The death of the short-form physics essay in the coming AI revolution

OPEN ACCESS PAPER

Phys. Educ. **58** (2023) 035027 (13pp) iopscience.org/ped

The death of the short-form physics essay in the coming AI revolution

Will Yeadon*, Oto-Obong Inyang, Arin Mizouri, Alex Peach and Craig P Testrow

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

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

Abstract

The latest AI language modules can produce original, high quality full short-form (300-word) Physics essays within seconds. These technologies such as ChatGPT and davinci-003 are freely available to anyone with an internet connection. In this work, we present evidence of AI generated short-form essays achieving First-Class grades on an essay writing assessment from an accredited, current university Physics module. The assessment requires students answer five open-ended questions with a short, 300-word essay each. Fifty AI answers were generated to create ten submissions that were independently marked by five separate markers. The AI generated submissions achieved an average mark of $71 \pm 2\%$, in strong agreement with the current module average of $71 \pm 5\%$. A typical AI submission would therefore most-likely be awarded a First Class, the highest classification available at UK universities. Plagiarism detection software returned a plagiarism score between $2 \pm 1\%$ (Grammarly) and $7 \pm 2\%$ (TurnitIn). We argue that these results indicate that current natural language processing AI represent a significant threat to the fidelity of short-form essays as an assessment method in Physics courses.

* Author to whom any correspondence should be addressed.

Keywords: pedagogy, language models, ethics, ChatGPT, NLP, authentic assessment

Supplementary material for this article is available online

1. Introduction

1.1. Background

AI text-completion technologies have undergone rapid development in recent years, outpacing commonplace perceptions of their quality. It is now possible to use these tools to reliably produce quality content that is accurate, clear and critical, on practically any topic of choice. These tools can produce content that is highly idiosyncratic, and which is nearly indistinguishable from human-produced content [1, 2]. Furthermore, these technologies are becoming increasingly accessible, fast, cheap and incredibly easy to use, essentially only requiring users to specify a brief text prompt. The effort required to produce such a prompt could entail simply copying and pasting a question from an assignment. Clearly, such technologies could pose a threat to the fidelity of existing forms of assessment, as AI written work could be submitted by students and both pass undetected through plagiarism prevention software and score higher marks than the student could have had they written the work themselves. In this work, we provide evidence that this threat is very significant. We demonstrate how high quality short-form essays, written by AI software in only a few seconds, can score a First Class for an assignment from an accredited university Physics module. This, we argue, effectively renders the short-form essay obsolete as an assessment tool. Despite popular reservations, the AI revolution has begun, and it is vital that educators remain proactive and aware of the capabilities of these technologies and their potential ramifications for teaching and assessment, particularly at this very early stage in the development of AI text-completion technologies.

Modern systems such GPT-3 are already extremely proficient at producing accurate, insightful responses on almost any topic, with applications in diverse areas such as healthcare [3], legal advice chat bots [4] and providing feedback for student's work [5]. They are thus potentially capable of answering the short-form essay questions which form the basis of assessments in many university modules. Whilst there are some technologies capable of detecting AI generated text [6], universities currently only look for plagiarism rather than whether text was AI generated. Previous technological innovations such as the widespread use of virtual learning within universities have had muted impact on teaching practices [7] yet the current capabilities of relatively earlier stage technologies such as ChatGPT and davinci-003 is likely to change this for the case of AI.

Given technologies such as davinci-003 and ChatGPT are only months old, there is limited prior work on AI written essays. An older language model, GPT-2, has been used by students to write essays in a co-piloting role [8]. Here, students were tasked with blending an essay they have written themselves with output from GPT-2. However, new models such as ChatGPT and davinci-003 can write entire essays using a single, user-defined sentence prompt. In fact, through stringing together several prompts GPT-3 has already demonstrated the potential to create an entire academic paper [9]. If one only cherry-picks the most exemplary outputs as examples of AI capabilities, this invariably implicates survivorship bias. However, a hypothetical student who wished to use an AI tool to answer their essay would most-likely choose the best output they could from a set of multiple, rapidly-generated options. Clearly it is prudent to appraise the capabilities of these technologies within the context of how they will most likely be used by students.

1.2. GPT-3: davinci-003 and ChatGPT

OpenAI's GPT-3 model is an autoregressive language model [10] that processes and generates text. Trained on a large dataset of human-generated text, GPT-3 can analyze a prompt (the input text) and use statistical techniques to predict the most likely next word or words that

Prompt: Explain why Shakespeare compared someone to a Summers' day instead of a Winters' day.

Output: Shakespeare most likely compared someone to a summers' day instead of a winters' day because a summers' day is typically seen as a day of happiness, while a winters' day is seen as a day of darkness.

Example 1. davinci-003 output showing relatively straightforward literary analysis.

Prompt: Is it ethical to use AI text-completion technologies for your homework assignments?

Output: There is no right answer to this question since it can be seen as ethical or unethical depending on one's perspective. Some people might see it as unethical because it is cheating to use AI technologies to complete homework assignments. Others might see it as ethical because it is simply using a tool to help complete the assignment. Ultimately, the decision of whether or not to use AI technologies for homework assignments is up to the individual.

Example 2. davinci-003 output that appears to take a moral position.

form a coherent and meaningful statement. Within the GPT-3 group of natural language processing (NLP) models, davinci-003 was released in late November 2022. This was followed by ChatGPT in December 2022. OpenAI's website features a free 'playground' web application which open allows a user to enter a short prompt from which davinci-003 will generate additional text. The 'playground' web application features various parameters a user can adjust such as the maximum length of the output or how random it is. Conversely, ChatGPT is a chatbot that seeks to converse with the user via text. Both technologies can produce text output but davinci-003 was principally used in this study.

As an example of the sophistication of the davinci-003 output, example 1 shows the output for a question demanding literary analysis of William Shakespeare. Note here how, whilst simple, this prompt still demands recognizing Shakespeare as a name of a person whilst Summer and Winter are recognized as names of seasons. Here, Summer is associated more closely to happiness and Winter is associated closer to darkness. Darkness is not an antonym of happiness.

Whilst these kind of NLP models have been around for some time, davinci-003 provides output which demonstrates the critical understanding and reason required to create excellent answers to essay questions. In example 2, davinci-003 appears to be considering the moral implications of using AI technologies to generate essays. This is despite how the text is generated based on learned patterns and structures.

(1) Is Physics based on facts that follow from observations?
(2) What was the most important advance in natural philosophy between 1100 and 1400?
(3) How did natural philosophers' understanding of electricity change during the 18th and 19th centuries?
(4) Does Kuhn or Popper give a more accurate description of physics?
(5) Was there a scientific revolution in 17th-century Europe?

Figure 1. The five questions used to generate submissions. Note a variety of historical, philosophical and factual themes are covered over the five questions.

(1) Is there a high academic content, at a suitably advanced level, thus indicating that the student is familiar with some of the key milestones in the history of physics, the philosophy of physics, science communication or ethics in academia?
(2) Has the student formed an appreciation of the physics underlying a particular topic?
(3) Does the student demonstrate a thorough grasp of the subject material?
(4) How well does the student address the specific question asked?
(5) Is the work written in a suitably authoritative, academic style, with the material presented in a logical, coherent and concise manner and supported by appropriate factual information?

Figure 2. The Physics in Society assessment proforma used to grade submissions.

2. Method

2.1. Outline of the module and exam

Physics in Society is a module offered by the Department of Physics at Durham University. The module content consists of the history and philosophy of Physics, the development of modern science and the ethical issues surrounding the use of technology in society. The major assessed component of the module is the exam which consists of a set of five short-form essay questions, each of no more than 300 words, on a selection of topics from the module. Figure 1 shows the five questions used for the AI submission generation. These questions serve to test a variety of aspects such as the history, philosophy, communication and ethics of Physics. This is reflected in the module assessment proforma, shown in figure 2, that outlines the five key criteria against which essays are marked.

Students are awarded a mark between 0 and 100 for each of the five categories specified in the module assessment proforma which are shown in figure 2; each of these categories is equally weighted when determining the students' final mark. The mark awarded for each category is based on the answers for all five questions; hence the marks must be balanced and reflect a general level of consistency within the answers. Students typically score 71 ± 5 on the Physics in Society module [11]. All markers were aware that they were marking AI generated scripts and marked the scripts solely for the purpose of this study.

2.2. Generating the AI scipts

A sample of $n = 10$ AI generated scripts, each containing five question-answer pairs, was compiled from a set of outputs of davinci-003, generated with suitably-chosen prompts based on the questions given in figure 1 entered into the OpenAI 'playground' web application. Repeatedly inputting prompts consisting of direct quotes the questions can lead to GPT-3 responding briefly and laconically. More discursive, original responses can be generated by slightly rephrasing prompts. For example the prompt, *'Was there a scientific revolution in 17th-century Europe?'* could be re-phrased as *'Did 17th-century Europe experience a scientific revolution?'*, or *'Was there a scientific revolution in seventeenth-century Europe?'*, which will generate more original responses. Rephrasing can also be combined with prefixing the prompt with phrases such as *'Write a long and detailed essay on'*, *'In 300 words or more, comment on'* or *'Explain in detail using more than 250 words'*. Such prompts are effective at instructing the AI to consistently generate more discursive responses. Here, the stated number of words used can be changed in tandem with the maximum output setting on the 'playground' to induce further novelty. Additionally, one can easily instruct GPT-3 to argue for (against) a certain proposition, by choosing a suitable prompts such as, *'Explain why there was (not) a scientific revolution in 17th-century Europe'*. This can be exploited effectively in cases where there are many viable answers to pick, but GPT-3 often favours a particular one. For example, the prompt *'What was the most important advance in natural philosophy between 1100 and 1400?'* typically generates a response from GPT-3 indicating that the scientific method was the most important advance. However, if the user explicitly states a relevant historical figure—such as Aquinas, Bacon or Buridan—within the prompt, as with the following example prompt; '*Explain why Thomas Aquinas' work was the most* ...' then GPT-3 will now provide a response that is consistent with the proposition specified by the user.

To obtain a consistent, minimal and fair benchmark of the quality of the davinci-003 essays, the output was not edited in any way. Only if the output was excessively similar to a prior generated one would it be rejected entirely, and a new one generated. It is reasonable to assume that a student who wished to use davinci-003 to write their answers would apply some minor edits, yet even without this the AI answers were of excellent quality. An example of davinci-003 output is presented in example 3.

3. Analysis and results

3.1. Overview

The ten AI generated submissions were marked independently by five different markers and achieved an average mark of $71 \pm 2\%$. This average is a high enough to be awarded a First Class grade; the highest available at UK universities. This compares to the average of $71 \pm 5\%$ in Physics in Society and an average score of $72 \pm 3\%$ for all second year Physics modules at Durham University [11]. These results strongly imply that davinci-003 and ChatGPT can write short-form Physics essays at a level of competency that is comparable to second year Physics students.

A significant trait of the AI generated answers were their consistent quality, as seen in the tight distribution of AI scores compared with student scores for the module. Figure 3 reveals that the very top performing students are able to outscore the AI essays, but the majority of the cohort performs either comparably or worse. Our results imply that students in the bottom third of their cohort would have a significant incentive to submit AI generated work for their assignments instead of their own original work.

Plagiarism detection tools cannot be relied upon to indicate the presence of AI generated work. This is shown in table 1, the AI essays score $2 \pm 1\%$ plagiarism on Grammarly and $7 \pm 2\%$ on TurnitIn. Both of these scores would be deemed sufficiently original if submitted in an actual university assessment; the only text highlighted as non-original were the questions given in figure 1.

In this study, the markers were aware that they were assessing AI generated scripts. However, prior research indicates that humans struggle to identify AI generated text consistently [12] thus specialized detection software is needed to

Prompt: Write a 300 word essay explaining why physics is not based on facts that follow from observations.

Output: Physics is not based purely on facts that follow from observations because some aspects of physics go beyond the reach of observation. At the root of this is the fact that physics is a field of study that attempts to uncover the fundamental workings of the universe. Physics strives to answer questions about how things work on the most basic level–questions about the behavior of particles and forces, for example. It is impossible to directly observe most of these phenomena; when an experiment is conducted, the results are generally observed indirectly through the measurements of related quantities. Therefore, some aspects of physics are based on theoretical models that do not rely directly on observable facts, but instead on the assumptions contained in the models.

Another reason why physics is not based on facts that follow from observations is that observations are always open to interpretation based on the observer's perspective. In general, different observers will have different interpretations of the same observations. This can lead to disagreement and debate over the interpretation of observed phenomena, which is not conducive to building a consensus on the truth of physical laws or theories.

Finally, even when multiple observers agree on the interpretation of an observation, there is no guarantee that the conclusion is correct or that it can be generalized to other physical phenomena. In other words, observations are often limited in terms of the conclusions that can be drawn from them. This means that while they are an important part of physics, they do not form the sole basis of the field.

In conclusion, physics is not based solely on facts that follow from observations because some aspects of physics go beyond the reach of direct observation. Additionally, observations are open to interpretation and are often limited in terms of the conclusions that can be drawn from them. Thus, while observations are an important part of the physics, they do not form the sole basis for the field.

Example 3. Unedited 328-word output from davinci-003 using a temperature (randomness) of 0.95 and a maximum output length of 700 tokens where one token is roughly four characters of English text.

identify AI generated text. There are some nascent technologies which claim to be able to detect AI generated work such as OpenAI's own AI text detection software [13] and GPTZero [14]. One of the key metrics that these technologies look for is the perplexity of the text. Intuitively, this is a measurement of uncertainty associated with the next words in a text passage given the prior words.

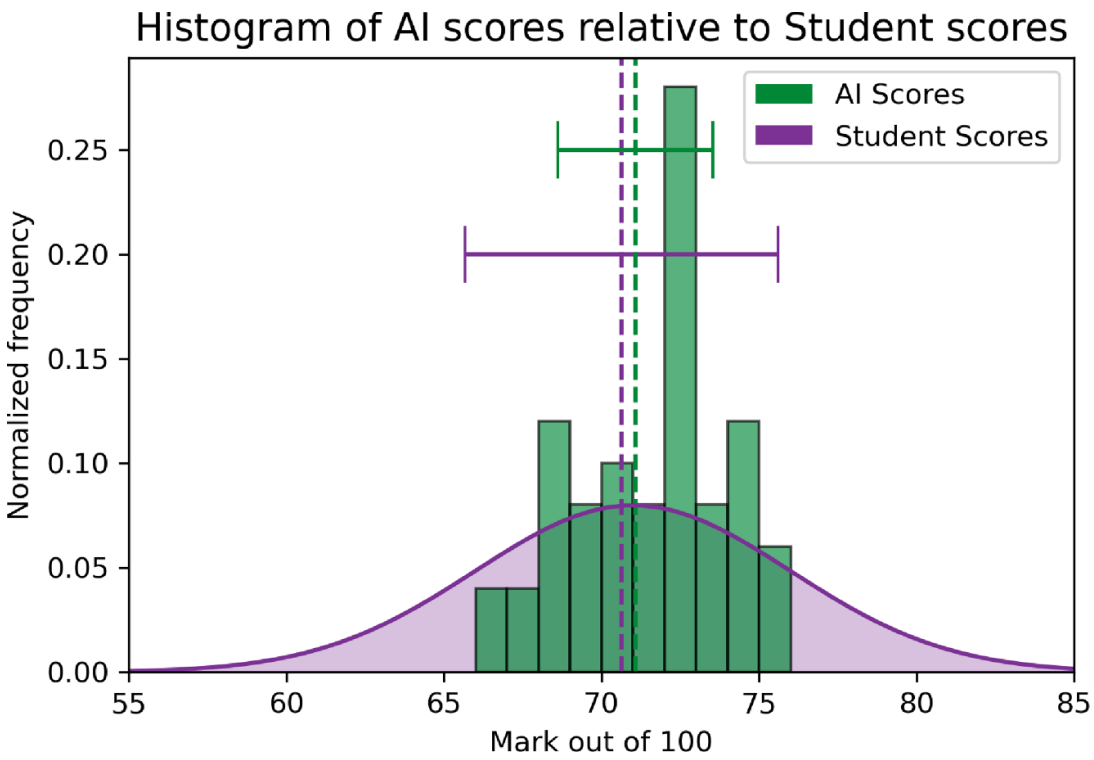

Figure 3. Histogram of AI scores (green) against a Gaussian distribution reflecting the average scores of students in Physics in Society [11].

Table 1. Table showing plagiarism percentage identified through both Grammarly and TurnitIn for each of the ten submissions.

Submission	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	Average
TurnitIn (%)	9	8	7	7	6	6	7	9	5	10	$7 \pm 2\%$
Grammarly (%)	1	2	1	0	2	3	1	2	1	2	$2 \pm 1\%$

The deterministic nature of much NLP software leads it to be more predictable than human-written text.

For the essays generated in this study, OpenAI's software characterized eight of the ten essays to be *'Very unlikely to be written by AI'* and the other two to be *'Likely to be written by AI'*. Whereas GPTZero found that nine of the ten essays *'May include parts written by AI'* whereas one was *'Likely to be entirely written by AI'*. Given that all the essays were in fact AI generated, this suggest that detecting AI written text is possible although GPTZero appears superior to OpenAI's software at the time of writing. Interestingly GPTZero and OpenAI's software disagreed on which essays were likely to be written by AI suggesting different approaches were taken. Nevertheless these results do suggest that in principle it is possible for software to identify AI generated text.

Figure 4 depicts how consistent the independent markers were with their scoring, whereby they awarded 73.0 ± 1.6, 72.6 ± 2.0, 69 ± 2, 70 ± 2 and 70.6 ± 1.9. These scores are in strong agreement with the module average of 71 ± 5 and—combined with passing the plagiarism check—suggest it would be challenging to identify the AI essays from those of students. Furthermore, the average scores awarded by each marker are all in reasonable agreement with one

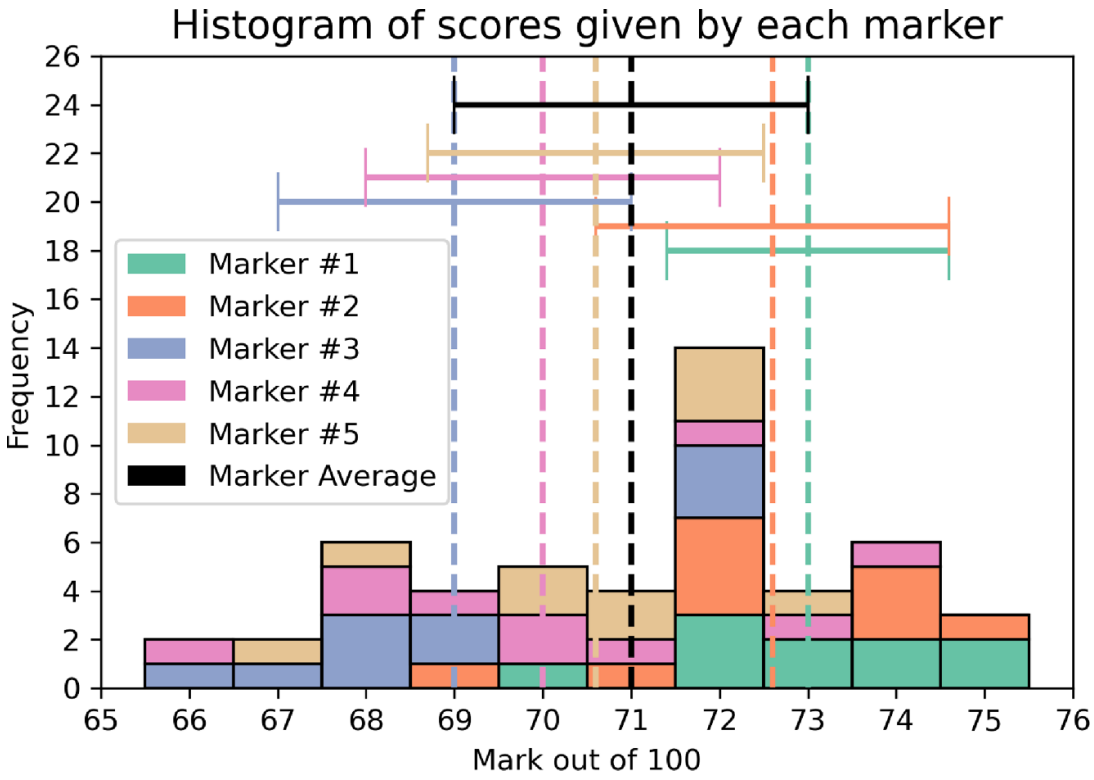

Figure 4. Stacked histogram of the independent marker's scores for the ten AI generated essays. The average and standard deviation for each marker is shown by the dashed line and error bars respectively. The overall average of the independent markers of 71 ± 2 is shown in black. Note the overlap between the different independent marker's scores.

another, indicating that the marking process is valid.

3.2. Written quality

Looking closely at example 3 reveals how the readability of certain phrases could be improved. As an illustration the sentence '*Physics strives to answer questions about how things work on the most basic level questions about the behavior of particles and forces, for example. It is impossible…*' could be rearranged to not finish with '*for example*'. Similarly, '*level–questions*' should be changed to '*level questions*'. In other essays often American English spelling was used—this would also require change for a UK context. These slight flaws show that the raw output is not necessarily semantically perfect. However, as prior mentioned, it is reasonable to assume that a student would apply at least some editing to the text if they wanted to use davinci-003 to write their submissions for them. Thus this semantic imperfection does not stop the use of davinci-003 as an essay writing tool.

4. Discussion & Conclusion

4.1. Impact on higher education

The results of this short study suggest modern AI systems such as davinci-003 and ChatGPT are capable of writing high quality Physics essays, capable of achieving a first-class score, for free and within a few seconds. These tools essentially represent a wide spread democratization of paid essay writing services whereby anyone can have davinci-003 or ChatGPT write an essay for them. We could argue that, in contrast to merely sounding a death-knell for certain kinds of assessments, this technology will once again force us to re-think assessment, and will confer far greater benefits in

the end than the comparatively small drawback of having to redesign existing assessments. Obviating the abuse of these technologies to undermine assessments could be as simple as requiring the students to produce the work within an invigilated setting. In such cases, the moral of the story is that while the threat to authentic assessment is profound, simple and practical solutions likely already exist that will not necessarily require a total upheaval of current assessments themselves, but may force a change of context at least.

Besides assessment, this technology allows users to generate innumerable, original examples, which can be shown by teachers within writing workshops. Similarly, as technologies such as ChatGPT mature we could see their use as one-on-one tutors. Although, the availability of knowledge is effectively already widely democratized. The full course contents for most degrees at Stanford and The Massachusetts Institute of Technology have been available on their respective Massive Online Open Courses (MOOC) platforms for a decade. Yet enrolling as a student at both institutions is still very competitive and a typical online MOOC can expect a completion rate of 10% [15] which is far below that of a typical university degree. Thus it is not often clear in advance how new technology will impact higher education.

4.2. Ease of sample generation

Getting davinci-003 or ChatGPT to respond in the desired manner often involves a process of trial-and-error, retrying or rewording prompts. Yet a transformational aspect of these technologies is how few times this needs to be done and how quickly the samples are generated. As covered in section 2.2 it is relatively straightforward to get a good quality answer to an essay question simply by preceding the question with *'Write a 300 word essay on'*. However, whilst 50 AI answers were created, it is unclear whether additional unique responses could realistically be created without some familiarity with the subject. Take for example the question *'What was the most important advance in natural philosophy between 1100 and 1400?'*, as covered in section 2.2 it was possible to create multiple unique answers through rephrasing the question with a sensible suggestion. Rephrasing the question as *'Explain why X was the most important advance in natural philosophy between 1100 and 1400.'* where *X* is Bacon's work on optics / Buridan's work on impetus / Wallingford's astronomical clock or similiar allows for many unique answers to be created. However, this is contingent on knowing who Bacon / Buridan / Wallingford are in the first place. Yet simply knowing Roger Bacon worked on optics is not equivalent to being able to write a high quality essay on his work. Even if the phase-space of AI answers for questions could be checked by assessors beforehand, through partially answering the question and incorporating AI as a co-pilot students could still have AI write their answers but with a reduced risk of being caught. Secondly, as with paid essay writing services, without clear conclusive proof it is unreasonable to accuse a submission of not have being written by its stated author.

4.3. Limitations and future work

In the present work, we have adopted a fairly rudimentary proxy, consisting of the average scores for the module, in order to contextualise the AI scores. More properly, we should compare the AI scores to those of the students for this exam. We aim to address this in a forthcoming paper in which we compare the performance of AI to student submissions, during which we also address the question of the extent to which markers can reliably distinguish between AI submissions and human submissions. Nevertheless, our key result, that AI can generate content that attains highly, in absolute terms, for short-form essays, is very significant in and of itself, despite the aforementioned limitations.

The focus of the present work are assignments consisting of short-form essay questions, which represent a minority of assessments within Science, Technology, Engineering, and Maths (STEM) teaching. However, future work will consider the capabilities of NLP AI with regard to scientific report writing and answering analytical questions which require the use of calculations, coding, symbolic manipulation and algebraic typesetting, which are staples of STEM assignments and examinations. Although, this is beyond the scope of the present work.

It is important to note that programs such as davinci-003 and ChatGPT are ultimately text

generation software. Therefore it may struggle with Physics questions that require complex numerical or graphical components in their answers. Example 4 is somewhat straightforward with the question at a level typical of a A-Level or a first year undergraduate Physics course. However, ChatGPT answers this question incorrectly stating the amplitude as A rather than $\frac{A}{2}(1-\cos(1))$ and the period to be $\frac{2\pi}{\omega}$ rather than $\frac{\pi}{\omega}$. Currently NLP programs still largely work by generating the most likely upcoming word. Therefore an intuition of what is happening here is that as out of all the times a non-harmonic oscillator is mentioned in the corpus of training materials used in ChatGPT the amplitude is typically A and the period is typically $\frac{2\pi}{\omega}$, for this question on non-harmonic oscillators A and $\frac{2\pi}{\omega}$ were the answers given in spite of the fact they are mathematically incorrect. Further, note how the Python code in example 4 for the plot is actually correct. This can be interpreted with this same intuition except that in this case the plot is typical of the training corpus thus the correct answer is given.

Whilst at the time of writing, ChatGPT and davinci-003 still struggle with more complex Physics questions suggesting it may be poor at answering Masters or PhD level questions. It is our view that due to the rapid progress and increasing interest in this area it is only a matter of time before these technologies develop to the point where they can not only pass, but excel in all of these aspects of STEM assignments too. The implication of this is that it is important educators simply do not resort to 'moving the goalposts' when adapting to their prevalence as such a strategy may only work for a short time before the next step change in AI capabilities.

4.4. AI generated feedback for AI essay

Given the ability of AI to create high quality essays the next logical step is to get an AI to mark the AI essays. ChatGPT is better suited to this task than davinci-003 and the output of a ChatGPT conversation is shown in example 5. Whilst when scoring itself the AI is a little generous—awarding 85/100—it does provide specific feedback comments that directly relate to both the input essay and the rubric.

This concept of AI marking is a non-trivial point. Automatic essay scoring (AES) is a mature technology in regular use for US High School essays [16, 17]. These AES technologies are trained on many examples of student essays and assign a grade based on various factors such as grammar, spelling, word choice, organization, and content. This is in contrast to a university module like Physics in Society that features questions relatively unique to the content taught and where a cohort typically ranges from 50 to 70 students so there are limited examples for comparison. Despite this, the score and feedback given in example 5 is clearly cogent and would likely pass undetected if given to the student as human-written feedback.

4.5. Concluding thoughts

We may be at the beginning of an AI revolution. In order to facilitate authentic assessment it is vital that we are aware of the capabilities of this technology and its ramifications on the way that credited work is assessed. In the present case, it is hard to avoid the conclusion that non-invigilated assessments based on short-form essay questions are already no longer fit for purpose; they are simply too vulnerable to current AI text-completion technologies, which can produce creditable content cheaply and quickly. We should note that this scenario, where an emerging technology threatens to undermine assessments, is not without some comparable historical counterparts. It was initially imagined that the advent of the internet posed an existential threat to certain kinds of assessments. However, what has emerged since then is a total revolution in the way that we teach, stemming from careful, evidence-based consideration of how we can use this platform to enhance the quality of our teaching and to improve the student experience. It would be hard to imagine delivering a modern course without using the internet in some way. On the basis of startling and exciting recent advances in the capabilities of AI technologies, notwithstanding the results of the present work, we would argue that AI is well-placed to instigate dramatic and profound changes to the way that we teach and assess students, forming an indispensable component of a new ethos within which we design and deliver teaching.

Prompt: What is the amplitude of the non-harmonic oscillation described by the function $x(t) = A\cos[\cos(\omega t)]$? Calculate the period of this oscillation.

Output: The amplitude of this non-harmonic oscillation can be calculated as follows: $A = max(x(t)) - min(x(t)) = max(A\cos[\cos(\omega t)]) - min(A\cos[\cos(\omega t)]) = A$.

So the amplitude is A. The period of this oscillation can be calculated as follows: $T = 2\pi/\omega$

However, since the frequency of the oscillation is not constant, it is not possible to find a single period for this function. The period depends on the value of ω and changes over time.

Prompt: Could you write some Python code to plot a graph of this function using matplotlib?

Output: Certainly! Here's some code to plot the function $x(t) = A\cos[\cos(\omega t)]$ using matplotlib in Python:

```
import numpy as np
import matplotlib.pyplot as plt

A = 1
ω = 2
t = np.linspace(0, 10, 1000)
x = A * np.cos(np.cos(ω * t))

plt.plot(t, x)
plt.xlabel('t')
plt.ylabel('x(t)')
plt.title('Graph of x(t) = A cos[cos(ω t)]')
plt.show()
```

Example 4. ChatGPT generated (incorrect) answer for a Physics question requiring both mathematical and graphical elements. The text has been lightly edited to replace the literal characters ω and π with \omega and \pi for illustration purposes.

Prompt: Using the following rubric: (*inserted rubric shown in Figure 2*). What score, out of 100, would you give the following essay (*inserted essay shown in Example 3*).

Output: Based on the provided rubric, I would give the essay a score of 85 out of 100.

The essay demonstrates a high level of academic content and an appreciation of the physics underlying the topic. The student thoroughly grasps the subject matter and addresses the specific question asked in a clear and concise manner. The essay is written in an academic style and presents the material in a logical, coherent manner, supported by appropriate factual information.

One potential area for improvement would be to further discuss the limitations of observations and their impact on the development of physical theories. The essay touches on this, but it could be further developed to provide a more in-depth analysis of the role of observations in physics. Additionally, citing sources or referencing relevant literature could also strengthen the essay and add credibility to the arguments presented. Overall, the essay is well-written and provides a thorough analysis of the topic.

Example 5. ChatGPT generated feedback for the AI essay in example 3.

Within the coming months, this area of research will likely undergo an explosion. Indeed, at the time of writing, we are on the cusp of the release of GPT-4. It is the view of the authors that within the next twelve months, the more pioneering students or *'Mavens'* [18] will realise the potential of these tools, and this news will steadily spread around student cohorts. Within two years, most higher education practitioners will be aware of how powerful these tools are and may have unwittingly marked work created in part by AI in a co-pilot manner. The potential for this technology to produce extremely accurate, high-quality content to almost any brief imaginable will surely lead to a new stage in the development of education. In the meantime, we must prepare for the AI revolution that is already underway, to ensure the fidelity of current forms of assessment.

Data availability statement

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

ORCID iDs

Will Yeadon https://orcid.org/0000-0002-9444-108X
Oto-Obong Inyang https://orcid.org/0000-0002-9001-0418
Craig P Testrow https://orcid.org/0000-0001-5639-3753

Received 21 December 2022, in final form 14 February 2023
Accepted for publication 20 March 2023
https://doi.org/10.1088/1361-6552/acc5cf

References

[1] Clark E, August T, Serrano S, Haduong N, Gururangan S and Smith N A 2021 All that's 'human' is not gold: evaluating human evaluation of generated text *Proc. 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int. Joint Conf. on Natural Language Processing* (*Long Papers*) vol 1 (Association for Computational Linguistics) pp 7282–96

[2] Metz C 2020 Meet GPT-3. It has learned to code (and blog and argue) *The New York Times* (available at: www.nytimes.com/2020/11/24/science/artificial-intelligence-ai-gpt3.html)

[3] Daniel J E, Brink W, Eloff R and Copley C 2019 Towards automating healthcare question answering in a noisy multilingual low-resource setting *Proc. 57th Annual Meeting of the Association for Computational Linguistics* (Florence: Association for Computational Linguistics) pp 948–53

[4] Queudot M, Charton É and Meurs M-J 2020 Improving access to justice with legal chatbots *Stats* **3** 356–75

[5] Malik A, Wu M, Vasavada V, Song J, Coots M, Mitchell J, Goodman N and Piech C 2019 Generative grading: near human-level accuracy for automated feedback on richly structured problems (arXiv:1905.09916)

[6] Pillutla K, Swayamdipta S, Zellers R, Thickstun J, Welleck S, Choi Y and Harchaoui Z 2021 MAUVE: measuring the gap between neural text and human text using divergence frontiers *Advances in Neural Information Processing Systems* vol 34, ed M Ranzato, A Beygelzimer, Y Dauphin, P Liang and J W Vaughan (Curran Associates, Inc.) pp 4816–28

[7] Blin F and Munro M 2008 Why hasn't technology disrupted academics' teaching practices? understanding resistance to change through the lens of activity theory *Comput. Educ.* **50** 475–90

[8] Fyfe P 2022 How to cheat on your final paper: assigning AI for student writing *AI Soc.* 1–11

[9] Sharples M 2022 Automated essay writing: an AIED opinion *Int. J. Artif. Intell. Educ.* **32** 1119–26

[10] Brown T 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

[11] Hall T 2021 Undergraduate module averages 2020 - a freedom of information request to University of Durham (available at: www.whatdotheyknow.com/request/undergraduate_module_averages_20_25)

[12] Köbis N and Mossink L D 2021 Artificial intelligence versus maya angelou: experimental evidence that people cannot differentiate ai-generated from human-written poetry *Comput. Hum. Behav.* **114** 106553

[13] OpenAI 2023 AI text classifier (available at: https://beta.openai.com/ai-text-classifier)

[14] Tian E 2023 Gptzero (available at: https://gptzero.me/)

[15] Jordan K 2015 Massive open online course completion rates revisited: assessment, length and attrition *Int. Rev. Res. Open Distrib. Learn.* **16** 341–58

[16] Shermis M D 2014 State-of-the-art automated essay scoring: competition, results and future directions from a United States demonstration *Assess. Writ.* **20** 53–76

[17] Ramesh D and Sanampudi S K 2021 An automated essay scoring systems: a systematic literature review *Artif. Intell. Rev.* **55** 2495–527

[18] Gladwell M 2003 *The Tipping Point: How Little Things Can Make a Big Difference* (*Compass Series*) (New York: Wheeler Publishing Inc.)