

# Multimodal large language models and physics visual tasks: comparative analysis of performance and costs

Giulia Polverini and Bor Gregorcic\*

Department of Physics and Astronomy, Uppsala University, Box 516,  
75120, Uppsala, Sweden.

## Abstract

Multimodal large language models (MLLMs) capable of processing both text and visual inputs are increasingly being explored for uses in physics education, such as tutoring, formative assessment, and grading. This study evaluates a range of publicly available MLLMs on a set of standardized, image-based physics research-based conceptual assessments (concept inventories). We benchmark 15 models from three major providers (Anthropic, Google, and OpenAI) across 102 physics items, focusing on two main questions: (1) How well do these models perform on conceptual physics tasks involving visual representations? and (2) What are the financial costs associated with their use? The results show high variability in both performance and cost. The performance of the tested models ranges from 81.5% to as low as 21%. We also found that expensive models do not always outperform cheaper ones and that, depending on the demands of the context, cheaper models may be sufficiently capable for some tasks. This is especially relevant in contexts where financial resources are limited or for large-scale educational implementation of MLLMs. By providing these analyses, our aim is to inform teachers, institutions, and other educational stakeholders so that they can make evidence-based decisions about the selection of models for use in AI-supported physics education.

**Keywords:** Multimodal large language models; Visual problem solving;  
Cost-performance analysis.

---

\*Corresponding author: [bor.gregorci@physics.uu.se](mailto:bor.gregorci@physics.uu.se)

# 1 Introduction

## 1.1 Large language models in physics education

Artificial intelligence (AI) has been playing an increasingly prominent role in education. Over the past decade, AI-driven tools have been integrated into a wide range of instructional applications, from intelligent tutoring systems to adaptive learning platforms [1–3]. AI tools promise greater scalability, personalization, and efficiency—especially in resource-constrained environments [4, 5].

*Large language models* (LLMs), which generate human-like text based on statistical patterns in large-scale datasets [6], have significantly accelerated this trend. Their improving performance has sparked growing interest in their application to subject-specific education [7, 8]—including physics, where researchers have assessed various LLMs capabilities using real-world physics and engineering exams from their educational contexts [9–13].

A growing body of research investigates how LLMs perform across a range of physics education areas. Much of this work has centered on ChatGPT [14], which has demonstrated notable capabilities across multiple physics-related tasks. Studies examining conceptual reasoning [15, 16] and problem solving [17–19] have shown that the chatbot can produce coherent and well-structured solutions, although it often still exhibits limitations in replicating human-like sensemaking. In essay-style assignments, ChatGPT-generated responses have reached grading levels comparable to high-performing university students [20], raising concerns about the validity of take-home examinations. It has also shown strengths in physics-related programming tasks [12, 21], performing reliably on structured coding tasks, and in lab-based problem solving [22, 23], analyzing experimental data and carrying out statistical analyses.

Very recently, a new class of models—often referred to as *reasoning language models* (RLMs) [24] (e.g., OpenAI’s o3 and GPT-5, DeepSeek-R1, Alibaba’s QwQ)—has been introduced to specifically enhance performance on complex, multi-step tasks. Unlike earlier chatbots, primarily developed for general linguistic fluency, RLMs are designed to imitate step-by-step reasoning, producing intermediate solution steps, applying domain-relevant procedures, and maintaining better consistency throughout the process [25]. Early evaluations suggest that these models outperform previous generations on benchmark assessments involving mathematical problem solving and STEM-focused conceptual reasoning, leading to a growing interest in their application to physics education [26, 27]. However, these capabilities do not come without important limitations, especially in high-complexity problem solving [28].

Since late 2023, LLMs have also been upgraded to *multimodal large language models* (MLLMs) [29, 30]—systems capable of processing inputs beyond text, such as images, video, and other data. While LLMs are built on a single transformer architecture optimized for sequential text processing, MLLMs integrate separate, modality-specific encoders (e.g., vision transformers for images, transformers for text). These encoders process their respective inputs and output embeddings, which are then aligned and fused—typically via cross-attention or token/feature-level fusion—into a shared representation [31]. This allows the model to process inputs (and often also generate outputs) across multiple modalities.

For physics education, the shift toward image processing is particularly relevant. Physics is a discipline that fundamentally relies on a variety of visual representations. Graphs, circuit diagrams, free-body diagrams, vector fields, sketches of experimental setups, and other representations are not merely supplementary, they are integral to physics conceptual reasoning, problem solving, and consequently also play a central role in physics education [32].

The first widely accessible model capable of processing images, ChatGPT-4, attracted early interest from the physics education research community. For instance, in one of the first published evaluations, ChatGPT-4 was tested on the Test of Understanding Graphs in Kinematics (TUG-K) [33]. While the model achieved a performance level comparable to high school students, detailed analyses revealed that its primary reason for failure was the visual misinterpretation of graphs.

Shortly after, with the rise of vision-capable models from a range of AI companies—offered at different prices and performance tiers—comparative evaluations have begun to emerge. These studies examine differences across model families and versions [34–38], languages [39], and compare free versus paid access levels [40]. These comparative studies consistently show that while models can perform well on text-only tasks, their primary bottleneck lies in interpreting visual inputs. In fact, their performance tends to degrade when tasks require image interpretation, regardless of the specific physics subdomain involved [39]. Instead, accuracy appears to hinge more on the type of demand posed by the visual representation. For example, Polverini et al. [37] highlight that ChatGPT-4o struggles with tasks involving spatial and embodied reasoning (e.g., the use of the right-hand rule).

In addition, a growing number of research studies in physics education have begun to focus on how AI-based systems can support instructors. One of the reasons is that the increasing student/teacher ratios, largely driven by the persistent shortage of qualified instructors [41], including in physics [42], have placed severe pressure on educational institutions and existing instructors. At the same time, there is growing demand for personalized learning: teachers are expected to deliver timely feedback [43], tailored guidance [44], and fair grading at scale [45], while their capacity is inherently limited [46]. Tasks such as grading [47–50], tips and feedback generation [51–53], and support for students with disabilities [54] often involve analysis of student-drawn representations, generation of explanations coupled to such visual representations, and assessing visually complex responses. MLLMs are increasingly considered as one possible solution to offload teachers from these tasks.

However, early findings indicate that while MLLMs are capable of spotting common mistakes and offering useful feedback, they often overlook the subtleties of student reasoning and struggle to deliver nuanced, differentiated assessment [55]. Addressing these shortcomings requires carefully designed prompting strategies and other methods aimed at minimizing the impact of their inherent unreliability [56], as well as identifying these models' weaknesses to avoid tasking them with operations that they are ill-suited for. Mok et al. [49] have found that an MLLM's quality of provided feedback and grading of student solutions of physics problems correlates with the system's own performance on the same problems. In other words, good subject-performance is

a prerequisite for educational value. For this reason, it is meaningful to test AIs on tasks that they will be expected to provide feedback on, grade, or tutor students on.

Despite all the recent progress, what we know about how MLLMs perform in different educational situations is still limited and often based on scattered or informal reports. Much of the available information comes from the companies that build these models using proprietary datasets and methods. As a result, it is difficult for educators to judge how well these models would actually work in real educational applications. At the same time, these tools are being widely marketed as ready-to-use solutions for education (e.g., [57, 58]), which can garner unrealistic expectations and lead to the use of AI tools in ways that may not be appropriate or effective. To use MLLMs responsibly in teaching and learning, we need independent studies of their performance in different domains of knowledge, including physics.

Alongside exploring the different models' performance, there is also the need to compare the actual cost of use for the tested models. Different MLLMs come with varying pricing, which, while often modest for individual queries, can become significant when scaled to institutional use. However, despite the growing interest in using AI tools in education, there is a lack of studies that systematically evaluate their cost-effectiveness or compare usage-based pricing across models. Understanding these costs is relevant for assessing the feasibility of deploying such models across entire classes or educational programs. Typically, model providers specify usage costs in terms of tokens (i.e., fragments of text or data units), with separate rates for input and output tokens, often expressed as cost per million tokens. However, estimating the number of tokens processed in a given query is not always straightforward [59]. Token count can be influenced not only by the length and complexity of the text but also by the presence of images (in the case of MLLMs) or extended chains of reasoning (in the case of RLMs). These factors make precise cost prediction difficult without detailed usage data, which underscores the need for transparent and accessible pricing analyses for educational planning.

This is relevant because the growing adoption of AI tools in education risks reinforcing existing technological divides [60, 61]. If the most capable models come with prohibitive costs, then only well-funded institutions will be able to afford them, leaving under-resourced schools and students at a disadvantage. This undermines efforts toward equitable access to quality education. By analyzing the performance–cost relationship, we can better understand whether lower-cost models offer sufficient educational value, and help ensure that effective AI-supported learning tools are accessible to a broader range of learners and institutions.

Multimodal AI tools thus display both promise and limitations in physics education. While early evidence suggests that MLLMs may support both teaching and learning physics, questions remain about their reliability and the cost implications of their deployment. These considerations motivate the present study.

## 1.2 Research Objectives

In this study, we benchmark several publicly available MLLMs—including some that also qualify as RLMs—on a set of conceptual physics questions that involve visual

interpretation. By evaluating models from multiple providers under the same conditions, and by pairing those results with the corresponding model usage costs, we aim to equip teachers, institutions, and other stakeholders with the information needed to choose the right tool for their needs and budgets.

We ask the following research questions:

1. *How do different MLLMs in mid-2025 perform on conceptual physics tasks that require interpretation of visual representations?*

Answering this question aims to provide an up-to-date and independent assessment of some of the most widely used MLLMs' capabilities in conceptual physics tasks where visual interpretation is essential. Understanding how the different models perform is critical for physics educators considering their use for tutoring, grading, or feedback generation purposes.

2. *What are the actual financial costs associated with running each of the tested models?*

Answering this question helps determine whether affordable models can offer sufficient performance for educational deployment. Cost is a key consideration for institutions serving large student populations and/or working within tight budgets, and exploring the performance–price trade-off is relevant for sustainable and equitable educational implementation of MLLMs.

## 2 Methodology

### 2.1 Model selection

To evaluate the performance of MLLMs on image-based physics tasks, we selected a sample of publicly available, vision-capable models from three major providers: Anthropic, Google, and OpenAI. Table 1 summarizes the models included in our study, along with their declared token-based pricing at the time of data collection.

Our selection was guided by the following criteria. First, we included only models that are multimodal and accessible via an application programming interface (API). Second, we aimed to cover a wide performance and pricing spectrum, including both premium-tier systems (e.g., Claude Opus 4, Gemini 2.5 Pro, GPT-5 and o3) and lighter alternatives (e.g., Claude Haiku 3.5, GPT-5 mini, GPT-5 nano, the GPT-4.1 series, and Gemini 2.0 Flash). This also allowed us to explore the cost-effectiveness of different models for potential educational deployment.

Considering potential trade-offs is relevant for accessibility and sustainability reasons. For example, in educational contexts, differences in access to more capable (and expensive) models may contribute to a technological divide, where organizations and individuals with the means to afford more expensive models may be in an advantageous position. Furthermore, if the performance of less expensive and resource-demanding models is as good as that of those that use more resources, it is environmentally responsible to use the less resource-hungry model.

**Table 1** Descriptions of the selected AI models. Prices are reported in USD per million tokens.

AI Company	Name	Model	Input Price	Output Price
Anthropic [62]	Claude Opus 4	claude-opus-4-20250514	15.00	75.00
	Claude Sonnet 4	claude-sonnet-4-20250514	3.00	15.00
	Claude Haiku 3.5	claude-3-5-haiku-20241022	0.80	4.00
Google [63, 64]	Gemini 2.5 Pro	gemini-2.5-pro-preview-06-05	1.25	10.00
	Gemini 2.5 Flash	gemini-2.5-flash-preview-05-20	0.30	2.50
	Gemini 2.0 Flash	gemini-2.0-flash	0.10	0.40
	Gemma 3-27b	gemma-3-27b-it	0	0
	Gemma 3-4b	gemma-3-4b-it	0	0
OpenAI [65]	GPT-5	gpt-5-2025-08-07	1.25	10.00
	GPT-5 mini	gpt-5-mini-2025-08-07	0.25	2.00
	GPT-5 nano	gpt-5-nano-2025-08-07	0.05	0.40
	o3	o3-2025-04-16	2.00	8.00
	o4 mini	o4-mini-2025-04-16	1.10	4.40
	GPT-4.1	gpt-4.1-2025-04-14	2.00	8.00
	GPT-4.1 mini	gpt-4.1-mini-2025-04-14	0.40	1.60
	GPT-4.1 nano	gpt-4.1-nano-2025-04-14	0.10	0.40
	GPT-4o	gpt-4o-2024-11-20	2.50	10.00

## 2.2 Task selection

To evaluate model performance on image-based conceptual physics tasks, we selected established concept inventories<sup>1</sup>. Physics concept inventories are research-based multiple-choice assessment tools, developed to probe students’ conceptual understanding of a topic. Each item of the tests typically presents a question and a number of response options. Their standardized structure, focus on conceptual reasoning, often coupled with physics visual representations, makes them well-suited for assessing the capabilities of AI systems in solving visually rich physics tasks. Fig. 1 shows an example of a test item as it was submitted.

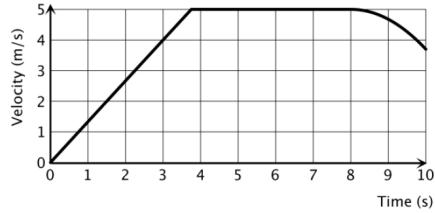
For this study, we opted for four tests that cover several areas of undergraduate physics: kinematics (TUG-K), electromagnetism (BEMA), quantum mechanics (QMVI), and geometrical optics (FTGOT). More details about the used inventories are summarized in Table 2. These tests are well-validated within the physics education research community and widely used across institutions. As such, they represent canonical measures of conceptual understanding, rather than institution-specific assessments. Additionally, they are almost entirely image-based, which aligns with our study’s focus of evaluating the visual processing abilities of MLLMs: 100 out of 102 total items require interpretation of physics visual representations.

However, it is important to clarify that this study does not aim to conduct a fine-grained analysis of model performance across the selected physics subdomains. We are not trying to compare, for instance, how well models perform in mechanics versus

<sup>1</sup>In order to protect the tests’ integrity, we do not share them. To facilitate the interpretation of this paper, we suggest readers access the tests from the [PhysPort](#) website.

4. An elevator moves from the basement to the tenth floor of a building. The mass of the elevator is 1000 kg and it moves as shown in the velocity-time graph below. How far does it move during the first three seconds of motion?

- A) 0.75 m
- B) 1.33 m
- C) 4.0 m
- D) 6.0 m
- E) 12.0 m



**Fig. 1** An example of one of the 102 test item screenshots submitted to the models. It consists of textual and image parts, as well as a multiple-choice answer list. Image adapted from [33] under the CC-BY 4.0 license.

**Table 2** Presentation of the selected concept inventories, including the number of items and a description of the types of visual representations involved.

Concept Inventory	Number of items	Description of visual representations
<b>BEMA</b> (Brief Electricity and Magnetism Assessment) [66]	31	Circuit diagrams, electric field lines, charge distributions, and force vectors. Requires interpretation of symbolic visuals and linking them to field concepts, often in three dimensions.
<b>FTGOT</b> (Four-Tier Geometrical Optics Test) [67]	20	Ray diagrams and schematic setups involving lenses, mirrors, and light paths. Demands geometrical and spatial reasoning, including estimating distances and angles, tracing ray behavior, and perspective shifts.
<b>QMVI</b> (Quantum Mechanics Visualization Instrument) [68]	25	Potential energy diagrams, wave functions, probability densities, and energy levels. Requires abstract conceptual mapping between representations. Main focus lies on integrating symbolic and visual representations of quantum phenomena.
<b>TUG-K</b> (Test of Understanding Graphs in Kinematics) [69]	26	Graphs of position, velocity, and acceleration over time. Items typically involve reading or interpreting relationships in line graphs. Involves translating between graphical and linguistic representations of motion, with a focus on slope and area interpretation.

electromagnetism. This is because our previous observations suggest that performance is not directly linked to the content area itself, but to the type of conceptual reasoning necessary to analyse the visual representations used in a task [40]. In other words, it is the nature of the visual input—how the model must interpret and reason through it—that plays a more central role than the specific topic.

That being said, there is often an indirect connection: certain types of visual representations tend to be associated with particular areas of physics. See Table 2 for further context: some tests rely heavily on spatial or geometric reasoning, while others emphasize graphical interpretation or symbolic-visual integration. This overlap can give the impression that performance is tied to the subject domain.

Disentangling this correlation is beyond the scope of our study. Instead, we treat the four test sets as components of a single benchmark that collectively represents a wide range of visual formats used in physics education. While we occasionally report results by individual test, this is done mainly to illustrate that performance differences do exist and may warrant further qualitative analysis. A deeper exploration of model behavior within each subdomain is an important next step—but is not the focus of this work.

### 2.3 Data collection

We captured a screenshot of each item from each test, including the question, the multiple-choice options, and any associated images. For most items, this process was straightforward. However, in the BEMA test a few items shared the same image across multiple questions (16 out of 31). In these cases, we manually separated the items and recreated them so each could stand alone. This adjustment was purely graphical: we simply duplicated the shared image and paired it with each corresponding question. Since none of the items depended on the answers to previous ones, this did not affect the integrity of the questions. Additionally, we edited all FTGOT items to standardize their format. Each original FTGOT item includes four parts: (1) the main question with a multiple-choice list, (2) a confidence rating for that response, (3) a follow-up multiple-choice question asking to explain the initial answer, and (4) another confidence rating related to the explanation. Since our analysis did not focus on reasoning or confidence levels, we excluded parts 2, 3, and 4 from our evaluation. As a result, each FTGOT item was edited to include only the core question, the answer options, and any accompanying images.

Each item of the selected inventories was presented to the models in the form of a screenshot, simulating a visual input scenario aligned with typical student-facing materials.

Every item was submitted 10 times independently in a new context window to each model. This repetition count was chosen based on prior experimentation indicating that MLLMs now mostly exhibit high response consistency across runs. Previous research have demonstrated that models tend to either consistently answer an item correctly or repeatedly select the same incorrect option [37, 40]. The use of 10 iterations thus balances statistical reliability with cost and environmental sustainability, as earlier protocols involving larger sample sizes are no longer necessary for capturing stable performance patterns.

Inputs were submitted through official APIs. The temperature parameter was set to 0.7 whenever possible to standardize response variability and reflect a moderate level of generative randomness across conditions [70]. For OpenAI’s reasoning models (o3 and o4-mini), the temperature parameter is not modifiable. For GPT-5 series models, the parameter “reasoning effort” was set to medium. A Python-based script

was used to automate the process and record responses into json files [71]. Answers (selected letter options appearing at the end of the responses) were extracted into csv files for easier processing. This ensured uniformity across trials and streamlined data collection across models.

To preserve the validity of performance comparisons across models, we deliberately avoided prompt engineering techniques. Each prompt consisted solely of a minimal instruction requesting a clear and structured response:

*Answer the question in the image. If none of the options are correct, answer with the letter N. In a separate last line of the response, restate the answer in the following form: Answer: letter*

The instruction to use the letter *N* when none of the options seem correct was added to reduce artificial inflation of accuracy due to guessing. While students in testing contexts might select an answer regardless of confidence, we sought to focus on the reasoning capabilities of the models rather than the selection of a letter by chance.

This approach was chosen to minimize potential confounding effects introduced by variations in prompt phrasing—a practice that remains largely empirical and difficult to standardize [6]. In particular, we avoided strategies such as few-shot examples [72] or explicit Chain-of-Thought (CoT) cues [73]. As current models often engage in CoT reasoning by default, we focused instead on capturing baseline model behavior.

## 2.4 Scoring and analysis

Model responses were coded as either correct or incorrect, based on the final selected answer option. We did not apply any conditional grading, even though it was suggested in a small number of items on one of the four assessments (i.e., BEMA). Answers containing the letter *N* were considered incorrect.

Thanks to the consistent output format requested in the prompt, final answers could be efficiently parsed using the last line of the model’s response (e.g., “Answer: B”). However, minor post-processing was required in a small number of cases where the model restated the full text of the chosen answer instead of the corresponding letter. These responses were manually reviewed and coded according to the appropriate option. In no cases did models return multiple selections or unrelated text.

To verify response quality, both authors independently reviewed a randomly selected subset (approximately 30%) of all responses. These checks confirmed that the models consistently followed the requested response format and that no systematic issues were present.

We computed the performance of each model on each concept inventory by first calculating the percentage of correct responses for each item (based on 10 repeated runs), and then averaging these per-item scores across the full inventory. Each concept inventory contained a fixed number of items (102 in total across all tests), and all models completed all items without submission failures. This item-level averaging approach ensures that each question contributes equally to the final score, regardless of individual item difficulty or model consistency. In addition to the average accuracy, we computed the standard deviation (SD) and standard error of the mean (SEM) for

each model on each test. These were calculated from the 10-run distribution for each item and then aggregated across items, providing a measure of response variability and confidence in the estimated performance levels. All results are reported as raw percentages, without normalization or scaling.

Importantly, this study does not analyze the reasoning or explanations provided in the model outputs. While responses often included extended justifications, our scoring procedure was based solely on the final selected answer. This means we did not assess whether the model’s reasoning was scientifically accurate, coherent, or aligned with its answer. As such, it is entirely possible that a model selected the correct option for incorrect reasons—or vice versa. This design choice reflects the quantitative nature of our analysis, which aims to establish baseline performance levels and relate them to practical considerations such as model cost and accessibility. A full qualitative analysis of reasoning accuracy, coherence, or potential biases would require a different methodological framework and falls outside the scope of this study. For the same reason, we do not offer detailed interpretations of why models performed differently across the four concept inventories. While we propose potential explanations in light of previous research, these remain speculative and are not derived directly from our results.

We determined the cost for running each model by tracking the number of input and output tokens for each model and multiplying them by the listed per-token prices. In reporting the model costs, we normalized the cost to represent the expected cost of running our benchmark on the model a single time—i.e., submitting all 102 items once and receiving one response for each item. The costs reported in this paper include both input and output costs. Calculating the expected input costs was straightforward; we simply multiplied the number of tokens for each input (which was the same for all 10 iterations of an item on a given model) with the model-specific input price per token (see Table 1 for token-based prices). For calculating the output costs, we took the average number of tokens outputted by a model for a given item across all 10 iterations, and summed these item averages across all items for each model to obtain the expected number of tokens for the entire benchmark. Multiplying this number with the price per output token for the model gave us the expected output cost of running the entire benchmark for that model. For reasoning models, reasoning tokens are counted and charged as output tokens, even though they are not directly accessible to the user. In these models, reasoning tokens represented most of the cost.

## 3 Results

### 3.1 Performance analysis

The results in Table 3 show performance outcomes for a set of MLLMs tested across four physics concept inventories. For each model, the table reports percentage accuracy (Perc), along with standard deviation (SD) and standard error of the mean (SEM) based on 10 repeated runs per item. Each item’s performance was treated as an independent Bernoulli variable with an experimentally determined success probability (item score). SD was calculated by taking the square root of the sum of item score variances, and SEM was calculated by dividing the SD by the square root of 10

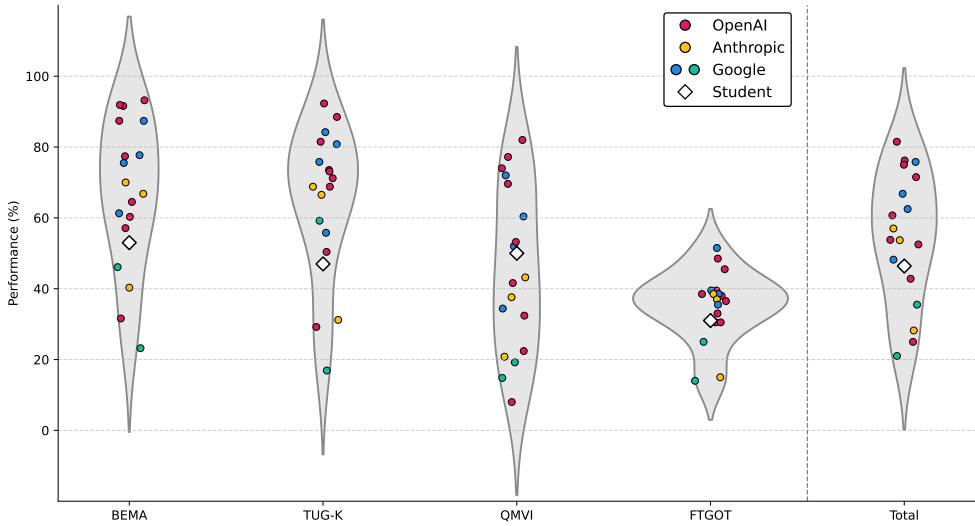
(number of repetitions of each item). Note that the SD of the performance of each model is a consequence of the variability of the model output and is not expected to decrease with further increasing number of repetitions of each item. The SEM, on the other hand, would further decrease.

Overall, there is substantial variation in performance across models. GPT-5 was the best performing model (81.5%). o3, Gemini 2.5 Pro and GPT-5 mini yielded high average scores (76.2%, 75.8%, and 75.0% respectively), followed closely by o4 mini (71.5%). Models such as Gemini 2.5 Flash (66.8%), GPT-5 nano (60.7%), Claude Opus 4 (57.0%), GPT-4.1 mini (53.8%), Claude Sonnet 4 (53.7%), and GPT-4.1 (52.5%) fall in the middle range. Others, like Claude Haiku 3.5 (28.2%), GPT-4.1 nano (25.0%), and Gemma 3-4b (21.0%) scored considerably lower. The range between the highest and lowest average model score exceeds 60 percentage points, highlighting the breadth of performance observed across the models.

**Table 3** Percentage performance (Perf), standard deviation (SD), and standard error of the mean (SEM) for each model on the selected concept inventories. Models are ordered by decreasing total score (Tot AI). The last row represents the average performance of all MLLMs on each concept inventory constituting the total benchmark.

Model		BEMA	TUG-K	QMVI	FTGOT	Tot AI
GPT-5	Perf	<b>93.2</b>	<b>92.3</b>	<b>82.0</b>	<b>48.5</b>	<b>81.5</b>
	SD	3.4	3.7	4.8	7.4	2.3
	SEM	1.1	1.2	1.5	2.3	0.7
o3	Perf	<b>87.4</b>	<b>88.5</b>	<b>74.0</b>	<b>45.5</b>	<b>76.2</b>
	SD	4.4	3.9	6.0	6.5	2.6
	SEM	1.4	1.2	1.9	2.1	0.8
Gemini 2.5 Pro	Perf	<b>87.4</b>	<b>84.2</b>	<b>72.0</b>	<b>51.5</b>	<b>75.8</b>
	SD	3.8	5.1	5.8	7.0	2.6
	SEM	1.2	1.6	1.8	2.2	0.8
GPT-5 mini	Perf	<b>91.6</b>	<b>81.5</b>	<b>77.2</b>	<b>38.0</b>	<b>75.0</b>
	SD	4.4	5.8	4.0	6.2	2.6
	SEM	1.4	1.8	1.2	2.0	0.8
o4 mini	Perf	<b>91.9</b>	<b>73.5</b>	<b>69.6</b>	<b>39.5</b>	<b>71.5</b>
	SD	4.0	4.6	4.2	7.1	2.4
	SEM	1.3	1.4	1.3	2.3	0.8
Gemini 2.5 Flash	Perf	<b>77.7</b>	<b>80.8</b>	<b>60.4</b>	<b>39.5</b>	<b>66.8</b>
	SD	4.7	6.1	7.0	5.7	3.0
	SEM	1.5	1.9	2.2	1.8	0.9
Gemini 2.5 Flash (no reasoning)	Perf	<b>75.5</b>	<b>75.8</b>	<b>52.0</b>	<b>38.5</b>	<b>62.5</b>
	SD	4.4	6.5	6.8	7.5	3.1
	SEM	1.4	2.1	2.1	2.4	1.0
GPT-5 nano	Perf	<b>77.4</b>	<b>71.2</b>	<b>53.2</b>	<b>30.5</b>	<b>60.7</b>
	SD	4.9	5.8	6.4	8.6	3.1
	SEM	1.5	1.8	2.0	2.7	1.0
Claude Opus 4	Perf	<b>70.0</b>	<b>68.8</b>	<b>43.2</b>	<b>38.5</b>	<b>57.0</b>
	SD	4.9	5.1	5.9	7.9	2.9
	SEM	1.5	1.6	1.9	2.5	0.9
GPT-4.1 mini	Perf	<b>64.5</b>	<b>68.8</b>	<b>41.6</b>	<b>33.0</b>	<b>53.8</b>
	SD	4.8	6.9	6.9	8.0	3.2
	SEM	1.5	2.2	2.2	2.5	1.0
Claude Sonnet 4	Perf	<b>66.8</b>	<b>66.5</b>	<b>37.6</b>	<b>37.0</b>	<b>53.7</b>
	SD	4.2	6.0	6.7	7.0	2.9
	SEM	1.3	1.9	2.1	2.2	0.9
GPT-4.1	Perf	<b>60.3</b>	<b>73.1</b>	<b>32.4</b>	<b>38.5</b>	<b>52.5</b>
	SD	4.2	5.5	6.5	5.0	2.7
	SEM	1.3	1.7	2.1	1.6	0.8
Gemini 2.0 Flash	Perf	<b>61.3</b>	<b>55.8</b>	<b>34.4</b>	<b>35.5</b>	<b>48.2</b>
	SD	4.6	6.4	6.3	6.9	3.0
	SEM	1.4	2.0	2.0	2.2	0.9
GPT-4o	Perf	<b>57.1</b>	<b>50.4</b>	<b>22.4</b>	<b>36.5</b>	<b>42.8</b>
	SD	5.0	5.6	6.4	7.5	3.0
	SEM	1.6	1.8	2.0	2.4	0.9
Gemma 3-27b	Perf	<b>46.1</b>	<b>59.2</b>	<b>14.8</b>	<b>14.0</b>	<b>35.5</b>
	SD	4.8	5.5	3.7	4.6	2.4
	SEM	1.5	1.7	1.2	1.5	0.8
Claude Haiku 3.5	Perf	<b>40.3</b>	<b>31.2</b>	<b>20.8</b>	<b>15.0</b>	<b>28.2</b>
	SD	6.0	5.4	5.8	5.3	2.9
	SEM	1.9	1.7	1.8	1.7	0.9
GPT-4.1 nano	Perf	<b>31.6</b>	<b>29.2</b>	<b>8.0</b>	<b>30.5</b>	<b>25.0</b>
	SD	6.0	7.5	5.2	7.6	3.3
	SEM	1.9	2.4	1.6	2.4	1.0
Gemma 3-4b	Perf	<b>23.2</b>	<b>16.9</b>	<b>19.2</b>	<b>25.0</b>	<b>21.0</b>
	SD	3.5	2.7	3.2	0.0	1.5
	SEM	1.1	0.9	1.0	0.0	0.5
Avg AI (per concept inventory)	Perf	<b>66.7</b>	<b>64.4</b>	<b>44.4</b>	<b>35.1</b>	—
	SD	4.6	5.5	5.7	6.3	—
	SEM	1.4	1.7	1.8	2.0	—

Performance also varied across concept inventories (Fig. 2). The average performance across all tested models on BEMA (66.7%) and TUG-K (64.4%) was higher than on QMVI (44.4%) and FTGOT (35.1%). These patterns are consistent across most models. For example, GPT-5 performed above 90% for BEMA and TUG-K, about 10 percentage points lower on QMVI, and less than 50% on FTGOT. Following a similar pattern, o3 and Gemini 2.5 Pro exceeded 84% on both BEMA and TUG-K, while none of them surpassed 75% on QMVI or 52% on FTGOT. Across inventories, the maximum difference between a given model’s highest and lowest test score ranged from 30 to 50 percentage points, suggesting some inventories (especially FTGOT) have higher difficulty even for high-performing models.



**Fig. 2** Distribution of model scores across the selected inventories, together with the average distribution (right-most violin). Individual models are color-coded according to their developer. The full set of data is available in Table 3. For reference, post-instruction university student performance is marked with a white diamond. Student data from [67, 68, 74, 75].

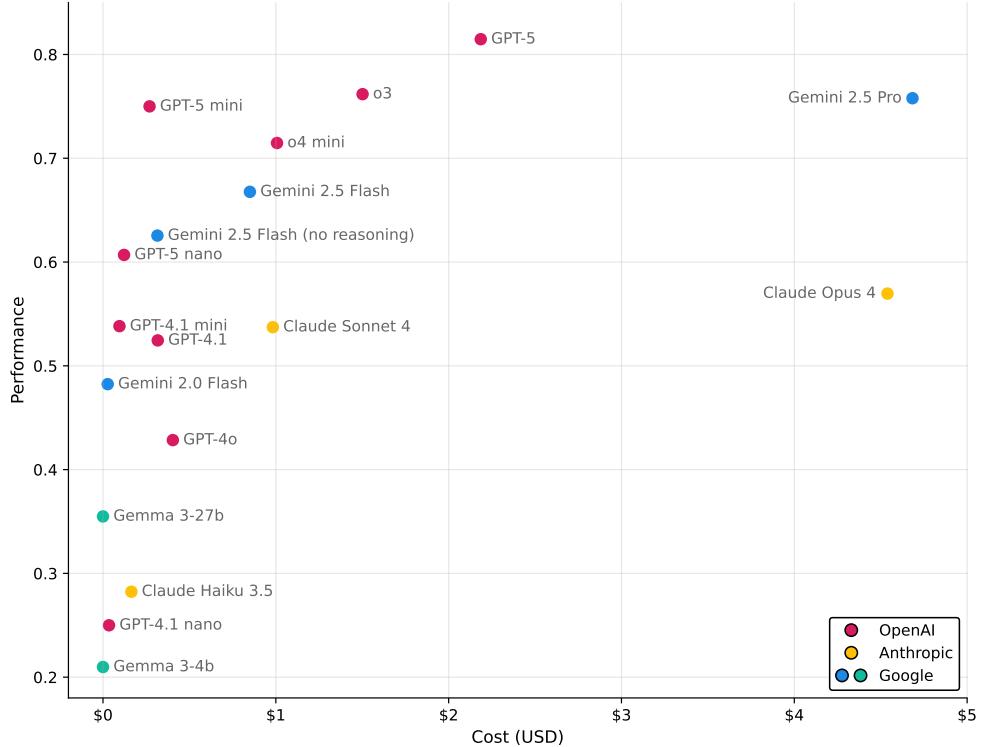
The statistical measures provide additional detail. SDs were mostly between 3 and 7 percentage points of the total available points, with most SEM values below 2.5%. In many cases, higher-scoring models exhibited both higher average performance and lower spread of scores, indicating greater consistency across runs. Some lower-performing models also showed relatively low variability of scores, suggesting consistent but inaccurate responses. In contrast, several mid-range models displayed greater variability, reflecting a mix of correct and incorrect answers across trials.

Looking at performance within model families, some patterns emerge. Among the OpenAI models, GPT-5, o3, GPT-5 mini, o4 mini and even GPT-5 nano, performed better than GPT-4.1 and GPT-4.1 mini, while GPT-4.1 nano scored lowest. Similarly, for Google’s models, Gemini 2.5 Pro outperformed Gemini 2.5 Flash, and Gemini 2.5 Flash outperformed the earlier Gemini 2.0 Flash, as well as Gemma. For Anthropic,

Claude Opus 4 performed better than Claude Sonnet 4, while Claude Haiku 3.5 scored lowest in that family. While exact differences vary, in most cases each successive tier within a provider correlates with a performance increase of 10–20 percentage points.

### 3.2 Cost analysis

To better understand the relationship between model performance and usage cost, we plotted each model’s average score against its cost. The resulting plot in Fig. 3 reveals a few patterns. For a more detailed breakdown of the costs, see Table 4.



**Fig. 3** Tested MLLMs’ distribution of their total performance on the benchmark (vertical axis) and the average cost of running the benchmark once for each model (horizontal axis).

First, while there is an overall correlation between cost and performance, the relationship is not proportional, and there are some interesting outliers. GPT-5 is the best performing model with an accuracy score of 81.5% at the cost of \$2.18. Some of the other relatively well-performing models, such as o3 and o4 mini, are positioned in the cost range between \$1.00 and \$1.50. Despite not being the most expensive, they achieved accuracy scores of 76.2% and 71.5%, respectively, which is comparable to the much more costly Gemini 2.5 Pro model (\$4.68), with a performance of 75.8%. The

clear outlier here is GPT-5 mini, with the accuracy score of 75% at an outstandingly low cost of \$0.26. On the other hand, Claude Opus 4, only marginally cheaper than Gemini 2.5 Pro, achieved only 57.0% on the benchmark, significantly underperforming relative to its cost.

Second, a cluster of low-cost, low-performing models is also evident. For instance, the Gemma models, which are freely available (no cost to the user), occupy the bottom of the performance scale (21.0–35.5%). Although they may be valuable in other contexts, they are currently not competitive for physics conceptual tasks requiring visual interpretation.

Third, several models strike a compelling balance between affordability and capability. Gemini 2.5 Flash, when the “reasoning is enabled”, performs in the high 60% range while maintaining the cost just below \$1.00. Even more interesting is the performance of Gemini 2.5 Flash with “reasoning disabled”. It achieved a performance in the lower 60% range at the cost of only \$0.31, making it an interesting contender for low-cost uses. For less demanding uses or tight budgets, GPT-5 nano appears to be a good candidate, with low cost and performance just above 60%. Note that this is 38 times cheaper than Claude Opus 4, which performs at only 57%. Interestingly, GPT-4.1 mini outperformed its bigger brother, GPT-4.1, and displayed performance in the mid-50% range at the price of only \$0.10. Once again, it is notable that GPT-5 mini is in a class of its own at this price point, performing about 10 percentage points better than other models at the cost of \$0.27.

Finally, grouping by provider reveals broader trends. OpenAI’s models tend to cluster in the upper-left quadrant—combining solid performance with reasonable costs. Google’s lineup is more varied, with models ranging from the very strong but expensive Gemini 2.5 Pro, to the underperforming but completely free Gemma models. Anthropic’s Claude models fall in the middle or lower performance tiers, despite being relatively expensive, suggesting a less favorable price-to-performance ratio for this particular task set.

Overall, these results suggest that model selection for educational applications should be guided by empirical performance data rather than assumptions based on provider reputation or listed per-token prices. Lower-cost models may, in some cases, offer equal, comparable, or even superior performance on physics conceptual tasks involving visual representations.

## 4 Discussion, limitations and future work

The analysis of performance shows that the best-performing MLLMs are already outperforming post-instruction university student averages on the four tested concept inventories. Some of them are coming close to expert-like performance on some of the inventories. The top-tier models from OpenAI and Google are now exceeding 80% and 75% accuracy, respectively, on our benchmark, suggesting real potential for educational deployment. However, the non-uniform performance across the four underlying concept inventories suggests that even the best models still struggle with many items. While the tested models collectively performed quite well on BEMA and TUG-K, their scores dropped significantly on QMVI and, most notably, FTGOT. These findings

**Table 4** Token usage and cost by model. Token usage gives the average number of tokens (input, output, and reasoning, where applicable) for one iteration of the 102 items in the benchmark test. Costs are given in USD. The final column shows the total cost of running all the 102 items once for each model.

Model	Total input		Total output		SUM		
	<i>Token<sub>in</sub></i>	<i>Cost<sub>in</sub></i>	<i>Token<sub>rea</sub></i>	<i>Cost<sub>rea</sub></i>	<i>Token<sub>out</sub></i>	<i>Cost<sub>out</sub></i>	
GPT-5	82622	0.103	204237	2.042	4016	0.040	2.186
o3	88152	0.176	159770	1.278	5927	0.047	1.502
Gemini 2.5 Pro	30600	0.038	403024	4.030	61533	0.615	4.684
GPT-5 mini	148923	0.037	112128	0.224	3949	0.008	0.269
o4 mini	211157	0.232	168288	0.741	7763	0.034	1.007
Gemini 2.5 Flash	30600	0.009	264592	0.662	71756	0.179	0.850
Gemini 2.5 Flash (no reasoning)	30600	0.009	—	—	122267	0.306	0.315
GPT-5 nano	184807	0.009	278515	0.111	2880	0.001	0.122
Claude Opus 4	138554	2.078	—	—	32812	2.461	4.539
GPT-4.1 mini	159056	0.064	—	—	19759	0.032	0.095
Claude Sonnet 4	138554	0.416	—	—	37799	0.567	0.983
GPT-4.1	99314	0.199	—	—	14777	0.118	0.317
Gemini 2.0 Flash	231738	0.023	—	—	9837	0.004	0.027
GPT-4o	99314	0.248	—	—	15585	0.156	0.404
Gemma 3-27b	30600	0	—	—	—	0	0
Claude Haiku 3.5	138702	0.111	—	—	13324	0.053	0.164
GPT-4.1 nano	299810	0.030	—	—	12009	0.005	0.035
Gemma 3-4b	30600	0	—	—	—	0	0

suggest a need for future research focused on analyzing how specific visual formats and task types impact model performance, as well as the need for qualitative studies examining the reasoning behind MLLM-selected answers.

The cost analysis reveals that performance does not scale linearly with costs. While some of the most capable models remain relatively cost-effective (e.g., GPT-5 mini, o4 mini and o3 mini), others—like Claude Opus 4—underperformed despite high costs. This suggests that institutions cannot rely solely on pricing tiers (cost per token) or provider reputation as proxies for quality across different educational settings. Conversely, certain mid-range or low-cost models offer a compelling balance of performance and affordability. In particular, GPT-5 mini stands out as a well-performing and low-cost model. GPT-5 nano and Gemini 2.5 Flash with “disabled reasoning” also achieved over 60% average accuracy at a fraction of the cost of more expensive models. Our results thus indicate that more affordable models can reach performance levels that are close to those of the best performing models. This has important implications for deploying AI in schools or universities operating under financial constraints. However, freely available or open-weight models (e.g., Gemma 3 series) currently perform well below acceptable thresholds for educational use on physics tasks containing images.

These models may be interesting for other roles, but are not yet suitable for student-facing educational applications or assistance with assessment or grading when physics images are involved.

While this study offers a broad and comparative view of MLLM performance on conceptual physics tasks involving images, it also has several limitations.

First, our analysis focused exclusively on multiple-choice items from well-established concept inventories. While this design allows for standardization and comparability, it does not capture how MLLMs perform on more open-ended or ill-structured physics tasks, including derivations, written explanations, or lab-based data analysis that better represent authentic assessment practices in physics classrooms. Future research should expand to include these formats, which are common in authentic classroom and assessment settings.

Second, our evaluation was quantitative in nature. We scored responses based solely on the selected answer choices, without analyzing the correctness or coherence of the generated output text. This leaves open the possibility that models selected correct answers for the wrong reasons—or, conversely, generated valid reasoning but selected incorrect options [33, 37]. A qualitative investigation of model-generated explanations would be a valuable next step to better understand reasoning quality and error patterns. Some models (e.g. OpenAI’s reasoning models) offer so-called “reasoning summaries” as an optional part of their output [76]. A reasoning summary is a condensed recap of the content of the reasoning that the model did in the background, as it employed the internal CoT process hidden from the user, to solve the task at hand. Future research could qualitatively investigate these “reasoning summaries” to get better insights into the strengths and weaknesses of these models. Such analyses could also help determine which types of visual tasks pose the greatest challenges, and why certain inventories consistently produce lower scores even for otherwise high-performing models. For example, follow-up studies could attempt to disentangle model strengths and weaknesses in spatial, symbolic, and graphical reasoning. This can be done by collecting new data, or re-analyzing the data collected in this project [71].

Third, all the models were evaluated using a static, minimal prompt. We intentionally avoided further prompt engineering to reflect default user scenarios and preserve comparability. However, this likely underestimates the full potential of certain models, particularly those responsive to CoT prompting or domain-specific scaffolding. Future studies could explore the impact of tailored prompts. It is important to note, though, that OpenAI explicitly advises against CoT prompting for its reasoning models (e.g., GPT-5 series, o3, o4-mini), stating that the approach is likely to be ineffective or can even worsen the models’ performance [77]. For “non-reasoning” models, on the other hand, performance might benefit from more specialised CoT prompting. However, this would likely also lead to an increased number of generated tokens and consequently a higher cost. Systematic evaluation is needed to determine if CoT prompting with non-reasoning models is economically preferable to running reasoning models. Another consideration that can be relevant from an educational perspective is the transparency of the model output. Using models that do not hide part of their output from the user can be preferred, especially when the focus lies on the process of getting to a solution, not just the final answer. In this respect, educators need to experiment with

different models and prompting approaches to achieve the desired behaviour for their particular use case (e.g., tutoring, grading, feedback).

Finally, this study represents a snapshot in time. The capabilities, pricing, and availability of MLLMs are evolving rapidly, and new model releases or fine-tuned educational variants may soon outperform those tested here. Maintaining up-to-date benchmarks and developing open testing protocols will continue to be important for tracking progress and supporting informed decision-making.

By addressing these limitations and building on our current findings, future research can deepen our understanding of how and when MLLMs can meaningfully contribute to physics education—and where caution, adaptation, or complementary approaches remain necessary.

## 5 Conclusion

This study provides a comparative evaluation of a selection of publicly available MLLMs on conceptual physics tasks requiring visual interpretation. By benchmarking both performance and cost across multiple concept inventories, we highlight critical differences among models that are not apparent from pricing or provider claims alone.

Our findings suggest that some MLLMs now approach expert-level accuracy on certain physics concept inventories in domains such as kinematics and electromagnetism. However, performance drops significantly on tasks involving complex spatial or abstract reasoning, particularly in geometrical optics. On the other hand, the analysis shows that high performance does not necessarily come with high cost. Several models offer favorable cost–performance ratios, making them viable options for educational deployment, including in resource-constrained settings. Conversely, some of the most expensive models underperformed, suggesting that informed model selection is crucial.

As MLLMs continue to improve and gain traction in educational applications, physics educators and institutions must continuously and critically evaluate them, and carefully consider the balance of capability, cost, and context-specific needs. This study offers an example of an evaluation that can help physics educators in this process.

## References

- [1] Gligore, I., Cioca, M., Oancea, R., Gorski, A.-T., Gorski, H., Tudorache, P.: Adaptive learning using artificial intelligence in e-learning: A literature review. *Education Sciences* **13**(12), 1216 (2023) <https://doi.org/10.3390/educsci13121216> EISSN 2227-7102
- [2] Ahmad, K., Iqbal, W., El-Hassan, A., Qadir, J., Benhaddou, D., Ayyash, M., Al-Fuqaha, A.: Data-driven artificial intelligence in education: A comprehensive review. *IEEE Transactions on Learning Technologies* **17**, 12–31 (2024)
- [3] Liu, V., Latif, E., Zhai, X.: Advancing education through tutoring systems: A systematic literature review (2025) [arXiv:2503.09748](https://arxiv.org/abs/2503.09748)

- [4] Aderibigbe, A.O., Ohenhen, P.E., Nwaobia, N.K., Gidiagba, J.O., Ani, E.C.: Artificial intelligence in developing countries: Bridging the gap between potential and implementation. *Computer Science & IT Research Journal* **4**(3), 185–199 (2023)
- [5] Božić, V.: Artificial intelligence as the reason and the solution of digital divide. *Language Education & Technology (LET Journal)* **3**(2), 96–109 (2023)
- [6] Polverini, G., Gregorcic, B.: How understanding large language models can inform the use of chatgpt in physics education. *European Journal of Physics* **45**(2), 025701 (2024) <https://doi.org/10.1088/1361-6404/ad1420>
- [7] Zeng, Z., Chen, P., Liu, S., Jiang, H., Jia, J.: Mr-gsm8k: A meta-reasoning benchmark for large language model evaluation (2023) [arXiv:2312.17080](https://arxiv.org/abs/2312.17080)
- [8] Xuan, W., Yang, R., Qi, H., Zeng, Q., Xiao, Y., Feng, A., Liu, D., Xing, Y., Wang, J., Gao, F., Lu, J., Jiang, Y., Li, H., Li, X., Yu, K., Dong, R., Gu, S., Li, Y., Xie, X., Juefei-Xu, F., Khomh, F., Yoshie, O., Chen, Q., Teodoro, D., Liu, N., Goebel, R., Ma, L., Marrese-Taylor, E., Lu, S., Iwasawa, Y., Matsuo, Y., Li, I.: Mmlu-prox: A multilingual benchmark for advanced large language model evaluation (2025) [arXiv:2503.10497](https://arxiv.org/abs/2503.10497)
- [9] Tschisgale, P., Maus, H., Kieser, F., Kroehs, B., Petersen, S., Wulff, P.: Evaluating gpt- and reasoning-based large language models on physics olympiad problems: Surpassing human performance and implications for educational assessment. *Physical Review Physics Education Research* **21**(2), 020157 (2025) <https://doi.org/10.1103/6fmnx-bsnl>
- [10] Yeadon, W., Hardy, T.: The impact of ai in physics education: A comprehensive review from gcse to university levels. *Physics Education* **59**(2), 025010 (2024) <https://doi.org/10.1088/1361-6552/ad1fa2>
- [11] Dao, X.-Q., Le, N.-B., Phan, X.-D., Ngo, B.-B., Vo, T.-D.: Evaluation of chatgpt and microsoft bing ai chat performances on physics exams of vietnamese national high school graduation examination (2023) [arXiv:2306.04538](https://arxiv.org/abs/2306.04538)
- [12] Kortemeyer, G.: Could an artificial-intelligence agent pass an introductory physics course? *Physical Review Physics Education Research* **19**(1), 010132 (2023) <https://doi.org/10.1103/PhysRevPhysEducRes.19.010132>
- [13] Frenkel, M., Emara, H.: Chatgpt & mechanical engineering: Examining performance on the fe mechanical engineering and undergraduate exams (2023) [arXiv:2309.15866](https://arxiv.org/abs/2309.15866)
- [14] OpenAI: ChatGPT: Overview. <https://openai.com/chatgpt/overview/>. Accessed 2025-08-19 (2025)
- [15] Gregorcic, B., Polverini, G., Sarlah, A.: Chatgpt as a tool for honing teachers'

- socratic dialogue skills. Physics Education **59**(4), 045005 (2024) <https://doi.org/10.1088/1361-6552/ad3d21>
- [16] Sirnoorkar, A., Zollman, D., Laverty, J.T., Magana, A.J., Rebello, S., Bryan, L.A.: Student and ai responses to physics problems examined through the lenses of sensemaking and mechanistic reasoning. Computers and Education: Artificial Intelligence **5**, 100318 (2024) <https://doi.org/10.1016/j.caear.2024.100318>
  - [17] Kestin, G., Miller, K., Klales, A., Milbourne, T., Ponti, G.: Ai tutoring outperforms in-class active learning: an rct introducing a novel research-based design in an authentic educational setting. Scientific Reports **15**(1), 17458 (2025) <https://doi.org/10.1038/s41598-025-97652-6>
  - [18] Wang, K.D., Burkholder, E., Wieman, C., Salehi, S., Haber, N.: Examining the potential and pitfalls of chatgpt in science and engineering problem-solving. Frontiers in Education **8** (2024) <https://doi.org/10.3389/feduc.2023.1330486>
  - [19] Kumar, T., Kats, M.A.: Chatgpt-4 with code interpreter can be used to solve introductory college-level vector calculus and electromagnetism problems. American Journal of Physics **91**(12), 955–958 (2023) <https://doi.org/10.1119/5.0182627>
  - [20] Yeadon, W., Inyang, O.-O., Missouri, A., Peach, A., Testrow, C.P.: The death of the short-form physics essay in the coming ai revolution. Physics Education **58**(3), 035027 (2023) <https://doi.org/10.1088/1361-6552/acc5cf>
  - [21] Yeadon, W., Peach, A., Testrow, C.: A comparison of human, gpt-3.5, and gpt-4 performance in a university-level coding course. Scientific Reports **14**, 23285 (2024) <https://doi.org/10.1038/s41598-024-73634-y>
  - [22] Kilde-Westberg, S., Johansson, A., Enger, J.: Generative ai as a lab partner: A case study. Phys. Rev. Phys. Educ. Res., (2025) <https://doi.org/10.1103/ggy1-3kjk>
  - [23] Low, A., Kalender, Z.Y.: Data dialogue with chatgpt: Using code interpreter to simulate and analyse experimental data (2023) [arXiv:2311.12415](https://arxiv.org/abs/2311.12415)
  - [24] Besta, M., Barth, J., Schreiber, E., Kubicek, A., Catarino, A., Gerstenberger, R., Nyczyk, P., Iff, P., Li, Y., Houlston, S., Sternal, T., Copik, M., Kwaśniewski, G., Müller, J., Flis, L., Eberhard, H., Chen, Z., Niewiadomski, H., Hoefer, T.: Reasoning language models: A blueprint (2025) [arXiv:2501.11223](https://arxiv.org/abs/2501.11223)
  - [25] Xu, F., Hao, Q., Zong, Z., Wang, J., Zhang, Y., Wang, J., Lan, X., Gong, J., Ouyang, T., Meng, F., Shao, C., Yan, Y., Yang, Q., Song, Y., Ren, S., Hu, X., Li, Y., Feng, J., Gao, C., Li, Y.: Towards large reasoning models: A survey of reinforced reasoning with large language models (2025) [arXiv:2501.09686](https://arxiv.org/abs/2501.09686)

- [26] Yoon, D., Kim, S., Yang, S., Kim, S., Kim, S., Kim, Y., Choi, E., Kim, Y., Seo, M.: Reasoning models better express their confidence (2025) [arXiv:2505.14489](https://arxiv.org/abs/2505.14489)
- [27] Zhang, X., Dong, Y., Wu, Y., Huang, J., Jia, C., Fernando, B., Shou, M.Z., Zhang, L., Liu, J.: Physreason: A comprehensive benchmark towards physics-based reasoning (2025) [arXiv:2502.12054](https://arxiv.org/abs/2502.12054)
- [28] Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., Farajtabar, M.: The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity (2025). <https://ml-site.cdn-apple.com/papers/the-illusion-of-thinking.pdf>
- [29] Wu, J., Gan, W., Chen, Z., Wan, S., Yu, P.S.: Multimodal large language models: A survey. In: Proceedings of the IEEE International Conference on Big Data (BigData), pp. 2247–2256 (2023). <https://doi.org/10.1109/BigData59044.2023.10386743>
- [30] Wang, J., Jiang, H., Liu, Y., Ma, C., Zhang, X., Pan, Y., Liu, M., Gu, P., Xia, S., Li, W., Zhang, Y., Wu, Z., Liu, Z., Zhong, T., Ge, B., Zhang, T., Qiang, N., Hu, X., Jiang, X., Zhang, X., Zhang, W., Shen, D., Liu, T., Zhang, S.: A comprehensive review of multimodal large language models: Performance and challenges across different tasks (2024) [arXiv:2408.01319](https://arxiv.org/abs/2408.01319)
- [31] Bewersdorff, A., Hartmann, C., Hornberger, M., Seßler, K., Bannert, M., Kasneci, E., Kasneci, G., Zhai, X., Nerdel, C.: Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in science education. Learning and Individual Differences **118**, 102601 (2025) <https://doi.org/10.1016/j.lindif.2024.102601>
- [32] Treagust, D.F., Duit, R., Fischer, H.E. (eds.): Multiple Representations in Physics Education vol. 10. Springer, Cham (2017). <https://doi.org/10.1007/978-3-319-58914-5>
- [33] Polverini, G., Gregorcic, B.: Performance of chatgpt on the test of understanding graphs in kinematics. Phys. Rev. Phys. Educ. Res. **20**, 010109 (2024) <https://doi.org/10.1103/PhysRevPhysEducRes.20.010109>
- [34] Bessas, N., Tzanaki, E., Vavougios, D., Plagianakos, V.P.: Comparative analysis of chatgpt and gemini; implications for junior high school physics education: Opportunities and ethical challenges. International Journal of Advanced Multi-disciplinary Research and Studies **5**(1), 7–18 (2025) <https://doi.org/10.62225/2583049X.2025.5.1.3610>
- [35] Jiang, Q., Gao, Z., Karniadakis, G.E.: Deepseek vs. chatgpt vs. claude: A comparative study for scientific computing and scientific machine learning tasks. Theoretical and Applied Mechanics Letters **15**(2), 100583 (2025) <https://doi.org/10.1016/j.taml.2025.100583>

- [36] Polverini, G., Gregorcic, B.: Performance of freely available vision-capable chatbots on the test for understanding graphs in kinematics. In: Proceedings of the Physics Education Research Conference (PERC), Boston, MA (2024). <https://doi.org/10.1119/perc.2024.pr.Polverini>
- [37] Polverini, G., Melin, J., Önerud, E., Gregorcic, B.: Performance of chatgpt on tasks involving physics visual representations: The case of the brief electricity and magnetism assessment. *Phys. Rev. Phys. Educ. Res.* **21**, 010154 (2025) <https://doi.org/10.1103/PhysRevPhysEducRes.21.010154>
- [38] Robledo-Rella, V., Gonzalez-Nucamendi, A., Neri, L., García-Castelán, R.M.G., Noguez, J., Valverde-Rebaza, J.: Can we trust ai chatbots to teach university physics? a performance comparison of ai chatbots. Artificial Intelligence in Physics Courses to Support Active Learning (ICSLT '24 proceedings), 68–75 (2024) <https://doi.org/10.1145/3678610.3678631>
- [39] Kortemeyer, G., Babayeva, M., Polverini, G., Widenhorn, R., Gregorcic, B.: Multilingual performance of a multimodal artificial intelligence system on multisubject physics concept inventories. *Phys. Rev. Phys. Educ. Res.* **21**, 020101 (2025) <https://doi.org/10.1103/98hg-rkrf>
- [40] Polverini, G., Gregorcic, B.: Evaluating vision-capable chatbots in interpreting kinematics graphs: a comparative study of free and subscription-based models. *Frontiers in Education* **9**, 1452414 (2024) <https://doi.org/10.3389/feduc.2024.1452414>
- [41] Organisation for Economic Co-operation and Development: Education at a Glance 2023: OECD Indicators. Technical report, OECD Publishing (2023). Accessed: 2025-08-19. [https://www.oecd.org/en/publications/2023/09/education-at-a-glance-2023\\_581c9602.html](https://www.oecd.org/en/publications/2023/09/education-at-a-glance-2023_581c9602.html)
- [42] Winter, J.: Educating pre-service physics teachers in england: The need for knowledge transformation. PhD thesis, Uppsala University, Department of Physics and Astronomy, Physics Didactics, Uppsala, Sweden (May 2025). <https://uu.diva-portal.org/smash/get/diva2:1945934/FULLTEXT01.pdf>
- [43] Paris, B.M.: Instructors' perspectives of challenges and barriers to providing effective feedback. *Teaching & Learning Inquiry* **10** (2022) <https://doi.org/10.20343/teachlearninqu.10.3>
- [44] Clark, R.E.: How much and what type of guidance is optimal for learning from instruction? In: Tobias, S., Duffy, T.M. (eds.) *Constructivist Instruction: Success or Failure?*, pp. 158–183. Routledge/Taylor & Francis Group, ??? (2009)
- [45] Tierney, R.D., Simon, M., Charland, J.: Being fair: Teachers' interpretations of principles for standards-based grading. *The Educational Forum* **75**(3), 210–227 (2011) <https://doi.org/10.1080/00131725.2011.577669>

- [46] Melnick, S.A., Meister, D.G.: A comparison of beginning and experienced teachers' concerns. *Educational Research Quarterly* **31**(3), 39–56 (2008)
- [47] Kortemeyer, G.: Toward ai grading of student problem solutions in introductory physics: A feasibility study. *Physical Review Physics Education Research* **19**(2), 020163 (2023) <https://doi.org/10.1103/PhysRevPhysEducRes.19.020163>
- [48] Kortemeyer, G., Nöhl, J., Onishchuk, D.: Grading assistance for a handwritten thermodynamics exam using artificial intelligence: An exploratory study. *Physical Review Physics Education Research* **20**(2), 020144 (2024) <https://doi.org/10.1103/PhysRevPhysEducRes.20.020144>
- [49] Mok, R., Akhtar, F., Clare, L., Li, C., Ida, J., Ross, L., Campanelli, M.: Using ai large language models for grading in education: A hands-on test for physics (2024) [arXiv:2411.13685](https://arxiv.org/abs/2411.13685)
- [50] Chen, Z., Wan, T.: Grading explanations of problem-solving process and generating feedback using large language models at human-level accuracy. *Physical Review Physics Education Research* **21**(1), 010126 (2025) <https://doi.org/10.1103/PhysRevPhysEducRes.21.010126>
- [51] Wan, T., Chen, Z.: Exploring generative ai assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning. *Physical Review Physics Education Research* **20**(1), 010152 (2024) <https://doi.org/10.1103/PhysRevPhysEducRes.20.010152>
- [52] Krupp, L., Bley, J., Gobbi, I., Geng, A., Müller, S., Suh, S., Moghiseh, A., Medina, A.C., Bartsch, V., Widera, A., Ott, H., Lukowicz, P., Karolus, J., Kiefer-Emmanouilidis, M.: Llm-generated tips rival expert-created tips in helping students answer quantum-computing questions. *EPJ Quantum Technology* **12**(1), 33 (2025) <https://doi.org/10.1140/epjqt/s40507-025-00334-5>
- [53] Guo, S., Latif, E., Zhou, Y., Huang, X., Zhai, X.: Using generative ai and multi-agents to provide automatic feedback (2024) [arXiv:2411.07407](https://arxiv.org/abs/2411.07407)
- [54] Clark, A.K., Hirt, A., Whitcomb, D., Thompson, W.J., Wine, M., Karvonen, M.: Artificial intelligence in science and mathematics assessment for students with disabilities: Opportunities and challenges. *Education Sciences* **15**(2), 233 (2025) <https://doi.org/10.3390/educsci15020233>
- [55] El-Adawy, S., MacDonagh, A., Abdelhafez, M.: Exploring large language models as formative feedback tools in physics. In: 2024 Physics Education Research Conference Proceedings, pp. 126–131. American Association of Physics Teachers, Boston, MA (2024). <https://doi.org/10.1119/perc.2024.pr.El-Adawy>
- [56] Kortemeyer, G., Nöhl, J.: Assessing confidence in ai-assisted grading of physics exams through psychometrics: An exploratory study. *Physical Review*

Physics Education Research **21**(1), 010136 (2025) <https://doi.org/10.1103/PhysRevPhysEducRes.21.010136>

- [57] OpenAI: GPT-4o (Omni) Math Tutoring Demo on Khan Academy. [https://www.youtube.com/watch?v=ivxzcocyu\\_m](https://www.youtube.com/watch?v=ivxzcocyu_m). Accessed June 18, 2025 (2024)
- [58] Google DeepMind: Math & Physics with AI — Gemini. <https://www.youtube.com/watch?v=k4px1vaxaaI>. Accessed June 18, 2025 (2023)
- [59] Zhang, X., Cao, J., You, C.: Counting ability of large language models and impact of tokenization (2024) [arXiv:2410.19730](https://arxiv.org/abs/2410.19730)
- [60] Ahmed, F.: The digital divide and ai in education: Addressing equity and accessibility. Journal of AI and Education (2025). Lahore University of Management Sciences (LUMS)
- [61] UNESCO: AI literacy and the new Digital Divide – A Global Call for Action. <https://www.unesco.org/en/articles/ai-literacy-and-new-digital-divide-global-call-action>. Last update: 28 February 2025 (2024)
- [62] Anthropic: Models overview. <https://docs.anthropic.com/en/docs/about-claude/models/overview>. Accessed: 2025-06-13 (2025)
- [63] Google AI: Gemini models. [https://ai.google.dev/gemini-api/docs/models?utm\\_source=chatgpt.com](https://ai.google.dev/gemini-api/docs/models?utm_source=chatgpt.com). Accessed: 2025-06-13 (2025)
- [64] Google DeepMind: Gemma models overview. <https://ai.google.dev/gemma/docs>. Last updated: 2025-03-04; Accessed: 2025-06-13 (2025)
- [65] OpenAI: Models. [https://platform.openai.com/docs/models?utm\\_source=chatgpt.com](https://platform.openai.com/docs/models?utm_source=chatgpt.com). Accessed: 2025-08-19 (2025)
- [66] Ding, L., Chabay, R., Sherwood, B., Beichner, R.: Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment. Physical Review Special Topics - Physics Education Research **2**(1), 010105 (2006) <https://doi.org/10.1103/PhysRevSTPER.2.010105>
- [67] Kaltakci-Gurel, D., Eryilmaz, A., McDermott, L.C.: Development and application of a four-tier test to assess pre-service physics teachers' misconceptions about geometrical optics. Research in Science & Technological Education **35**(2), 238–260 (2017) <https://doi.org/10.1080/02635143.2017.1310094>
- [68] Cataloglu, E., Robinett, R.W.: Testing the development of student conceptual and visualization understanding in quantum mechanics through the undergraduate career. American Journal of Physics **70**(3), 238–251 (2002) <https://doi.org/10.1119/1.1405509>

- [69] Beichner, R.J.: Testing student interpretation of kinematics graphs. *American Journal of Physics* **62**(8), 750–762 (1994) <https://doi.org/10.1119/1.17449>
- [70] Garn, D.: Understanding the Role of Temperature Settings in AI Output. TechTarget SearchEnterpriseAI (2025). <https://www.techtarget.com/searchenterpriseai/tip/Understanding-the-role-of-temperature-settings-in-AI-output>
- [71] Gregorcic, B., Polverini, G.: Responses of Multimodal Large Language Models on BEMA, TUG-K, QMVI and FTGOT. Dataset on Zenodo (2025). <https://doi.org/10.5281/zenodo.15719827>
- [72] 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., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners (2020) [arXiv:2005.14165](https://arxiv.org/abs/2005.14165)
- [73] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E.H., Le, Q.V., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models (2022) [arXiv:2201.11903](https://arxiv.org/abs/2201.11903)
- [74] Wheatley, C., Wells, J., Stewart, J.: Applying module analysis to the brief electricity and magnetism assessment. *Physical Review Physics Education Research* **20**(1), 010104 (2024) <https://doi.org/10.1103/PhysRevPhysEducRes.20.010104>
- [75] Zavala, G., Tejeda, S., Barniol, P., Beichner, R.J.: Modifying the test of understanding graphs in kinematics. *Physical Review Physics Education Research* **13**(2), 020111 (2017) <https://doi.org/10.1103/PhysRevPhysEducRes.13.020111>
- [76] OpenAI: Reasoning. <https://platform.openai.com/docs/guides/reasoning>. Accessed: 2025-08-19 (2025)
- [77] OpenAI: Reasoning Best Practices. <https://platform.openai.com/docs/guides/reasoning-best-practices>. Accessed: 2025-08-19 (2025)