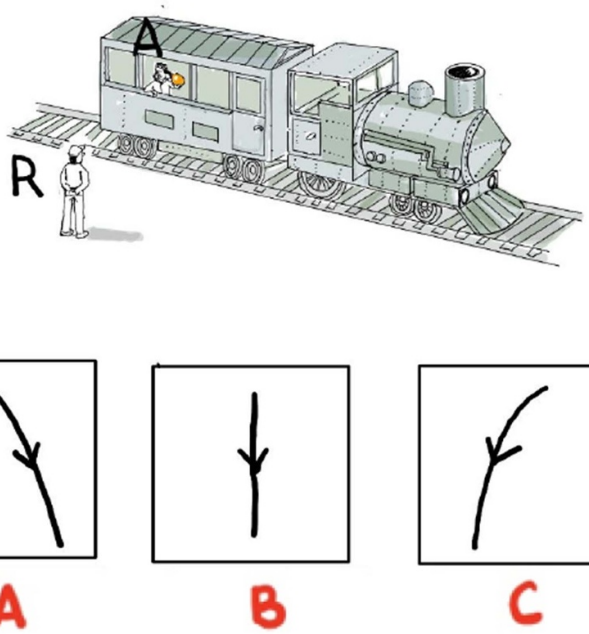


**Figure 1.** Item Q7 of the classical relativity concept inventory from [11].

Although the CRCI is administered as a single-response multiple-choice instrument, item Q21 has two acceptable answers, C and E. Both students and LLMs were awarded one point if the selected answer was C or E, and zero otherwise; no partial credit was used. The presence of two acceptable responses was not pointed out to respondents, since the diagnostic purpose of the question, when administered to students, focused precisely on the preferred choice between the two physically valid representations [11].

The development of the CRCI followed established procedures in the concept inventory literature, including cognitive interviews, expert review, pilot studies, and rigorous statistical validation using classical test theory and the Rasch model. The instrument was administered to 267 first-year physics students at the University of Trento, pooled across three academic years, during which the course structure and instructional approach remained substantially stable across the three cohorts. The inventory was administered in Italian, as a post-test after instruction. Three authors of the present study are also authors of the CRCI validation paper [11], the present article does not revalidate the instrument, but uses the already validated CRCI and its student dataset as a reference point for investigating LLM performance and failure modes.

A crucial feature for the present study is that the CRCI was not publicly available before or during our data collection. At the time the LLMs were tested, the CRCI had not appeared online in any form—neither the questions nor the answer key. This differs substantially from the situation for established instruments such as the FCI or BEMA, whose items have been widely reproduced and discussed in publicly accessible sources for years.

### 2.2. The models

We tested three widely accessible LLMs available in December 2025 [12]: GPT-5.2 (OpenAI), Gemini 3 Pro (Google), and Gemini 3 Flash (Google).

For the Google family, we set the temperature $T = 0.7$, consistent with the settings used in previous studies [6, 9], as both Gemini 3 Pro and Gemini 3 Flash allow temperature control. We note, however, that this choice deviates from Google's own recommendations, which report $T = 1.0$ as the recommended value [15]. We adopted $T = 0.7$ to maintain comparability with prior work, but we have

also done a robustness check at $T = 1.0$, as discussed in sections 3.1 and 4.1. We included both Pro and Flash models in order to compare a larger, more capable model (Gemini 3 Pro) with a lighter and faster one (Gemini 3 Flash). As we report in section 4.1, Gemini 3 Flash slightly outperforms the larger Gemini 3 Pro on most items.

For GPT-5.2, the study uses the baseline configuration reasoning_effort = 'none' with temperature $T = 0.7$, consistent with [6, 9]. The choice of disabling the reasoning mode reflects an API constraint: OpenAI's reasoning models do not accept the temperature parameter when reasoning_effort is set to 'medium' or 'high' [16]. Setting reasoning_effort = 'none' is therefore the only configuration in which $T$ can be explicitly fixed to a value comparable across studies and across the Gemini models tested here. We note that enabling reasoning improves GPT-5.2's performance on specific items; a sensitivity analysis at reasoning_effort = 'medium' and 'high' is reported in section 5, where the trade-off between configuration parity and reasoning depth is discussed in detail.

All three models are multimodal, meaning they can process both text and images as input. This is essential because the survey items contain visual scenarios that must be interpreted to answer the questions.

### 2.3. Research questions

The following research questions guide our analysis:

**RQ1: How do multimodal LLMs perform on the CRCI, a recently developed concept inventory unlikely to be present in their training data?**
The CRCI was not publicly available at the time of testing. This allows us to assess LLM performance in a setting where high accuracy cannot be attributed to memorization of items from training data.

**RQ2: When the models answer incorrectly, what types of errors emerge, and how do error profiles differ across GPT-5.2, Gemini 3 Pro, and Gemini 3 Flash?**
Following Polverini *et al* [6], we classify errors observed in the chain-of-thought output as arising from: (a) incorrect interpretation of the visual content of the item (visual errors), (b) incorrect application of physics principles, evaluated independently of the correctness of the visual interpretation (reasoning errors), or (c) inconsistencies between the reasoning and the selected answer (coordination errors). By testing three models under identical conditions, we can both characterize the nature of the difficulties and identify systematic differences across architectures. This multi-model comparison adds a dimension that is generally absent from the existing literature, which has predominantly focused on OpenAI models.

**RQ3: How does LLM performance compare to that of university students, and do the models gravitate toward the same distractors?**
We compare the models' overall scores and item-level performance to the student data reported in [11] ($N = 267$ first-year physics students). Beyond aggregate performance, we examine whether the incorrect answer options preferred by the LLMs correspond to the distractors most frequently selected by students. This comparison probes whether LLM errors reflect patterns similar to student misconceptions or arise from qualitatively different mechanisms. As noted in [9], care is needed in interpreting such comparisons, since the cognitive processes underlying student and LLM responses are fundamentally different—a point reinforced by the case of Q1 in the CRCI, where GPT-5.2 fails systematically due to a terminological sensitivity, see section 4.1.

## 3. Methods

### 3.1. Data collection

Each of the 21 CRCI items was submitted as a high-resolution screenshot of the English translation [11], replicating the format a student would encounter on paper. The items were submitted individually—one item per API call—to each of the three models via their respective APIs.

The prompt was designed to elicit a structured chain-of-thought response. Each model was instructed to: (1) describe the visual content of the image (Visual Phase), (2) generate a step-by-step derivation about the physics of the problem (Reasoning Phase), and (3) provide a final answer letter (Answer Phase). The output was returned as a structured JSON file using the API's structured output functionality, ensuring consistent parsing across all responses, see the supplementary material. This three-phase structure, consistent with the chain-of-thought approach used in [6, 9], was essential for our qualitative analysis, as it allows the source of errors to be traced to a specific stage of the response.

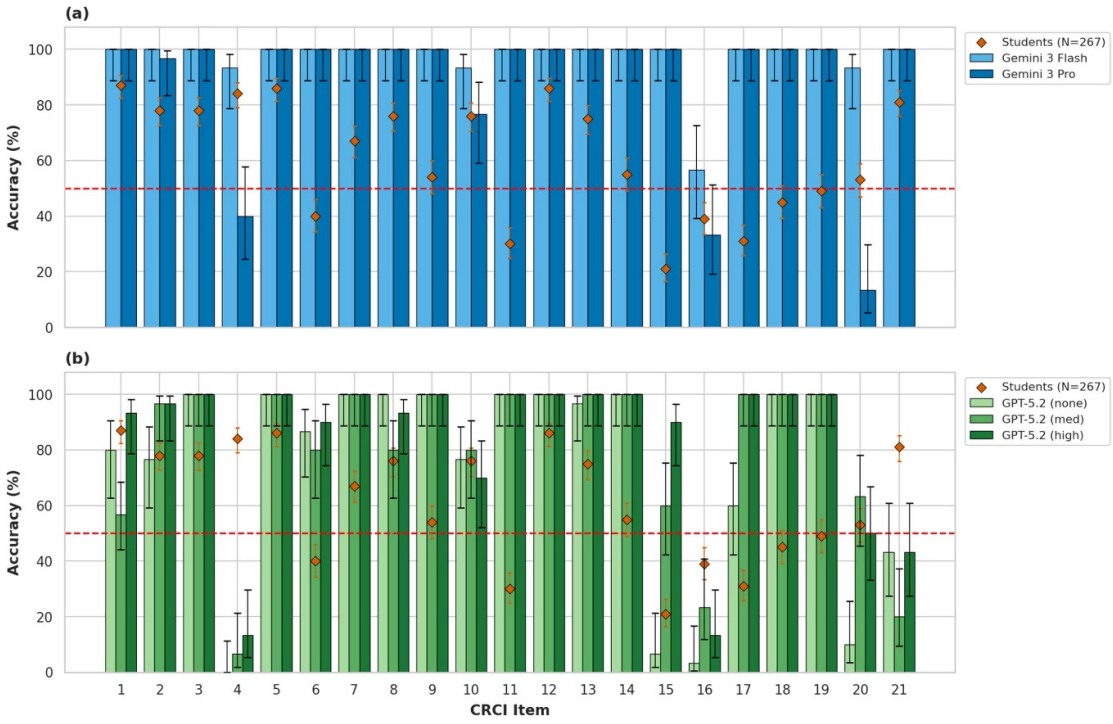


**Figure 2.** Item-level accuracy on the 21 CRCI items. Results are shown in two panels because the two model families—Gemini and ChatGPT—are not run under comparable configurations. (a) Gemini 3 Flash and Gemini 3 Pro, both operating with reasoning enabled by default (thinking_level = 'high'). (b) GPT-5.2 at three reasoning settings (reasoning_effort = 'none', 'medium', 'high'); the 'none' configuration is the baseline used for the qualitative error analysis in sections 4.2 and 4.3. For each LLM, accuracy is defined as the percentage of correct responses out of 30 independent iterations. The GPT-5.2 'none' baseline was run at temperature $T = 0.7$; setting the temperature is only possible with reasoning disabled, so the GPT-5.2 'medium' and 'high' runs were obtained with the API's default sampling and their temperature is not specified. The Gemini runs were conducted at temperature $T = 0.7$. Error bars are 95% Wilson confidence intervals on the per-item proportion of correct responses—over the 30 iterations for each LLM and over the $N = 267$ students for the student sample [11]; where accuracy reaches 100% the interval is one-sided (bounded above by 100%). The student sample is plotted as diamonds to keep it visually distinct from the models' bars. The dashed line marks 50%. Numerical values are reported in table S1. For Q1, the value shown refers to the post-hoc wording check (modified prompt) attached; the original comoving wording is reported separately in table 1 and table S1, and discussed below.

The temperature was set to 0.7 for all three models, consistent with the default used in [6, 9]. For GPT-5.2, the reasoning mode was disabled (reasoning_effort = 'none'), since the OpenAI API does not allow temperature control and reasoning to be active simultaneously (see section 2.2). Each API call was independent: no conversation history was carried over between iterations.

Google's Gemini 3 Developer Guide recommends $T = 1.0$ as the default for the Gemini 3 family [15]. To verify the robustness of our findings against this alternative configuration, we performed a full replication at $T = 1.0$ for both Gemini 3 Flash and Gemini 3 Pro; the results are reported in supplementary material and indicate that the qualitative error patterns described in the following sections are unaffected by this choice.

Each item was submitted 30 times per model, yielding 630 responses per model, for a grand total of 1890 responses. Data were collected between 20 December 2025 and 10 January 2026.

The choice of 30 repeated submissions per item, following the repeated-trial protocol of Polverini *et al* [6], is conservative relative to current practice: earlier studies used 60 iterations [5, 6] reduced this to 30 after finding that the largest item-level uncertainties increased by only about three percentage points; and more recent work uses as few as 10, noting that models tend to answer an item either consistently correctly or with a repeatedly preferred incorrect option, so that larger samples are no longer necessary to capture stable performance patterns [7]. We report per-item proportions with 95% Wilson confidence intervals (figure 2).

### 3.2. Qualitative coding of responses

All responses—not only incorrect ones—were coded qualitatively. The decision to code all responses, including those with a correct final answer, was motivated by the well-documented phenomenon of spurious correct responses in LLM outputs: cases where the model arrives at the correct answer despite

flawed reasoning [6]. Conversely, spurious incorrect responses—correct reasoning leading to an incorrect answer—can also occur, for example, when the model correctly interprets a visual scenario, but then selects the wrong letter.

The coding protocol was developed starting from the framework introduced by Polverini *et al* [6] for the BEMA, and adapted to the specific characteristics of the CRCI. Each response was evaluated on three binary dimensions:

- **Visual score (V0/V1):** Does the model correctly describe the visual content of the item? This includes correctly parsing the physical scenario, the spatial arrangement of objects, the reference frames involved, and any graphical elements (trajectories, arrows, diagrams).
- **Physics reasoning score (P0/P1):** Does the model apply the relevant physics correctly? This evaluates the identification and application of physical principles (e.g. Galilean transformations, fictitious forces, the equivalence principle), independent of whether the visual interpretation was correct.
- **Coordination score (C0/C1):** Is the final answer logically consistent with the stated reasoning? This parameter assesses the internal coherence of the response—whether the selected letter is the logical consequence of what the model claimed to see and to know, regardless of whether the reasoning itself is correct.

This structured output allows us to decouple visual encoding from physical reasoning—a separation that is not possible when only the final answer is evaluated—and constitutes the methodological basis for the error analysis in section 4.

Moreover, to make the coding procedure transparent and reproducible, we report here two examples of borderline coding decisions. The first case comes from items Q15 and Q16, where the models had to interpret images of containers with differently inclined water surfaces. In most of the responses, the initial visual description remained too generic to determine whether the relevant configuration had been correctly identified. It was just in the relative physics reasoning that the models explicitly described the selected image, this evidence was used to assign the visual score: V1 if the configuration was correctly recognized, and V0 otherwise. Instead, in Q2–Q4, responses describing the segment PQ as 'diagonal' were coded as V0 whenever this apparent diagonality was treated as a real physical feature rather than as an effect of perspective, since this interpretation directly shaped the subsequent reasoning and led to errors, as discussed in section 4.3.

To establish the reliability of the coding scheme, a calibration phase was carried out before coding the full dataset. Two authors independently coded a calibration subset of 180 responses, comprising six item-condition combinations—Q1 in both wording conditions ('comoving' and 'attached', see section 4.1), together with Q3, Q15, Q16, and Q20—each with their 30 independent iterations. All calibration responses were drawn from GPT-5.2 in its baseline configuration (reasoning_effort = 'none').

The subset was chosen to span the range of coding situations rather than as a random sample. GPT-5.2 in its baseline configuration was selected because it was the lowest-performing of the three models and therefore the most likely to produce ambiguous or borderline responses. Q1 was included because it showed a systematic error on an item that students found easy, compounded by a translation artifact (section 4.1). Q3 served as an anchor item, since it was answered correctly in all runs. Q15 and Q16 were included as the lowest-accuracy items, and Q20 was included because a visual misinterpretation was suspected. Because these are among the most difficult and ambiguous items, the resulting agreement is a conservative estimate of reliability over the full inventory.

The two coders disagreed on only two of the 180 responses, both on the physics-reasoning dimension. Agreement was therefore 100% for the visual dimension, 98.9% for the physics-reasoning dimension, and 100% for the coordination dimension, corresponding to 538 of 540 binary coding decisions.

After calibration, the remaining responses, including those from the other models, were coded by a single coder. We acknowledge this single-coder design as a limitation in section 5.

We emphasize that our coding classifies the model's textual output, not its internal processes. A 'visual error' means the written description of the scene is incorrect, not that the model's perceptual mechanisms failed. This is standard practice in the LLM evaluation literature [6], where chain-of-thought outputs are treated as observable behavior, without claims about underlying cognition.

## 4. Results and discussion

The structured three-phase output of each response (Visual Phase, Reasoning Phase, Answer Phase) allows us to go beyond binary accuracy and distinguish the model's ability to extract features from the item's visual content, from its ability to apply physics principles. This separation—which is not possible when only the final answer is evaluated—is the basis of the analysis that follows.

### 4.1. Overall performance

Figure 2 shows the accuracy for each CRCI item—expressed as percentage of correct responses out of 30 iterations—for all three models alongside the student sample from [11].

The most striking feature is that all three models achieve near-perfect accuracy on a large subset of items. Items Q3, Q5, Q7, Q8, Q9, Q11, Q12, Q13, Q14, Q18, and Q19 are answered correctly in close to 100% of iterations by all models. These items span different conceptual areas of the CRCI (inertial frames, Galilean velocity composition, free-fall scenarios, orbital motion), indicating that the models possess robust knowledge of the underlying physics when the visual component is straightforward or does not affect the response.

However, a number of items reveal clear and systematic difficulties. When analyzing the behavior of Gemini Flash, Gemini Pro and ChatGPT-5.2 with reasoning_effort = 'none'—LLMs on which this article primarily focuses—we find that the pattern of failure differs across models, and three clusters can be identified:

(a) Items involving fluids on inclined planes (Q15, Q16): These items involve the position of a fluid on an inclined plane viewed from a non-inertial reference frame. Q16 (with friction) is the only item on which all three models show low accuracy (Flash 57%, Pro 33%, GPT-5.2 3%), suggesting that the visual interpretation of this scenario—which requires coordinating a 3D perspective with a frame transformation—poses a fundamental challenge for current multimodal architectures. Q15 (frictionless case), by contrast, reveals a model-specific failure: both Gemini models answer correctly in all 30 iterations, while GPT-5.2 selects distractor E in 93% of cases—a failure rate even higher than the student average of 21% on the same item. Interestingly, this failure is partially resolved when reasoning is enabled: GPT-5.2's accuracy on Q15 rises from 7% to 90% at reasoning_effort = 'high' (table S1).

(b) Items where specific models fail (Q4, Q17, Q20, Q21): Q4 (Galilean velocity composition) is particularly revealing: Gemini 3 Flash answers correctly in nearly all iterations, while GPT-5.2 scores 0%, with all incorrect responses clustering on the same wrong option (see section 4.3 table 3). Q17 (pendulum in free fall) shows a different pattern: both Gemini models answer correctly in all iterations, while GPT-5.2 selects the correct answer in only 60% of cases (see section 4.3 table 3). Q20 and Q21 also show model-specific failures, with GPT-5.2 scoring 10% and 43% respectively, while Flash remains above 93% on both.

(c) Post-hoc wording check for Q1: Q1 required a separate post-hoc check because the English translation initially used *comoving* for the Italian expression 'solidale al corpo'. With this wording, in the no-reasoning condition, GPT-5.2 answered correctly in only 1 out of 30 iterations (accuracy of 3%), while both Gemini models answered correctly in all iterations, see table 1. Replacing *comoving* with *attached* increased GPT-5.2 accuracy to 80%, without affecting the Gemini results (table 1). We therefore report the modified Q1 result in the main item-level comparison, but explicitly treat it as a post-hoc wording check rather than as part of the original standardized administration.
Qualitative inspection of no-reasoning GPT-5.2's justifications suggests a case of terminological sensitivity: with *comoving*, the model often interpreted the frame as merely instantaneously velocity-matching, rather than as a continuously body-fixed frame; with *attached*, it more often recognized that the body has fixed coordinates in its own frame, so its coordinate acceleration is zero. However, residual errors remained, involving a shift from coordinate acceleration in the attached frame to physical or inertial-frame acceleration.

Table S1 in the supplementary material reports Item-level accuracy (%) for each model and for the student sample.

Across the 21 items, Gemini 3 Flash slightly outperforms the larger Gemini 3 Pro, with mean accuracies of 97.0% and 88.6%, respectively (see table 2). There is no item on which Pro performs better than Flash (see table S1). We describe this as an inverse performance ordering rather than inverse scaling [17]: with only two models from the same family, and without access to their parameter counts or training

**Table 1.** Item-level accuracy (%)—percentage of correct responses out of 30 independent iterations—for each model for Q1 when the term comoving appears in the question text compared to those where the term attached appears.

| Model | Gemini 3 Flash | Gemini 3 Pro | GPT-5.2 reasoning_effort = 'none' |
|---|---|---|---|
| Q1 (*comoving*) | 100.0 | 100.0 | 3.0 |
| Q1 (*attached*) | 100.0 | 100.0 | 80.0 |

**Table 2.** Overall performance on the CRCI, summarized across the 21 items. For each LLM, the accuracy on a given item is the percentage of correct responses out of 30 independent iterations. 'Items at 100%' denotes the number of items answered correctly in all iterations, whereas 'Items below 50%' counts those below the 50% threshold. (a) Results for the two Gemini models, both run with reasoning enabled by default. (b) Results for GPT-5.2 under three reasoning_effort settings; the 'none' configuration is the baseline used for the qualitative error analysis in sections 4.2 and 4.3. **The values in (a) and (b) are not directly comparable, because the Gemini models and the GPT-5.2 baseline were run under different reasoning configurations** (see sections 2.2 and 3.1). The student sample ($N = 267$ [11],) averaged 61.5% (mean total score ≈ 12.9/21; SD across students 4.0 points); the full distribution of student scores is reported in [11], figure 1.

(a)

| Model | Mean accuracy (%) | Items at 100% | Items below 50% |
|---|---|---|---|
| Gemini 3 Flash | 97.0 | 17 | 0 |
| Gemini 3 Pro | 88.6 | 16 | 3 |

(b)

| GPT-5.2 configuration | Mean accuracy (%) | Items at 100% | Items below 50% |
|---|---|---|---|
| Reasoning_effort = 'none' (baseline) | 73.3 | 10 | 5 |
| Reasoning_effort = 'medium' | 84.9 | 11 | 2 |
| Reasoning_effort = 'high' | 85.9 | 11 | 2 |

**Student sample:** $N = 267$ [11]: students averaged 61.5% (mean total score ≈ 12.9/21; SD across students = 4.0 points); full score distribution in [11], figure 1.

details, we cannot attribute it to model size; a more parsimonious explanation is that Flash and Pro are separately optimized systems. The broader inverse-scaling phenomenon reported in the AI benchmarks literature [17] is mentioned here only as background. A robustness check at $T = 1.0$, Google's recommended default for Gemini 3 (see table S2 in the supplementary material), reduces the gap but does not eliminate it, indicating that the difference is not an artifact of the temperature setting.

When compared to students, the LLMs outperform the student average on most items. However, the exceptions are significant. On Q4, students score approximately 84% while GPT-5.2 scores 0%, due to a systematic visual error, as better discussed in section 4.3.

On Q16, both students and all three models perform poorly, but students' errors are distributed across several distractors while the models tend to concentrate on a single wrong option. On Q17, the student average is modest (∼31%), and GPT-5.2 also shows difficulty (60% accuracy), while both Gemini models answer correctly in all iterations.

These cases illustrate that high overall AI accuracy can coexist with complete failure on specific items—a concern for instructors who might rely on chatbots as tutoring tools.

We note that GPT-5.2 was tested with reasoning disabled (see section 2.2). When reasoning is enabled, its mean accuracy rises to 85%–86%, approaching that of Gemini 3 Pro (see table S1 and section 5).

### 4.2. Error profiles

Figure 3 presents the error composition for each item and each model—Gemini Flash, Gemini Pro and ChatGPT-5.2 with reasoning_effort = 'none'—, classified through qualitative coding into visual errors, reasoning errors, and coordination errors (the coding scheme and the inter-coder agreement on a calibration subset are described in section 3.2).

**Gemini 3 Flash (panel a)** produces few incorrect responses which are all classified as visual errors. When Flash answers incorrectly, it does so because it misinterprets the image, not because it lacks the relevant physics knowledge. Items showing elevated visual error rates in the coding—including cases where the

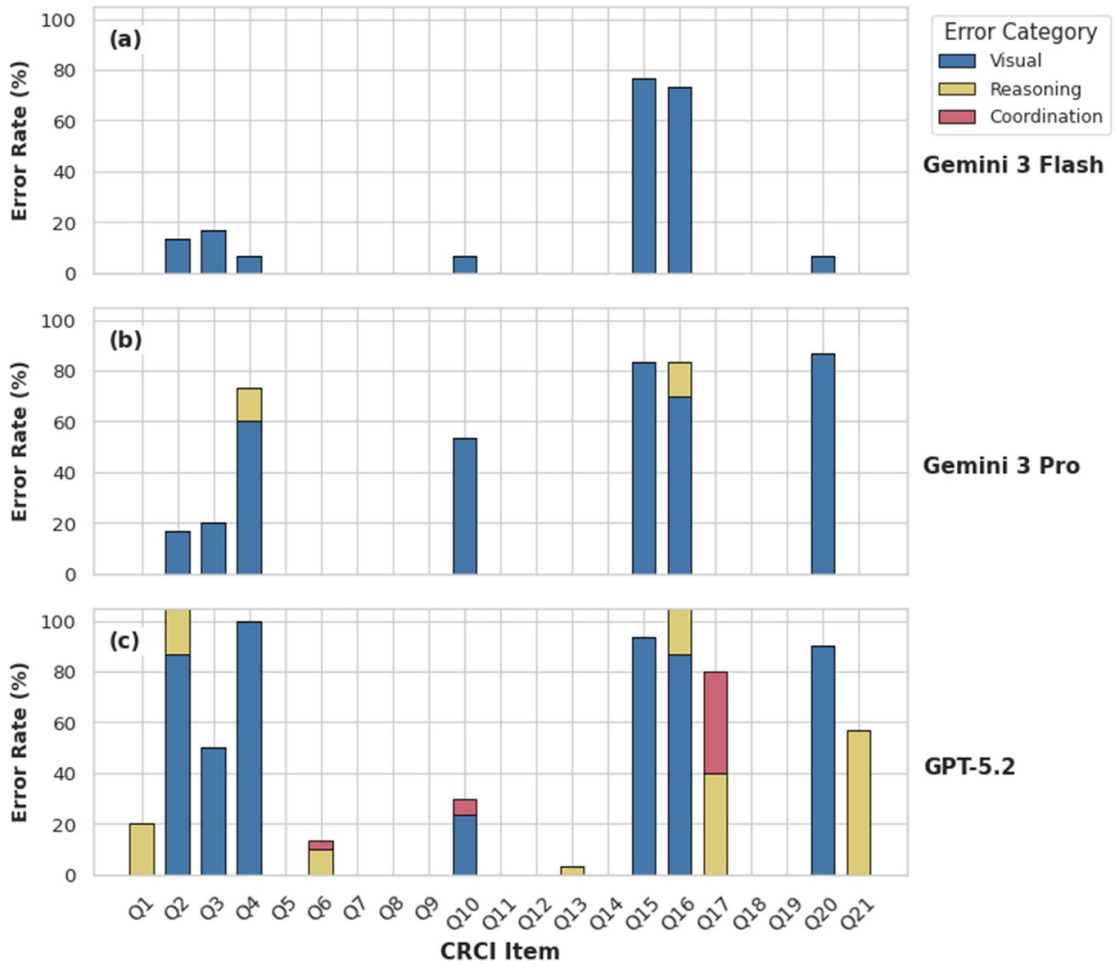


**Figure 3.** Error composition for each CRCI item, by model. Stacked bars show the breakdown of coded responses into visual errors only (V0, P1), reasoning errors only (V1, P0), combined visual and reasoning errors (V0, P0), and coordination errors (V1, P1, C0), based on the qualitative coding described in section 3.2. Panels: (a) Gemini 3 Flash, (b) Gemini 3 Pro, (c) GPT-5.2 (baseline configuration, reasoning_effort = 'none').

model reaches the correct answer despite a flawed visual description—are Q2, Q3, Q4, Q10, Q15, Q16, and Q20, all of which involve diagrams with spatial or perspective complexity. Particularly, for Q15, Flash is coded with a visual error in 23 out of 30 responses, yet it answers correctly every time because the visual error affects one of the wrong answers. Reasoning and coordination errors are absent.

**Gemini 3 Pro (panel b)** shows a different pattern. Visual errors remain the dominant category, but reasoning errors appear on two items: Q4 and Q16. On Q4, 14 out of 18 incorrect responses are classified as pure visual errors (V0, P1), while the remaining 4 combine visual and reasoning errors (V0, P0). Q16 shows a similar pattern, with 16 visual errors and 4 visual-plus-reasoning errors. On Q2, Q3, Q10, and Q20, errors are exclusively visual. As with Flash, Q15 shows a high visual error rate in the coding (25 out of 30 responses coded V0) despite 100% accuracy—a compensatory pattern even more pronounced than in Flash. The overall error rate is substantially higher than for Flash, consistent with the inverse performance ordering described in section 4.1 and here visible also at the qualitative level; we retain [17] only as a reference to the broader phenomenon.

**GPT-5.2, reasoning_effort = 'none' (panel c)** presents the most complex error profile, with errors across all three categories. Reasoning errors dominate on Q1, where a form of semantic sensitivity induces the model to mix up coordinate acceleration in an attached frame to physical acceleration in an inertial frame, and on Q21, where all 17 errors are coded (V1, P0). In Q2, visual errors are dominant, with 22 out of 29 coded errors classified as V0. Seven responses involve reasoning errors, including 4 coded (V0, P0) and 3 coded (V1, P0). Q2 also includes two spurious correct responses, in which the chatbot selects the correct final answer despite producing reasoning that is not fully coherent with that answer. These two responses are coded (V0, P1, C0) and (V0, P0, C0). Although they have different

codes, both cases appear to stem from the same underlying pattern: the chatbot initially fails to interpret the schematic arrangement of the diagram correctly, but then redescribes it correctly at the end of the physics reasoning phase, leading it to select the right final answer. On Q4, GPT-5.2 fails entirely, but the qualitative coding reveals that all 30 errors are just visual (V0, P1, C1): the model misreads the spatial arrangement of the scenario and then applies correct physics to its flawed geometric premise. Visual errors also dominate on Q15, Q16, and Q20. In particular, on Q16 a spurious correct response occurs, in which flawed reasoning does not affect the selection of the correct answer. Q17 stands apart: here, GPT-5.2's 12 errors are classified as combined reasoning and coordination failures (V1, P0, C0), meaning the model correctly interprets the visual scenario, but applies incorrect physics and then selects an answer inconsistent even with its own flawed reasoning. Pure coordination errors appear on Q6 (1 out of 4 errors) and Q10 (2 out of 7 errors). For example, in Q6, GPT-5.2 explains why the motion of an apple falling from a tree—from the perspective of an observer on a train that is accelerating at that very moment—should be uniformly accelerated linear motion. However, at the end of the reasoning, it becomes inconsistent by stating that since this is the motion of a projectile, it must be parabolic, thus contradicting what was described just moments earlier.

From this analysis, four main failure modes can be identified:

- **Semantic misalignment:** The model's response is driven by a linguistic interpretation that diverges from the intended physical meaning. The clearest example is Q1 (see section 4.1 and table S1), where GPT-5.2 interprets the terms in a specialized terminological sense found in certain treatments of special relativity [18].
- **3D spatial vision failure:** The model fails to correctly parse diagrams involving perspective, inclined planes, or reference frame transformations represented visually. This is the dominant failure mode on Q16 across all models, and on Q15 for GPT-5.2.
- **Spurious and compensatory correct answers:** The model arrives at the correct answer despite flawed visual interpretation or incorrect reasoning and a compensatory correct response, in which the visual interpretation is incorrect but the error does not affect the option selected, so the answer is still correct. These were detected by coding all responses, not only incorrect ones.
- **Coordination failure:** The model describes the scenario correctly, generates an adequate physics derivation, but selects a letter that does not match its own conclusion. This mode is rare but was observable in GPT-5.2.

Repeating the protocol on GPT-5.2 at higher reasoning effort (table S1) shows that the effect of reasoning depends on the error type, and three patterns emerge. Reasoning-fixable items improve as reasoning increases, because deeper reasoning removes faulty physics logic (e.g. Q17: 60% → 100% → 100%; Q15; Q21). Reasoning-resistant items stay low regardless of effort, because the failure is perceptual rather than logical (Q4; Q16, which also peaks at medium). Non-monotonic items peak at medium and decline at high (e.g. Q20: 10% → 63% → 50%), where additional reasoning becomes counterproductive—an over-thinking effect in which excessive scrutiny discards correct options (clearest on Q21 and Q16). Overall, reasoning removes baseline logical errors but cannot compensate for visual misinterpretation, and beyond a point it can degrade performance.

The overall picture supports the conclusion that, on the CRCI, multimodal failures stem primarily from visual misinterpretations rather than from deficits in physics knowledge. This is most clearly demonstrated by the Flash model, whose errors are almost entirely visual, but the pattern holds for the other models as well: reasoning errors, when they occur, are concentrated on a small number of items and are often entangled with visual difficulties.

One plausible explanation is architectural. Current vision-language models encode the image with CLIP-style features that capture the gist of a scene better than its exact geometry. Visually different configurations can thus receive almost identical representations—the 'CLIP-blind pairs' of [14]—leading to systematic visual errors even in strong models. In a physics concept inventory, this matters because diagrams, for example, are not merely illustrative. When the diagram is essential to the item, these errors lead to wrong answers even when the physics is correct.

### 4.3. Distractor analysis and comparison with students

Figures S1 and S2 in supplementary material show the distribution of selected answer options for each CRCI item, comparing the three models with the student population from [11]. The correct answer is

**Table 3.** Most frequently selected distractor for each CRCI item, by model and by students. For each LLM, the letter indicates the most selected incorrect option across 30 iterations, with the percentage of all responses in parentheses. For students ($N = 267$), the most selected incorrect option is reported with the percentage of students who chose it [11]. A dash (—) indicates that all responses were correct. The correct answer for each item is listed in the key column. The full response distributions are available in the supplementary material.

| Item | Key | Gemini 3 Flash | Gemini 3 Pro | GPT-5.2 | Students |
|---|---|---|---|---|---|
| Q1 | A | — | — | B (20%) | B (13%) |
| Q2 | B | — | A (3%) | D (23%) | A (15%) |
| Q3 | C | — | — | — | A (18%) |
| Q4 | A | C (7%) | C (40%) | C (100%) | B (9%) |
| Q5 | B | — | — | — | A (14%) |
| Q6 | C | — | — | B (13%) | B (43%) |
| Q7 | B | — | — | — | C (26%) |
| Q8 | A | — | — | — | B (19%) |
| Q9 | C | — | — | — | A (38%) |
| Q10 | B | A (7%) | A (23%) | A (17%) | A (12%) |
| Q11 | A | — | — | — | C (40%) |
| Q12 | B | — | — | — | A (10%) |
| Q13 | D | — | — | B (3%) | B (24%) |
| Q14 | C | — | — | — | B (25%) |
| Q15 | B | — | — | E (93%) | A (30%) |
| Q16 | E | D (27%) | D (43%) | A (70%) | C (24%) |
| Q17 | C | — | — | B (40%) | A (27%) |
| Q18 | E | — | — | — | A (32%) |
| Q19 | C | — | — | — | B (19%) |
| Q20 | C | B (3%) | B (87%) | B (47%) | A (42%) |
| Q21 | C/E | — | — | D (57%) | A (10%) |

Q4. Roberto is sitting on a bridge and sees Alice passing by on a boat traveling at a constant speed along the river. Just as Alice is passing by, a drone D starts from point P and flies at a constant speed to point Q on the opposite bank. How intense will the drone's speed be in Alice's frame of reference?

(A) Greater than the magnitude of the drone's velocity in Roberto's frame of reference.

(B) Equal to the magnitude of the drone's velocity from Roberto's frame of reference.

(C) Impossible to determine without knowing the magnitude of the boat's and drone's velocities.

(D) Less than the magnitude of the drone's velocity from Roberto's frame of reference.

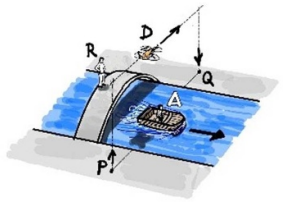


**Figure 4.** Item Q4 of the concept inventories from [11].

marked with an asterisk. Table 3 shows the most selected distractor for each question and each model and students.

On items where all models perform well (e.g. Q5, Q7, Q8, Q9, Q11, Q12, Q13), the distractor distributions are highly concentrated on the correct option: models select the correct option in nearly all iterations, and students show a broader but still predominantly correct distribution.

We note that in Q3, models correctly select C as their final answers (see table S1), even though all three LLMs exhibit a visual error (figure 3), because the misinterpretation of the spatial arrangement does not affect the answer to this question.

The most informative cases are those where errors occur.

We underscore five items:

**Q4 (Galilean velocity composition):** This is the most impressive divergence between models. The item tests conceptual understanding of Galilean relative velocity by requiring a vector—rather than scalar—interpretation of motion and a clear geometric recognition of perpendicular directions: subtracting the boat's velocity from the drone's velocity produces a resultant whose magnitude necessarily increases, since the vector difference introduces an additional orthogonal component (figure 4). The result can therefore be inferred qualitatively, without any numerical calculation.

As reported in table S1, Gemini 3 Flash selects the correct answer (A) in 93% of iterations, closely matching students' performance (about 85%). By contrast, GPT-5.2 with reasoning disabled selects option C in 100% of iterations—a complete failure with zero variance—and Gemini 3 Pro is split across A (40%), C (40%), and D (20%). In particular, the distractor C corresponds to the claim that the drone's speed in Alice's frame cannot be determined without knowing the numerical values of

the velocities. The correct reasoning requires recognizing that the drone flies perpendicularly to the boat's direction, so the Galilean velocity sum always has a greater magnitude than the drone's velocity alone—regardless of the specific values. GPT-5.2's behavior here resembles a systematic error rather than stochastic error. In fact, qualitative analysis (section 4.2) reveals it stems from a visual error: the model misreads the spatial arrangement in the diagram, interpreting the drone's trajectory as having a component along the boat's direction, and then applies the Galilean transformation correctly to its faulty geometric premise. Gemini 3 Pro's selection of D stems from the same visual error; in two-thirds of those cases, however the model also introduces a hallucinated assumption that, in physics problems, boats are generally slower than a flying drone.

**Q15 (fluid on frictionless inclined plane):** shows a striking divergence: both Gemini models answer correctly in all iterations, while GPT-5.2 selects distractor E in 93% of cases.

**Q16 (fluid on inclined plane with friction):** is the only item where all three models struggle, suggesting that this specific visual scenario poses a fundamental challenge for current architectures.

**Q17 (pendulum in free fall):** Both Gemini models select the correct answer (C, uniform circular motion) in all iterations. GPT-5.2 selects C in 60% of cases, correctly recognizing that the rigid rod provides centripetal acceleration in the zero-gravity frame. In the remaining 40%, the same model dismisses this possibility—arguing that centripetal acceleration 'cannot be sustained'—and concludes the trajectory is a straight line, yet selects 'parabolic motion' (B) as 'the closest description'. This 60/40 split on identical input reveals that GPT-5.2's grasp of mechanical constraints under effective weightlessness is fragile rather than absent.

**Q21 (orbital motion):** This item accepts two correct answers (C and E). Flash gravitates strongly toward C, Pro toward E, and GPT-5.2 distributes across C, D, and E. The student distribution is broadly similar, with the majority on C and E. Notably, GPT-5.2 places a substantial fraction on D, a distractor that students rarely select, suggesting a model-specific error pattern unrelated to student reasoning.

Several confounds constrain the interpretation of the model-student comparison, which we therefore treat as descriptive and exploratory. First, the two groups received the items in different languages: students took the CRCI in the original Italian, whereas the models received the English translation; as Q1 shows (section 4.1), translation choices can substantially affect model responses, and some distractor wordings may not be equivalent across languages. Second, the student data are secondary: they come from a single first-year cohort assessed on paper in a separate study [11], with one response per student, whereas each model was queried 30 times via an API on screenshots of the items. The two distributions are therefore constructed differently—across 267 individuals each answering once versus one model resampled 30 times—and the student results may not generalize beyond this cohort. Third, and most fundamentally, the cognitive processes underlying student and model responses differ in kind [9]: student distractor choices reflect conceptual difficulties, whereas model choices largely reflect visual misinterpretation and terminological sensitivity. We therefore do not read item-level agreements or disagreements between model and student distractors as evidence about model mechanisms.

Overall, the distractor analysis indicates that LLM error profiles do not reproduce students' common difficulties patterns. When models err, they tend to concentrate on a single distractor with high consistency (low variance), whereas students distribute their errors more broadly. Moreover, the specific distractor preferred by a failing model is often different from the one most frequently selected by students. This finding is consistent with the observation that model errors are predominantly visual or semantic in origin (as established by the qualitative coding in section 4.2), rather than reflecting the conceptual difficulties that drive student misconceptions; this is also in line with [9]. The most robust quantitative result of this comparison is at the aggregate level: using Shannon entropy [19] as a per-item dispersion measure, a paired Wilcoxon signed-rank test across the 21 items shows that LLM response distributions are systematically more concentrated than students' ($p < 0.005$ for all three models). At the item level, by contrast, the comparisons are descriptive: with 30 iterations per model per item the per-item distributions do not support robust hypothesis testing, consistent with [6].

### 4.4. Implications for teaching

The results presented above have direct implications for the use of AI chatbots in physics education, which we organize around three audiences: developer of concepts questionnaires, instructors, and students.

A common thread is that the high overall accuracy of all three models on the majority of CRCI items is misleading: a chatbot that answers 18 of 21 CRCI items correctly appears competent, yet its few failures are severe, confident, and unpredictable—in our survey, on Q4, for example, GPT-5.2 gives a confidently wrong answer in every iteration.

For **developers** of concept questionnaires, the key finding is that chatbot performance is item-dependent and is not predictable from the difficulty an item poses to human students. Developers should therefore be aware that surface manipulations—item rotation, shuffling, or re-rendering a figure—do not guarantee that an item becomes resistant to chatbots: an item that a model misreads can remain vulnerable in a different format. Where it matters that an item cannot simply be answered by a chatbot—for example, when an instrument is used for assessment—this resistance must be checked empirically against current models rather than assumed, and representation-heavy items, which we found to be the most error-prone, deserve particular attention.

For **instructors**, three points follow. First, the difference in performance between the lighter and the more expensive model—Gemini 3 Flash performs slightly better than the larger Gemini 3 Pro across both temperature configurations tested—suggests that instructors should not assume that a more advanced or more expensive model is necessarily more reliable for physics tasks. This is consistent with the cost-performance analysis reported by Polverini and Gregorcic [7], who found that higher-priced models do not always yield better results on physics concept inventories; model selection should be guided by empirical evaluation on the specific domain, not by marketing claims or general benchmarks. Second, the prevalence of visual errors limits the usefulness of these tools as tutors: many CRCI items that challenge the models involve hand-drawn illustrations, spatial perspectives, and reference-frame diagrams—precisely the representations that are central to physics instruction—so their value as tutors on topics that require interpreting experimental setups, free-body diagrams, or frame transformations is limited [6], and this weakness is not easy to detect, since the model's textual reasoning may appear sound even when its visual interpretation is wrong. Third, on assessment: because chatbots achieve near-perfect scores on most CRCI items, take-home administration of the CRCI is no longer a reliable measure of student understanding, and supervised, in-person settings are preferable. More broadly, the high accuracy reported across many concept inventories by Kortemeyer *et al* [9] suggests that this concern is not specific to the CRCI. While supervised administration is not itself a new recommendation, the specific point here is that chatbot failure is item-dependent and format-resistant, so rotating or reformatting items may not make a take-home instrument safe. Instructors should be aware of these patterns and communicate them to students.

For **students**, our findings argue for developing epistemic vigilance toward AI outputs—a scientific habit of mind, the disposition to be alert to how a claim was generated and to check it against the physics [20]. Concretely, students should know that high average performance does not guarantee reliability on any specific item, and that diagrams, spatial perspectives, and reference-frame transformations are where these tools are weakest; on such items they are better served by re-deriving or independently verifying the result than by accepting the answer. The risk is twofold. A confident wrong answer, as on Q4, is one. The other is a correct answer produced by flawed reasoning: a student who asks a chatbot to explain its reasoning on a question it answered correctly may receive a plausible but physically incorrect explanation, which could reinforce misconceptions rather than address them. An example is the response to Q16 by GPT-5.2 with reasoning disabled, which asks for the inclination of the water surface inside a container sliding down a rough inclined plane: according to the chatbot's reasoning, the water surface should be even steeper than the plane itself—a physical impossibility, since only gravity and kinetic friction act on the system—yet the model selects the correct option, owing to a visual error.

A related concern arises from coordination failures, where the model provides a correct visual description and a sound physics argument but then selects an answer that contradicts its own reasoning, as on Q6 (see section 4.2). Such cases are rare, but they show that a fluent, internally consistent-looking explanation can still end in a wrong answer.

Overall, these findings call for critical awareness of chatbot limitations rather than for avoiding the tools entirely, particularly on tasks involving visual interpretation and conceptual reasoning about reference frames and relative motion.

## 5. Limitations and robustness

Several limitations of this study should be acknowledged.

First, our results represent a snapshot of the current state of AI technology. LLMs are evolving rapidly, and future models may overcome the visual and reasoning difficulties documented here. The

models we tested may themselves be deprecated after publication. This study should therefore be read as a historical reference point, against which future progress can be measured—much as earlier evaluations of ChatGPT-3.5 on the FCI now serve as benchmarks for how quickly the technology has advanced [21].

This time-dependent nature of the study also frames the choice of the models. We focused on GPT-5.2 and the Gemini 3 models because, at the time of data collection, they were among the most widely accessible multimodal chatbots [12]. In addition, querying several models 30 times per item through paid APIs carries a substantial computational and monetary cost, which in practice limits the number of systems that can be evaluated [7]. This selection was therefore pragmatic rather than exhaustive. Other widely used systems, including Anthropic's Claude and Microsoft Copilot, were not included, and the results should not be generalized to all available chatbots.

Second, several configuration choices warrant attention.

The temperature parameter was set to $T = 0.7$ across all models, while Google's recommended default for the Gemini 3 family is $T = 1.0$. A sensitivity analysis at this alternative setting is reported in table S2 of the supplementary material, and confirms that the qualitative findings of this study are not affected.

Furthermore, the three models were not tested under strictly equivalent configurations. All models were queried at $T = 0.7$, but for GPT-5.2 this required disabling the reasoning mode (reasoning_effort = 'none'), as the OpenAI API does not allow temperature control and reasoning to be active simultaneously [16].

The Gemini 3 models were queried without specifying a thinking level, which defaults to 'high' in the Google API [22]. This means that the Gemini models operated with their full reasoning capabilities enabled, while GPT-5.2 did not—an asymmetry that reflects each provider's API design rather than a methodological oversight. Nevertheless, cross-model accuracy comparisons should be interpreted with this asymmetry in mind; our primary focus is on characterizing error types rather than ranking models.

A related limitation concerns the rapid evolution of the LLM landscape. The original gemini-3-pro-preview model used in our main data collection (released 18 November 2025) was deprecated by Google on 9 March 2026, less than four months later [23]. The robustness check at $T = 1.0$ reported in table S2 therefore uses the successor model gemini-3.1-pro-preview (released 19 February 2026), with thinking_level set to 'medium', the closest available equivalent to the original model configuration. While the Gemini 3 Flash model used in this study remains available, the deprecation of the Pro model after less than four months of availability illustrates a broader methodological concern for PER-on-LLM research: studies of this kind are inevitably time-stamped, and exact replication is sometimes materially impossible because the systems being characterized cease to exist. To enable post-hoc verification, the raw model outputs from this study will be made available in a public repository upon publication [24]. We expect the qualitative error taxonomy (visual, reasoning, coordination), being grounded in the structure of the task rather than in a particular model, to transfer to successor systems, even though the specific accuracy values are likely to change. Readers should therefore treat the accuracy figures as a time-stamped snapshot, and the error taxonomy and the dominance of visual errors as the more durable contribution.

To address the configuration asymmetry previously described, we assessed the sensitivity of GPT-5.2's performance to reasoning depth by repeating the 30-iteration protocol at reasoning_effort = 'medium' and 'high' (summarized in table 2 and reported item by item in table S1). At these settings the temperature parameter is not user-configurable [16]. Items whose baseline errors were classified as physics failures improved substantially at 'medium' and held at 'high' (e.g. Q17: 60% → 100% → 100%; Q21: 43% → 97% → 93%). Items dominated by visual errors showed a more complex pattern: Q15 improved monotonically (7% → 60% → 90%), while Q4 remained near zero across all settings, confirming a visual bottleneck that reasoning cannot overcome. On several items (Q16, Q20), accuracy peaked at 'medium' and decreased at 'high', indicating that excessive reasoning can degrade performance on visually demanding items. Overall, the sensitivity analysis reveals a clear interaction between error type and reasoning depth: items dominated by physics errors are resolved by moderate reasoning, items with pure visual errors are unaffected regardless of reasoning level, and items with complex visual demands show non-monotonic behavior where excessive reasoning can degrade performance.

Indeed, when reasoning is enabled, GPT-5.2's mean accuracy (85%–86%) approaches that of Gemini 3 Pro (89%) (see table S1), suggesting that much of the apparent performance gap in the baseline comparison reflects the configuration asymmetry rather than a fundamental difference in physics capabilities.

At the same time, raising the reasoning effort does not recover the items that fail for visual reasons: across the none, medium, and high settings, the visually driven failures (for example Q4 and Q16) remain low. The limiting factor on these items is therefore perception rather than reasoning. Progress is more likely to come from strengthening the models' visual grounding—for example by integrating

self-supervised visual features, as in [14]—than from more elaborate prompting. This is a direction for model development, not an intervention available to users of current chatbots.

Third, we tested a single concept inventory covering a specific area of physics (classical relativity). While the CRCI spans a range of conceptual areas—reference frames, Galilean transformations, and the equivalence principle—our findings may not generalize to other topics or to other types of assessment instruments. In particular, the strong role of visual errors in our results is likely related to the specific nature of the CRCI diagrams, which involve 3D perspectives, inclined planes, and reference frame transformations. Instruments with different visual demands may produce different error profiles.

Fourth, the qualitative coding of responses—which was limited for the majority of the analysis to one coder—, while grounded in the framework of Polverini *et al* [6], inevitably involves a degree of subjectivity. The binary scoring (0/1) on each dimension simplifies a continuous spectrum of quality, and borderline cases require judgment calls.

Fifth, we did not explore the effect of prompt engineering on model performance. Our three-phase prompt—visual description, physics reasoning, final answer—follows the chain-of-thought coding methodology of Polverini *et al* [6] and Kortemeyer *et al* [9], and is what allows visual and reasoning errors to be separated. As any prompt is an intervention, this structure may influence the error distribution: requiring an explicit visual description first could either propagate an early misreading into the subsequent reasoning or, conversely, help the model attend to relevant details. We did not test alternative prompts, so the direction of this effect remains open. The visual bottleneck, however, is independently supported—by the broader vision-language-model literature and by the fact that increasing the reasoning effort does not rescue the visually driven failures—so it is unlikely to be an artifact of the prompt alone.

Finally, while the CRCI items were almost certainly absent from the training data, the underlying physics is standard textbook material. LLMs do not retrieve stored answers but generate responses from learned patterns, so familiarity with analogous problems may contribute to the high accuracy observed on most items. What the CRCI tests is the models' ability to handle novel formulations—new scenarios, diagrams, and distractor structures—of familiar physics. This is a qualitatively different situation from established instruments such as the FCI, whose specific items and answer keys have been widely reproduced online for decades.

## 6. Conclusions

This study addressed three research questions concerning the performance (RQ1), error profiles (RQ2), and comparison with student data (RQ3) of widely accessible LLMs on the CRCI. We evaluated three large multimodal models, GPT-5.2, Gemini 3 Pro, and Gemini 3 Flash, on the CRCI, a recently developed and validated instrument that was not publicly available at the time of testing. By collecting 30 independent iterations per item per model and qualitatively coding all responses on three dimensions (visual interpretation, physics reasoning, and coordination), we obtained a detailed picture of how current AI chatbots handle conceptual physics tasks on an uncontaminated instrument.

All three models achieved high accuracy on the majority of CRCI items, demonstrating robust physics knowledge on topics such as inertial reference frames, Galilean velocity composition, and free-fall scenarios. However, specific items, particularly those involving 3D spatial interpretation of diagrams (Q16 for all models, Q15 for GPT-5.2) or requiring careful semantic parsing (Q1), revealed systematic and often complete failures. The qualitative analysis showed that these failures stem predominantly from visual misinterpretations rather than from deficits in physics knowledge: when the models successfully extract the visual features, they almost always generate a correct physical derivation.

Across the 21 items, Gemini 3 Flash slightly outperformed the larger Gemini 3 Pro on most items. A robustness check at the alternative temperature setting recommended by Google for the Gemini 3 family ($T = 1.0$, see table S2 in the supplementary material) reduces but does not eliminate this gap, indicating that the difference reflects a model-level rather than a configuration effect. This observation is relevant for instructors and students making choices about which tools to use, as it suggests that a more advanced or expensive model is not necessarily more reliable for physics tasks.

A sensitivity analysis on GPT-5.2 at three reasoning_effort levels ('none', 'medium', 'high') confirmed quantitatively what the qualitative error analysis showed for all three models: enabling reasoning substantially improves performance on items dominated by physics errors, but has little effect on items dominated by visual errors. Visual misinterpretation, rather than insufficient reasoning, is the primary bottleneck for current multimodal models on the CRCI.

The comparison with student data ($N = 267$) revealed that LLM error profiles do not reproduce the patterns of student misconceptions. When models err, they tend to converge on a single distractor with

high consistency, whereas students distribute their errors more broadly across multiple options. This indicates that AI errors and student misconceptions arise from qualitatively different mechanisms. This difference is not a limitation but a crucial point: students' reasoning processes are typically heterogeneous, context-dependent, and often internally inconsistent, and it is precisely this variability that makes their thinking diagnostically accessible and pedagogically relevant. By contrast, the higher internal coherence of AI errors reflects optimization processes that do not necessarily map onto students' cognitive structures. This point should inform both the design of assessments and the pedagogical use of chatbots.

The divergence between AI errors and student errors should be regarded as a noticeable feature rather than a gap to be closed. It emphasizes that learning relies on processes—uncertainty, partial reasoning, revision, errors and even inconsistency—that are intrinsic to human cognition. This distinction should inform both the design of assessments and the cautious, well-structured integration of AI tools in educational contexts.

For physics instructors, the practical message is clear: current chatbots can be impressively accurate on conceptual physics tasks, but their reliability is item-dependent and unpredictable. High average accuracy can mask complete failures on specific questions, and correct answers may be accompanied by flawed reasoning. These tools should be used with critical awareness, and concept inventories should be administered in supervised settings rather than as take-home assignments.

Given the rapid speed of LLM development, follow-up evaluations on the same instrument will be needed to track progress. This study and the use of a novel instrument provide a baseline for tracking future AI progress on conceptual physics tasks.

## Acknowledgment

The PhD program attended by the author C.G. is financed by the European Union—Next Generation EU, Mission 4 Component 2 CUP E66E24000030008.

## Data availability statement

The data that support the findings of this study are openly available at: the following URL/DOI: https://github.com/etufino/LLMs-performances.

Supplementary material available at: https://doi.org/10.1088/1361-6404/ae8c90/data1.

## Institutional review board statement and informed consent statement

The research conducted by Zamboni *et al* [11]—cited here and whose data have been reused in this article—was conducted in accordance with the Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects and the ICMJE guidelines on Protection of Research Participants. Approval by the local research ethics committee was not required since no medical treatment was carried out and participants were anonymous. Informed consent forms were signed by the participants before the beginning of the activities, in order to fulfil the requirements of Italian law.

## ORCID iDs

Eugenio Tufino 0000-0002-6784-3761
Caterina Giovanzana 0009-0008-0316-6029
Andrea Zamboni 0009-0006-4085-1178
Pasquale Onorato 0000-0002-2110-7582
Stefano Oss 0000-0001-9427-2151

## References

[1] Hestenes D, Wells M and Swackhamer G 1992 Force concept inventory *Phys. Teach.* **30** 141–58
[2] Madsen A, McKagan S B and Sayre E C 2017 Best practices for administering concept inventories *Phys. Teach.* **55** 530–6
[3] Kieser F, Wulff P, Kuhn J and Küchemann S 2023 Educational data augmentation in physics education research using ChatGPT *Phys. Rev. Phys. Educ. Res.* **19** 020150

[4] Wheeler S and Scherr R E 2023 ChatGPT reflects student misconceptions in physics. *2023 Physics Education Research Conf. Proc.* (American Association of Physics Teachers) pp 386–90
[5] Polverini G and Gregorcic B 2024 Performance of ChatGPT on the test of understanding graphs in kinematics *Phys. Rev. Phys. Educ. Res.* **20** 010109
[6] Polverini G, Melin J, Önerud E and Gregorcic B 2025 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
[7] Polverini G and Gregorcic B 2025 Multimodal large language models and physics visual tasks: comparative analysis of performance and costs *Eur. J. Phys.* **46** 055708
[8] Ding L, Chabay R, Sherwood B and Beichner R 2006 Evaluating an electricity and magnetism assessment tool: brief electricity and magnetism assessment *Phys. Rev. Spec. Top. Phys. Educ. Res.* **2** 010105
[9] Kortemeyer G, Babayeva M, Polverini G, Widenhorn R and Gregorcic B 2025 Multilingual performance of a multimodal artificial intelligence system on multisubject physics concept inventories *Phys. Rev. Phys. Educ. Res.* **21** 020101
[10] Beichner R J 1994 Testing student interpretation of kinematics graphs *Am. J. Phys.* **62** 750–62
[11] Zamboni A, Marzari A, Malgieri M, Onorato P and Oss S 2026 A diagnostic multiple-choice questionnaire on student misconceptions about relativity in classical mechanics *Eur. J. Phys.* **47** 015708
[12] First Page Sage 2026 First Page Sage (available at: https://firstpagesage.com/reports/top-generative-ai-chatbots/) (Accessed 26 June 2026)
[13] Larkin J H and Simon H A 1987 Why a diagram is (sometimes) worth ten thousand words *Cogn. Sci.* **11** 65–100
[14] Tong S, Liu Z, Zhai Y, Ma Y, LeCun Y and Xie S E Eyes Wide Shut? 2024 Exploring the visual shortcomings of multimodal LLM *2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)* (IEEE) pp 9568–78
[15] Google AI for Developers 2026 Gemini 3 Developer Guide. Gemini API (available at: https://ai.google.dev/gemini-api/docs/gemini-3) (Accessed 26 April 2026)
[16] OpenAI GPT-5.2 Model. OpenAI API (available at: https://developers.openai.com/api/docs/models/gpt-5.2) (Accessed 26 April 2026)
[17] McKenzie I R *et al* 2023 Inverse scaling: when bigger is not better
[18] Drory A 2019 Comoving frames and the Lorentz–Fitzgerald contraction *Am. J. Phys.* **87** 5–9
[19] Dahl F A and Østerås N 2010 Quantifying information content in survey data by entropy *Entropy* **12** 161–3
[20] Nehm R H and Kubsch M 2026 AI in contemporary scientific practice: implications for AI-integrated science education *Advancing AI in Science Education* (Springer) pp 27–48
[21] West C G 2023 AI and the FCI: can ChatGPT project an understanding of introductory Physics?
[22] Google AI for Developers 2026 Gemini thinking Google (available at: https://ai.google.dev/gemini-api/docs/thinking) (Accessed 26 April 2026)
[23] Google AI for Developers Gemini deprecations Gemini API (available at: https://ai.google.dev/gemini-api/docs/deprecations) (Accessed 26 April 2026)
[24] etufino LLMs-performances. GitHub (available at: https://github.com/etufino/LLMs-performances)