Physics Education

PAPER • OPEN ACCESS

The impact of AI in physics education: a comprehensive review from GCSE to university levels

To cite this article: Will Yeadon and Tom Hardy 2024 *Phys. Educ.* **59** 025010

View the article online for updates and enhancements.

You may also like

- Leveraging large language models to monitor climate technology innovation
Malte Toetzke, Benedict Probst and Stefan Feuerriegel
- Frequency multiplication with toroidal mode number of kink/fishbone modes on a static HL-2A-like tokamak
Zhihui ZOU, , Ping ZHU et al.
- Exploring the role of large language models in radiation emergency response
Anirudh Chandra and Abinash Chakraborty

This content was downloaded from IP address 81.156.86.109 on 12/03/2024 at 17:38

OPEN ACCESS
PAPER
Phys. Educ. **59** (2024) 025010 (18pp)
iopscience.org/ped

The impact of AI in physics education: a comprehensive review from GCSE to university levels

Will Yeadon* **and Tom Hardy**

Department of Physics, Durham University, Lower Mountjoy, South Rd, Durham DH1 3LE, United Kingdom

E-mail: will.yeadon@durham.ac.uk

CrossMark

Abstract

With the rapid evolution of artificial intelligence (AI), its potential implications for higher education have become a focal point of interest. This study delves into the capabilities of AI in physics education and offers actionable AI policy recommendations. Using openAI's flagship gpt-3.5-turbo large language model (LLM), we assessed its ability to answer 1337 physics exam questions spanning general certificate of secondary education (GCSE), A-Level, and introductory university curricula. We employed various AI prompting techniques: Zero Shot, in context learning, and confirmatory checking, which merges chain of thought reasoning with reflection. The proficiency of gpt-3.5-turbo varied across academic levels: it scored an average of 83.4% on GCSE, 63.8% on A-Level, and 37.4% on university-level questions, with an overall average of 59.9% using the most effective prompting technique. In a separate test, the LLM's accuracy on 5000 mathematical operations was found to be 45.2%. When evaluated as a marking tool, the LLM's concordance with human markers averaged at 50.8%, with notable inaccuracies in marking straightforward questions, like

---

* Author to whom any correspondence should be addressed.

Original Content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

© 2024 The Author(s). Published by IOP Publishing Ltd

multiple-choice. Given these results, our recommendations underscore caution: while current LLMs can consistently perform well on physics questions at earlier educational stages, their efficacy diminishes with advanced content and complex calculations. LLM outputs often showcase novel methods not in the syllabus, excessive verbosity, and miscalculations in basic arithmetic. This suggests that at university, there's no substantial threat from LLMs for non-invigilated physics questions. However, given the LLMs' considerable proficiency in writing physics essays and coding abilities, non-invigilated examinations of these skills in physics are highly vulnerable to automated completion by LLMs. This vulnerability also extends to pysics questions pitched at lower academic levels. It is thus recommended that educators be transparent about LLM capabilities with their students, while emphasizing caution against overreliance on their output due to its tendency to sound plausible but be incorrect.

Keywords: ChatGPT, prompt engineering, higher education, AI benchmark, Zero Shot, Few Shot, AI in physics education

1. Introduction

1.1. Background

Since OpenAI released ChatGPT, there has been a burgeoning interest in the higher education (HE) sector regarding the potential impact of artificial intelligence (AI) on learning and teaching [1–3]. The transformative potential of AI, particularly large language models (LLMs) - neural networks trained on vast amounts of text—has captivated educators. Reinforcing its significance in the educational realm, OpenAI even released a 'tips for educators' blog post[1]. Chatbots like ChatGPT, built on the transformer architecture [4], use a decoder-only design to predict subsequent words, equipping them to handle intricate queries. Following the prominence of ChatGPT, tech behemoths such as Meta, Google, and Baidu have launched their own AI-driven chatbots: LLama, Gemini [5], and Ernie. While these models excel in various tasks, preliminary research indicates that they may not consistently meet the rigorous academic standards of university settings [6], with GPT-4, the latest iteration of the GPT series, outperforming its counterparts.

Research within physics specifically has begun to assess the capabilities and implications of AI, largely focusing on ChatGPT. For instance, at the secondary school level, a pilot study led by Bitzenbauer engaged students in asking ChatGPT physics questions as a learning exercise and elicited their feedback on the generated responses [7]. Moreover, Yeadon and Halliday, when examining a set of physics exams administered at Durham university, found that GPT-4 typically achieved scores around the 50% mark [8]. Interestingly, the markers frequently noted the plausible sounding nature of the responses from GPT-4, despite them not always being correct. This phenomenon was also highlighted in a study by Dahlkemper *et al* [9]. They observed that responses from ChatGPT to challenging physics questions—ones that were more difficult than the students tested had previously encountered—were rated comparably to researcher-written responses. This was the case even though, for simpler questions, incorrect ChatGPT responses could be identified more easily by the students. Similarly, Gregorcic and Pendrill [10] found that a conversation with ChatGPT could yield intricate dialogue but incorrect physics concepts. This issue of complex yet plausible-sounding language masking incorrect

---

[1] Available at https://openai.com/blog/teaching-with-ai.

content is a hallmark of ChatGPT completions. Focusing on essay-based physics assignments, researchers discovered that ChatGPT's performance was generally on par with the average student's score on short-form physics essay assignments [11].

A growing body of research suggests the importance of equipping students with skills and experience to interact effectively with AI [7, 12, 13]. While this idea seems practical, it is crucial to acknowledge the continuous evolution of AI and computational technologies. As these systems become more user-friendly, the requirement for extensive technical knowledge decreases. This trend is evident in the rise of intuitive development environments like Replit and design tools like Figma, both of which have simplified many complexities in software creation. A similar trend towards readability is seen in programming languages, with Python being a prime example. Supporting this trend, ChatGPT has shown the capability to convert natural language into functional source code that can solve Leetcode problems [14]. Here, natural language can be seen as an even higher-level representation of source code, which itself is a higher-level representation of machine code. This suggests a future where specialized knowledge in areas like prompt engineering could become less important, replaced by more intuitive and direct interaction with AI systems. Indeed, the interaction techniques used in this research might soon be outdated due to the rapid development in AI technology.

While these studies underscore AI's prowess in solving physics problems, they represent just one facet of physics education. Assessment for learning, with its focus on continuous, formative assessments to enhance teaching and learning, contrasts starkly with summative assessments aimed at mere problem-solving. This distinction becomes crucial as AI, particularly modern LLMs, are advancing in ability and becoming closer to student performance in exams [8]. In global contexts, such as Singapore [15], where high-stakes exams—tests that constitute over 50% of the final grades—dominate, there's a tendency towards an exam-driven learning approach, perhaps at the expense of a deeper grasp of the subject. It's imperative, therefore, when considering AI's role in assessments, to look more broadly at technology-enhanced assessment (TEA). French *et al* [16] emphasize the need for a comprehensive re-evaluation of traditional assessment methods in education, arguing for the integration of TEA to foster deeper, more authentic learning experiences that align with the demands of a digitally evolving world. Their analysis underscores the importance of moving beyond rote memorization and exam-centric approaches, advocating for assessments that encourage critical thinking, creativity, and the application of knowledge in diverse contexts.

Understanding the effects of different interaction methods on AI performance, along with grasping AI's current capabilities, is vital for the physics education community. This study aims to delve into these topics to provide educators with a better understanding of how to adapt to the AI evolution and to suggest practical ways to adjust to this rapid change.

1.2. Interaction with AI models

There's a growing recognition that the conventional back-and-forth messaging inherent in chat-style interactions may not be entirely representative of the full potential of LLMs. The interaction quality and the outcome largely depend on not just the proper formulation of the prompt, but also on the application of various frameworks like reflection and chain of thought reasoning. Furthermore, integrating external tools such as Wolfram Alpha can significantly enhance the performance of the LLMs.

Prompting techniques, including Zero Shot and Few Shot, are among the primary factors influencing the output quality. The Zero Shot approach entails asking a question directly and expecting an answer, without providing any prior context or examples. On the other hand, Few Shot involves presenting examples to the model before posing the question [17], thereby giving the model a context to generate a more informed response. This study uses OpenAI's ChatGPT thus in order to implement Few Shot prompting in context learning (ICL) is used whereby the examples are presented within the prompt sent to the LLM.

The Confirmatory Check technique is an implementation that combines elements of the chain of thought [18] and reflection [19] methods.

It encourages a LLM to reconsider its previous outputs, removing excess content if appropriate. This method prompts the model to evaluate its initial response, thereby mitigating the problem where the LLM becomes 'stuck' with a mistake in its produced answer. Additionally, LLM outputs can sometimes be long, rambling, and inconsistent with the complexity of the question. The confirmatory check technique provides an opportunity for the LLM to avoid these issues. This approach offers valuable insights in educational contexts, where it mimics a more conversational interaction between students and the LLM. This back-and-forth dialogue can lead to more refined and accurate answers.

The idea of equipping LLMs with external tools to handle challenging tasks has also gained traction recently. A notable instance is the integration of Wolfram Alpha with ChatGPT, allowing the LLM to leverage Wolfram Alpha's capabilities to tackle complex mathematical tasks that are typically difficult for LLMs [20].

In exploring these different techniques, our aim is not only to provide a broader understanding of how these models can be utilized but also to evaluate their efficacy within the context of physics education. This serves the larger goal of this study—to benchmark these cutting-edge LLMs thoroughly and provide educators with a clearer picture of AI's strengths and weaknesses. By doing so, we hope to offer a comprehensive resource to understand AI's current capabilities and thereby inform educators about effective ways to integrate AI into their teaching practices.

2. Method

2.1. Question sources

To ensure a comprehensive examination of the AI's capabilities across various difficulty levels, we sourced questions that spanned from general certificate of secondary education (GCSE) to A-Level, as well as textbooks used in introductory university courses, often covering first-year content and occasionally extending into second-year material. These questions were obtained from a wide array of educational boards and institutions, culminating in a diverse and robust dataset. The GCSE and A Levels are key academic stages in the UK education system. GCSEs, undertaken at ages 14–16, mark the end of compulsory education and encompass a range of subjects. Post-GCSE, students aged 16–18 take A Levels, choosing three to four subjects aligned with their future goals. A Levels are critical for university admission, with their grades significantly influencing university entrance decisions.

To transfer questions from their original sources into a digital, machine-readable format, we utilized a combination of regular expression matching and manual transcription. Special emphasis was placed on maintaining the accuracy of the transcription process, preserving the original complexity and structure of each question. However, due to the requirement of sending API requests in Latin-1 encoding (ISO/IEC 8859-1), mathematical notations such as the square root or integral symbols were unavailable. We adapted to this constraint by using natural language shorthands, such as 'sqrt(x)' or 'integrate(x)', which proved to be an effective solution. Further, when questions incorporated tables or figures, we adopted specific strategies. Tables were reformatted to resemble nested Python lists. As for figures, we provided detailed descriptions. However, this approach for figures was seldom practical. The questions were organized into three distinct categories: numerical, where calculations such as 'find the acceleration' were required; multiple choice, involving selection from a list of options; and written descriptive answers, where textual responses were needed. The sources of the questions are detailed in table 1.

Building on the extensive research focusing on university-level physics exam questions [8], the current study narrows its scope to introductory level questions. The textbooks from which these questions were sourced are shown in table 1. To ensure fair evaluation, the scoring procedures for these questions were standardized across the different educational levels. For GCSE and A-Level questions, we adhered to the respective mark schemes provided. On the other hand, university-level questions, being derived from textbooks, lacked a standard mark scheme. To address this, a customized scoring rubric was developed. Specifically, questions from *University physics with modern physics* were found to be more elaborate and were consequently

Table 1. Question sources used for the evaluation.

Level	Source	Number of questions
GCSE	OCR physics A—Gateway physics 2017–2021	263
	OCR physics B—21st Century physics 2017–2021	91
A-Level	OCR physics A 2017–2021	244
	OCR physics B (Advancing physics) 2017–2021	130
University	University physics with modern physics 15th Ed.	175
	Physics for scientists and engineers 4th Ed.	180
	College physics 2e 2nd Ed.	112
	Physics principles with applications 7th Ed.	142
Total		1337

marked on a 2-point scale. A score of 2/2 was awarded for completely accurate answers, 1/2 for answers demonstrating an understanding of the physics concept being tested but with minor numerical or procedural errors, and 0/2 for all other responses. For questions sourced from the other university textbooks, a simpler 1-point scale was utilized, wherein each question was marked as either correct or incorrect. This approach aimed to strike a balance between accommodating the inherent complexity of questions from different sources and maintaining an equitable evaluation framework.

2.2. *Generating the AI answers*

We utilized the OpenAI API, specifically the gpt-3.5-turbo language model, to generate AI responses from an array of message objects [21]. By altering the format of the message object array, we implemented various interaction techniques: Zero Shot, ICL and Confirmatory Check. Each message object has a role of either *system*, *user*, or *assistant*. The *system* message objects guide the behavior of the LLM. The OpenAI default *system* message is 'You are a knowledgeable assistant' [22], which was retained for the Zero-shot prompting interactions. It was followed by a *system* message reading 'Please answer the following question.' to ensure the question was answered, and then a *user* message containing the actual question content.

For the ICL prompt implementation, the system messages were modified to include a series of example question-answer pairs before the target question, as shown in figure 1. These examples served to establish the context for the expected responses. Studies have shown that beyond five examples, the benefits of additional examples become negligible [23]. We found the LLM would often provide lengthy responses, so the examples were deliberately concise. Although a word-based example was initially included, it was determined to be unnecessary since the model is trained principally on long text passages.

The confirmatory check technique was implemented by sending the ICL message object to the API with the ICL response appended as an *assistant* message. It was followed by a *user* message reading, *'Please check the previous answer to ensure you are happy with it. If you feel that you can express it more succinctly, then please do so. For reference, this was the original question: <question inserted>'*. This approach allowed the LLM an opportunity to refine its ICL answer.

While the OpenAI API does not directly provide a confidence score or probability with each response, the 'temperature' parameter was set at 0 to eliminate randomness in the generated responses [24]. The 'max tokens' parameter was set at 2000, suitable for extensive answers. After processing each question, the result was saved with the new answers in an excel workbook to prevent data loss in case of program termination. The grading and interpretation of the AI's responses are discussed in the subsequent sections.

2.3. *Automated grading*

To assess the LLM's capability in evaluating its own responses, the answer from each question—spanning different prompting styles—was submitted to the API. This submission

```
[{"role": "user", "content": """
     A 30W light bulb uses 600J of electrical energy in time t
        to produce 450 J of light energy. What is the efficiency
         eta of the light bulb?
    """ },
    {"role": "assistant", "content": """
     The efficiency is the ratio of useful energy output to the
        total energy input expressed as a percentage so in this
        case:
         eta = (450 J / 600 J) * 100%
         eta = 75 %
     """ },
    {"role": "user", "content": """
     Interference fringes, produced by monochromatic light are
        viewed on a screen placed a distance L from a double
        slit system with slit separation S. The distance between
         the centres of two adjacent fringes (the fringe
        separation) is W. If both S and L are doubled, what will
         be the new fringe separation?
         A) 2W
         B) W/2
         C) W
         D) 4W
    """ },
    {"role": "assistant", "content": """
    C
     """ },
    {"role": "user", "content": """
     A car accelerates from 12 m s to 21 m s in 6.0 s. How far
        did it travel in this time? Assume constant acceleration
        .
    """ },
    {"role": "assistant", "content": """
    First, we'll find the acceleration using the equation: a = (
       v - u) / t = (21 - 12) / 6 = 1.5 m/s^2

    Next, we'll use the equation for the distance traveled: s =
       ut + (1/2) * a * t^2 = 12 * 6 + (1/2) * 1.5 * 36 = 99 m

    So, the car traveled a distance of 99 meters during this
       time.
     """ },]
```

Figure 1. Message array used to implement the Few-shot prompting via ICL, illustrating how context is provided to guide the language model's responses.

included the solution, marking guidance, available marks, and the original question[2]. Tasking the LLM with marking its answers emulates a human marker's role. Subsequent comparisons between LLM-assigned scores and human evaluations provided insights into the LLM's efficacy. Given the potential for the LLM to assign improbable scores, like values below zero or exceeding available marks, checks were put in place. If an invalid score was provided thrice consecutively, it was recorded as '-1' signifying a marking failure. Due to the comprehensive marking guidance availability, only GCSE and A-Level sources were utilized.

2.4. *Mathematical capabilities*

Answering physics questions often requires mathematical calculation, something LLMs are not trained on. To evaluate the LLM's utility in physics problem-solving, where mathematical calculations are often essential, we assessed its computational accuracy through two datasets totaling 5000 numbers. The first dataset, with 2500 integer pairs ranging from 1 to 5 digits, tested basic arithmetic operations. The second dataset involved single operand operations (like squaring or calculating natural logarithms) on 2500 integers. This approach aimed to understand the LLM's potential errors in numerical computations often encountered in physics questions. The accuracy was judged based on perfect match criteria, with a focus on the LLM's ability to handle numerical complexity relevant to physics applications.

3. Analysis and results

3.1. *Overview*

Figure 2 illustrates the overall scores achieved by different AI prompting techniques across three academic levels: GCSE, A Level, and Introductory University. The three techniques represented are Zero Shot (blue), ICL (red), and confirmatory check (green). It shows that the performance of the three prompting techniques remains relatively consistent across the three academic levels, while the overall performance decreases as the academic level increases. Although there are slight variations in the percentage of correct answers, none of the techniques consistently outperforms the others across all levels.

To determine if these observed differences were statistically significant, an analysis of variance (ANOVA) test was conducted, with the results summarized in table 2. ANOVA is particularly apt for this analysis as it allows for a comparison of means across more than two groups. The null hypothesis for the ANOVA test states that there is no significant difference between the group means. The alternative hypothesis posits that at least one group mean is different. For the GCSE, A Level, and introductory university levels, the p-values were 0.5429, 0.1310, and 0.8828, respectively, indicating that we fail to reject the null hypothesis for all three academic levels. This suggests that the choice of prompting technique does not play a pivotal role in the AI's performance.

For a more nuanced analysis, each question was categorized as either multiple choice, numerical, or word-based. However, at the introductory university level, the dataset is overwhelmingly composed of numerical questions (>99%). This dominance renders a detailed, segregated analysis by question type challenging for this academic level. Nevertheless, the ANOVA test results for the GCSE and A Level, as showcased in table 2, indicate a statistically significant difference in the performance of the three prompting techniques across the various question types. Yet the differences are not consistent between academic levels with the LLM performing best on numerical questions at GCSE but best on word based at A-Level. Further word based questions were the worse performing type for the LLM at GCSE. The nature of the question can notably affect the LLM's accuracy. For example, in multiple choice questions, the LLM frequently settled on an answer that was not among the provided options. In these scenarios, it either refrained from answering altogether or selected the option that was closest to its often incorrect answer. Beyond these question types, it was observed that questions with tables scored similarly to those without, indicating that tables do not hinder LLM performance.

---

[2] A overview of the prompt instructing the AI to mark its own work can be found in the appendix figure A1.

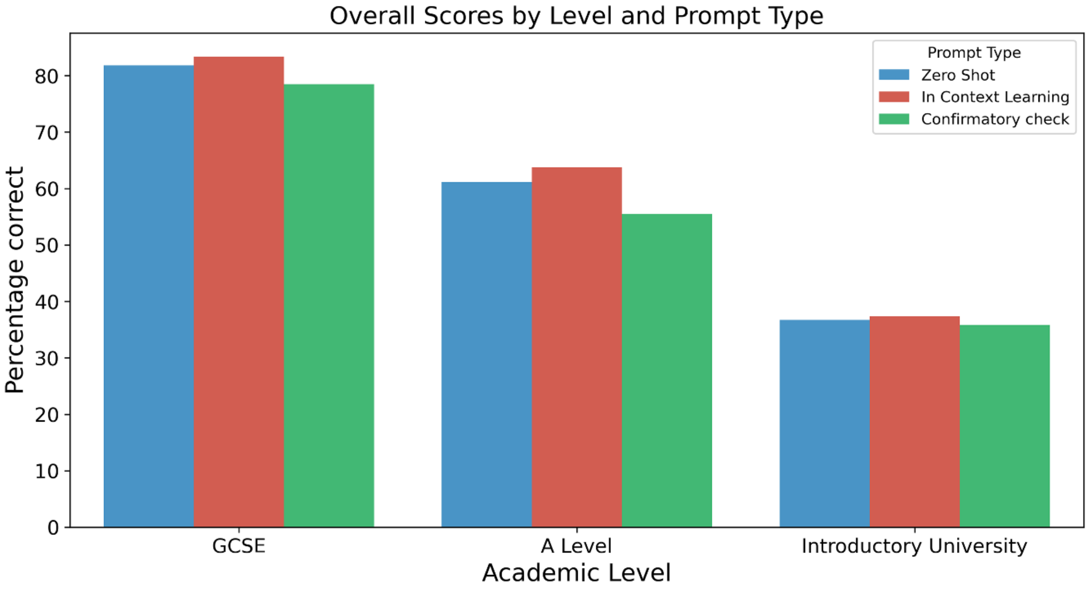

Figure 2. Comparative analysis of overall scores achieved by different AI prompting techniques (Zero Shot, Few Shot, confirmatory check) across three academic levels: GCSE, A Level, and introductory university.

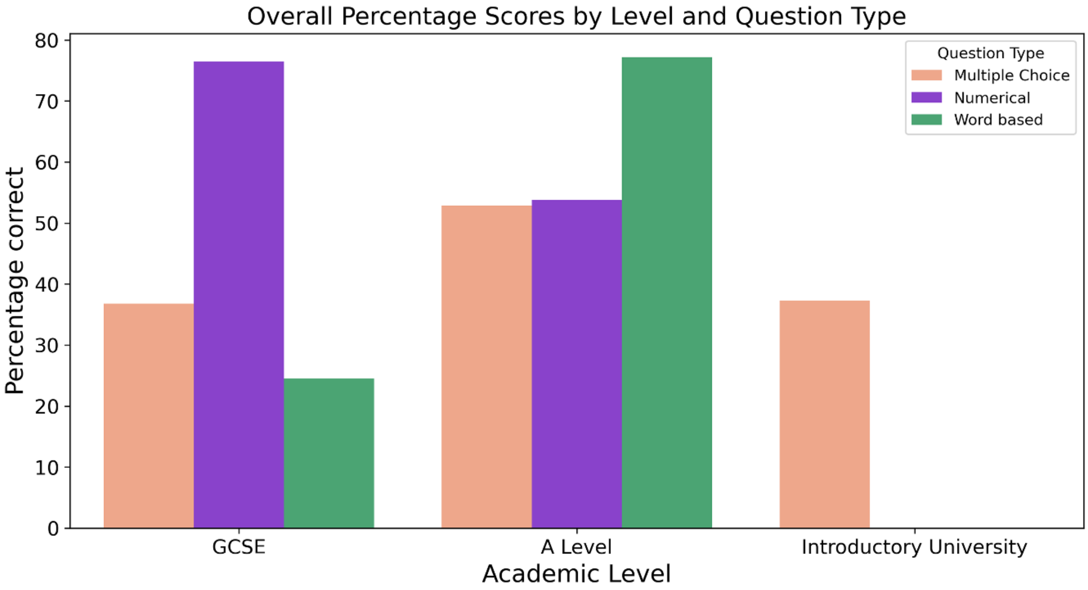

Figure 3. A detailed breakdown of the AI's performance in terms of percentage correct for various question types (multiple choice, numerical, word-based) at different academic levels, highlighting areas of success and potential improvement.

3.2. Example question answer

Looking at specific examples offers a clear perspective on the influence of prompt engineering. As depicted in figure 4, the nuances of different prompting styles can lead to varied responses. Given the physics question '*Write a decay equation in terms of a quark model for beta-minus decay*' the Zero-Shot prompt failed to appreciate the question asked about the quark

Table 2. ANOVA results for different prompting techniques and question types.

Category	F-statistic	p-value
Overall ANOVA Results		
GCSE Level	0.6111	0.5429
A Level	2.0366	0.1310
Introductory University Level	0.1246	0.8828
GCSE Level ANOVA Results for Different Question Types		
Zero Shot	80.7864	1.3781×10^{-29}
In Context Learning	88.5268	7.4370×10^{-32}
Confirmatory Check	92.3540	5.9499×10^{-33}
A Level ANOVA Results for Different Question Types		
Zero Shot	46.5909	8.8761×10^{-19}
In Context Learning	57.8985	1.3057×10^{-22}
Confirmatory Check	41.2386	6.7268×10^{-17}

model instead detailing β^- in a nucleus. The ICL prompting got the question completely correct but confirmatory check approach lost marks due to it stating an electron neutrino rather than an electron antineutrino in the answer. Interesting this may have been because a actual ν character was returned instead of the words 'anti-v' but $\bar{\nu}$ is not available in the Latin-1 character set.

The Zero-shot approach, while thorough, often yielded verbose answers, averaging 427 characters in length. In contrast, the ICL method trimmed responses to an average of 405 characters. The confirmatory check approach stood out as the most concise, with answers averaging just 228 characters. Additionally, while some mathematical content in the responses mirrored conventional formats, there were instances where the representations, though appearing correct, were mathematically inaccurate.

3.3. *AI marking*

For this evaluation, only instances where both human and the LLM successfully assigned a grade were included. Out of 3486 AI-generated answers to 1162 questions[3], the LLM only successfully graded 2209 instances, achieving a 63.4% rate of successful evaluations. All scores were normalized to facilitate a fair comparison across questions with different maximum marks. Human and LLM evaluation showed a concordance in scores for Zero-shot, ICL, and confirmatory checking with rates of 49.82%, 51.96%, and 50.54%, respectively. This means that for approximately half of the questions, the LLM gave the same score as the human markers. Among these 2209 graded instances, human markers assigned an average normalized score of 0.515, with a standard deviation of 0.448. The LLM's average normalized score was a lot higher at 0.952 but had a lower standard deviation of 0.167.

The observed correlations in table 3 show that human markers often grade ICL and confirmatory checks in a correlated manner, evidenced by a strong internal correlation of 0.913. In contrast, AI markers displayed a slightly weaker internal correlation of 0.662 between these same methods. Comparing human and AI grading reveals a moderate level of agreement, particularly for ICL and confirmatory checking (CC) with correlation values of 0.241 and 0.257, respectively. Zero-shot prompting shows a weaker correlation of 0.189.

Understanding these correlation values alongside the concordance rates suggests that the agreement is higher for straightforward questions with single correct answers. Meanwhile, more complex questions are likely sources of

---

[3] Questions from *University physics with modern physics* had a separate solutions booklet so could not be marked by the AI.

Prompt Type	Response
Zero-shot	In the quark model, beta-minus decay can be represented by the following decay equation: n → p + e- + anti(v)e where “n” represents a neutron, “p” represents a proton, “e-” represents an electron, and “anti(v)e” represents an antineutrino.
In Context Learning	In terms of a quark model, beta-minus decay can be represented by the following decay equation: d → u + e- + anti-v(e) In this equation, the down quark (d) decays into an up quark (u), an electron (e-), and an electron antineutrino (anti-v(e)).
Confirmatory Check	In terms of a quark model, beta-minus decay can be represented by the following decay equation: d → u + e- + νe

Figure 4. Comparison of responses for the given question based on different prompting styles in response to the question ‘*Write a decay equation in terms of a quark model for beta-minus decay*’.

Table 3. Correlation matrix for grades assigned by humans and the LLM. Human-ZS: Zero-shot prompted answers; human-ICL: In context learning prompted answers; human-CC: confirmatory check prompted answers; LLM-ZS, LLM-ICL, and LLM-CC are analogous for the LLM.

	Human-ZS	Human-ICL	Human-CC	LLM-ZS	LLM-ICL	LLM-CC
Human-ZS	1.000	0.800	0.754	0.189	0.156	0.139
Human-ICL	0.800	1.000	0.913	0.110	0.241	0.201
Human-CC	0.754	0.913	1.000	0.102	0.197	0.257
AI-ZS	0.189	0.110	0.102	1.000	0.334	0.284
AI-ICL	0.156	0.241	0.197	0.334	1.000	0.662
AI-CC	0.139	0.201	0.257	0.284	0.662	1.000

disagreement. These discrepancies may arise from the LLM’s different interpretation of the marking guidance or its emphasis on different parts of the response. The LLM-assigned scores also have a lower standard deviation, indicating a more consistent but potentially less nuanced grading approach.

3.4. Mathematical capabilities

In assessing the LLM’s mathematical abilities for physics applications, it achieved a 52.3% exact accuracy rate in basic arithmetic operations (addition, subtraction, multiplication, division) with two integers. This accuracy improved to 75.8% when allowing a ±5% margin of error. Notably, the LLM performed better in addition and subtraction than in multiplication and division, particularly with larger numbers. For single operand operations, such as squaring and calculating logarithms, the LLM’s exact accuracy was 38.1%, rising to 63.2% within a 5% error margin. Its performance was relatively weaker in handling trigonometric functions with multi-digit numbers. These findings indicate the LLM’s varying proficiency in mathematical calculations commonly encountered in physics, suggesting careful consideration for tasks requiring high numerical precision.

4. Discussion and conclusion

4.1. Overview

AI, especially in the realm of LLMs, continues to draw attention in academic circles. Within this landscape, this study set out to evaluate the proficiency of AI in physics education. The results presented in this study and elsewhere allow us to make general conclusions about LLM use within physics education and to provide recommendations for educators.

For the characteristics of LLM output, one notable aspect is that without a specific syllabus to adhere to, LLMs often introduced innovative methods, leading to novel approaches in answering. While this can be a fresh perspective, it does not always align with the traditional academic evaluations. Contrary to prior work emphasizing the importance of good prompting [25, 26], our investigation revealed statistically insignificant difference between different interaction techniques. We found that AI struggles with harder physics, as shown in figure 3. As the academic level increased, the amount of correct responses decreased. Previous research has highlighted how AI can often struggle with more complex physics; beyond introductory textbooks, Yeadon and Halliday [8] demonstrated how GPT-3.5 typically fails to pass most physics exams at Durham university. However the latest foundation model GPT-4 consistently outperforms GPT-3.5 and often scores nearly 50% on exams, this is shown in figure 5. Given these results, and as highlighted at the end of [6], the current potential threat of AI in non-invigilated online exams at university level seems to be relatively contained. In fact, it would be prudent to warn students that AI performance at GCSE and A Level may not transfer to university assessments. This leads to the conclusion that whilst non-invigilated GCSE and A-Level assessments should be wary of how good the latest foundational AI models are, at university level the threat is not as dire. The score of the best AI systems seems to, on average, peak at around 50% for physics questions meaning currently only the weaker students would benefit.

As a part of a physics degree, often there are written elements and computational work. Here the threat to assessment fidelity is more pronounced. There are LLMs specifically trained on coding examples which can excel at complex coding tasks found in a computer science focused degree where the complexity would typically be beyond that found in a physics degree [27, 28]. Further, research looking at physics essays specifically found AI excels here [11]. It is important that educators are aware of the capabilities in these areas and it is recommended that for coding and essay work, if the assessment is non-invigilated educators should enter their assignments into GPT-4 and see the capabilities themselves. The wide availability and capability of modern LLMs may be irreconcilable with with take home short essays or typical physics coding tasks.

LLMs often produce verbose outputs, the AI's proclivity to produce extensive responses, often not proportional to the question's complexity, is not only a hallmark but seems to be an integral part of quality answers. Whilst not statistically significant, there was a decline in performance with the confirmatory checking raises concerns about the AI's current capacity for iterative, conversational interaction, resonating with the observations by [10]. Interestingly, looking at the linguistics of the output much prior research has highlight how AI generated content is both difficult to detect [29, 30] and potentially bias against non-native english speakers [31]. Curiously there are simple techniques to get the AI to reveal itself such as asking 'Do you agree with this statement?' will often get the LLM to state 'As an AI assistant I do not have personal opinions, emotions, or preferences'. Similarly the use of zero-width spaces or hidden prompt injection attacks [32] within questions can also foil LLM effectiveness.

The present work also highlighted how LLMs can struggle with mathematical computations as the lengths of digits involved increases. Of the 5000 mathematical questions asked only 45.2% were answered correctly. The difference here however is that modern computers already have sophisicated mathematical capabilities meaning it

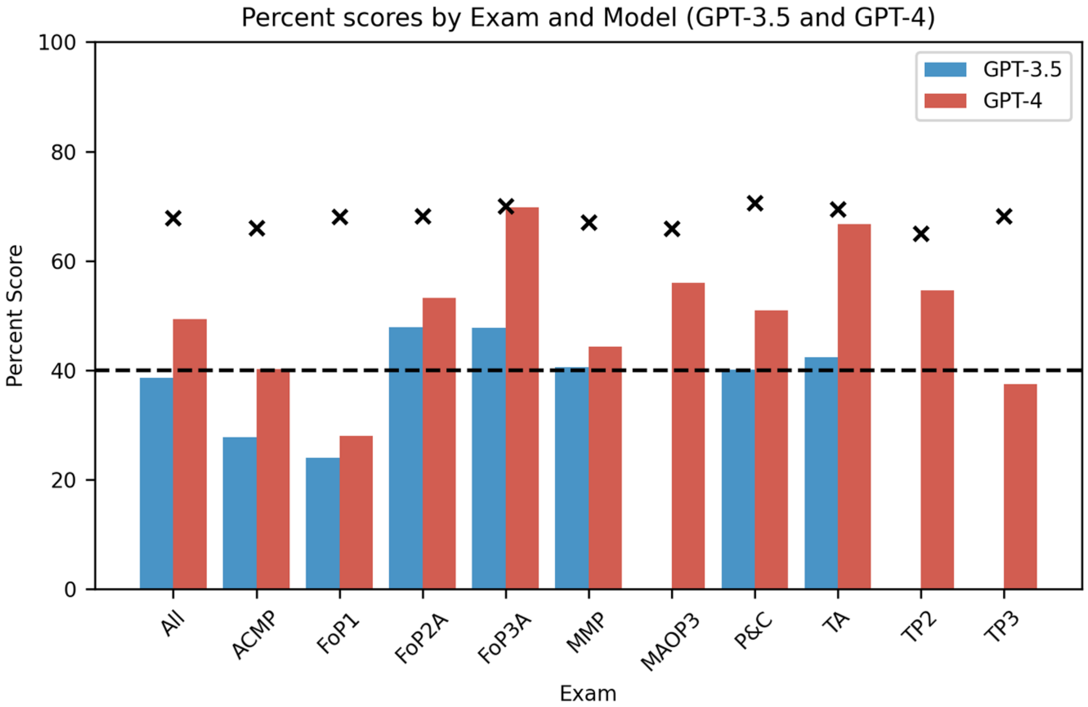

Figure 5. Performance of GPT-4 and GPT-3.5 on different physics exams as presented by Yeadon and Halliday [8]. The black crosses indicate the average student mark from 2018 to 2021 on the modules for the exam and the dashed black line shows the 40% score required to pass the exam. Critically, these exams were marked by the same academics who mark student exams. Acronym definitions are provided below. Reproduced with permission from [8]. The modules acronyms are for: ACMP : advanced condensed matter physics, FoP1 : foundations of physics 1, FoP2A : foundations of physics 2A, FoP3A : foundations of physics 3A, MMP : mathematical methods in physics, MAOP3 : modern atomic and optical physics 3, P&C : planets and cosmology, TA : theoretical astrophysics, TP2: theoretical physics 2, TP3: theoretical physics 2.

would be inapt to use a LLM to work out the cosine of a number when calculators are available. The AI's grading capability further supports this viewpoint, when marking multiple choice questions the AI often struggled to do this simple task correctly, a case of over engineering / using the AI for the wrong task. In fact when extending the marking to all questions a congruence rate of only 50.8% with human evaluations was found, indicating clear limitations in certain areas.

On a positive note, during our interactions, the AI maintained a respectful tone without displaying any abusive or exclusionary language, reflecting advancements in ethical AI design. While premium versions of some technologies might be inaccessible to some due to cost, educators should ensure that no student is mandated to use paid resources. To summarize, while AI has made significant strides, limitations persist in its application to physics. The key conclusions from our study are outlined in figure 6.

4.2. Recommendations

The swift progress in AI technology raises numerous ethical dilemmas, especially regarding its potential misuse in academia, its inherent biases, and its overarching societal repercussions. Echoing the concerns raised by [33], the incorporation of AI into the educational realm warrants a balanced mix of skepticism and meticulous scrutiny. As AI models continuously advance, a shared responsibility falls upon educators, developers, and policymakers to maintain vigilance, ensuring that AI tools are harnessed ethically and judiciously. In light of the current state of affairs, specific recommendations are posited, as depicted in figure 7.

1. **Novel Approaches:** LLMs, unbound by a specific syllabus, frequently adopt innovative methods when addressing questions. Such distinct approaches can serve as an indication that LLMs might have been employed in crafting the answer.
2. **Mathematical Hurdles:** Pure LLMs, lacking computational tools, face challenges when handling mathematical operations, particularly with numbers exceeding three digits.
3. **Prompting Limitations:** Contrary to initial beliefs about best-practice prompt engineering, this study revealed its apparent limited efficacy in the realm of Physics questions. While ICL enhancements did yield improved results, the advantage over Zero-shot approaches was marginal.
4. **Verbose Outputs:** The AI consistently produces verbose answers, often misaligned with the question's complexity. Notably, when provided an opportunity to refine its outputs, the LLM frequently produced content of diminished quality.
5. **Graphical Challenges:** In this study, an effective method for LLMs to address graphical questions was not identified. With minimal reformatting, LLMs handled questions involving tabled data comparably to other queries. Moreover, for 'sketch' / 'diagram' tasks, LLMs frequently used rows of symbols, offering reasonable attempts.

Figure 6. Key conclusions derived from this study's assessment of LLM responses to physics questions.

We also express caution about the limits of what this research shows. This study specifically looks at AI's ability to solve physics problems but there is a wider context of digitization of education. We echo the themes highlighted by Timmis *et al* [36] in their review of assessment in a digital age, where digital tools have the potential to provide bespoke education that adapts to student needs, but this comes with risks. The two principle risks applicable to this study are the misuse of student data, a risk that depends on the exact AI resource used, and the exacerbation of the 'digital divide' in the age of AI. Celik identified that 'computational thinking' (the skill of solving problems in a similar way to how computers work) and the access to information and communication technologies as key determinants of AI literacy [37]. We thus stress educators should be wary of recommending the use of commercial generative AI systems where the use of confidential information is required or state-of-the-art AI capabilities are required.

As previously mentioned, it is important to view the impact of AI in the broader context of increased digitization over the past decades. The effects of AI in education may be more extreme, but they can be viewed as similar to earlier technological advancements that have transformed educational practices. Just as earlier innovations brought their own challenges and opportunities, the integration of AI and digital tools in education continues this trend. It demands a careful consideration of the balance between technological potential and ethical responsibility, ensuring that advancements in educational technology contribute positively to the learning experience without compromising equity or privacy.

(1) **Transparency About Capabilities:** With over 100 million users, ChatGPT's influence is undeniable, and it's frequently highlighted in the news. Educators should openly discuss its strengths and weaknesses with students, especially its propensity to produce plausible yet occasionally incorrect or incomplete answers.

(2) **Caution Students Against Overreliance:** While AI may prove valuable at GCSE levels, its effectiveness can diminish in university settings, as illustrated in Figure 2. Students should be reminded that relying heavily on AI can deprive them of genuine learning experiences.

(3) **Avoid Teaching AI Interaction Techniques:** The study found no significant variance in performance across different prompt engineering methods for Physics questions. This was surprising as effective prompting techniques are subject of much research and reported improved performance [34, 35, 18]. However, from a Physics teaching perspective there is not enough clear benefit in improving Physics question answering abilities. Further, given the rapid advancements in AI, previously effective techniques can soon become outdated.

(4) **Change Some Assessment Methods:** Non-invigilated coding and short form essays are very vulnerable to automated completion by LLMs [11]. Further, as AI-written text is difficult to discern [29, 30] and potentially bias against non-native English speakers [31]. Advertised AI detectors should not be trusted.

(5) **Anticipation of Evolving Capabilities:** Educators should stay updated with the latest in AI advancements. As Yeadon and Halliday's study [8] illustrates, there's a discernible improvement from GPT-3.5 to GPT-4. However, it remains uncertain whether future models will improve further still or approach an asymptote.

(6) **Ethical Considerations in AI Use:** AI's interaction has shown a respectful tone without exclusionary language, highlighting advancements in ethical AI design. However, educators should ensure equitable access by not mandating the use of premium, potentially inaccessible technologies for students.

Figure 7. Recommendations for educators in addressing AI.

4.3. Concluding thoughts

AI is set to change how we approach education. Drawing from the findings of this study and the broader literature, it is clear that within the realm of physics education, AI presents a spectrum of threats and opportunities that vary based on context. Assessments at earlier educational stages, such as GCSE and A-Level, are notably susceptible when they are open-book. In contrast, when addressing advanced topics—especially at the university level and in textbook work—AI does not consistently provide correct answers, regardless of the prompting style. Moreover, students producing a high volume of quality work should not be unwelcome. The primary concern should be the active and meaningful involvement of students in creating such work. The path ahead remains uncertain; forthcoming foundation models might bring about marginal enhancements or represent substantial breakthroughs in capabilities. With sustained research, assessment, and collaboration, the academic community has the opportunity to channel the potential of AI, ensuring it enhances, rather than diminishes, physics education.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Appendix. AI Marking Prompt

```
"""
Your task is to assess an answer based on the provided format:
    Question: []
    Marks: []
    Solution: []
    Guidance: []
    Answer: []

Square brackets contain necessary information. Based on the question,
solution, and any guidance, assess the answer’s correctness. Return
only a number indicating the marks.

Example:
Question: [Interference fringes, produced by monochromatic light...]
Marks: [1]
Solution: [C]
Guidance: []
Answer: [C]

Correct Response: 1
Incorrect Response: This answer is correct.

Always return a number from 0 to 9. Responses are tested using Python
’.isdigit()’ method. Any non-numeric answer will be sent back for re-
evaluation.

To assess:
Question: [{question}]
Marks: [{marks}]
Solution: [{solution}]
Guidance: [{guidance}]
Answer: [{answer}]
"""
```

Figure A1. Condensed system prompt for AI self-marking. The AI was programmed to return a numerical score based on the question’s solution and guidance. The full prompt, with multiple detailed examples, is abbreviated here for brevity. The AI accurately marked questions 58.8% of the time.

ORCID iD

Will Yeadon https://orcid.org/0000-0002-9444-108X

Received 11 September 2023, in final form 13 December 2023
Accepted for publication 17 January 2024
https://doi.org/10.1088/1361-6552/ad1fa2

References

[1] Baidoo-Anu D and Owusu Ansah L 2023 Education in the era of generative artificial intelligence (AI): understanding the potential benefits of ChatGPT in promoting teaching and learning *SSRN* 4337484 (http://dx.doi.org/10.2139/ssrn.4337484)

[2] Rudolph J, Tan S and Tan S 2023 ChatGPT: bullshit spewer or the end of traditional assessments in higher education? *J. Appl. Learn. Teach.* **6** 1

[3] Rospigliosi P 2023 Artificial intelligence in teaching and learning: what questions should we ask of ChatGPT? *Interact. Learn. Environ.* **31** 1–3

[4] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I 2017 Attention is all you need *Advances in Neural Information Processing Systems* vol 30 (available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf)

[5] Gemini Team Google 2023 Gemini: a family of highly capable multimodal models (available at: https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf) (Accessed 06 December 2023)

[6] Rudolph J, Tan S and Tan S 2023 War of the chatbots: Bard, bing chat, ChatGPT, ernie and beyond. The new AI gold rush and its impact on higher education *J. Appl. Learn. Teach.* **6** 1

[7] Bitzenbauer P 2023 ChatGPT in physics education: a pilot study on easy-to-implement activities *Contemp. Educ. Technol.* **15** ep430

[8] Yeadon W and Halliday D P 2023 Exploring durham university physics exams with large language models (arXiv:2306.15609)

[9] Dahlkemper M N, Lahme S Z and Klein P 2023 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

[10] Gregorcic B and Pendrill A-M 2023 ChatGPT and the frustrated socrates *Phys. Educ.* **58** 035021

[11] Yeadon W, Inyang O-O, Mizouri A, Peach A and Testrow C P 2023 The death of the short-form physics essay in the coming AI revolution *Phys. Educ.* **58** 035027

[12] Kohnke L, Moorhouse B L and Zou D 2023 ChatGPT for language teaching and learning *RELC J.* **54** 00336882231162868

[13] Adiguzel T, Kaya M H and Cansu F K 2023 Revolutionizing education with AI: exploring the transformative potential of ChatGPT *Contemp. Educ. Technol.* **15** ep429

[14] Tian H, Lu W, Li T O, Tang X, Cheung S-C, Klein J and Bissyandé T F 2023 Is ChatGPT the ultimate programming assistant–how far is it? (arXiv:2304.11938)

[15] Wong H M, Kwek D and Tan K 2020 Changing assessments and the examination culture in singapore: a review and analysis of singapore's assessment policies *Asian Pac. J. Educ.* **40** 433–57

[16] French S, Dickerson A and Mulder R A 2023 A review of the benefits and drawbacks of high-stakes final examinations in higher education *Higher Educ.* **1** 1–26

[17] Brown T *et al* 2020 Language models are few-shot learners *Advances in Neural Information Processing Systems* vol 33, ed H Larochelle, M Ranzato, R Hadsell, M Balcan and H Lin (Curran Associates, Inc) pp 1877–901 (available at: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf)

[18] Wei J *et al* 2022 Chain-of-thought prompting elicits reasoning in large language models *Advances in Neural Information Processing Systems* vol 35, ed S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho and A Oh (Curran Associates, Inc) pp 24824–37 (available at: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf)

[19] Shinn N, Labash B and Gopinath A 2023 Reflexion: an autonomous agent with dynamic memory and self-reflection (arXiv:2303.11366)

[20] Frieder S, Pinchetti L, Griffiths R-R, Salvatori T, Lukasiewicz T, Petersen P C, Chevalier A and Berner J 2023 Mathematical capabilities of ChatGPT (arXiv:2301.13867)

[21] OpenAI 2020 OpenAI GPT-3 models (available at: https://platform.openai.com/docs/guides/gpt)

[22] OpenAI 2023 Best practices for prompt engineering with OpenAI API (available at: https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-openai-api)

[23] Dong Q, Li L, Dai D, Zheng C, Wu Z, Chang B, Sun X, Xu J and Sui Z 2022 A survey for in-context learning (arXiv:2301.00234)

[24] OpenAI 2020 OpenAI API documentation: parameters (available at: https://platform.openai.com/docs/api-reference/completions/create)

[25] Giray L 2023 Prompt engineering with ChatGPT: a guide for academic writers *Ann. Biomed. Eng.* **51** 1–5

[26] Lo L S 2023 The clear path: a framework for enhancing information literacy through prompt engineering *J. Acad. Librariansh.* **49** 102720

[27] Nijkamp E, Pang B, Hayashi H, Tu L, Wang H, Zhou Y, Savarese S and Xiong C, Codegen: an open large language model for code with multi-turn program synthesis (arXiv:2203.13474)

[28] Chen M *et al* 2021 Evaluating large language models trained on code (arXiv:2107.03374)

[29] Liu Z, Yao Z, Li F and Luo B 2023 Check me if you can: detecting ChatGPT-generated academic writing using checkGPT (arXiv:2306.05524)

[30] Perkins M, Roe J, Postma D, McGaughran J and Hickerson D 2023 Game of tones: faculty detection of GPT-4 generated content in university assessments (arXiv:2305.18081)

[31] Liang W, Yuksekgonul M, Mao Y, Wu E and Zou J 2023 Gpt detectors are biased against non-native English writers *Patterns* **4** 100779

[32] Greshake K, Abdelnabi S, Mishra S, Endres C, Holz T and Fritz M 2023 More than you've asked for: a comprehensive analysis of novel prompt injection threats to application-integrated large language models (arXiv:2302.12173)

[33] Cooper G 2023 Examining science education in ChatGPT: an exploratory study of generative artificial intelligence *J. Sci. Educ. Technol.* **32** 444–52

[34] Wang X, Wei J, Schuurmans D, Le Q V, Chi E H, Narang S, Chowdhery A and Zhou D 2023 Self-consistency improves chain of thought reasoning in language models *ICLR 2023* (arXiv:2203.11171)

[35] Kojima T, Gu S S, Reid M, Matsuo Y and Iwasawa Y 2022 Large language models are zero-shot reasoners *Advances in Neural Information Processing Systems* vol 35, ed S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho and A Oh (Curran Associates, Inc) pp 22 199–22 213 (available at: https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf)

[36] Timmis S, Broadfoot P, Sutherland R and Oldfield A 2016 Rethinking assessment in a digital age: opportunities, challenges and risks *Br. Educ. Res. J.* **42** 454–76

[37] Celik I 2023 Exploring the determinants of artificial intelligence (AI) literacy: digital divide, computational thinking, cognitive absorption *Telemat. Inform.* **83** 102026