

Multilingual performance of a multimodal artificial intelligence system on multisubject physics concept inventories

Gerd Kortemeyer^{1,2}, Marina Babayeva³, Giulia Polverini⁴,
Ralf Widenhorn⁵ and Bor Gregorcic⁴

¹Rectorate and AI Center, *ETH Zurich*, 8092 Zurich, Switzerland

²*Michigan State University*, East Lansing, Michigan 48823, USA

³Department of Physics Education, *Charles University*, Prague, 8, Czech Republic

⁴Department of Physics and Astronomy, *Uppsala University*, Box 516, 75120 Uppsala, Sweden

⁵Department of Physics, *Portland State University*, Portland, Oregon 97207, USA



(Received 13 January 2025; accepted 3 June 2025; published 8 July 2025)

We investigate the multilingual and multimodal performance of a large language model-based artificial intelligence (AI) system, GPT-4o, using a diverse set of physics concept inventories spanning multiple languages and subject categories. The inventories, sourced from the PhysPort website, cover classical physics topics such as mechanics, electromagnetism, optics, and thermodynamics, as well as relativity, quantum mechanics, astronomy, mathematics, and laboratory skills. Unlike previous text-only studies, we uploaded the inventories as images to reflect what a student would see on paper, thereby assessing the system's multimodal functionality. Our results indicate variation in performance across subjects, with laboratory skills standing out as the weakest. We also observe differences across languages, with English and European languages showing the strongest performance. Notably, the relative difficulty of an inventory item is largely independent of the language of the test. When comparing AI results to existing literature on student performance, we find that the AI system outperforms average postinstruction undergraduate students in all subject categories except laboratory skills. Furthermore, the AI performs worse on items requiring visual interpretation of images than on those that are purely text-based. While our exploratory findings show GPT-4o's potential usefulness in physics education, they highlight the critical need for instructors to foster students' ability to critically evaluate AI outputs, adapt curricula thoughtfully in response to AI advancements, and address equity concerns associated with AI integration.

DOI: [10.1103/98hg-rkrf](https://doi.org/10.1103/98hg-rkrf)

I. INTRODUCTION

A. Generative artificial intelligence in physics education

The public availability of large language models (LLMs), like those built on the architecture introduced by Vaswani *et al.* [1], has unlocked new possibilities across various domains, including education [2,3]. Since the release of ChatGPT in the fall of 2022 [4], LLMs have surged in popularity, with scholars showcasing their remarkable capabilities. Beyond the wave of enthusiasm generated by humanlike responses that have been shown to pass the Turing Test [5] with a majority of human test subjects [6], the OpenAI's GPT series demonstrated proficiency in academic fields such as physics in a number of

benchmarks [7]. Both the initial version and later iterations, particularly GPT-4 [8], have achieved impressive results in physics, such as passing standardized exams, excelling in introductory courses, and even coming close to passing entire degrees [9–13].

The technology is increasingly being embraced in physics education [14], offering promising applications for both teaching and learning physics. LLMs have proven valuable for teachers, helping in the creation of tailored materials and tasks, student assessments [15], and personalized feedback [16–21], as well as for their training [22]. For students, these models represent ever-available, patient, and knowledgeable resources [23,24]. However, these opportunities come with significant risks, particularly the potential for users to overly trust AI when assessing its scientific accuracy [25,26].

Early publicly available generative AI systems have been limited to processing and generating text-based content only. Thus, earlier studies of AI's performance on physics tasks, such as concept inventories, were restricted to text-based materials or textual descriptions of visual elements [9,27–30]. More recently, multimodal systems, which can

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

also input and output auditory and visual data, have broadened the scope of such studies [31].

In this study, we confront GPT-4o [32], a popular AI model from OpenAI, with screenshots of tasks from physics concept inventories, which reflect the test items as seen by learners, including the accompanying diagrams, sketches, graphs, and illustrations. Earlier studies using GPT-4 and GPT-4o with the FCI [33], TUG-K [34], and BEMA [35] concept inventories not only demonstrated promising behavior but also revealed significant challenges, mainly related to the AI system's limited ability to interpret visual information [30,36,37]. We considered the English version of each selected concept inventory and all its available translations.

B. Concept inventories in physics

Examinations in physics generally involve symbolic and numerical calculations, whereas concept inventories are typically different, focusing primarily on conceptual understanding [38,39]. Scores on these different types of assessments do not necessarily correlate, as for example, the FCI would underpredict success in a calculus-based physics course [40]. On the surface, this would work in AI's favor, as AI systems used to be notoriously "bad at math," which hampered their performance on traditional physics exam questions [9]; this is remedied in newer models that generate Python code for calculations or are explicitly "reasoning," such as GPT-o1 and GPT-o3-mini [41,42]. However, early investigations of AI's performance on conceptual tasks suggest that these, too, can present several challenges [10,43].

Concept inventories have played and continue to play an important role in physics education research [44], and some of the most influential studies have been based on their outcomes, most notably with respect to active engagement [45]. Unlike traditional assessments that focus on individual learners, concept inventories are primarily designed to evaluate instructional methods, oftentimes with a focus on learning gains rather than absolute scores. As we embark on assessing AI on the base of absolute scores, we deviate from this practice.

Arguably, while many concept inventory items assess students' understanding of core ideas and concepts [39], they do not necessarily capture evidence of scientific practices or crosscutting concepts [46] (even though this distinction is debatable [47]). It is thus important to emphasize that our study assesses AI's performance on physics conceptual tasks, but does not evaluate if it has the qualities of a physicist.

C. The language problem

Language plays a crucial role in learning physics. Expert physicists structure their knowledge using layered metaphorical systems and specific grammatical

frameworks, which shape how they communicate complex ideas [48,49]. These unconscious linguistic patterns can influence how students understand physics [48,50]. Although English has become the *lingua franca* of most scientific communications, undergraduate students tend to be instructed in their native language, which is also the language in which they will tend to tackle physics problems.

LLMs hold the promise of facilitating language-related learning tasks [51], yet they currently do not function equally well in all languages; for example, OpenAI's research on the GPT-4's performance shows inconsistent results across different languages [52]. Because of the disparities in the prevalence, quantity, and quality of information available across languages, there exists a disparity in the resources available for LLM training, which can have an impact on model performance. Nicholas and Bhatia highlight that although LLMs are designed to mitigate the issue of underrepresentation of certain languages in learning data, as of 2023, early LLMs were still predominantly trained on English materials [53]. While this may be changing as governments and companies attempt to strengthen AI capabilities in their countries (e.g., models like DeepSeek [54] or Qwen [55] developed by companies in China, or the Swiss AI Initiative [56]), it is likely that LLMs remain biased toward the needs of major economies with the financial resources needed to train AI systems. Similar issues are also discussed in the analysis conducted by "Cohere for AI" company in their report titled "The AI Language Gap" [57]. Feng *et al.* also acknowledge the disparity in LLM performance across different languages when it comes to abstaining from hallucinations, resulting in a gap of approximately 20% between high-resource and low-resource languages [58]; "hallucinations" refer to instances where an LLM generates plausible-sounding but inaccurate or entirely fabricated information, indicative of shortcomings in calibration and reasoning. Particularly relevant to our study is a recent finding that a Chinese-trained model performed better on the FCI when prompted in Chinese rather than English [29].

Understanding and using physics-specific language is essential for physics literacy and learning. To address both the linguistic disparities of LLMs and the specialized nature of physics language, it is important to examine how LLMs perform across diverse languages in the context of physics education. While current research highlights the general challenges of multilingual use of LLMs, the intersection of language and subject-specific terminology, such as physics, is understudied and lacks a clear understanding of the current situation.

D. Relevance to physics education research

While AI's potential as a learning tool, assessment assistant, and research aid is widely acknowledged [12,59–62], the performance of LLMs on established and validated physics conceptual assessments, particularly in multilingual

contexts, and on multiple subject domains, remains underexplored.

As AI assumes an increasingly large role in the educational process, it is also important for physics education stakeholders to develop a sense of its capabilities in physics-related tasks. Assessing AI's ability to solve the kind of problems we use for assessing student understanding of physics concepts is a necessary step if we want to make appropriate and responsible use of these systems in physics education. The physics education research community, with its wealth of research-based assessment instruments, so-called concept inventories, is well positioned to engage in such evaluation. Concept inventories are developed to be robust tools for assessing university students' conceptual understanding. Comparing AI's performance to that of students can thus provide a student-centered reference point that is more meaningful for PER researchers than AI-facing benchmark assessments (which are oftentimes designed to assess physical reasoning tasks and inform machine-learning engineering [7,63–66]), and they are more standardized and universal than exams from individual courses [13].

Comparing AI's average performance to university students' postinstruction performance can thus provide a rough student-centric measure of the level of performance of AI systems. There is a need for caution here, however. There are important differences in information processing between students and AI models. This means that, for example, when AI reaches a numerical performance similar to that of an average student, it does not necessarily mean that its strengths and difficulties are similar to those of an average student. In fact, previous studies have shown important differences in the profile of students' and AI's difficulties—i.e., AI's displayed difficulties have been found to be uncharacteristic of typical student difficulties [36,37].

One of the motivations for our work is exploring what types of tasks might be difficult for GPT-4o to solve. For instructors, knowing this would allow them to communicate to their students the strengths and weaknesses of the AI system, so that students may use it in responsible and productive ways when working unsupervised (e.g., while doing homework or project assignments) or in contexts beyond formal education. Exploring the different facets of an AI's performance can also inform developers about AI's potential as an educational tool for automated grading [20,67,68], generating feedback [69,70], and personalizing instruction [71]. However, for AI to be genuinely useful in these applications, it must demonstrate consistency and reliability across diverse educational contexts. For example, a recent study suggests that an AI system's ability to grade student answers on a topic is correlated to its problem-solving ability [68] on that topic. Analyzing GPT-4o's performance on structured, standardized conceptual assessments, can therefore also provide insights into whether AI

can reliably assist instructors in evaluating student understanding and be integrated into automated feedback and grading systems.

Another major motivation driving this research is the impact of AI's increasing accessibility on student learning and assessment validity in physics education [11]. With generative AI now readily available, students are already using it for study purposes, sometimes in ways that challenge traditional expectations of academic integrity [72]. More broadly, this issue ties into an ongoing discussion in PER regarding how assessments need to evolve in an AI-assisted learning environment and how physics curricula should adapt to the fact that many tasks can now be outsourced to AI.

Another key aspect of this study is its multilingual approach. While much of PER is conducted in English [73,74], physics is taught in a wide range of languages, and the educational impact of AI should be considered beyond English-speaking classrooms. LMMs like GPT-4o have been trained primarily on English-language data [53], which raises concerns about whether their accuracy is consistent across different linguistic versions of physics assessments. If AI models perform significantly better in one language than in another, this could reinforce existing inequities in educational technology [75]. Understanding these potential disparities is necessary to ensure equitable access to AI-assisted learning tools and to assess whether AI can serve as a reliable resource for students who learn physics in languages other than English.

Finally, our study contributes to a growing body of research on the multimodal performance of LLMs, particularly on tasks involving physics visual representations. This ties into the question of what types of tasks are especially difficult for AI to solve. While it is known that GPT-4o struggles with tasks involving kinematics graphs [31], it remains unclear if it experiences similar limitations across other types of physics problems and subject categories, and whether these weaknesses persist when the visual data are embedded in diverse real-world assessment formats.

This study is exploratory in nature; rather than aiming to provide definitive answers to broad questions about AI's role in physics education, it seeks to identify emerging trends in AI performance on physics concept inventories. By analyzing GPT-4o's accuracy on multiple-choice questions across different languages, with and without accompanying images, we aim to highlight patterns that warrant further investigation rather than making strong claims about AI's conceptual reasoning abilities. Any broader generalizations about AI's impact on assessment, instruction, or research would require a more detailed analysis of its reasoning processes and interactions with learners, which is beyond the scope of this paper. Nevertheless, by mapping out these initial trends, this study contributes to a growing conversation in PER about how AI is reshaping

physics education, helping to inform future research directions and the responsible development and integration of AI into learning environments and educational processes.

E. Research questions

Despite a growing number of exploratory studies on LLMs in physics education [12,15–17,25,26,30], including some of our own [9,31,37], there remains a lack of a broad, systematic analysis of how multimodal AI models perform on validated, research-based physics assessments, particularly when these assessments include diverse subject domains, visual components, and multiple languages. Prior studies have often focused on individual inventories or English-only text-based inputs. This study extends that work by (i) synthesizing performance trends across inventories and physics subject categories, (ii) exploring performance in different languages, (iii) comparing AI performance to student benchmarks, and (iv) reevaluating AI’s visual interpretation challenges in a multilingual and multisubject framework. The following research questions guide our exploratory analysis:

- RQ1. How does GPT-4o perform across different physics concept inventories in English?
- RQ2. How does language influence the performance of the AI system?
- RQ3. How does its performance compare to student performance at the undergraduate level?
- RQ4. How does the presence of images influence the performance of the AI system?

These questions are, in turn, addressed in Sec. IV.

II. DATASET

As our dataset, we used concept inventories published in PhysPort [76], containing a comprehensive collection of research-based inventories in multiple translations [77]. We included inventories that had at least a “bronze star” classification assigned on PhysPort. This indicates that they had been studied with respect to at least three of the seven research validation categories that the platform employs (research into student thinking; studied with student interviews, expert review, and appropriate statistical analysis; research conducted at multiple institutions, by multiple research groups, and with peer-reviewed publication). We excluded inventories for which we could not obtain an answer key. Figure 1 shows inventory items from the Force Concept Inventory in Persian and the Heat and Temperature Conceptual Evaluation in Chinese. These inventories cover a broad spectrum of physics subject categories, listed in Table I.

A total of 35 languages are currently represented in PhysPort. While the authors are familiar with some of the languages, the quality of the majority of the non-English translations could not be evaluated.

Tables II–V show the concept inventories included in our investigation, along with their available translations and literature references. For some inventories, more than one version was available. The column “%Post” lists some postinstruction inventory scores reported in the literature for undergraduate-level courses; these are collected on a best-effort basis and not necessarily representative (typically, inventory gains are reported much more systematically [78]). We omitted results with very small sample sizes, as well as reported scores at the graduate-student or postdoctoral level. For subsequent comparisons, the scores found for each inventory were averaged.

The rightmost column shows the subject-category identifier under which we classified the test. Explanations for the abbreviations used in this column are provided in Table I. These labels are derived from the PhysPort classification, however, we divided electricity and magnetism into EM-F (primarily dealing with fields and potentials), and EM-C (primarily focusing on dc and ac circuits). The multisubject Next Gen Physical Science Diagnostic (NGPSD) was given no classification.

We cannot exclude the possibility that some of these inventories appeared in the text corpus used to train the model, which could be seen as “teaching to the test.” However, GPT’s training corpus is closed source and its contents remain mostly speculative with only very few hints given [79]. While popular inventories like the FCI might have been included, it is less likely that many of the more obscure inventories were present. Furthermore, in the scientific literature, the solution keys are typically provided separately from the inventory items. This means that the model is unlikely to associate correct answers to their corresponding questions on the basis of proximity in the training data.

III. METHODOLOGY

A. Data preparation

The items from the concept inventories were captured using screenshots like the ones shown in Fig. 1. If an item had multiple parts referring to the same scenario or each other, these were combined in one image; at times, this required manual image editing to close page breaks. The inventories in Tables II–V resulted in 3662 separate image files that were submitted to the model.

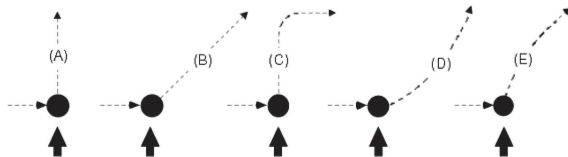
The solution keys were transcribed from PhysPort, where we transformed the various option values such as B), 2., b., β), and non-Arabic, non-Latin, non-Greek characters into lower-case Latin characters for easier automatic processing.

For the MUQ, which assigns partial credit for incorrect answers, we simplified the evaluation by only giving credit for correct answers; this would underestimate the model’s performance. We did the same for BEMA, which has conditional grading rules for 4 out of 31 items, depending

با توجه به متن و شکل زیر به چهار سؤال بعد (۸ تا ۱۱) پاسخ دهید.
 شکل یک توپ هاکی را که با سرعت ثابت v در خطی مستقیم از نقطه "a" به نقطه "b" روی یک سطح بدون اصطکاک لیز می خورد را نشان می دهد. نیروهای اعمال شده بوسیله هوا قابل چشم پوشی است. شما در حال نگاه کردن به توپ از بالا هستید. هنگامی که توپ به نقطه "b" می رسد، یک ضربه افقی سریع در جهت پیکان پر رنگ دریافت می کند. اگر توپ در نقطه "b" ساکن می بود، ضربه آن را به حرکت افقی با سرعت ثابت v_k در جهت ضربه وا می داشت.



۸- کدام یک از مسیرهای زیر به مسیر توپ پس از دریافت ضربه نزدیکتر است؟



۹- سرعت توپ دقیقاً پس از دریافت ضربه:

(الف) برابر با سرعت v است که پیش از ضربه داشته است.

(ب) برابر با سرعت v_k در نتیجه ضربه و مستقل از v است.

(پ) برابر با جمع جبری سرعت های v و v_k است.

(ت) کوچکتر از هر دو سرعت v یا v_k است.

(ث) بزرگتر از هر دو سرعت v یا v_k است، اما کوچکتر از جمع جبری این دو سرعت می باشد.

۱۰- در طول مسیر بدون اصطکاک که در سؤال ۸ انتخاب کرده بودید، سرعت توپ پس از دریافت ضربه:

(الف) ثابت است.

(ب) به طور پیوسته افزایش می یابد.

(پ) به طور پیوسته کاهش می یابد.

(ت) برای لحظه ای افزایش و پس از آن کاهش می یابد.

(ث) برای لحظه ای ثابت است و پس از آن کاهش می یابد.

۱۱- در طول مسیر بدون اصطکاک که در سؤال ۸ انتخاب کرده بودید، نیروی (های) اصلی عمل کننده بر توپ پس

از دریافت ضربه عبارتند از:

(الف) نیروی پایین سوی جاذبه

(ب) نیروی پایین سوی جاذبه، نیروی افقی در جهت حرکت.

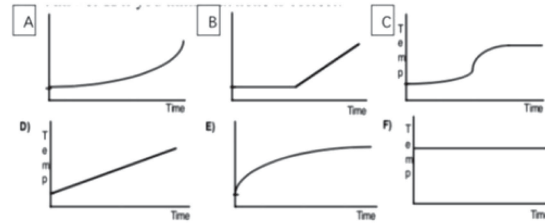
(پ) نیروی پایین سوی جاذبه، نیروی بالاسوی اعمال شده بوسیله سطح و نیروی افقی در جهت حرکت.

(ت) نیروی پایین سوی جاذبه، نیروی بالاسوی اعمال شده بوسیله سطح.

(ث) هیچکدام. (هیچ نیرویی بر توپ اعمال نمی شود)

问题 16 至 19 涉及一个处在室温中装有水的杯子，这个杯子是完全隔热的，即没有热量能传给杯子，也没有热量从杯子流出。杯中放入一小加热器，可用来加热杯中的水，水不处于沸腾状态。

16. 如果热量以恒定的速率传递到杯子里，下面哪个图形最能代表热量传递时水的温度随时间变化的形状？（如果你觉得没有一个选项正确，就选择 H）（ ）



从 17 题至 19 题，基于第 16 题（即理想绝热水杯，水中有一加热器）情况下发生的改变，对于每一题，选择一个最佳答案来描述基于这些变化，温度是如何上升的。

17. 水量相同，但传递的热量加倍（ ）

A) 温度的升高是以前的四倍

B) 温度的升高是以前的两倍

C) 温度的升高和以前的一样

D) 温度的升高是以前的二分之一

E) 温度的升高是以前的四分之一

H) 以上答案都不对

18. 水量减半，传递的热量相同（ ）

A) 温度的升高是以前的四倍

B) 温度的升高是以前的两倍

C) 温度的升高和以前的一样

D) 温度的升高是以前的二分之一

E) 温度的升高是以前的四分之一

H) 以上答案都不对

19. 水被替换成质量相等、但比热容仅为一半的液体，传递的热量相同（ ）

A) 温度的升高是以前的四倍

B) 温度的升高是以前的两倍

C) 温度的升高和以前的一样

D) 温度的升高是以前的二分之一

E) 温度的升高是以前的四分之一

H) 以上答案都不对

FIG. 1. Examples of uploaded problem images: FCI, items 8–11, in Persian (left panel) and HTCE, items 16–19, in Chinese (right panel).

on answers from earlier items. Here, for two items, this simplification favors the model, while for the other two items, it works to the model's disadvantage. In the FTGOT, we excluded the items assessing certainty in the answer and compared results for a two-tier test version. Finally, we skipped the free-response items that were included in some of the inventories, for example, to explain reasoning, as those were not scored in the original inventories.

For each individual item, we coded whether it was text-only or included an image, graph, or scenario illustration. Additionally, we coded if consulting the image was necessary for correctly answering the question (required image) or if all the required information was already present in the text (unneeded image). These codes will be used in analyses of subclasses of problems in Sec. IV D.

TABLE I. Subject categories of the concept inventories we investigated, as well as their abbreviations used throughout.

Abbreviation	Description
AST	Astronomy
EM-F	Electricity/magnetism—fields
EM-C	Electricity/magnetism—circuits
MATH	Mathematics
LAB	Laboratory skills
MECH	Mechanics
OPT	Optics
QP	Quantum physics
REAS	Reasoning
RELA	Relativity
THERM	Thermodynamics

TABLE II. Concept inventories under consideration. Descriptions are taken from PhysPort [76,77].

Title	Full title	Description	References	Version	%Post	Languages	Category
ADT	Astronomy Diagnostic Test	Astronomy content knowledge (apparent motion of the sun, scale of the solar system, phases of the moon, linear distance scales, seasons, global warming, nature of light, gravity, stars, cosmology)	[80]	2.0	41 [81] 54 [81]	English, Spanish, Swedish	AST
BEMA	Brief Electricity and Magnetism Assessment	Electricity/magnetism content knowledge (circuits, electrostatics, magnetic fields, and forces)	[35]	1	42 [35] 43 [82] 61 [83]	Chinese, English, Japanese, Portuguese, Spanish, Swedish	EM-F
CCI	Calculus Concept Inventory	Mathematics content knowledge (functions, derivatives, limits, ratios, the continuum)	[84]	5	50 [85] 52 [85]	Czech, English	MATH
CDPA	Concise Data Processing Assessment	Lab skills (uncertainty in measurement, relationship between functions graphs and numbers)	[86]	2	39 [86]	English, Spanish	LAB
CSEM	Conceptual Survey of Electricity and Magnetism	Electricity/magnetism content knowledge (electrostatics, magnetic fields and forces, Faraday's law)	[87]	H	61 [88] 66 [83]	English, Indonesian, Malay, Spanish, Swedish	EM-F
CTSR	Lawson Classroom Test of Scientific Reasoning	Scientific reasoning (proportional thinking, probabilistic thinking, correlational thinking, hypothetico-deductive reasoning)	[89]	2	54 [90] 75 [90]	English, Spanish, Swedish	REAS
DIRECT	Determining and Interpreting Resistive Electric Circuit Concepts Test	Electricity/magnetism content knowledge (dc circuits)	[91]	1.2	44 [91] 63 [92]	Chinese, English, Finnish, German, Greek, Spanish, Swedish	EM-C
DS	Density Survey	Mechanics content knowledge (density)	[93]	1	57 [94]	English, German	MECH
ECA	Energy Concept Assessment	Mechanics content knowledge (energy principle, forms of energy, work and heat, absorption/emission spectrum, specifying appropriate systems)	[95]	2	47 [95] 50 [95]	Croatian, English	MECH
ECCE	Electric Circuits Conceptual Evaluation	Electricity/magnetism content knowledge (dc and ac circuits)	[96]	1	37 [97] 42 [96] 64 [96]	English	EM-C
EMCA	Electricity and Magnetism Conceptual Assessment	Electricity/magnetism content knowledge (electrostatics, electric fields and force, circuits, magnetism, induction)	[98]	1	49 [98] 58 [98]	English, Indonesian	EM-F
EMCS	Energy and Momentum Conceptual Survey	Mechanics content knowledge (energy, momentum)	[99]	1	49 [99] 52 [99] 54 [100] 69 [100]	English, Finnish, Indonesian, Swedish	MECH
FCI	Force Concept Inventory	Mechanics content knowledge (forces, kinematics)	[33]	v95	38 [101] 56 [102] 66 [103]	Arabic, Bengali, Catalan, Chinese, Croatian, Czech, Dutch, English, Filipino, Finnish, French, German, Greek, Hebrew, Hindi, Hungarian, Icelandic, Italian, Japanese, Malay, Norwegian, Persian, Polish, Portuguese, Punjabi, Russian, Slovak, Spanish, Swedish, Tamil, Thai, Turkish	MECH
FMCE	Force and Motion Conceptual Evaluation	Mechanics content knowledge (kinematics, forces, energy, graphing)	[104]	v99	55 [105]	English, Indonesian, Japanese, Spanish	MECH

TABLE III. Concept inventories under consideration (cont.). Descriptions are taken from PhysPort [76,77].

Title	Full title	Description	References	Version	%Post	Languages	Category
FORT	Montana State University Formal Reasoning Test	Scientific reasoning (hypothesis testing, correlational reasoning, probability, control of variables, proportional reasoning)	[106]	1	55 [106]	English	REAS
FTGOT	Four-Tier Geometrical Optics Test	Waves/optics content knowledge (plane mirrors, spherical mirrors, lenses); used as two-tier test	[107]	1	18 [107]	English, Turkish	OPT
FVA	Force, Velocity, and Acceleration Test	Mechanics content knowledge (forces, velocity, acceleration)	[108]	3.2.3a	38 [108] 70 [108]	English	MECH
GECI	Greenhouse Effect Concept Inventory	Astronomy content knowledge (types of greenhouse gases, types of electromagnetic energy, energy equilibrium balance, greenhouse effect mechanisms, global warming vs greenhouse effect)	[109]	vC	55 [109]	English, Japanese	AST
HTCE	Heat and Temperature Conceptual Evaluation	Thermal/statistical content knowledge (temperature, phase change, heat transfer, thermal properties of materials)	[110]	1	78 [110]	Chinese, English	THERM
IBCDC	Inventory of Basic Conceptions—dc Circuits	Electricity/magnetism content knowledge (dc circuits)	[111]	F06	English: 40 French: 35 [111]	English, French	EM-C
IBCM	Inventory of Basic Conceptions—Mechanics	Mechanics content knowledge (forces, kinematics)	[111]	F06	English: 26 French: 31 [111]	English, French	MECH
LPCA	Light Phenomena Conceptual Assessment	Waves/optics content knowledge (reflection, refraction, Snells law, wavelength and frequency, light scattering, electromagnetic spectrum, the human eye)	[112]	1	41 [112]	English	OPT
LPCI	Lunar Phases Concept Inventory	Astronomy content knowledge (phases of the moon)	[113]	3	42 [113] 55 [113]	English, Spanish	AST
LSCI	Light and Spectroscopy Concept Inventory	Astronomy content knowledge (light, waves, spectroscopy)	[114]	1	47 [115] 51 [115] 52 [115]	English	AST
MBT	Mechanics Baseline Test	Mechanics content knowledge (kinematics, forces, momentum, energy)	[116]	97	35 [117] 48 [118] 66 [119] 73 [119]	English, Finnish, French, German, Greek, Italian, Japanese, Malay, Persian, Portuguese, Spanish, Turkish	MECH
MCS	Magnetism Conceptual Survey	Electricity/magnetism content knowledge (magnetic fields and forces, Faraday's law)	[120]	1	41 [120] 44 [120]	English	EM-F
MUQ	Measurement Uncertainty Quiz	Lab skills (calculating error from measurements, accuracy and precision, sources of error)	[121]	1	—	English	LAB
MWCS	Mechanical Wave Conceptual Survey	Waves/optics content knowledge (mechanical waves, wave propagation, wave superposition, wave reflection, standing waves)	[122]	1 2	47 [123]	English, Spanish, Thai English, Spanish	OPT
NGCI	Newtonian Gravity Concept Inventory	Astronomy content knowledge (directionality of gravity, force law, thresholds related to gravity, independence of forces)	[124]	3	50 [124] 55 [124]	Arabic, English	AST
NGPSD	Next Gen Physical Science Diagnostic	Mechanics content knowledge (magnetism, static electricity, energy, forces, waves and sound, light)	[125]	2	...	English	...

TABLE IV. Concept inventories under consideration (cont.). Descriptions are taken from PhysPort [76,77].

Title	Full Title	Description	References	Version	%Post	Languages	Category
PIQL	Physics Inventory of Quantitative Literacy	Mathematics scientific reasoning (proportional reasoning, reasoning with signed quantities, covariational reasoning)	[126]	2.2	55 [126]	English	REAS
QMCA	Quantum Mechanics Concept Assessment	Modern/quantum content knowledge (wave functions, measurement, time dependence, probability, infinite square well, 1D tunneling, energy levels, spins)	[127]	5.5.7 6.6.2	54 [127]	English, Portuguese English	QP
QMCS	Quantum Mechanics Conceptual Survey	Modern/quantum content knowledge (wave functions, probability, wave-particle duality, uncertainty principle, infinite square well, one-dimensional tunneling, energy levels)	[128]	2.0	51 [128] 69 [128]	English, Finnish, Japanese	QP
QMFPS	Quantum Mechanics Formalism and Postulates Survey	Modern/quantum content knowledge (quantum mechanics formalism, quantum mechanics postulates)	[129]	29	32 [130] 37 [130]	English, Spanish	QP
QMS	Quantum Mechanics Survey	Modern/quantum content knowledge (wave functions, measurement, expectation values, Hamiltonian, time dependence, probability, infinite square well, finite square well, harmonic oscillator, 1D tunneling)	[131]	18	38 [131]	English	QP
QMVI	Quantum mechanics visualization instrument	Modern/quantum content knowledge (wave functions, probability, infinite square well, 1D tunneling, time dependence, momentum space, 2D potentials, visualization of the relationship between potentials and wave functions)	[132]	0.4	28 [132] 29 [132] 45 [132] 58 [132]	English	QP
QPCS	Quantum Physics Conceptual Survey	Modern/quantum content knowledge (de Broglie wavelength, double slit interference, uncertainty principle, photoelectric effect, wave particle duality)	[133]	1	59 [133] 75 [133]	English, Thai	QP
RAPT	Rate and Potential Test	Electricity/magnetism content knowledge (electric potential, rate of change)	[134]	1A	61 [134] 73 [134]	English	EM-F
RCI	Relativity Concept Inventory	Modern/quantum content knowledge	[135]	1	71 [135]	English	RELA
RFCI	Representational Variant of the Force Concept Inventory	Mechanics content knowledge (kinematics, forces, graphing, multiple representations)	[136]	2007	60 [136] 61 [136]	English, Finnish	MECH
RKI	Rotational Kinematics Inventory	Mechanics content knowledge (part 1: rotational kinematics of a particle, part 2: rotational kinematics of a particle in rectilinear motion, part 3: rotational kinematics of a rigid body about a fixed axis)	[137]	1	38 [138] 55 [138]	English, French	MECH
RRMCS	Rotational and Rolling Motion Conceptual Survey	Mechanics content knowledge (rotational kinetic energy, torque, rotational kinematics, moment of inertia)	[139]	1	75 [139]	English	MECH
SGCE	Symmetry and Gauss's Law Conceptual Evaluation	Electricity/magnetism content knowledge (symmetry, electric field, electric flux)	[140]	1	49 [140]	English	EM-F
SPCI	Star Properties Concept Inventory	Astronomy content knowledge (stellar properties, nuclear fusion, star formation)	[141]	4	30 [141] 51 [141]	English, Japanese, Spanish	AST

TABLE V. Concept inventories under consideration (cont.). Descriptions are taken from PhysPort [76,77].

Title	Full title	Description	References	Version	%Post	Languages	Category
STPFaSL	Survey of Thermodynamic Processes and First and Second Laws	Thermal/statistical content knowledge (first law of thermodynamics, second law of thermodynamics, PV diagrams, reversible processes, irreversible processes)	[142]	short	37 [142]	Chinese, English, Indonesian, Portuguese	THERM
TCE	Thermal Concept Evaluation	Thermal/statistical content knowledge (temperature, heat transfer, phase change, thermal properties of materials)	[143]	1	78 [143]	Chinese, English, Japanese, Portuguese	THERM
TCS	Thermodynamic Concept Survey	Thermal/statistical content knowledge (temperature, heat transfer, ideal gas law, first law of thermodynamics, phase change, thermal properties of materials)	[144]	2	43 [144] 46 [144] 58 [144]	Chinese, English, Thai	THERM
TOAST	Test of Astronomy Standards	Astronomy content knowledge (gravity, electromagnetic radiation, fusion and formation of heavy elements, evolution of the universe, star and stellar evolution, evolution and structure of the solar system, seasons, scale, yearly patterns, daily patterns, moon phases)	[145]	vf	44 [145]	English, Japanese	AST
TUG-K	Test of Understanding Graphs in Kinematics	Mechanics content knowledge (kinematics, graphing)	[34]	2.6	59 [146]	Arabic, Finnish, French, German, Hebrew	MECH
				3.0–4.0		Chinese, English, Greek, Portuguese, Spanish, Swedish, Ukrainian	
TUV	Test of Understanding of Vectors	Mathematics content knowledge (magnitude, direction, components, unit vector, addition, subtraction, multiplication, dot and cross product)	[147]	1	68 [147]	Arabic, English, Spanish	MATH

B. AI processing

We used GPT-4o [32] Version 2024-08-06 via Microsoft Azure AI Services [148] at ETH Zurich. The university's contract includes provisions that any data submitted will not be used for training purposes; this provision is crucial to avoid compromising the confidentiality and validity of the concept inventories.

The model was prompted to extract from each submitted image the number and the text of the inventory item, followed by written-out reasoning steps (explanation), and finally the letter option corresponding to its selected answer. In cases where multiple items appeared in the same screenshot, the model was instructed to repeat this process for each item. To process such outputs effectively, we had to prompt the model to return structured outputs in the form of a JSON schema. We found that prompting for structured outputs in languages other than English was unreliable and hindered the type of analysis we aimed to

perform. As a result, we decided to keep the prompt in English. Each submitted image, however, contained text in one of the different languages of the concept inventories. The API calls and prompts used in this study are available in Appendix A.

LLMs are probabilistic systems, and thus, responses to the same prompts vary. To obtain some statistics, each screenshot was independently submitted 3 times, and the resulting three outputs were combined into a solution array.

Altogether, we obtained 14 022 solutions for 4674 items (1498 for English and 3176 for non-English language inventories).

C. Analysis methods

The LLM's answer choices were normed to the same lower-case Latin characters as the solution keys. In cases where the AI did not provide a valid answer (e.g., where it claimed that there was no correct answer specified), the

answer was counted as incorrect. Some concept inventories had images or scenarios labeled with roman characters and answer choices such as “(A) I, B) II, C) III, D) IV, E) none of the above”; in this case, if the AI picked “IV,” this was manually converted to “d”; this happened for 0.4% of the responses. Another source of possible error was numbered scenarios within numbered multipart problems; in this case, the AI at times ignored the problem numbers provided in the prompt and instead used the scenario numbers; this was fairly obvious during the evaluation, but had to be fixed manually. In 0.6% of the cases, the LLM provided no valid response; these were counted as incorrect.

We counted each answer in an inventory—three answers per item—as either correct or incorrect, and considered the percentage of correct answers across each inventory and language as the performance measure for that test and language. For each answer, we coded the language of both the inventory description and the answer explanation, primarily using *langdetect* [149] and manual determination in some cases. Responses were categorized as fully in the language of the test, fully in English, or mixed. The most common mixed scenario occurred when the problem description was still in the language of the test, but the explanation was in English. However, there were also rare cases where the language switched midstream for either of those. We refer to this behavior as language switching.

D. Use of AI

While obviously being the subject of this study, AI (GPT-o1 [41]) has also been used for the following aspects of the study: initial drafts of analysis programs in R and Python, exploratory data analysis, LaTeX formatting of manuscript components, and improving the grammar and readability of manuscript passages.

IV. RESULTS

A. How does GPT-4o perform across different physics concept inventories in English?

Notably, English was the language with the most concept inventories: 53 out of 54 tested concept inventories were available in English. It is also the only language that has inventories across all the subject categories. In Fig. 2, the English results for individual inventories stand out as the most populated column. The average performance across all inventories in English is 71.1%, with results on individual inventories ranging from as low as 27% (FTGOT) to as high as 97% (DS and STPFaSLlo). Figure 3 shows the distribution of GPT-4o’s performance across the different concept inventories in English.

Examining the performance within specific subject categories, it is noticeable that for some subjects, the performance varies widely across inventories. As Fig. 4

shows, this includes outliers in the 30% range (that is, only slightly better than randomly picking answer options):

- Within quantum physics (QP), this outlier is the Quantum Mechanics Visualization Inventory (QMVI) at 32%. This inventory heavily focuses on graphical visualizations of wave functions.
- Within optics (OPT), the outlier is the Four-tier Geometrical Optics Test (FTGOT) at 26%. This inventory deals with ray optics, another graphical visualization topic.

However, the extreme values on individual inventories mostly average out—except in categories with few inventories (e.g., relativity and laboratory skills).

The AI performs best in thermodynamics (85.2%), followed by astronomy (80.4%), reasoning (78.7%), and relativity (75.0%). Inventories in astronomy tend to focus on factual knowledge, while those in thermodynamics, reasoning, and relativity tend to use exact language. The AI’s performance is weaker in laboratory skills (35.0%), which include strategies for data collection and analysis.

B. How does language influence the performance of the AI system?

Figure 2 also shows the scores on each inventory in its respective nominal language (we use “nominal” to denote the language in which the inventory was presented). One immediate observation is the uneven coverage of assessments across languages. Many inventories are available in only a handful of languages, making broad comparisons challenging. Some assessments, however—most notably FCI—are available in numerous languages.

In the FCI, performance ranges from as low as 20% in Punjabi and 22% in Tamil to as high as 74% in Portuguese and Polish, suggesting different performance across languages. Since the FCI uses five answer options, a 20% performance is equivalent to random guessing.

Similarly, MBT scores fluctuate widely—e.g., 27% in Persian, 53% in Finnish, and 44% in Italian and Hungarian ($SD = \pm 6\%$; range = 26%)—with Finnish performance nearly double that of Persian. In contrast, some inventories show more consistent performance across multiple languages. For example, QMCS shows high and relatively stable scores (78%–86%) in the available languages ($SD = \pm 4\%$; range = 8%), indicating that the AI handles this conceptual domain and its translations fairly well. TUV, ADT, and FMCE also show relatively stable performance in the 60%–80% range. Therefore, these results suggest that performance stability across languages is inventory dependent.

Out of 36 inventories available in both English and at least one other language, English was the best-performing language in 27 cases (75%). Figure 5 shows the relative performance by language compared to English. For some languages, only a single inventory was available—usually

	Arabic	Bengali	Catalan	Chinese	Croatian	Czech	Dutch	English	Filipino	Finnish	French	German	Greek	Hebrew	Hindi	Hungarian	Icelandic	Indonesian	Italian	Japanese	Malay	Norwegian	Persian	Polish	Portuguese	Punjabi	Russian	Serbian	Slovak	Spanish	Swedish	Tamil	Thai	Turkish	Ukrainian
ADT	-	-	-	-	-	-	-	78	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
BEMA	-	-	-	44	-	-	-	66	-	-	-	-	-	-	-	-	-	-	-	53	-	-	-	-	57	-	-	-	-	81	83	-	-	-	-
CCI	-	-	-	-	-	79	-	83	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	49	51	-	-	-	-
CDPA	-	-	-	-	-	-	-	33	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	40	-	-	-	-	-
CSEM	-	-	-	-	-	-	-	54	-	-	-	-	-	-	-	-	-	38	-	-	53	-	-	-	-	-	-	-	-	53	45	-	-	-	-
CTSR	-	-	-	-	-	-	-	86	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	62	-	74	76	-	-	-	-	-
DIRECT	-	-	-	59	-	-	-	49	-	52	-	45	53	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	33	49	-	-	-	-
DS	-	-	-	-	-	-	-	97	-	-	-	79	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ECA	-	-	-	-	72	-	-	79	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ECCE	-	-	-	-	-	-	-	61	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
EMCA	-	-	-	-	-	-	-	78	-	-	-	-	-	-	-	-	-	78	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
EMCS	-	-	-	-	-	-	-	79	-	75	-	-	-	-	-	-	-	75	-	-	-	-	-	-	-	-	-	-	-	71	-	-	-	-	-
FCI	58	39	67	57	67	62	62	69	59	67	70	68	66	50	49	67	66	-	72	60	57	69	52	73	74	20	56	-	62	68	64	22	47	66	
FMCE	-	-	-	-	-	-	-	72	-	-	-	-	-	-	-	-	-	74	-	39	-	-	-	-	-	-	-	-	-	81	-	-	-	-	-
FORT	-	-	-	-	-	-	-	70	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
FTGOT	-	-	-	-	-	-	-	27	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	23	-
FVA	-	-	-	-	-	-	-	76	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GECI	-	-	-	-	-	-	-	82	-	-	-	-	-	-	-	-	-	-	-	82	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
HTCE	-	-	-	68	-	-	-	76	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
IBCDC	-	-	-	-	-	-	-	62	-	-	64	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
IBCM	-	-	-	-	-	-	-	74	-	-	66	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
LPCA	-	-	-	-	-	-	-	84	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
LPCI	-	-	-	-	-	-	-	65	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	62	-	-	-	-	-
LSCI	-	-	-	-	-	-	-	73	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
MBT	-	-	-	-	-	-	-	44	-	53	41	41	44	-	-	-	-	-	38	44	38	-	27	-	35	-	-	-	40	-	-	-	37	-	-
MCS	-	-	-	-	-	-	-	58	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
MUQ	-	-	-	-	-	-	-	37	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
MWCS1	-	-	-	-	-	-	-	79	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	76	-	-	35	-	-
MWCS2	-	-	-	-	-	-	-	68	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	62	-	-	-	-	-
NGCI	76	-	-	-	-	-	-	87	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
NGPSD	-	-	-	-	-	-	-	95	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PIQL	-	-	-	-	-	-	-	80	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
QMCA5	-	-	-	-	-	-	-	74	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	67	-	-	-	-	-	-	-	-	-	-
QMCA6	-	-	-	-	-	-	-	81	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
QMCS	-	-	-	-	-	-	-	86	-	81	-	-	-	-	-	-	-	-	-	78	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
QMFPS	-	-	-	-	-	-	-	60	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	49	-	-	-	-	-
QMS	-	-	-	-	-	-	-	69	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
QMVI	-	-	-	-	-	-	-	32	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
QPCS	-	-	-	-	-	-	-	71	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	40	-	-
RAPT	-	-	-	-	-	-	-	81	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
RCI	-	-	-	-	-	-	-	75	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
RFCI	-	-	-	-	-	-	-	80	-	80	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
RKI	-	-	-	-	-	-	-	70	-	-	67	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
RRMCS	-	-	-	-	-	-	-	74	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SGCE	-	-	-	-	-	-	-	60	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SPCI	-	-	-	-	-	-	-	95	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
STPFaSLSh	-	-	-	59	-	-	-	86	-	-	-	-	-	-	-	-	70	-	86	-	-	-	-	72	-	-	-	-	-	94	-	-	-	-	-
STPFaSLlo	-	-	-	-	-	-	-	97	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
TCE	-	-	-	59	-	-	-	88	-	-	-	-	-	-	-	-	-	-	-	85	-	-	-	-	78	-	-	-	-	-	-	-	-	-	-
TCS	-	-	-	69	-	-	-	79	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	59	-	-
TOAST	-	-	-	-	-	-	-	83	-	-	-	-	-	-	-	-	-	-	-	80	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
TUG-K2.6	48	-	-	-	-	-	-	-	-	60	67	60	-	43	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
TUG-K3.0-4.0	-	-	-	45	-	-	-	55	-	-	-	-	44	-	-	-	-	-	-	-	-	-	-	65	-	-	-	-	60	50	-	-	-	-	45
TUV	58	-	-	-	-	-	-	60	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	62	-	-	-	-	-	-

FIG. 2. Scores in percent on each of the inventories in each of the available languages. A green background indicates that the AI score is higher than the average student post-test scores found in literature for undergraduate-level courses (see column “%Post” in Tables II–V); red indicates lower AI-performance. A blue background indicates that no student score was available.

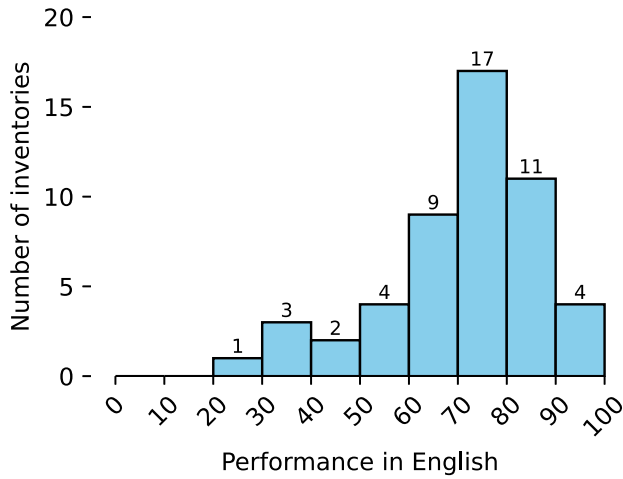


FIG. 3. Distribution of scores achieved by GPT-4o on physics concept inventories administered in English.

the FCI. Notably for Finnish and French, where we have several translated inventories available, the performance is similar to that in English, with a number of other languages, mostly European, trailing closely behind. On the other hand, performance is much lower in languages such as Persian and Thai.

Merely examining performance on an inventory in a given language is insufficient to determine whether the AI struggles with the same items across languages. Although a detailed item-level analysis is beyond the scope of this manuscript, we provide a brief investigation to assess whether items that the AI finds challenging in English remain so in their non-English versions. One indicator of item difficulty is the agreement among the three independent responses and how consistently the AI system provides the correct answers. Since all inventories—except TUG-K2.6—were presented to the AI in English, we can directly compare GPT-4o’s performance on English items to their translated counterparts. Our findings indicate that items deemed difficult in English tend to remain challenging in other languages, and vice versa. When all three English responses were incorrect, the correct response rate for the non-English versions was only 18%. This increased to 33% when one English response was correct, 48% for two, and 82% for three. Table VIII in Appendix C illustrates that this trend—where the same items are consistently challenging across various languages—holds true for most languages in our dataset.

An interesting finding is that GPT-4o often exhibits language-switching behavior with non-English inventory items. Portuguese and Spanish were the only two languages where the majority of answers (56% and 59%, respectively) were entirely in the nominal language. In all other cases, the model predominantly switched to English, either fully or in part. Table VII in Appendix B provides additional details on this language-switching behavior.

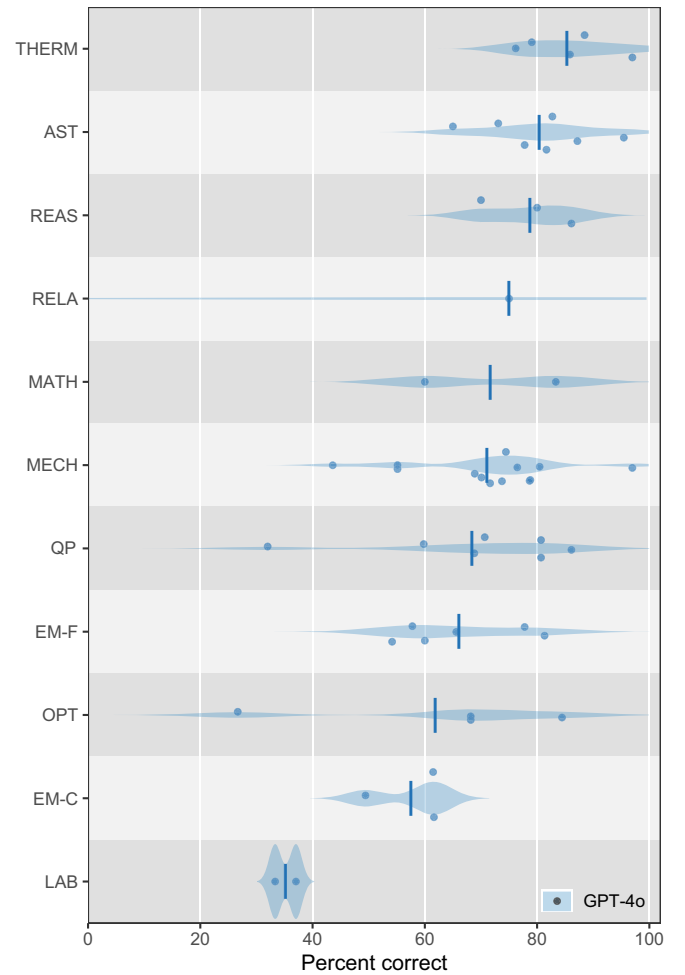


FIG. 4. Sina plot [150] of GPT-4o’s performance on English-language concept inventories, grouped by category (see Table I). The categories are sorted by average performance, indicated by the vertical markers.

C. How does the AI’s performance compare to student performance at the undergraduate level?

In Fig. 2, we use color to indicate the relative performance of GPT-4o compared to published postinstruction scores at the undergraduate student level found in the literature (see column “%Post” in Tables II–V). For most inventories in most languages, the AI system outperforms the student average. Where student data were available, the AI outperformed average undergraduate postinstruction scores in 68.9% of cases. To obtain this value, we compared student averages on an inventory to AI’s performance on the inventory for each available language and summed across all inventories.

As we have pointed out in the Introduction (Sec. ID), the comparison of the average performance of GPT-4o to that of students provides a rough proxy measure for its capabilities in solving physics conceptual tasks in relation to student capabilities. However, caution is needed when

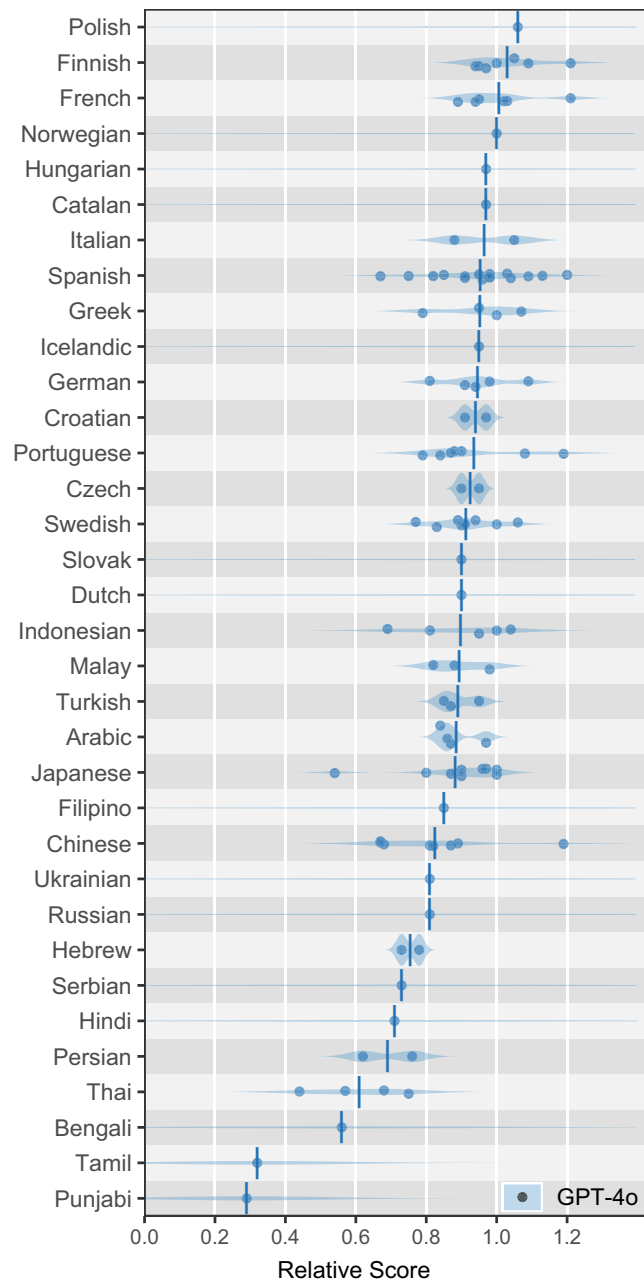


FIG. 5. Sina plots [150] of GPT-4o's performance on inventories in different nominal languages, relative to performance on the same tests in English. The English performance on each inventory was normed as unity, and the plots show the distribution of other-language relative performance. The nominal languages are sorted by average relative performance, indicated by the vertical markers.

interpreting these results, as the profile of GPT-4o's and the "average" student's strengths and difficulties can differ.

Figure 6 shows the distributions of post-test student scores and AI scores across all languages, grouped by subject category. The averages are indicated by vertical markers. With the exception of laboratory skills (LAB), GPT-4o outperforms the average undergraduate student

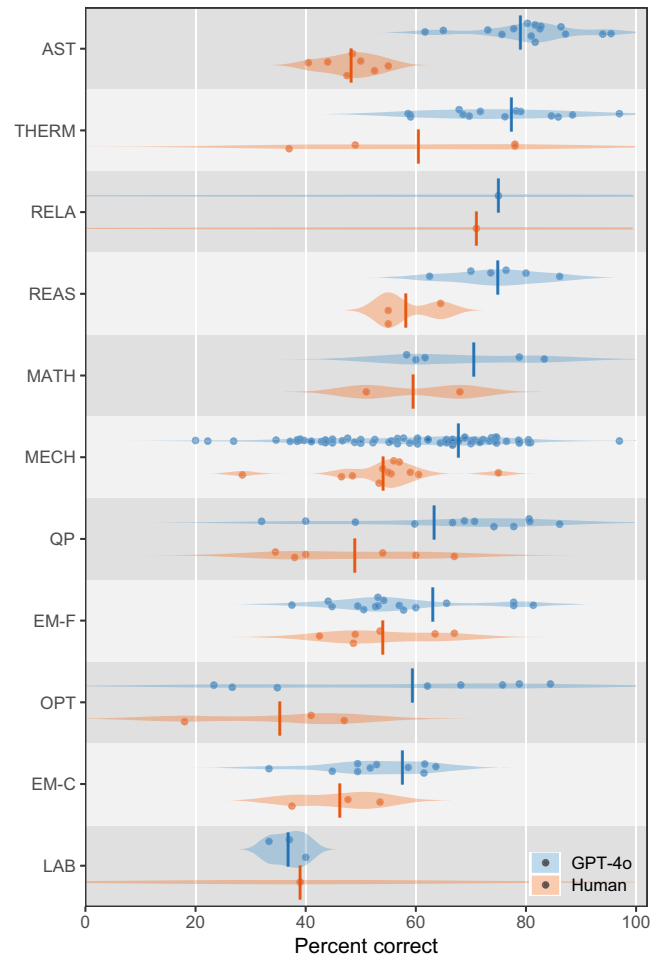


FIG. 6. Sina plots [150] of GPT-4o (all languages) and student scores on physics concept inventories grouped by subject category. The categories are arranged in descending order of the average GPT-4o score; average scores are indicated by vertical markers.

in every subject category of concept inventories, with the most significant differences in astronomy and reasoning. The extremely wide distribution of the mechanics (MECH) scores is mostly because of the fact that the Force Concept Inventory (FCI) is available in a wide variety of languages, including languages such as Punjabi and Tamil, which GPT-4o does not appear to adequately master (see Fig. 5). The wide distributions in quantum physics and optics are again because of inventories with mostly graphical, visual representations.

Well-designed assessment instruments for teaching typically include multiple-choice distractors that probe specific student misconceptions. Since we drew our questions from the PhysPort library, we can assume that most concept inventories in our dataset were designed carefully. Therefore, we expect students to gravitate toward certain incorrect answers that reflect common misconceptions. This raises the question of whether the AI similarly gravitates toward specific incorrect answers. Although

not all the assessments analyzed in this study had five possible choices, the majority did, allowing us to use this as a reference point (see Appendix C and Table IX for more detail). If the AI were selecting incorrect answers randomly, we would expect a 75% probability that two incorrect answers differ and a 25% chance that they would be the same. However, for our data, this was reversed: 66% of all items had the same two incorrect answers, while 34% had different ones. The same pattern held for cases where all three answers were incorrect. In a random scenario, we would expect 6.25% of items to have three identical answers, 37.5% to have three different ones, and 56.25% to have two identical answers (see Appendix C for more details). In reality, AI responses were far from random and tended to gravitate toward specific incorrect answers: 53% of such items had three identical incorrect answers, while only 8% had three different ones. Similar trends were observed for both English and non-English items. This analysis suggests that the AI system exhibits some consistency in its choice of incorrect answers. However, further research is necessary to investigate how students and AI systems may differ in how they respond to different types of multiple-choice options.

D. How does the presence of images influence the performance of the AI system?

Here, we calculate the performance based on the sum across individual inventory items appearing in the different inventories (all languages combined). For each item, we coded if it contained or referred to an image—that is, a visual representation such as a sketch, graph, or diagram—and whether interpreting the image was required for correctly solving the task (“required image”), or if the image was redundant, meaning all information required for solving the task was already provided in the text (“unneeded image”). The percentages shown in Fig. 7 were obtained by dividing the number of correct responses in each category by the total number of submissions in that category.

Overall, we found that the performance on text-only tasks was 81%, compared to 79% on tasks containing unneeded images and just 49% on tasks with required images (see Fig. 7).

When examining individual subject categories, GPT-4o consistently performs worse on items that require image interpretation than on text-only items. Relativity, optics, mechanics, and mathematics (which had no unneeded image items), as well as astronomy, exhibited especially large performance gaps. Furthermore, while QP as a whole was not among the lowest-performing subject categories on image-based problem types, QMVI—entirely image-based and composed predominantly of required-image items—was the second worst-performing inventory in English (32%). In line with previous research on the topic [36],

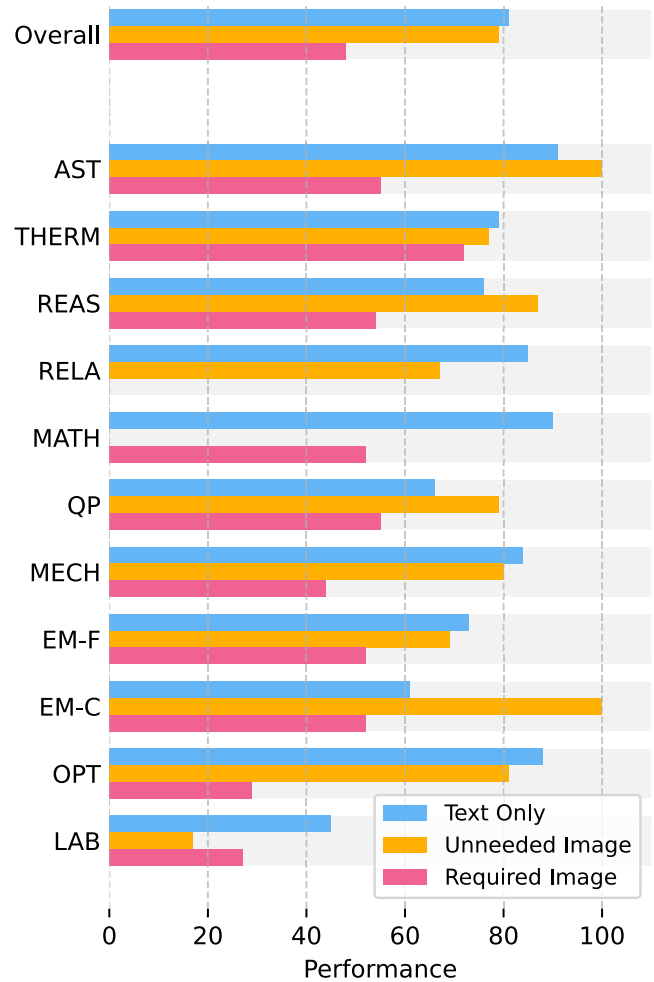


FIG. 7. Performance on inventory items categorized by their usage of images and grouped by their subject category. Note that in contrast to the inventory-level analysis presented in previous sections, the values on this histogram present the summative performance on all submissions of inventory items belonging to a given image category, collected across the concept inventories in each subject category.

these findings suggest that visual interpretation is one of GPT-4o’s major weaknesses.

V. DISCUSSION AND FUTURE RESEARCH DIRECTIONS

Unlike many previous studies where inventories were provided as text-only materials, this study provided the AI with a screenshot of the item as it would appear to a student. Using authentic screenshots rather than isolated text offers a closer approximation of the model’s behavior in realistic educational settings, especially in multilingual and multi-modal contexts. However, this approach also increases the complexity of the input and the analytical demand for interpreting the model outputs.

Across the majority of inventories and subject categories (with laboratory skill being the exception), GPT-4o

achieves higher scores and outperforms the average reported postinstruction undergraduate-student score. It needs to be emphasized that the instruments are research-based and thus carefully designed with regard to psychometric properties, which is reflected in their student score distributions: (i) The inventory items aim for medium difficulty, resulting in average student scores around 50%, and (ii) The items are designed for high discrimination, leading to broad score distributions. Thus, the finding that the AI performs better than the average student on these inventories broadly means that it performs above 50% on these instruments. It also means that in almost all cases, some students in the dataset still outperformed GPT-4o. Furthermore, in studies where results for graduate or postdoctoral populations were available, these outperformed GPT-4o.

It is important to note that even when GPT-4o outperformed the average student, this does not mean that it necessarily exhibited a strength and difficulty profile similar to that of a well-performing student. Our findings support previous research [37] and suggest that image interpretation is one common GPT-4o difficulty that is not typically seen in students.

This highlights the importance of carefully considering the role of AI in physics assessment: while it can match or exceed undergraduate averages on concept inventories, its performance should not be mistaken for human expertlike reasoning or deep conceptual understanding. For example, a test on which AI performs poorly might not necessarily be difficult for students, and vice versa. Much depends on the type of tasks involved and how they align with the AI's and students' strength and difficulty profiles.

If AI systems like GPT-4o play a role in the pipeline of assessment design and validation, the role will likely be different from that of a student or a domain expert. While there exists some research on making an LLM behave like a student with a certain difficulty profile [151], more work is needed to establish the feasibility of such use at scale and across different subject areas. On the other hand, given the increasingly important role AI systems will likely play in physics education, it is worthwhile to consider developing assessments specifically for those AI systems that are going to be used in educational contexts. Such assessment would provide curriculum developers and instructors with even more pertinent information about these systems' capabilities and potential in educational contexts. Exactly what such an assessment would look like remains unclear. However, it is reasonable to expect that it will be developed in response to the practical needs and demands of the new educational landscape.

On the linguistic side, our results reveal a complex relationship between performance and language. The overall pattern that emerged is that GPT-4o performed better in most cases when tasks were given in English. While some other languages—mostly European and Latin script—

exhibited roughly similar performance, others showed significantly lower results. One contributing factor for this difference could be that GPT-4o had to handle English language prompts with screenshots using other language scripts, a task that is presumably harder than just dealing with one language or script. Understanding this better would need to be explored further in a future study. This is particularly important as it has implications for accessibility and utility of AI tools for non-Western language-speaking populations and may risk perpetuating or even exacerbating disparities in access to educational and technological resources worldwide.

Our results also indicate that tasks found to be challenging in English tend to remain so when posed in other languages. Notably, the model often produces consistent patterns of incorrect responses, even though its training data likely differs across languages. This observation raises questions about whether these errors stem from biases in the training data or reflect deeper limitations in the model's inference process. While our study does not resolve these issues, it highlights that certain items consistently challenge the model across languages. These findings may inform future research on assessment designs that account for the limitations of multimodal AI systems, whether the items are visual or text-based.

Examining the presence and function of visual representations in inventory items reveals a clear pattern. The AI performed better on text-only items and items with redundant visual representations, compared to items where visual interpretation of images was necessary for solving the task correctly. This corroborates previous findings on the topic, suggesting that the AI's visual reasoning remains a major weakness, hampering its ability to engage effectively with graphical or picture formats common in physics tasks [31,37]. Anecdotal evidence in recent months suggests that, compared to GPT-4o, GPT-o3, and o4-mini [152] perform better on physics tasks, including numerical ones, but often still struggle with the interpretation of figures.

There may be other variables, beyond subject area, language, and image presence, which influence the performance of GPT-4o and other AI systems on physics tasks. The mathematical complexity of a task and the presence of redundant information are just two of the many possible aspects that could influence AI's ability to solve a task. More research is needed to develop a better sense of what makes a physics task difficult for an AI system to solve.

Based on our findings, we suggest that future efforts to improve AI performance on physics tasks should focus on enhancing the multimodal processing components and achieving more balanced performance across languages and conceptual domains.

VI. IMPLICATIONS FOR INSTRUCTION

Our study is decidedly exploratory and does not aim to provide direct advice to instructors on how to implement

GPT-4o in education. Still, there are some broad findings that emerged from our exploration of GPT-4o's performance on physics concept inventories, which can be informative for physics instructors and curricular developers.

We expect that in the future, multimodal AI systems like GPT-4o will continue to influence the space of physics education through at least two mechanisms. First, their accessibility and availability and their relatively high capabilities make them attractive to learners. They will likely continue to be used by students to help them with physics tasks. Second, AI tools are likely to see increased uptake by educational institutions to support students in their learning as well as to support instructors in their teaching and administrative tasks [153–158].

It is important to note once more that while GPT-4o, on the surface, exhibits performance that is numerically better than university students' postinstruction average, a closer look reveals important caveats with this simple interpretation. If students want to use GPT-4o productively and responsibly, they should be aware of its limitations. We believe that instructors should inform students of these drawbacks to mitigate the risks of over-reliance on AI tools, and foster a critical perspective on the outputs these tools generate. This should arguably become one of the newly emerging instructional goals because such skills will remain useful even as students leave education and enter the workforce. As AI becomes part of our everyday lives and work, evaluating its output becomes an important skill, which cannot be learned by always outsourcing physics reasoning to an AI. Exposing AI's drawbacks to students can thus also serve as a motivation for students to engage more deeply in learning physics. Our research can help inform such efforts. For example, we have shown that GPT-4o's performance is not equally good across all subject areas of physics and that it often struggles when prompted in non-Western languages. Furthermore, a major drawback is its ability to interpret images.

However, because of the incredible pace of AI progress, sooner or later, physics educators will have to contend with the question of what remains meaningful to teach, when AI systems perform well on many tasks that were previously squarely in the domain of human physics experts. The physics education community will likely need to seriously reflect on whether physics curricula, which have in many cases remained nearly unchanged for decades, should evolve to better reflect the new reality. If we conclude that the kind of conceptual understanding that is being tested by research-based concept inventories is still valuable and important for our students, then these assessments will likely continue to play an important role in evaluating whether our students have reached the desired learning objectives. In such a case, they should arguably be administered so that students cannot use AI to help them.

In making these important decisions, the physics education community will likely also need to address the following (and other) questions on a continuous basis:

What are the foundational skills that we should not routinely outsource to AI? What are the central tasks that students first need to master themselves, so that they can later judiciously and responsibly outsource them to AI? How do we ensure that access to top-performing AI does not generate or worsen existing divides based on students' economic or ethnic background?

VII. LIMITATIONS

This study is decidedly exploratory and empirical. The preparation of item images was done manually, and some manual cleanup of the data was required. Given more than 3600 images and the random oddities occurring in over 14 000 solutions generated by a probabilistic system, clerical errors cannot be excluded. Additionally, each screenshot of inventory items was iterated only 3 times: given the stochastic nature of LLM outputs, this introduces variability that limits the results' generalizability. The study also did not consider the quality of the concept inventories' translations, and lower scores may be because of incorrect or confusing translations. AI's performance might be dependent on prompts, and the ones used for our study (shown in Figs. 9 and 10 of Appendix A) may not be the best choices. Future studies may look into whether and how prompt-engineering techniques might improve (or worsen) AI's performance.

Furthermore, it is unclear to what extent the use of an English prompt, combined with inventory text in the nominal language of each inventory, influenced performance across languages and contributed to language-switching behavior; Appendix B discusses some of the preliminary observations. Future research may explore this further by varying the prompting approach, for example, by prompting the system entirely in the nominal language. However, because of the unreliability of structured outputs with non-English prompts (see Appendix A), such an approach would likely require alternative output formats, more human involvement in the coding of the answers, as well as accurate translations of the prompts into multiple languages.

The model's visual encoder transforms the pixel data into a latent representation that encapsulates both textual and graphical information. This has the side effect that there is no standalone raw OCR (optical character recognition) output that we could compare to the original text in the various Latin and non-Latin scripts used in the study. One possible explanation for the difference in performance between languages could be incorrect recognition of non-Latin characters.

Furthermore, it should be noted that we did not score the correctness of the physics reasoning in the AI-written explanations. Future studies could explore this in more detail, potentially using the data collected in this study [159] (as an example, see Fig. 12 in Appendix A) to evaluate the model's reasoning separately from its final answers. An example of how this could be done can be found in Ref. [36].

A possible stumbling block for the AI can be the processing of graphical information unrelated to a physics concept or language. For example, for item 7 of the FCI, the multiple-choice options are embedded in a graphic showing a ball swinging in a circular path and are not listed separately in the text. Although the AI frequently gave physically reasonable explanations, across 32 languages and 96 total answers, it selected the correct embedded option only once. For no obvious reason, it chose instead the same incorrect answer an eyebrow-raising 88 times. Once again, this finding aligns with previous studies showing that the graphical layout of images and the spatial arrangement of answer options can play an important role in the AI's selection of answers [37].

The performance of GPT-4o, as measured in our study represents a momentary snapshot of one model's capabilities in early 2025. It is very likely that future models will outperform GPT-4o. However, to keep track of such developments, more studies similar to ours will be needed.

Finally, human postinstruction scores were gathered on a best-effort basis, which may introduce additional variability into the comparisons. Moreover, most of the human data came from English-speaking students taking the English versions of the inventories.

Against the background of these limitations, it is clear that our study only scratches the surface of this exciting area of research and there are many open questions that invite further exploration.

VIII. CONCLUSION

The results of this study underscore the complexity and variability inherent in using multimodal large language models for physics assessment tasks across multiple languages, subject categories, and formats.

The marked differences in performance across languages highlight that GPT-4o is not equally competent in all tongues. Anecdotally, this is true for many LLMs. This suggests a risk of generating new, as well as maintaining or exacerbating existing inequities in the access to educational resources and technologies across the world.

Based on published student scores on the tested inventories, we found that GPT-4o outperforms undergraduate student postinstruction averages on most inventories, and in all subject categories except laboratory skills. The reasons for the subject's dependence on its performance are not entirely clear. Possible explanations include different levels of representation in the training data and its varying quality across subjects, or potential differences in the inherent difficulty of tasks in the assessments covering different categories.

We have also found that the presence of non-redundant visual representations negatively influences AI's performance across all subject categories. This suggests that the AI's vision abilities still present a major weakness and consequently limit its utility for some educational uses.

In sum, this exploratory study demonstrates that the studied AI system exhibits significant variations in performance depending on the language, conceptual domain, and presence of visual information. The work points toward the need for future improvements in training data diversity, model fine-tuning, and prompt engineering to enhance its performance. It also highlights the need for careful consideration when implementing such AI systems in educational contexts, ensuring that their use is both equitable and aligned with pedagogical goals.

ACKNOWLEDGMENTS

We would like to thank Sam McKagan for her invaluable support in making this study possible, as well as for her contributions to PhysPort and other educational portals, and for her ongoing efforts to advance physics education research. We also thank Lyle Barbato for enabling the publication of our data on PhysPort.

DATA AVAILABILITY

The data that support the findings of this article are openly available [159].

APPENDIX A: IMPLEMENTATION OF THE LLM-API CALLS

The screenshots were submitted via the deployment's API in base64 encoding as shown in Fig. 8, setting a

```
responseGPT = client.chat.completions.create(
    model="EthelOmni",
    temperature=0.7,
    messages=[
        {"role": "system", "content": role},
        {
            "role": "user",
            "content": [
                {"type": "text", "text":
                    prompt},
                {
                    "type": "image_url",
                    "image_url": {
                        "url": image_data
                    }
                }
            ]
        }
    ],
    max_tokens=1000,
    response_format={
        "type": "json_schema",
        "json_schema": {
            "name": "problem_response",
            "strict": True,
            "schema": json_schema
        }
    }
)
```

FIG. 8. The API call used in this study. “EthelOmni” is a GPT-4o Azure deployment.

You are a physics and mathematics expert.
 You are given images of multiple-choice test questions,
 which you will answer correctly in JSON format.
 You work very carefully, and you strongly favor correct
 answers over rapid responses.
 You do mathematical calculations and derivations
 step-by-step.
 If there are graphics, you consult them more than once if
 needed in order to get information for your reasoning.

FIG. 9. The role used for this study.

temperature of 0.7 (which is the default in the chat clients); newer reasoning models like GPT-o1 or GPT-o3-mini do not accept temperature anymore. As a role, we used the text shown in Fig. 9, and as prompt the text shown in Fig. 10. Role (job description) and prompt (task description) tend to be slightly redundant; newer models do neither expect nor accept the role parameter anymore.

To facilitate further processing and evaluation of the answers, we provided the structured JSON output schema shown in Fig. 11; this forces the model to provide its output with the given data structure instead of in narrative form. We included the “problem_description” and “explanation” fields in the output structure to force the AI system to describe the problem “in its own words” and to provide reasoning for the response, thereby triggering chain-of-thought [10].

You are tasked with correctly solving the problem or
 problems from the image.
 The image should contain problem(s) with the problem
 identifiers: [Problems]
 For this multiple-choice problem or these multiple-choice
 problems, carefully consider the situation and document
 your reasoning, and only then pick the answer choice that
 aligns with your reasoning.

FIG. 10. The prompt used for this study. Before submission, [Problems] is replaced by the problem numbers contained in the image, e.g., “3, 4, 5.” The image is submitted alongside in base64-encoding.

As LLMs are probabilistic systems, each screenshot was evaluated 3 times; this resulted in output like the one shown in Fig. 12. Each run is independent, and the model may provide different interpretations of the problem, different reasoning steps, and potentially different final answers every time. The firmer the required concept for a problem is anchored in the model’s parameters, the less its answers will vary.

The role and prompt we used can result in mixed-language outputs. While this issue could probably have been avoided by translating the role and prompts into the language of the inventories, doing so for languages using non-ASCII characters would likely have made the

```
# Define the JSON schema for the structured response
json_schema = {
  "type": "object",
  "properties": {
    "problems": {
      "type": "array",
      "items": {
        "type": "object",
        "properties": {
          "problem_description": {"type": "string"},
          "questions": {
            "type": "array",
            "items": {
              "type": "object",
              "properties": {
                "question_number": {"type": "integer"},
                "explanation": {"type": "string"},
                "correct_answer": {"type": "string"}
              },
              "required": ["question_number", "explanation", "correct_answer"],
              "additionalProperties": False
            }
          },
          "required": ["problem_description", "questions"],
          "additionalProperties": False
        }
      },
      "required": ["problems"],
      "additionalProperties": False
    }
  },
  "required": ["problems"],
  "additionalProperties": False
}
```

FIG. 11. The data structure used for this study.

```

{
  "solutions": [
    {
      "problems": [
        {
          "problem_description": "What happens to a positive charge that is placed at
            rest in a uniform magnetic field? (A uniform field is one whose strength and
            direction are the same at all points.)",
          "questions": [
            {
              "question_number": 21,
              "explanation": "The force on a charge moving in a magnetic field is
                given by the Lorentz force,  $F = q(\mathbf{v} \times \mathbf{B})$ , where  $q$  is the charge,  $v$ 
                is the velocity, and  $B$  is the magnetic field. If the charge is at
                rest, then  $v = 0$ , and thus  $F = 0$ . Therefore, there is no force
                acting on the charge, and it will remain at rest.",
              "correct_answer": "e"
            }
          ]
        }
      ]
    },
    {
      "problems": [
        {
          "problem_description": "What happens to a positive charge that is placed at
            rest in a uniform magnetic field? (A uniform field is one whose strength and
            direction are the same at all points.)",
          "questions": [
            {
              "question_number": 21,
              "explanation": "A magnetic field exerts a force on a moving charge,
                given by the Lorentz force equation:  $F = q(\mathbf{v} \times \mathbf{B})$ , where  $q$  is
                the charge,  $v$  is the velocity, and  $B$  is the magnetic field. If the
                charge is at rest,  $v = 0$ , and thus the force  $F = 0$ . Therefore, there
                is no force acting on the charge to cause any acceleration or
                motion. The charge remains at rest as there is no initial velocity
                and no force to change its state.",
              "correct_answer": "e"
            }
          ]
        }
      ]
    },
    {
      "problems": [
        {
          "problem_description": "What happens to a positive charge that is placed at
            rest in a uniform magnetic field? (A uniform field is one whose strength and
            direction are the same at all points.)",
          "questions": [
            {
              "question_number": 21,
              "explanation": "A magnetic field exerts a force on a moving charge
                according to the Lorentz force law, which states that the force is
                proportional to the velocity of the charge and the magnetic field.
                The force is given by  $\mathbf{F} = q(\mathbf{v} \times \mathbf{B})$ . If the charge is initially at rest, then the
                velocity  $\mathbf{v}$  is zero, resulting in zero force.
                Therefore, a charge placed at rest in a uniform magnetic field will
                experience no force and will remain at rest.",
              "correct_answer": "e"
            }
          ]
        }
      ]
    }
  ]
}

```

FIG. 12. Typical output; each problem is independently solved 3 times (three “problems”-blocks inside of “solutions”).

mapping between the prompt and the fields in the JSON structure unreliable. Colloquially speaking, “the AI was allowed to think in the language of its choice.” Some observations on this behavior are discussed in Appendix B.

APPENDIX B: LANGUAGE-SWITCHING BEHAVIOR

While the model transforms the input into abstract latent representations rather than processing it in any particular human language, the language of the output may provide some insights into these underlying representations. Table VII shows that, for certain languages, overall accuracy increases when responses are generated entirely in English or as a mix with English. For example, for Bengali, Persian, and Punjabi, accuracy is noticeably higher when English is used in the output. This pattern may reflect differences in the effectiveness of the model’s representations for these languages compared to English. In contrast, for languages such as English, Spanish, and French, high correctness is maintained even when responses remain solely in the original language.

Furthermore, when performance on a given item in English is treated as a baseline for difficulty, Table VI indicates that for non-English inputs the output remains in the original language in 24% of cases that are easier in English (i.e., those with 3/3 correct answers). This proportion is higher than the 15%–17% observed for items where at least one response in English was incorrect. These observations suggest a correlation between the response language and the relative difficulty as measured by performance in English.

Although GPT-4o does not engage in extended chain-of-thought reasoning similar to GPT-o1, DeepSeek, or GPT-o3-mini, its selection of output language—whether remaining in the nominal language, switching entirely to English, or producing a mixed-language response—emerges from patterns learned during training rather than from explicit instructions. The prompt was provided in English, while the item content was often in another language and accompanied by multimodal elements such as diagrams and sketches. Consequently, the variation in output language and the resulting solution accuracy offer insights into the model’s internal processing of diverse input formats without implying deliberate reasoning.

TABLE VI. Language switching of a non-English item based on level of difficulty for the same item in English.

Language switch	Answers correct in English			
	0/3	1/3	2/3	3/3
Switch to English (%)	69	69	69	63
Mixed language (%)	14	17	15	12
Nominal language (%)	17	15	17	24

TABLE VII. Performance by language and language switching in percent.

Language	%Correct	Nominal language		Switch to English		Mixed languages	
		%	%Correct	%	%Correct	%	%Correct
Arabic	60	6	89	92	58	2	83
Bengali	39	2	100	98	38	0	
Catalan	67	19	94	60	57	21	68
Chinese	58	2	57	97	58	1	20
Croatian	69	11	81	64	69	25	66
Czech	69	3	100	58	54	40	89
Dutch	62	34	68	39	46	27	79
English	72	100	72	...	...	...	...
Finnish	66	4	86	60	65	36	64
Filipino	59	0		100	59	0	
French	63	37	74	43	55	20	60
German	56	19	70	40	49	41	56
Greek	52	8	63	80	50	11	61
Hebrew	47	1	100	97	46	2	67
Hindi	49	9	75	90	47	1	0
Hungarian	67	9	100	70	65	21	58
Icelandic	66	0		100	66	0	
Indonesian	67	26	71	59	67	15	60
Italian	57	37	71	53	48	10	47
Japanese	63	8	87	83	60	8	72
Malay	50	0		100	50	0	
Norwegian	69	13	75	71	62	16	93
Persian	40	1	0	98	40	1	100
Polish	73	0		100	73	0	
Portuguese	64	56	73	34	50	10	63
Punjabi	20	2	100	93	19	4	0
Russian	56	8	57	84	55	8	57
Serbian	62	0		100	62	0	
Slovak	62	0		76	62	24	64
Spanish	62	59	63	33	62	8	52
Swedish	60	13	65	56	59	31	58
Tamil	22	1	100	87	19	12	36
Thai	47	7	67	86	44	7	55
Turkish	40	18	51	64	39	18	35
Ukrainian	45	3	0	79	45	18	50
Overall	63	46	71	45	55	9	60

APPENDIX C: ITEM LEVEL ANALYSIS

Table VIII can be used to assess whether the AI struggles with the same items across different languages. To illustrate how to read Table VIII, let us consider the second most common language, Spanish. There were 416 items (with 1248 responses) across all inventories in Spanish. For the corresponding items in English, there were 102 items with 0/3 correct, 42 items with 1/3 correct, 45 with 2/3 correct, and 227 with 3/3 correct in English. The percentages in the table show how the AI system performed on these items in Spanish. For the 102 items (306 responses) for which all three English responses were incorrect, 56 of the responses, corresponding to 18%, were correct in Spanish. Hence, items that were extremely difficult in English (no correct

TABLE VIII. Percentage of correct AI responses in various languages, grouped by the corresponding difficulty level of the same items in English. Items are categorized based on how many English responses (out of three) were correct. Each cell shows the percentage of correct responses in the given language, with the number of items for each category provided in parentheses.

Language	Answers correct in English			
	0/3	1/3	2/3	3/3
Arabic	17% (16)	8% (4)	50% (6)	85% (50)
Bengali	5% (7)	0% (2)	11% (3)	61% (18)
Catalan	14% (7)	0% (2)	78% (3)	93% (18)
Chinese	21% (46)	43% (24)	40% (15)	73% (153)
Croatian	3% (11)	33% (6)	67% (4)	92% (42)
Czech	17% (8)	8% (4)	76% (7)	88% (33)
Dutch	14% (7)	0% (2)	56% (3)	89% (18)
Finnish	24% (33)	48% (22)	48% (11)	90% (85)
Filipino	10% (7)	17% (2)	44% (3)	85% (18)
French	23% (42)	42% (12)	43% (21)	89% (86)
German	14% (26)	36% (15)	41% (9)	87% (46)
Greek	14% (36)	43% (17)	52% (9)	83% (49)
Hebrew	10% (7)	17% (2)	67% (3)	67% (18)
Hindi	14% (7)	0% (2)	11% (3)	74% (18)
Hungarian	19% (7)	17% (2)	56% (3)	93% (18)
Indonesian	25% (29)	48% (16)	44% (15)	84% (107)
Icelandic	14% (7)	0% (2)	67% (3)	93% (18)
Italian	17% (18)	22% (6)	56% (6)	92% (26)
Japanese	16% (47)	31% (14)	35% (17)	83% (163)
Malay	9% (29)	37% (10)	52% (9)	82% (40)
Norwegian	10% (7)	33% (2)	56% (3)	98% (18)
Persian	20% (18)	33% (6)	28% (6)	59% (26)
Punjabi	10% (7)	0% (2)	56% (3)	20% (18)
Polish	29% (7)	17% (2)	67% (3)	98% (18)
Portuguese	26% (44)	37% (20)	48% (14)	84% (125)
Russian	5% (7)	17% (2)	67% (3)	78% (18)
Serbian	67% (1)	33% (3)	0% (1)	70% (19)
Slovak	14% (7)	17% (2)	67% (3)	85% (18)
Spanish	18% (102)	32% (42)	56% (45)	88% (227)
Swedish	18% (49)	36% (29)	44% (18)	84% (122)
Tamil	5% (7)	50% (2)	22% (3)	26% (18)
Thai	17% (21)	25% (8)	24% (7)	59% (76)
Turkish	15% (41)	13% (13)	45% (11)	84% (31)
Ukrainian	17% (10)	0% (2)	33% (1)	74% (13)
Overall non-English	18% (725)	33% (301)	48% (274)	82% (1771)
Overall non-English text only	16% (51)	52% (18)	65% (51)	84% (759)

answers at all) were also difficult in Spanish. This increases to 32% (40/126) for the 42 items with 1/3 correct in English, 56% (75/135) for the 45 items with 2/3 correct in English, and 88% (601/681) for the 227 items where all three English responses were correct. The data show that as the items become easier in English, they also become easier in Spanish. Similar trends can be observed for other languages. The only notable exceptions results for Punjabi

TABLE IX. Incorrect answer analysis. Percentages with the corresponding number of items in parenthesis for which the incorrect answers were the same or different.

Incorrect answers	Different or same	All	English	Non-English
3	3 different	8% (98)	6% (18)	9% (80)
3	3 the same	53% (617)	61% (173)	50% (444)
3	2 the same	38% (448)	33% (93)	40% (355)
2	2 different	34% (197)	28% (37)	36% (160)
2	2 the same	66% (376)	72% (94)	64% (282)

and Tamil, which have only 20% and 26% correctness for the items where the AI performed well in English (3/3).

Table IX shows the items that had two or three incorrect answers for the set of the three AI responses. It then shows if the selected multiple-choice items were the same or different. The percentages for the full dataset as well as the dataset split into English and non-English responses show that the AI frequently picked the same incorrect multiple-choice items.

If picking a particular incorrect answer was random, the theoretical probabilities would be different from the percentages in Table IX. Hence, the AI system gravitates toward particular incorrect answers. Here are the calculations to get to those probabilities:

- When two of the three answers are incorrect, we can have two possible outcomes.

For the two incorrect answers to be the same, we need to choose which of the four incorrect options appears twice, giving us four possibilities. The total number of ways two incorrect answers can be the same is therefore 4, while the total possible combinations of 2 incorrect answers from 4 options is $4 \times 4 = 16$. This gives us a probability of $4/16 = 1/4 = 0.25$. For the two incorrect answers to be different, the first incorrect answer can be any of the four options, and the second incorrect answer must be different from the first (three possibilities). The total number of ways this can happen is $4 \times 3 = 12$, and the total possible combinations remains 16, giving us a probability of $12/16 = 3/4 = 0.75$.

- When all three picks are incorrect, we have three possible scenarios.

For all three incorrect answers to be the same, we need to pick the same incorrect option 3 times. For any specific incorrect option, the probability is $(1/4) \times (1/4) \times (1/4) = 1/64$. Since we have four different incorrect options to choose from, the probability becomes $4 \times (1/64) = 4/64 = 1/16 = 0.0625$.

For all three incorrect answers to be different, the first incorrect answer can be any of the four options, the second must be different from the first (three possibilities), and the third must be different from both previous picks (two possibilities). The total number of ways this can happen is $4 \times 3 \times 2 = 24$.

With total possible outcomes when picking from four options 3 times being $4^3 = 64$, the probability is $24/64 = 6/16 = 0.375$.

Finally, for exactly two incorrect answers to be the same, we can have three patterns: $(1 = 2 \neq 3)$, $(1 = 3 \neq 2)$, or $(1 \neq 2 \neq 3)$. For each pattern, we

choose which of four options is repeated (four possibilities) and which option appears for the non-repeating position (three possibilities), giving $4 \times 3 = 12$ ways for each pattern. The total ways across all three patterns is $12 + 12 + 12 = 36$, resulting in a probability of $36/64 = 9/16 = 0.5625$.

-
- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, in NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems (Curran Associates Inc., Red Hook, NY, 2017), Vol. 30.
 - [2] T. H. Kung, M. Cheatham, A. Medinilla, ChatGPT, C. Sillos, L. De Leon, C. Elepano, M. Madriaga, R. Aggabao, and G. Diaz-Candido *et al.*, Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models, MedRxiv (2022), [10.1101/2022.12.19.22283643](https://doi.org/10.1101/2022.12.19.22283643).
 - [3] Samantha Murphy Kelly, ChatGPT passes exams from law and business schools, <https://edition.cnn.com/2023/01/26/tech/chatgpt-passes-exams/index.html> [accessed January 2023].
 - [4] OpenAI, ChatGPT, <https://chat.openai.com/> [accessed April 2024].
 - [5] A. M. Turing, Computing machinery and intelligence, *Mind* **49**, 433 (1950).
 - [6] C. R. Jones and B. K. Bergen, People cannot distinguish GPT-4 from a human in a Turing Test, [arXiv:2405.08007](https://arxiv.org/abs/2405.08007).
 - [7] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altschmidt, S. Altman S. Anadkat *et al.*, GPT-4 technical report, [arXiv:2303.08774](https://arxiv.org/abs/2303.08774).
 - [8] OpenAI, ChatGPT, <https://openai.com/research/gpt-4> [accessed April 2024].
 - [9] G. Kortemeyer, Could an artificial-intelligence agent pass an introductory physics course?, *Phys. Rev. Phys. Educ. Res.* **19**, 010132 (2023).
 - [10] G. Polverini and B. Gregorcic, How understanding large language models can inform the use of ChatGPT in physics education, *Eur. J. Phys.* **45**, 025701 (2024).
 - [11] G. Kortemeyer and W. Bauer, Cheat sites and artificial intelligence usage in online introductory physics courses: What is the extent and what effect does it have on assessments?, *Phys. Rev. Phys. Educ. Res.* **20**, 010145 (2024).
 - [12] W. Yeadon and T. Hardy, The impact of AI in physics education: A comprehensive review from GCSE to university levels, *Phys. Educ.* **59**, 025010 (2024).
 - [13] K. A. Pimblett and L. J. Morrell, Can ChatGPT pass a physics degree? Making a case for reformation of assessment of undergraduate degrees, *Eur. J. Phys.* **46**, 015702 (2024).
 - [14] A. Sperling and J. Lincoln, Artificial intelligence and high school physics, *Phys. Teach.* **62**, 314 (2024).
 - [15] S. Küchemann, M. Rau, A. Schmidt, and J. Kuhn, ChatGPT's quality: Reliability and validity of concept inventory items, *Front. Psychol.* **15**, 1426209 (2024).
 - [16] P. Bitzenbauer, ChatGPT in physics education: A pilot study on easy-to-implement activities, *Contemp. Educ. Technol.* **15**, ep430 (2023).
 - [17] S. Küchemann, S. Steinert, N. Revenga, M. Schweinberger, Y. Dinc, K. E. Avila, and J. Kuhn, Can ChatGPT support prospective teachers in physics task development?, *Phys. Rev. Phys. Educ. Res.* **19**, 020128 (2023).
 - [18] G. Kortemeyer, Using artificial-intelligence tools to make LaTeX content accessible to blind readers, *TUGboat* **44**, 390 (2023).
 - [19] T. Wan and Z. Chen, Exploring generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning, *Phys. Rev. Phys. Educ. Res.* **20**, 010152 (2024).
 - [20] Z. Chen and T. Wan, Grading explanations of problem-solving process and generating feedback using large language models at human-level accuracy, *Phys. Rev. Phys. Educ. Res.* **21**, 010126 (2025).
 - [21] R. K. Fussell, M. Flynn, A. Damle, M. F. Fox, and N. Holmes, Comparing large language models for supervised analysis of students' lab notes, *Phys. Rev. Phys. Educ. Res.* **21**, 010128 (2025).
 - [22] B. Gregorcic, G. Polverini, and A. Sarlah, ChatGPT as a tool for honing teachers' socratic dialogue skills, *Phys. Educ.* **59**, 045005 (2024).
 - [23] J. Crawford, M. Cowling, and K.-A. Allen, Leadership is needed for ethical ChatGPT: Character, assessment, and learning using artificial intelligence (AI), *J. Univ. Teach. Learn. Pract.* **20**, 045005 (2023).
 - [24] M. A. R. Vasconcelos and R. P. Dos Santos, Enhancing STEM learning with ChatGPT and Bing chat as objects to think with: A case study, *Eurasia J. Math. Sci. Technol. Educ.* **19**, em2296 (2023).
 - [25] M. N. Dahlkemper, S. Z. Lahme, and P. Klein, How do physics students evaluate artificial intelligence responses on comprehension questions? A study on the perceived scientific accuracy and linguistic quality of ChatGPT, *Phys. Rev. Phys. Educ. Res.* **19**, 010142 (2023).
 - [26] L. Ding, T. Li, S. Jiang, and A. Gapud, Students' perceptions of using ChatGPT in a physics class as a virtual tutor, *Int. J. Educ. Technol. Higher Educ.* **20**, 63 (2023).

- [27] C. G. West, AI and the FCI: Can ChatGPT project an understanding of introductory physics?, [arXiv:2303.01067](#).
- [28] S. Wheeler and R. E. Scherr, ChatGPT reflects student misconceptions in physics, in *Presented at PER Conference 2023, Sacramento, CA*, [10.1119/perc.2023.pr.Wheeler](#).
- [29] N. Cho, An investigation of using Spark generative AI in solving physics concept inventories in English and Chinese: Performance and issues, *Discover Artif. Intell.* **4**, 1 (2024).
- [30] S. Aldazharova, G. Issayeva, S. Maxutov, and N. Balta, Assessing AI's problem solving in physics: Analyzing reasoning, false positives and negatives through the force concept inventory, *Contemp. Educ. Technol.* **16**, ep538 (2024).
- [31] G. Polverini and B. Gregorcic, Evaluating vision-capable chatbots in interpreting kinematics graphs: A comparative study of free and subscription-based models, *Front. Educ.* **9**, 1452414 (2024).
- [32] OpenAI, Hello GPT-4o, <https://openai.com/index/hello-gpt-4o/> [accessed June 2024].
- [33] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, *Phys. Teach.* **30**, 141 (1992).
- [34] R. J. Beichner, Testing student interpretation of kinematics graphs, *Am. J. Phys.* **62**, 750 (1994).
- [35] L. Ding, R. Chabay, B. Sherwood, and R. Beichner, Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment, *Phys. Rev. ST Phys. Educ. Res.* **2**, 010105 (2006).
- [36] G. Polverini and B. Gregorcic, Performance of ChatGPT on the test of understanding graphs in kinematics, *Phys. Rev. Phys. Educ. Res.* **20**, 010109 (2024).
- [37] G. Polverini, J. Melin, E. Önerud, and B. Gregorcic, Performance of ChatGPT on tasks involving physics visual representations: The case of the brief electricity and magnetism assessment, *Phys. Rev. Phys. Educ. Res.* **21**, 010154 (2025).
- [38] J. I. Smith and K. Tanner, The problem of revealing how students think: Concept inventories and beyond, *CBE Life Sci. Educ.* **9**, 1 (2010).
- [39] D. Sands, M. Parker, H. Hedgeland, S. Jordan, and R. Galloway, Using concept inventories to measure understanding, *Higher Educ. Pedagog.* **3**, 173 (2018).
- [40] C. Henderson, Common concerns about the force concept inventory, *Phys. Teach.* **40**, 542 (2002).
- [41] OpenAI, Introducing GPT-o1, <https://openai.com/o1/> [accessed January 2025].
- [42] T. Geisler, Quality metrics for automated evaluation of exercises within student-LLM dialogues, M.Sc. thesis, ETH Zurich, 2025 (unpublished).
- [43] B. Gregorcic and A.-M. Pendrill, ChatGPT and the frustrated Socrates, *Phys. Educ.* **58**, 035021 (2023).
- [44] A. Madsen, S. B. McKagan, and E. C. Sayre, Best practices for administering concept inventories, *Phys. Teach.* **55**, 530 (2017).
- [45] R. R. Hake, Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, *Am. J. Phys.* **66**, 64 (1998).
- [46] J. T. Lavery and M. D. Caballero, Analysis of the most common concept inventories in physics: What are we assessing?, *Phys. Rev. Phys. Educ. Res.* **14**, 010123 (2018).
- [47] S. M. Stoen, M. A. McDaniel, R. F. Frey, K. M. Hynes, and M. J. Cahill, Force concept inventory: More than just conceptual understanding, *Phys. Rev. Phys. Educ. Res.* **16**, 010105 (2020).
- [48] D. T. Brookes and E. Etkina, Using conceptual metaphor and functional grammar to explore how language used in physics affects student learning, *Phys. Rev. ST Phys. Educ. Res.* **3**, 010105 (2007).
- [49] E. Euler, E. Rådahl, and B. Gregorcic, Embodiment in physics learning: A social-semiotic look, *Phys. Rev. Phys. Educ. Res.* **15**, 010134 (2019).
- [50] D. T. Brookes, The role of language in learning physics, Ph.D. thesis, Rutgers University, 2006.
- [51] P. Wulff, Physics language and language use in physics—what do we know and how AI might enhance language-related research and instruction, *Eur. J. Phys.* **45**, 023001 (2024).
- [52] OpenAI, GPT-4, <https://openai.com/index/gpt-4-research/> [accessed December 2024].
- [53] G. Nicholas and A. Bhatia, Lost in translation: Large language models in non-English content analysis, [arXiv:2306.07377](#).
- [54] DeepSeek, <https://www.deepseek.com/> [accessed December 2024].
- [55] Alibaba Cloud, Qwen, <https://qwen.ai.com/> [accessed December 2024].
- [56] Swiss AI initiative, Leveraging the world's most AI-capable supercomputer, <https://www.swiss-ai.org/> [accessed January 2025].
- [57] Cohere for AI, The AI language gap, <https://cohere.com/research/papers/the-ai-language-gap.pdf> [accessed December 2024].
- [58] S. Feng, W. Shi, Y. Wang, W. Ding, O. Ahia, S. S. Li, V. Balachandran, S. Sitaram, and Y. Tsvetkov, Teaching LLMs to abstain across languages via multilingual feedback, in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (2024), pp. 4125–4150, [10.18653/v1/2024.emnlp-main.239](#).
- [59] K. T. Kotsis, ChatGPT as teacher assistant for physics teaching, *EIKI J. Eff. Teach. Methods* **2** (2024).
- [60] P. Tschisgale, P. Wulff, and M. Kubsch, Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory, *Phys. Rev. Phys. Educ. Res.* **19**, 020123 (2023).
- [61] P. Tschisgale, P. Wulff, and M. Kubsch, Erratum: Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory [Phys. Rev. Phys. Educ. Res. **19**, 020123 (2023)], *Phys. Rev. Phys. Educ. Res.* **21**, 019901 (2025).
- [62] T. O. B. Odden, H. Tyseng, J. T. Mjaaland, M. F. Kreutzer, and A. Malthe-Sørenssen, Using text embeddings for deductive qualitative research at scale in physics education, *Phys. Rev. Phys. Educ. Res.* **20**, 020151 (2024).

- [63] A. Bakhtin, L. van der Maaten, J. Johnson, L. Gustafson, and R. Girshick, PHYRE: A new benchmark for physical reasoning, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Vancouver, 2019), Vol. 32.
- [64] C. Xue, V. Pinto, C. Gamage, E. Nikonova, P. Zhang, and J. Renz, Phy-Q: A benchmark for physical reasoning, *arXiv:2108.13696*.
- [65] A. Melnik, R. Schiewer, M. Lange, A. Muresanu, M. Saeidi, A. Garg, and H. Ritter, Benchmarks for physical reasoning AI, *arXiv:2312.10728*.
- [66] PhysBench: Physical reasoning benchmark, <https://physbench.com/> [accessed: May 7, 2025].
- [67] G. Kortemeyer and J. Nöhl, Assessing confidence in AI-assisted grading of physics exams through psychometrics: An exploratory study, *Phys. Rev. Phys. Educ. Res.* **21**, 010136 (2025).
- [68] R. Mok, F. Akhtar, L. Clare, C. Li, J. Ida, L. Ross, and M. Campanelli, Using large language models for grading in education: An applied test for physics, *Phys. Educ.* **60**, 035006 (2025).
- [69] S. Guo, E. Latif, Y. Zhou, X. Huang, and X. Zhai, Using generative AI and multi-agents to provide automatic feedback, *arXiv:2411.07407*.
- [70] L. Krupp, J. Bley, I. Gobbi, A. Geng, S. Müller, S. Suh, A. Moghiseh, A. C. Medina, V. Bartsch, and A. Widera *et al.*, LLM-generated tips rival expert-created tips in helping students answer quantum-computing questions, *Eur. Phys. J. Quantum Technol.* **12**, 33 (2025).
- [71] J. R. Aguilar-Mejía, S. Tejeda, C. V. Ramirez-Lopez, and C. L. Garay-Rondero, Design and use of a chatbot for learning selected topics of physics, in *Technology-Enabled Innovations in Education*, Transactions on Computer Systems and Networks (Springer, Singapore, 2022), p. 175, 10.1007/978-981-19-3383-7.
- [72] V. R. Lee, D. Pope, S. Miles, and R. C. Zarate, Cheating in the age of generative AI: High school survey study of cheating behaviors before and after the release of ChatGPT, *Comput. Educ.* **7**, 100253 (2024).
- [73] J. L. Docktor and J. P. Mestre, Synthesis of discipline-based education research in physics, *Phys. Rev. ST Phys. Educ. Res.* **10**, 020119 (2014).
- [74] D. E. Meltzer and V. K. Otero, A brief history of physics education in the United States, *Am. J. Phys.* **83**, 447 (2015).
- [75] S. Bulathwela, M. Pérez-Ortiz, C. Holloway, M. Cukurova, and J. Shawe-Taylor, Artificial intelligence alone will not democratise education: On educational inequality, techno-solutionism and inclusive tools, *Sustainability* **16**, 781 (2024).
- [76] American Association of Physics Teachers. PhysPort, <https://www.physport.org> (2017) [retrieved November 2024].
- [77] S. B. McKagan, L. E. Strubbe, L. J. Barbato, B. A. Mason, A. M. Madsen, and E. C. Sayre, PhysPort use and growth: Supporting physics teaching with research-based resources since 2011, *Phys. Teach.* **58**, 465 (2020).
- [78] J. Von Korff, B. Archibeque, K. A. Gomez, T. Heckendorf, S. B. McKagan, E. C. Sayre, E. W. Schenk, C. Shepherd, and L. Sorell, Secondary analysis of teaching methods in introductory physics: A 50 k-student study, *Am. J. Phys.* **84**, 969 (2016).
- [79] OpenAI, How ChatGPT and our foundation models are developed, <https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed> [accessed March 2025].
- [80] B. Hufnagel, Development of the astronomy diagnostic test, *Astron. Educ. Rev.* **1**, 47 (2002).
- [81] E. Brogt, D. Sabers, E. E. Prather, G. L. Deming, B. Hufnagel, and T. F. Slater, Analysis of the astronomy diagnostic test, *Astron. Educ. Rev.* **6**, 25 (2007).
- [82] N. O. Koca, N. alhuda Al Saqri, H. Al Hamrashdi, and N. Al Kindi, Evaluating the students' learning on the electricity and magnetism using a conceptual survey BEMA, *Phys. Educ.* **60**, 015022 (2024).
- [83] S. J. Pollock, Comparing student learning with multiple research-based conceptual surveys: CSEM and BEMA, *AIP Conf. Proc.* **1064**, 171 (2008).
- [84] J. Epstein, The calculus concept inventory, in *National STEM Assessment Conference*, Washington, DC (Academia, San Francisco, CA, 2006), p. 60.
- [85] W. Maciejewski, Flipping the calculus classroom: An evaluative study, *Teach. Math. Appl.* **35**, 187 (2016).
- [86] J. Day and D. Bonn, Development of the concise data processing assessment, *Phys. Rev. ST Phys. Educ. Res.* **7**, 010114 (2011).
- [87] D. P. Maloney, T. L. O'Kuma, C. J. Hieggelke, and A. Van Heuvelen, Surveying students' conceptual knowledge of electricity and magnetism, *Am. J. Phys.* **69**, S12 (2001).
- [88] R. Tapping, G. Lepage, and N. Holmes, Visualizing patterns in CSEM responses to assess student conceptual understanding, in *Presented at PER Conference 2019*, Washington, DC, 10.1119/perc.2018.pr.Tapping.
- [89] A. E. Lawson, The development and validation of a classroom test of formal reasoning, *J. Res. Sci. Teach.* **15**, 11 (1978).
- [90] J. C. Moore and L. J. Rubbo, Scientific reasoning abilities of nonscience majors in physics-based courses, *Phys. Rev. ST Phys. Educ. Res.* **8**, 010106 (2012).
- [91] P. V. Engelhardt and R. J. Beichner, Students' understanding of direct current resistive electrical circuits, *Am. J. Phys.* **72**, 98 (2004).
- [92] D. Sangam and B. K. Jesiek, Conceptual understanding of resistive electric circuits among first-year engineering students, in *Proceedings of the 2012 ASEE Annual Conference & Exposition* (American Society for Engineering Education, Washington, DC, 2012), pp. 25–339.
- [93] R. Yeend, M. Loverude, and B. Gonzales, Student understanding of density: A cross-age investigation, in *Proceedings of the 2001 Physics Education Conference*, Rochester, NY (AIP, New York, 2001).
- [94] T. Zenger and P. Bitzenbauer, Exploring German secondary school students' conceptual knowledge of density, *Sci. Educ. Int.* **33**, 86 (2022).
- [95] L. Ding, R. Chabay, and B. Sherwood, How do students in an innovative principle-based mechanics course understand energy concepts?, *J. Res. Sci. Teach.* **50**, 722 (2013).
- [96] D. R. Sokoloff, Teaching electric circuit concepts using microcomputer-based current/voltage probes, in

- Microcomputer-Based Labs: Educational Research and Standards* (Springer, New York, 1996), pp. 129–146.
- [97] G. Kortemeyer, D. Anderson, A. M. Desrochers, A. Hackbardt, K. Hoekstra, A. Holt, A. Iftekhar, T. Kabaker, N. Keller, and Z. Korzecke *et al.*, Using a computer game to teach circuit concepts, *Eur. J. Phys.* **40**, 055703 (2019).
- [98] M. W. McColgan, R. A. Finn, D. L. Broder, and G. E. Hassel, Assessing students' conceptual knowledge of electricity and magnetism, *Phys. Rev. Phys. Educ. Res.* **13**, 020121 (2017).
- [99] C. Singh and D. Rosengrant, Multiple-choice test of energy and momentum concepts, *Am. J. Phys.* **71**, 607 (2003).
- [100] M. Sahin, The impact of problem-based learning on engineering students' beliefs about physics and conceptual understanding of energy and momentum, *Eur. J. Eng. Educ.* **35**, 519 (2010).
- [101] A. J. Mason, Learning goals and perceived irrelevance to major within life science majors in introductory physics, [10.1119/perc.2019.pr.Mason](https://doi.org/10.1119/perc.2019.pr.Mason) (2020).
- [102] G. Kortemeyer, Gender differences in the use of an online homework system in an introductory physics course, *Phys. Rev. ST Phys. Educ. Res.* **5**, 010107 (2009).
- [103] J. Han, L. Bao, L. Chen, T. Cai, Y. Pi, S. Zhou, Y. Tu, and K. Koenig, Dividing the force concept inventory into two equivalent half-length tests, *Phys. Rev. ST Phys. Educ. Res.* **11**, 010112 (2015).
- [104] R. K. Thornton and D. R. Sokoloff, Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula, *Am. J. Phys.* **66**, 338 (1998).
- [105] K. Cummings, J. Marx, R. Thornton, and D. Kuhl, Evaluating innovation in studio physics, *Am. J. Phys.* **67**, S38 (1999).
- [106] S. T. Kalinowski and S. Willoughby, Development and validation of a scientific (formal) reasoning test for college students, *J. Res. Sci. Teach.* **56**, 1269 (2019).
- [107] D. Kaltakci-Gurel, A. Eryilmaz, and L. C. McDermott, Development and application of a four-tier test to assess pre-service physics teachers' misconceptions about geometrical optics, *Res. Sci. Technol. Educ.* **35**, 238 (2017).
- [108] R. Rosenblatt and A. F. Heckler, Systematic study of student understanding of the relationships between the directions of force, velocity, and acceleration in one dimension, *Phys. Rev. ST Phys. Educ. Res.* **7**, 020112 (2011).
- [109] J. M. Keller, Part I. Development of a concept inventory addressing students' beliefs and reasoning difficulties regarding the greenhouse effect, part II. Distribution of chlorine measured by the Mars Odyssey Gamma Ray Spectrometer, Doctoral thesis, The University of Arizona, 2006.
- [110] C. Tanahoung, M. D. Sharma, I. D. Johnston, R. Chitaree, and C. Soankwan, Surveying Sydney introductory physics students' understandings of heat and temperature, in *Proceedings of the Australian Institute of Physics 17th National Congress, Brisbane* (Australian Institute of Physics, Clayton, 2006), Paper No. WC0233.
- [111] I. Halloun, Evaluation of the impact of the new physics curriculum on the conceptual profiles of secondary students (2007), pp. 1–25, <https://www.halloun.net/wp-content/uploads/2016/10/LU-Summative-Report-10-07.pdf>.
- [112] K. Ndiokubwayo, J. Uwamahoro, I. Ndayambaje, and M. Ralph, Light phenomena conceptual assessment: An inventory tool for teachers, *Phys. Educ.* **55**, 035009 (2020).
- [113] R. S. Lindell and J. P. Olsen, Developing the lunar phases concept inventory, in *Proceedings of the 2002 Physics Education Research Conference* (PERC Publishing, New York, 2002).
- [114] E. M. Bardar, E. E. Prather, K. Brecher, and T. F. Slater, Development and validation of the light and spectroscopy concept inventory, *Astron. Educ. Rev.* **5**, 103 (2007).
- [115] C. S. Wallace, T. G. Chambers, and E. E. Prather, Item response theory evaluation of the light and spectroscopy concept inventory national data set, *Phys. Rev. Phys. Educ. Res.* **14**, 010149 (2018).
- [116] D. Hestenes and M. Wells, A mechanics baseline test, *Phys. Teach.* **30**, 159 (1992).
- [117] C. P. Millán and S. Otranto, Thirty-six years of the forced concept inventory and the mechanics baseline test: Is Aristotle still playing hide and seek in our classrooms?, *Latin-Am. J. Phys. Educ.* **15**, 2309 (2021), http://www.lajpe.org/jun21/15_2_09.pdf.
- [118] V. Antwi, R. Hanson, A. Sam, E. Savelsbergh, and H. Eijkelhof, The impact of interactive-engagement (IE) teaching on students understanding of concepts in mechanics: The use of force concept inventory (FCI) and mechanics baseline test (MBT), *Int. J. Educ. Plann. Adm.* **1**, 81 (2011), https://www.ripublication.com/ijepav/ijepav1n1_9.pdf.
- [119] C. Kádár and P. Tasnádi, The knowledge of Hungarian students in the light of the mechanics baseline test, *J. Phys. Conf. Ser.* **1286**, 012026 (2019).
- [120] J. Li and C. Singh, Developing and validating a conceptual survey to assess introductory physics students' understanding of magnetism, *Eur. J. Phys.* **38**, 025702 (2016).
- [121] D. L. Deardorff, Introductory physics students' treatment of measurement uncertainty, Doctoral thesis, North Carolina State University, 2001.
- [122] A. Tongchai, M. D. Sharma, I. D. Johnston, K. Arayathanitkul, and C. Soankwan, Developing, evaluating and demonstrating the use of a conceptual survey in mechanical waves, *Int. J. Sci. Educ.* **31**, 2437 (2009).
- [123] P. H. Santoso, E. Istiyono, and H. Haryanto, Principal component analysis and exploratory factor analysis of the mechanical waves conceptual survey, *JP3I* **11**, 209 (2022).
- [124] K. E. Williamson, S. D. Willoughby, and E. E. Prather, Development of the Newtonian gravity concept inventory, *Astron. Educ. Rev.* **12**, 1 (2013).
- [125] P. V. Engelhardt, S. Robinson, E. P. Price, P. S. Smith, and F. Goldberg, Developing a conceptual assessment for a modular curriculum, in *Presented at PER Conference*

- 2018, Washington, DC, [10.1119/perc.2018.pr.Engelhardt](https://arxiv.org/abs/10.1119/perc.2018.pr.Engelhardt).
- [126] S. White Brahmia, A. Olsho, T. I. Smith, A. Boudreaux, P. Eaton, and C. Zimmerman, Physics inventory of quantitative literacy: A tool for assessing mathematical reasoning in introductory physics, *Phys. Rev. Phys. Educ. Res.* **17**, 020129 (2021).
- [127] H. R. Sadaghiani and S. J. Pollock, Quantum mechanics concept assessment: Development and validation study, *Phys. Rev. ST Phys. Educ. Res.* **11**, 010110 (2015).
- [128] S. McKagan, K. Perkins, and C. Wieman, Design and validation of the quantum mechanics conceptual survey, *Phys. Rev. ST Phys. Educ. Res.* **6**, 020121 (2010).
- [129] E. Marshman and C. Singh, Validation and administration of a conceptual survey on the formalism and postulates of quantum mechanics, *Phys. Rev. Phys. Educ. Res.* **15**, 020128 (2019).
- [130] E. M. Marshman. Improving the quantum mechanics content knowledge and pedagogical content knowledge of physics graduate students, Ph.D. thesis, University of Pittsburgh, 2015.
- [131] G. Zhu and C. Singh, Surveying students' understanding of quantum mechanics in one spatial dimension, *Am. J. Phys.* **80**, 252 (2012).
- [132] E. Cataloglu and R. Robinett, Testing the development of student conceptual and visualization understanding in quantum mechanics through the undergraduate career, *Am. J. Phys.* **70**, 238 (2002).
- [133] S. Wuttirom, M. D. Sharma, I. D. Johnston, R. Chitaree, and C. Soankwan, Development and use of a conceptual survey in introductory quantum physics, *Int. J. Sci. Educ.* **31**, 631 (2009).
- [134] R. J. Allain, Investigating the relationship between student difficulties with the concept of electric potential and the concept of rate of change, Doctoral thesis, North Carolina State University, 2001.
- [135] J. Aslanides and C. M. Savage, Relativity concept inventory: Development, analysis, and results, *Phys. Rev. ST Phys. Educ. Res.* **9**, 010118 (2013).
- [136] P. Nieminen, A. Savinainen, and J. Viiri, Force concept inventory-based multiple-choice test for investigating students' representational consistency, *Phys. Rev. ST Phys. Educ. Res.* **6**, 020109 (2010).
- [137] K. Mashood and V. A. Singh, An inventory on rotational kinematics of a particle: Unravelling misconceptions and pitfalls in reasoning, *Eur. J. Phys.* **33**, 1301 (2012).
- [138] M. Suárez, S. Pandiella, and J. Benegas, Tutorials+ PhET: A simple and efficient active-learning approach for the teaching of kinematics of circular motion in a technically-oriented high school, *Phys. Educ.* **58**, 035005 (2023).
- [139] L. G. Rimoldini and C. Singh, Student understanding of rotational and rolling motion concepts, *Phys. Rev. ST Phys. Educ. Res.* **1**, 010102 (2005).
- [140] C. Singh, Student understanding of symmetry and Gauss's law of electricity, *Am. J. Phys.* **74**, 923 (2006).
- [141] J. M. Bailey, B. Johnson, E. E. Prather, and T. F. Slater, Development and validation of the star properties concept inventory, *Int. J. Sci. Educ.* **34**, 2257 (2012).
- [142] B. Brown and C. Singh, Development and validation of a conceptual survey instrument to evaluate students' understanding of thermodynamics, *Phys. Rev. Phys. Educ. Res.* **17**, 010104 (2021).
- [143] S. Yeo and M. Zadnik, Introductory thermal concept evaluation: assessing students' understanding, *Phys. Teach.* **39**, 496 (2001).
- [144] P. Wattanakasiwich, P. Taleab, M. D. Sharma, and I. D. Johnston, Construction and implementation of a conceptual survey in thermodynamics, *Int. J. Innov. Sci. Math. Educ.* **21**, 29 (2013), <https://openjournals.library.sydney.edu.au/CAL/article/view/6459>.
- [145] S. J. Slater, The development and validation of the test of astronomy standards (TOAST), *J. Astron. Earth Sci. Educ.* **1**, 1 (2014).
- [146] P. Klein, A. Lichtenberger, S. Küchemann, S. Becker, M. Kekule, J. Viiri, C. Baadte, A. Vaterlaus, and J. Kuhn, Visual attention while solving the test of understanding graphs in kinematics: An eye-tracking analysis, *Eur. J. Phys.* **41**, 025701 (2020).
- [147] P. Barniol and G. Zavala, Test of understanding of vectors: A reliable multiple-choice vector concept test, *Phys. Rev. ST Phys. Educ. Res.* **10**, 010121 (2014).
- [148] Microsoft, Azure AI services, <https://azure.microsoft.com/en-us/products/ai-services> [accessed June 2024].
- [149] M. Danilák, Langdetect, <https://pypi.org/project/langdetect/> [accessed December 2024].
- [150] N. Sidiropoulos, S. H. Sohi, T. L. Pedersen, B. T. Porse, O. Winther, N. Rapin, and F. O. Bagger, SinaPlot: An enhanced chart for simple and truthful representation of single observations over multiple classes, *J. Comput. Graph. Stat.* **27**, 673 (2018).
- [151] F. Kieser, P. Wulff, J. Kuhn, and S. Küchemann, Educational data augmentation in physics education research using ChatGPT, *Phys. Rev. Phys. Educ. Res.* **19**, 020150 (2023).
- [152] OpenAI, OpenAI o3 and o4-mini, <https://openai.com/index/introducing-o3-and-o4-mini/> [accessed May 2025].
- [153] Y. Liang, D. Zou, H. Xie, and F. L. Wang, Exploring the potential of using ChatGPT in physics education, *Smart Learn. Environ.* **10**, 52 (2023).
- [154] E. Latif, R. Parasuraman, and X. Zhai, Physicsassistant: An LLM-powered interactive learning robot for physics lab investigations. in *Proceedings of the IEEE International Conference on Robot and Human Interactive Communication (ROMAN)* (IEEE, New York, 2024), pp. 864–871.
- [155] A. Lieb and T. Goel, Student interaction with NewtBot: An LLM-as-tutor chatbot for secondary physics education, in *Proceedings of the CHI EA '24: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, Honolulu HI* (Association for Computing Machinery, New York, NY, 2024), pp. 1–8, [10.1145/3613905.3647957](https://doi.org/10.1145/3613905.3647957).
- [156] R. Kahaleh and V. Lopez, Evaluating large language models in high school physics education: Addressing misconceptions and fostering conceptual understanding, *Phys. Educ.* **60**, 025013 (2025).
- [157] K. E. Avila, S. Steinert, S. Ruzika, J. Kuhn, and S. Küchemann, Using ChatGPT for teaching physics, *Phys. Teach.* **62**, 536 (2024).

-
- [158] Y. Zhu, Z.-Y. Khoo, J.S.C. Low, and S. Bressan, A personalised learning tool for physics undergraduate students built on a large language model for symbolic regression, in *Proceedings of the 2024 IEEE Conference on Artificial Intelligence* (IEEE, New York, 2024), pp. 38–43.
- [159] G. Kortemeyer, M. Babayeva, G. Polverini, R. Widenhorn, and B. Gregorcic, Data for the paper “multilingual performance of a multimodal artificial intelligence system on multisubject physics concept inventories”, <https://per-central.org/i/17053> (2025).