CCSE
Center for Computing in Science Education

Overview of research on the use of AI in science teaching and learning

Tor Ole Odden
Center for Computing in Science Education
AIMSC Network Meeting
January 23, 2026

UiO : Det matematisk-naturvitenskapelige fakultet

Centre for Excellence in Education

Overview of the research on GenAI in Science Education

Journals:

- **Science Education**:
  JRST, Science Education, IJSE, SSE, S&E, JSET
- **Physics Education**:
  PRPER, TPT, PE, EJP

Main topics:

1. Assessment
2. Classroom use
3. Teaching use
4. AI Knowledge
5. Research Applications

AI Use Declaration

Tor:

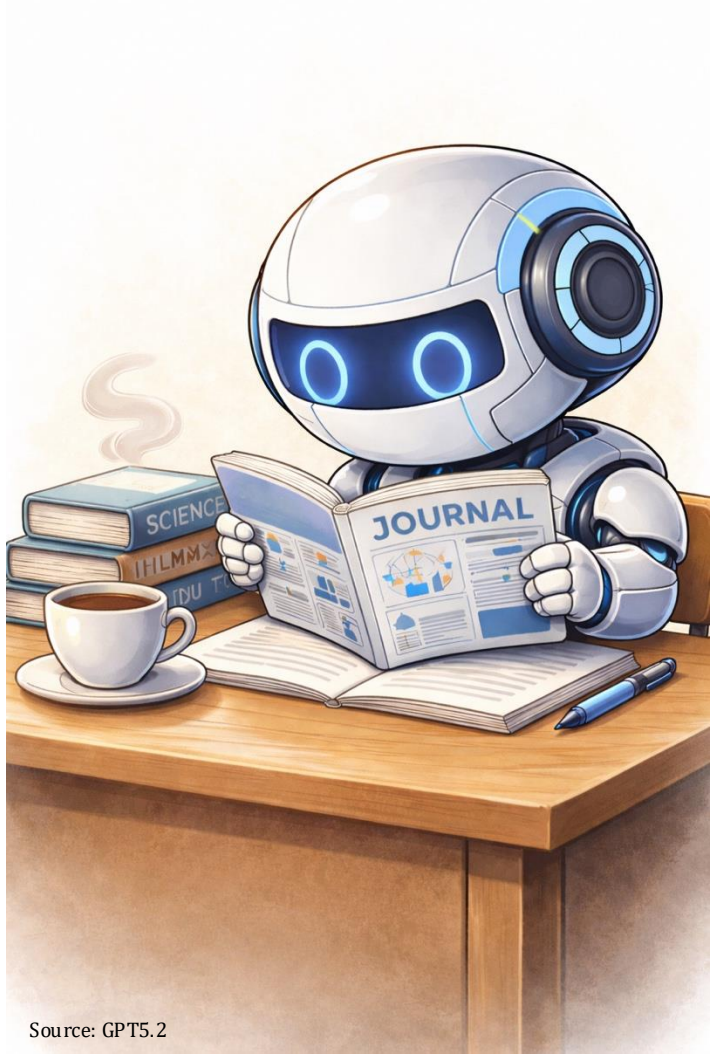
- Gathered the articles
- Read some of the articles
- Made the categories
- Rough-sorted articles into categories

ChatGPT:

- Reviewed articles in each category, provided overview (with references)

Tor:

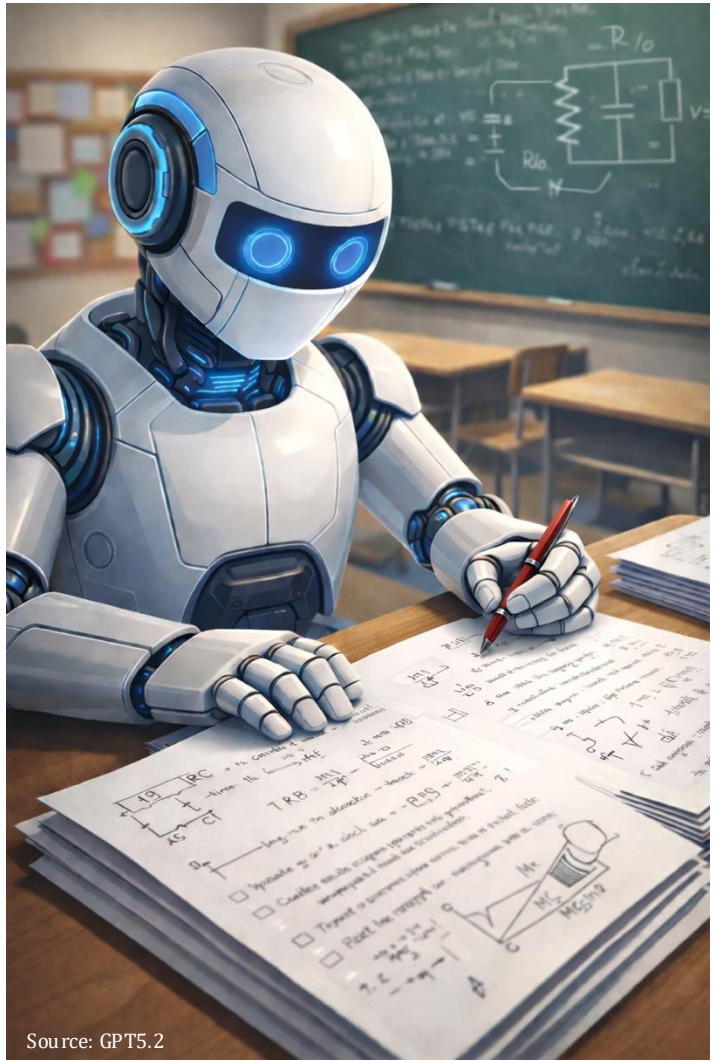
- Double-checked overview and highlights



Source: GPT5.2

1. GenAI for Assessment

- GenAI can grade short responses **near human level**
  - Works best with **human-designed prompts & rubrics**
- Most useful for making **low-stakes, formative assessments**
  - Making high-stakes tests requires psychometric validation
- **Assessment design is shifting** now that AI can pass exams



Source: GPT5.2

PHYSICAL REVIEW PHYSICS EDUCATION RESEARCH **21,** 010126 (2025)

Featured in Physics

Grading explanations of problem-solving process and generating feedback using large language models at human-level accuracy

Zhongzhou Chen* and Tong Wan

Department of Physics, University of Central Florida, Orlando, Florida 32816, USA

(Received 17 December 2024; accepted 21 February 2025; published 27 March 2025)

[This paper is part of the Focused Collection in Artificial Intelligence Tools in Physics Teaching and Physics Education Research.] This study examines the feasibility and potential advantages of using large language models, in particular GPT-4o, to perform partial credit grading of large numbers of student written responses to introductory level physics problems. Students were instructed to write down verbal explanations of their reasoning process when solving one conceptual and two numerical calculation problems on two exams. The explanations were then graded according to a three-item rubric with each item graded as binary (1 or 0). We first demonstrate that machine grading using GPT-4o with no examples or reference answers can reliably agree with human graders in 70%–80% of all cases, which is equal to or higher than the level at which two human graders agree with each other. Two methods are essential for achieving this level of accuracy: (i) Adding explanation language to each rubric item that targets the errors of initial machine grading. (ii) Running the grading process 5 times and taking the most frequent outcome. Next, we show that the variation in outcomes across five machine grading attempts can serve as a grading confidence index. The index allows a human expert to identify ~40% of all potentially incorrect gradings by reviewing just 10%–15% of all responses with the highest variation. Finally, we show that it is straightforward to use GPT-4o to write a clear and detailed explanation of the partial credit grading outcome. Those explanations can be used as feedback for students, which will allow students to understand their grades and raise different opinions when necessary. Almost all feedback messages generated were rated three or above on a five-point scale by two instructors who had taught the course multiple times. The entire grading and feedback generating process costs roughly $5 per 100 student answers, which shows immense promise for automating labor-intensive grading process through a combination of machine grading with human input and supervision.

DOI: 10.1103/PhysRevPhysEducRes.21.010126

I. INTRODUCTION

The rapid recent advancement in both capability and availability of Generative Artificial Intelligence (GenAI), in particular, large language models (LLMs) such as GPT-4, has spurred a growing body of research exploring their applications across many different areas of education [1,2]. One of those areas that has gained a substantial amount of recent interest is Automated Short Answer Grading (ASAG) [3,4] and Automated Long Answer Grading (ALAG) [5]. Both ASAG and ALAG utilize LLMs ability to process and generate natural language to grade large numbers of student written responses, especially for problems in STEM domains including physics. Using GenAI to assist in the assigning of students' written responses has the clear benefit of relieving instructors from time consuming and largely repetitive work of grading [6,7], and could potentially provide students with more detailed and timely feedback on their understanding, which is highly beneficial for learning [8].

Most ASAG or ALAG studies published prior to 2023 used smaller language models that need to first be trained on a corpus of domain-specific data (for example, see Refs. [3,9]) to achieve satisfactory performance. Since 2023, many studies have shifted toward using general purpose language models, also called *foundational models* [10], with the most popular foundational models being GPT-3.5 and GPT-4 [11]. The most significant advantage of using foundational LLMs is that they require no pretraining on existing data, which greatly simplifies the grading process. In most cases, grading one student's response simply involves sending a piece of natural language text, called a *prompt*, to the LLM. A prompt usually consists of a problem body, a grading rubric and grading requirements, and a student response. Sometimes, one or more example response-grading pairs could also be included in the prompt.

It is important to clarify that the term GenAI refers to a family of artificial intelligence models that generate content ranging from text, computer code, images, and videos [12]. LLMs are a type of GenAI specialized in processing and

*Contact author: Zhongzhou.Chen@ucf.edu

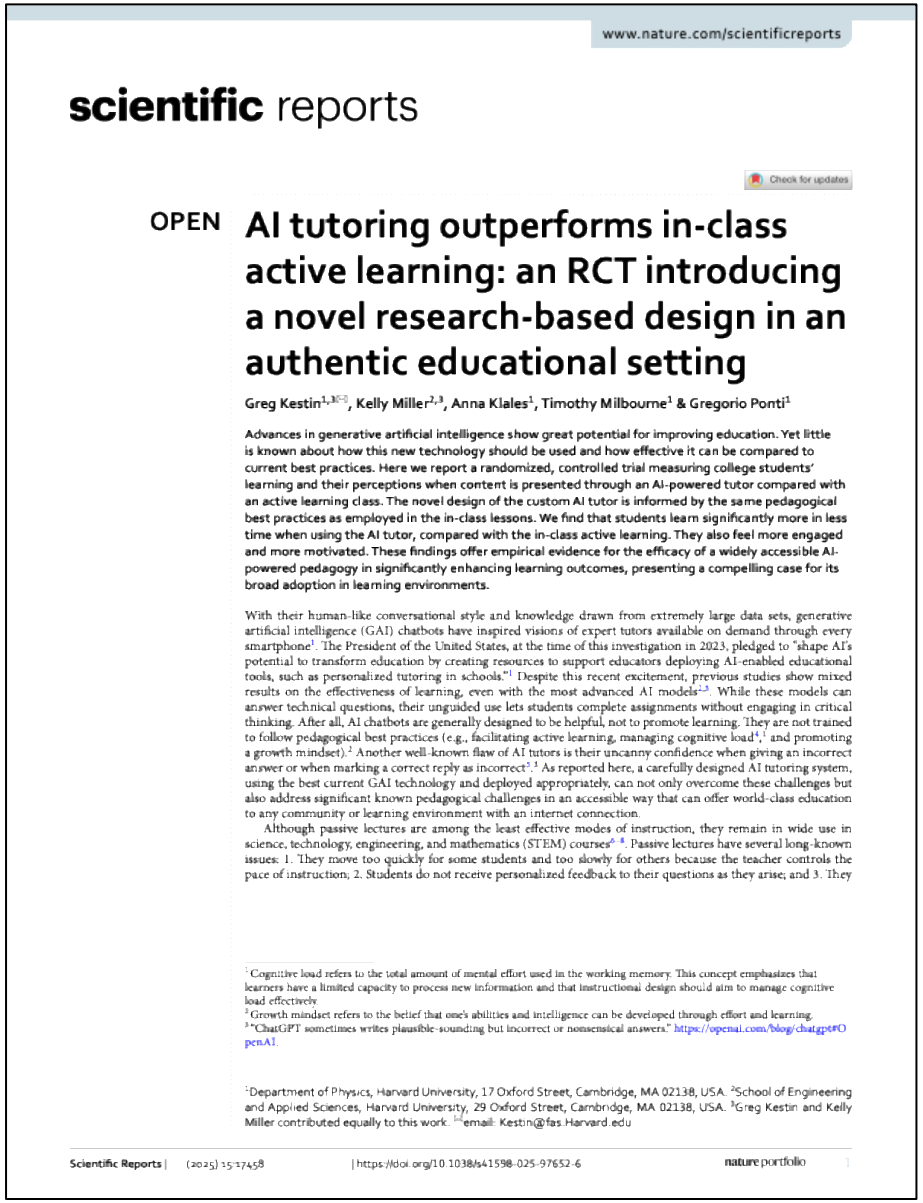
Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

- Used GPT-4 to grade university physics exam responses
- achieved human-level agreement in scoring (around 70–80%, on par with instructors)
- AI also generated explanations for each grade as feedback to students
- Whole AI grading process cost only about $0.05 per response

2. GenAI for Classroom Use

- GenAI works best as **structured scaffolding**, not open-ended use
  - Most effective for **hints, feedback, dialogue, and tutoring**
- Learning gains are real but strongly **design-dependent**
  - **Metacognitive support** often outperforms content-only help
- Equity and affective impacts vary with **prior knowledge** and **AI literacy**

Source: GPT5.2



www.nature.com/scientificreports

scientific reports

OPEN

AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting

Greg Kestin[1,*], Kelly Miller[2,3], Anna Klales[1], Timothy Milbourne[1] & Gregorio Ponti[1]

Advances in generative artificial intelligence show great potential for improving education. Yet little is known about how this new technology should be used and how effective it can be compared to current best practices. Here we report a randomized, controlled trial measuring college students' learning and their perceptions when content is presented through an AI-powered tutor compared with an active learning class. The novel design of the custom AI tutor is informed by the same pedagogical best practices as employed in the in-class lessons. We find that students learn significantly more in less time when using the AI tutor, compared with the in-class active learning. They also feel more engaged and more motivated. These findings offer empirical evidence for the efficacy of a widely accessible AI-powered pedagogy in significantly enhancing learning outcomes, presenting a compelling case for its broad adoption in learning environments.

With their human-like conversational style and knowledge drawn from extremely large data sets, generative artificial intelligence (GAI) chatbots have inspired visions of expert tutors available on demand through every smartphone[1]. The President of the United States, at the time of this investigation in 2023, pledged to "shape AI's potential to transform education by creating resources to support educators deploying AI-enabled educational tools, such as personalized tutoring in schools."[1] Despite this recent excitement, previous studies show mixed results on the effectiveness of learning, even with the most advanced AI models[2,3]. While these models can answer technical questions, their unguided use lets students complete assignments without engaging in critical thinking. After all, AI chatbots are generally designed to be helpful, not to promote learning. They are not trained to follow pedagogical best practices (e.g., facilitating active learning, managing cognitive load[4],[†] and promoting a growth mindset).[‡] Another well-known flaw of AI tutors is their uncanny confidence when giving an incorrect answer or when marking a correct reply as incorrect[2,3]. As reported here, a carefully designed AI tutoring system, using the best current GAI technology and deployed appropriately, can not only overcome these challenges but also address significant known pedagogical challenges in an accessible way that can offer world-class education to any community or learning environment with an internet connection.

Although passive lectures are among the least effective modes of instruction, they remain in wide use in science, technology, engineering, and mathematics (STEM) courses[6–8]. Passive lectures have several long-known issues: 1. They move too quickly for some students and too slowly for others because the teacher controls the pace of instruction; 2. Students do not receive personalized feedback to their questions as they arise; and 3. They

[†] Cognitive load refers to the total amount of mental effort used in the working memory. This concept emphasizes that learners have a limited capacity to process new information and that instructional design should aim to manage cognitive load effectively.

[‡] Growth mindset refers to the belief that one's abilities and intelligence can be developed through effort and learning.

[3] "ChatGPT sometimes writes plausible-sounding but incorrect or nonsensical answers." https://openai.com/blog/chatgpt#OpenAI.

[1]Department of Physics, Harvard University, 17 Oxford Street, Cambridge, MA 02138, USA. [2]School of Engineering and Applied Sciences, Harvard University, 29 Oxford Street, Cambridge, MA 02138, USA. [3]Greg Kestin and Kelly Miller contributed equally to this work. [*]email: Kestin@fas.Harvard.edu

- Randomized controlled trial in undergraduate physics
- AI tutor provided adaptive, step-by-step problem-solving support
- Students using AI tutoring **outperformed peers in active-learning sections**
- Largest gains for procedural and quantitative problem-solving tasks

3. GenAI for Teaching, Lesson Planning, and Teacher Training

- GenAI works best as a **co-planner** and **pedagogical partner**, not an autonomous teacher
  - Lesson quality is **highly variable** and mediated by teachers' PCK, AI literacy, and prompting skill
- **Inquiry-based teaching is hard for GenAI** without human framing and differentiation
- **Teacher education is shifting** toward AI literacy, critical evaluation, and agency, not tool mastery alone
  - Adoption depends more on **perceived usefulness, norms, and institutional support** than technical skill

Source: GPT5.2

INTERNATIONAL JOURNAL OF SCIENCE EDUCATION
https://doi.org/10.1080/09500693.2025.2567509

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Lesson planning with ChatGPT for inquiry-based biology instruction – A(I) roll of the dice?

Leroy Großmann [a*], Maren Koberstein-Schwarz [b], Dirk Krüger [a] and Moritz Krell [b]

[a]Biology Education, Freie Universität Berlin, Berlin, Germany; [b]Biology Education, IPN – Leibniz Institute for Science and Mathematics Education, Kiel, Germany

ABSTRACT

This study investigates the capability of ChatGPT-4o to generate high-quality inquiry-based science lesson plans, that is, aligning all elements of a written lesson plan to students' learning about procedural and epistemic aspects of science instead of gaining subject matter knowledge. Using an exploratory sequential mixed methods design, we analysed $N = 60$ biology lesson plans generated by the research team across four key topics (cell biology, genetics/evolution, human biology, ecology) and five scientific inquiry practices (microscopy, observing, experimenting, modelling, reflecting the nature of science) from the German curriculum . First, lesson plans were quantitatively evaluated using a modified version of Großmann and Krügers' [(2024). Assessing the quality of science teachers' lesson plans: Evaluation and application of a novel instrument. *Science Education, 108*(1), 153–189. https://doi.org/10.1002/sce.v108.1] scoring rubric, which assessed ten quality criteria with substantial interrater agreement (Cohen's $\kappa = .67$). Second, based on the score distribution, we conducted qualitative analyses to identify strengths and weaknesses in ChatGPT-generated inquiry-based lesson plans. Results revealed considerable variation in lesson plan quality, with $n = 22$ lesson plans achieving less than 50% of the maximum possible score and only $n = 5$ lesson plans reaching 75% or higher. While the lesson plans demonstrated particular strengths in content accuracy and learning outcome alignment, they exhibited significant weaknesses in addressing students' inquiry-related conceptions and maintaining consistent focus on scientific inquiry. While ChatGPT-4o successfully generated some high-quality lessons plans for inquiry-based instruction, many lesson plans shifted inappropriately towards subject matter knowledge acquisition rather than scientific inquiry processes. This discrepancy between our initial prompts and the resulting lesson plans indicates that while large language models like ChatGPT-4o demonstrate potential as preliminary planning tools, they

ARTICLE HISTORY
Received 13 November 2024
Accepted 17 September 2025

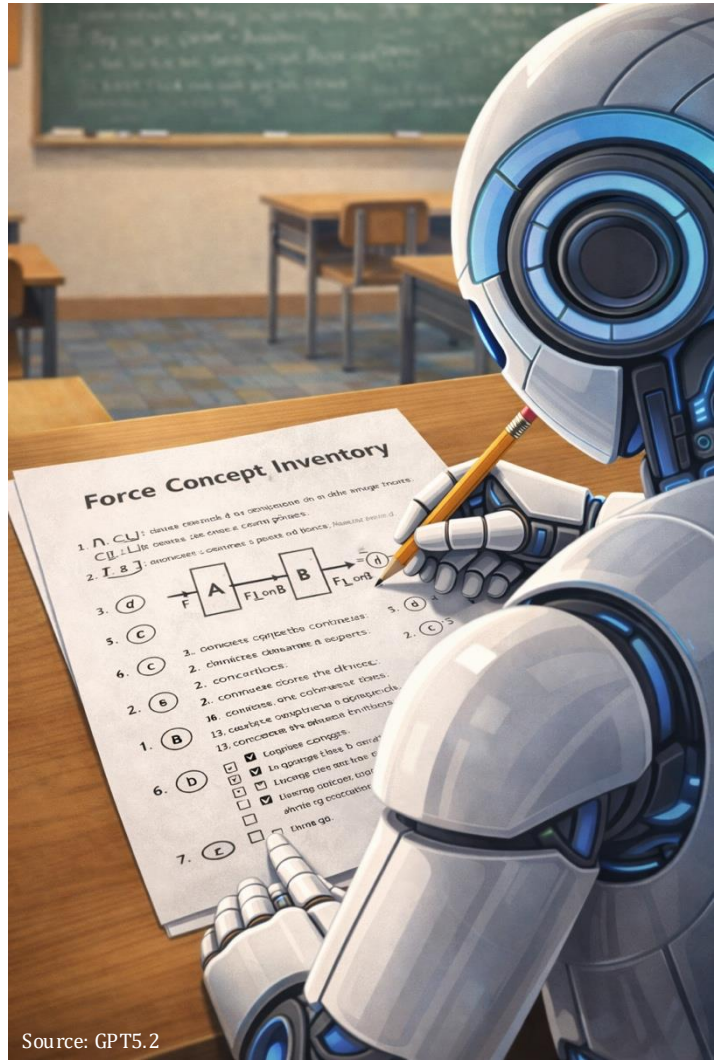
KEYWORDS
lesson planning; artificial intelligence; ChatGPT; inquiry; biology education

CONTACT Leroy Großmann leroy.grossmann@ruhr-uni-bochum.de Biology Education, Ruhr University Bochum, 44801 Bochum, Germany
*Present Address: Biology Education, Ruhr University Bochum, Bochum, Germany
Supplemental data for this article can be accessed online at https://doi.org/10.1080/09500693.2025.2567509
© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

- GenAI can generate structurally sound lesson plans, but **quality is inconsistent**
- Performs well on **content accuracy and alignment**, weak on epistemic goals
- Not great at inquiry-based teaching
- GenAI shifts planning work toward **evaluation, refinement, and redesign**

4. Evaluating AI Knowledge and Skills

- **Strong on text-based questions**, often at or above student level
  - **Weak on visuals & representations** (graphs, diagrams, spatial reasoning)
  - **Errors mirror student misconceptions**, not random failures
- **Multistep reasoning is fragile** unless using reasoning-optimized models
- **Reasoning models rival experts**, challenging unsupervised assessment



Source: GPT5.2

PHYSICAL REVIEW PHYSICS EDUCATION RESEARCH **21,** 020115 (2025)

Featured in Physics

Evaluating GPT- and reasoning-based large language models on Physics Olympiad problems: Surpassing human performance and implications for educational assessment

Paul Tschisgale,[1] Holger Maus,[1] Fabian Kieser,[2] Ben Kroehs,[3] Stefan Petersen,[1] and Peter Wulff[4]

[1]*Department of Physics Education, Leibniz Institute for Science and Mathematics Education, 24118 Kiel, Germany*
[2]*Department of Physics-Physics Education Research, Free University of Berlin, 14195 Berlin, Germany*
[3]*Kirchhoff Institute for Physics, Ruprecht Karl University of Heidelberg, 69117 Heidelberg, Germany*
[4]*Department of Physics and Physics Education Research, Ludwigsburg University of Education, 71602 Ludwigsburg, Germany*

(Received 15 May 2025; accepted 27 June 2025; published 13 August 2025)

Large language models (LLMs) are now widely accessible, reaching learners across all educational levels. This development has raised concerns that their use may circumvent essential learning processes and compromise the integrity of established assessment formats. In physics education, where problem solving plays a central role in both instruction and assessment, it is therefore essential to understand the physics-specific problem-solving capabilities of LLMs. Such understanding is key to informing responsible and pedagogically sound approaches to integrating LLMs into instruction and assessment. This study therefore compares the problem-solving performance of a general-purpose LLM (*GPT-4o*, using varying prompting techniques) and a reasoning-optimized model (*o1-preview*) with that of participants in the German Physics Olympiad, based on a set of well-defined Olympiad problems. In addition to evaluating the correctness of the generated solutions, the study analyzes the characteristic strengths and limitations of LLM-generated solutions. The results of this study indicate that both tested LLMs (*GPT-4o* and *o1-preview*) demonstrate advanced problem-solving capabilities on Olympiad-type physics problems, on average outperforming the human participants. Prompting techniques had little effect on *GPT-4o*'s performance, and *o1-preview* almost consistently outperformed both *GPT-4o* and the human benchmark. The main implications of these findings are twofold: LLMs pose a challenge for summative assessment in unsupervised settings, as they can solve advanced physics problems at a level that exceeds top-performing students, making it difficult to ensure the authenticity of student work. At the same time, their problem-solving capabilities offer potential for formative assessment, where LLMs can support students in evaluating their own solutions to problems.

DOI: 10.1103/6frmx-bsml

I. INTRODUCTION

Large language models (LLMs) are now widely accessible to the public, including students across all educational levels. Since the release of ChatGPT in November 2022, there has been rapid progress in LLM development, accompanied by striking improvements in their apparent capabilities. In parallel, concerns have been raised about students using these tools to circumvent meaningful learning processes [1,2] and to undermine the integrity of unsupervised exams or written assignments [3,4]. More broadly, LLMs have become central to ongoing discussions about the role of artificial intelligence (AI) in education [5–8]. It is therefore essential for educators and educational researchers to stay informed about the rapid advancements in AI—particularly with regard to LLMs—in order to investigate how these technologies can be integrated into education in effective and responsible ways.

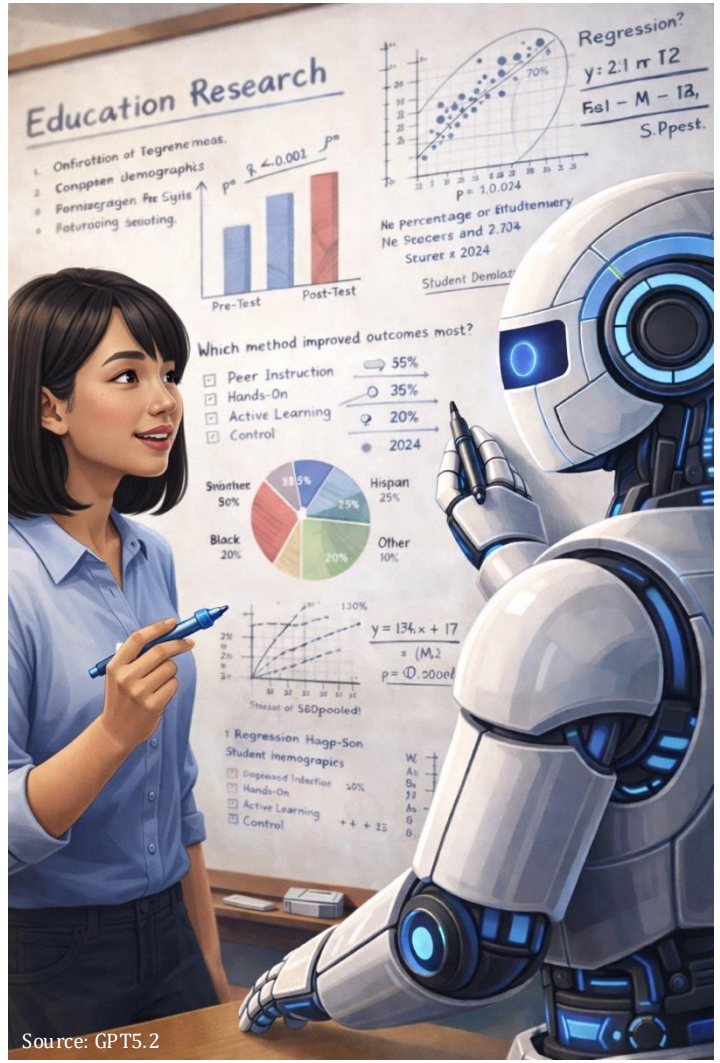
A step in this direction is to examine how LLMs relate to specific forms of instruction and assessments that are central in different disciplines. In physics education, problem solving is central to instruction and assessment, as it relates to both conceptual understanding and the application of physics-specific knowledge in structured and goal-oriented ways. Physics problem-solving abilities are pivotal to master for students planning to engage in a physics-related career [9,10]. Supporting the development of this ability is therefore a central goal of physics education and physics education research [11,12]. Understanding the apparent capabilities of LLMs in this context is important, particularly in light of emerging evidence connecting their problem-solving and assessment capabilities [13].

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

- GPT-4o and reasoning model o1-preview outperform Olympiad students on average
- Reasoning-optimized models show more structured, reliable solution paths
- Prompting has limited impact compared to model architecture
- LLMs excel at well-defined problems, still struggle with ill-defined ones
- Raises serious concerns for unsupervised summative assessment

5. AI in Science Education Research

- Scales **qualitative analysis** to datasets impossible for humans alone
- **Embeddings + ML** reliably reproduce expert coding patterns without GenAI black-box issues
- Human-in-the loop is important for validity, transparency, and trust (computational grounded theory)



PHYSICAL REVIEW PHYSICS EDUCATION RESEARCH **20,** 020151 (2024)

Using text embeddings for deductive qualitative research at scale in physics education

Tor Ole B. Odden,[1,*] Halvor Tyseng,[2,*] Jonas Timmann Mjaaland,[2,*]
Markus Fleten Kreutzer,[2,*] and Anders Malthe-Sørenssen[1,2]
[1]*Center for Computing in Science Education, University of Oslo, NO-0316 Oslo, Norway*
[2]*Center for Interdisciplinary Education, University of Oslo, NO-0316 Oslo, Norway*

(Received 26 February 2024; accepted 13 November 2024; published 20 December 2024)

We propose a technique for performing deductive qualitative data analysis at scale on text-based data. Using a natural language processing technique known as text embeddings, we create vector-based representations of texts in a high-dimensional meaning space within which it is possible to quantify differences in meaning as vector distances. To apply the technique, we build off prior work that used topic modeling via latent Dirichlet allocation to thematically analyze 18 years of the Physics Education Research Conference Proceedings literature. We first extend this analysis through 2023. Next, we create embeddings of all texts and, using representative articles from the 10 topics found by the LDA analysis, define centroids in the meaning space. We calculate the distances between every article and centroid and use the inverted, scaled distances between these centroids and articles to create an alternate topic model. We benchmark this model against the LDA model results and show that this embeddings model recovers most of the trends from that analysis. Finally, to illustrate the versatility of the method, we define eight new topic centroids derived from a review of the physics education research literature by Docktor and Mestre and reanalyze the literature using these researcher-defined topics. Based on these analyses, we critically discuss the features, uses, and limitations of this method and argue that it holds promise for flexible deductive qualitative analysis of a wide variety of text-based data that avoids many of the drawbacks inherent to prior NLP methods.

DOI: 10.1103/PhysRevPhysEducRes.20.020151

I. INTRODUCTION

Qualitative analysis is not easy. It is time-consuming and often difficult to replicate. In physics education research (PER), specifically, many qualitative studies aim to infer students' or instructors' understandings, framings, emotions, or social dynamics, and therefore, require fine-grained analysis of subtle details of dialogue, gesture, or speech. Categorizations can hinge on the use of specific phrases or words, the lack thereof, or even require researchers to go beyond surface content and context to focus on more gestalt aspects of the data.

Despite these challenges, qualitative analysis is often the best method we have for many of the research problems in a field as diverse, multifaceted, and complex as PER. Ours is a fairly new field, and as such is still in the process of identifying, exploring, and drawing in various methods and research questions [1]. It also addresses a sufficiently complex topic—human learning—that the methods used often look less like standard physics than biology or social sciences research, where qualitative methods are much more prevalent [2].

At the same time, there is plenty of room for methodological improvement. As theoretical paradigms and research questions become established, researchers often move from theoretical papers, small-N case studies, and existence proofs (e.g., [3,4]) to large-scale validation studies (e.g., [5,6]). But large qualitative studies often come at the cost of either significantly higher time investment or lower analytical grain size.

Over the past several years, the field of natural language processing (NLP) has made huge gains in the mathematical representation, manipulation, and analysis of text. These tools hold significant promise in supporting qualitative research on text-based data since computers are now able to reliably perform many of the types of content-based analyses that were previously only possible to do by a human researcher. Thus, these methods have the potential to help researchers scale up their qualitative analyses. In this article, we show a proof-of-concept of how one of these current, cutting-edge NLP methods can be used as an aid in performing deductive qualitative analysis at scale in PER.

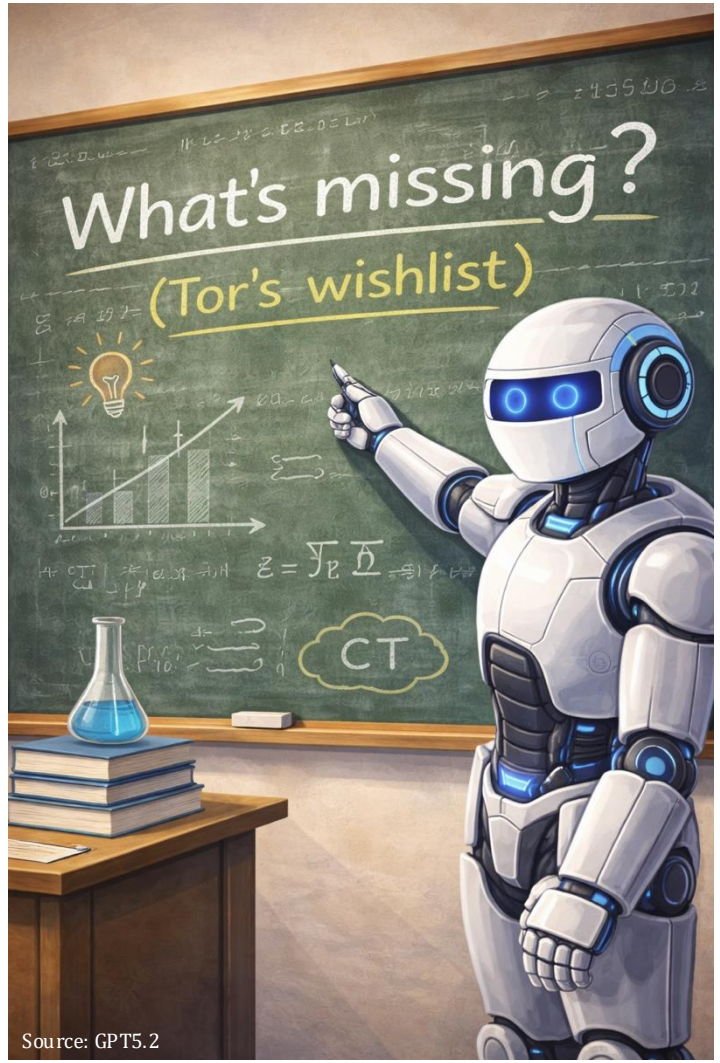
*These authors contributed equally to this work.
†Contact author: t.o.b.odden@fys.uio.no

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

- Used **embeddings** to quantify meaning of physics education research conference papers
- Humans **deductively** define categories based on examples
- Good **agreement** with prior analyses
- Avoids "black box" issues with chat-based models

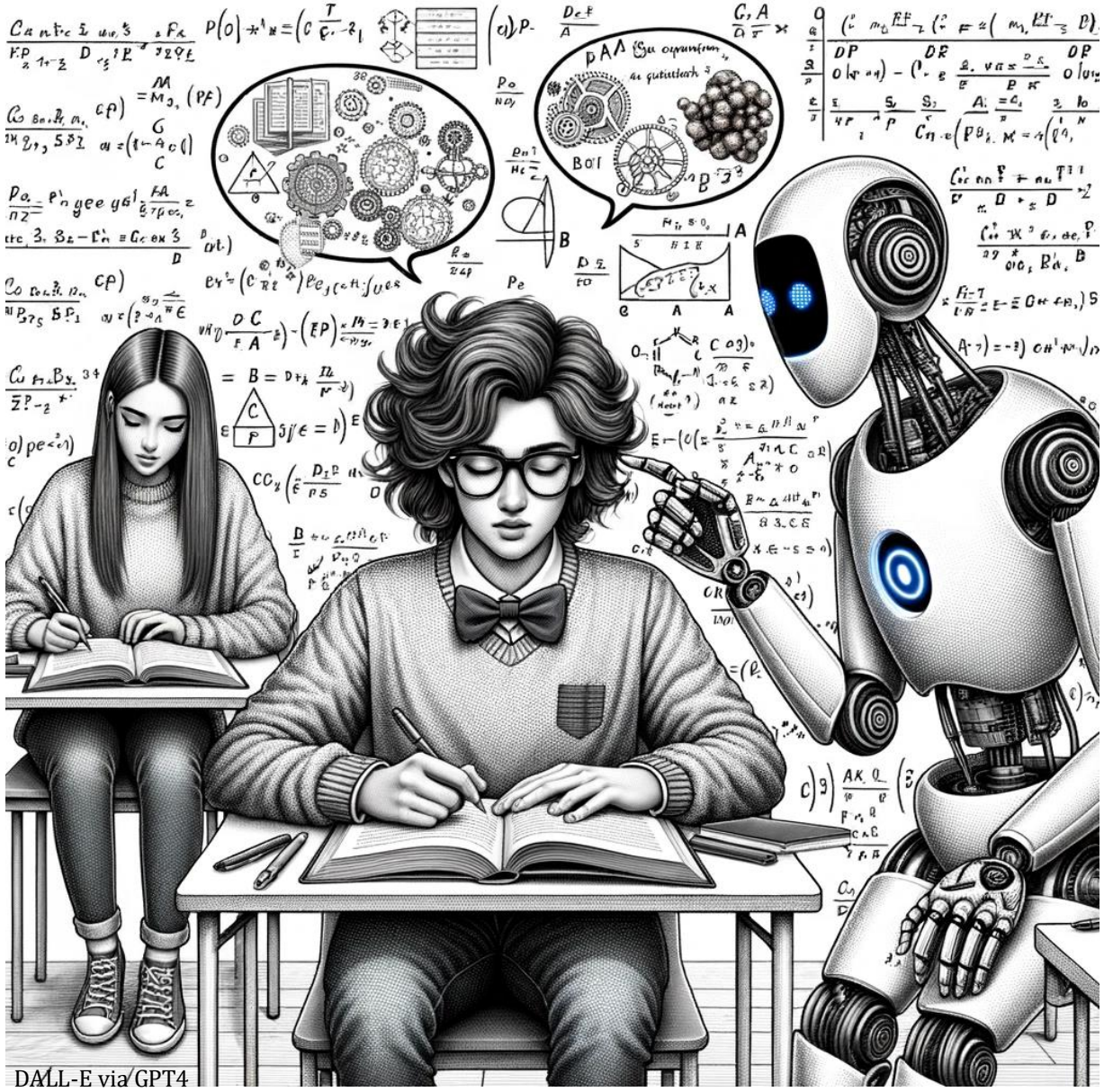
What's missing? (Tor's wishlist)

1. **Theory-driven studies** of AI-mediated learning
   - Build on the theory we already have!
2. How AI can address enduring **"friction points"** in our education systems
   - CT, active learning, differentiation etc.
3. Documenting and addressing **learning-loss** due to AI



Source: GPT5.2

THANK YOU!



DALL-E via GPT4