Evaluating AI and human authorship quality in academic writing through physics essays

OPEN ACCESS
IOP Publishing
European Journal of Physics
Eur. J. Phys. **45** (2024) 055703 (17pp)
https://doi.org/10.1088/1361-6404/ad669d

Evaluating AI and human authorship quality in academic writing through physics essays

Will Yeadon, Elise Agra, Oto-Obong Inyang, Paul Mackay and Arin Mizouri

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

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

Received 15 April 2024, revised 17 June 2024
Accepted for publication 23 July 2024
Published 2 September 2024

Abstract

This study aims to compare the academic writing quality and detectability of authorship between human and AI-generated texts by evaluating $n = 300$ short-form physics essay submissions, equally divided between student work submitted before the introduction of ChatGPT and those generated by OpenAI's GPT-4. In blinded evaluations conducted by five independent markers who were unaware of the origin of the essays, we observed no statistically significant differences in scores between essays authored by humans and those produced by AI (p-value $= 0.107$, $\alpha = 0.05$). Additionally, when the markers subsequently attempted to identify the authorship of the essays on a 4-point Likert scale—from 'Definitely AI' to 'Definitely Human'—their performance was only marginally better than random chance. This outcome not only underscores the convergence of AI and human authorship quality but also highlights the difficulty of discerning AI-generated content solely through human judgment. Furthermore, the effectiveness of five commercially available software tools for identifying essay authorship was evaluated. Among these, ZeroGPT was the most accurate, achieving a 98% accuracy rate and a precision score of 1.0 when its classifications were reduced to binary outcomes. This result is a source of potential optimism for maintaining assessment integrity. Finally, we propose that texts with $\leqslant 50\%$ AI-generated content should be considered the upper limit for classification as human-authored, a boundary inclusive of a future with ubiquitous AI assistance whilst also respecting human-authorship.

Keywords: AI, academic writing, ChatGPT, benchmark

1. Introduction

1.1. Background

The year 2023 marked a pivotal moment in the integration of AI text-completion technologies within educational settings worldwide. Educators were confronted with the reality that students could use tools like ChatGPT to instantly complete assignments, sparking fears of academic dishonesty and the undermining of the educational process. In response to these challenges, some institutions ceased assigning traditional homework, opting instead for in-depth preparatory work that fosters discussion and assessment within the classroom [1].

The proficiency of modern Large Language Models (LLMs) in generating text across a broad spectrum of topics, from science [2] to finance [3], is well-documented. The advancements in these models are evident when comparing their current outputs to those from five years ago, showcasing a significant improvement in quality. This progression has prompted academics to investigate the capabilities of AI in composing essays on a wide range of subjects. Notably, studies have demonstrated that AI-written documents excel in tasks ranging from the analysis of general legal principles [4] to Old English Poetry, including intricate analyses of works like Beowulf [5].

In response to the rapid advancements in AI and their potential impact on academic integrity, this study investigates the quality of AI-authored compared to human-authored essays in an essay writing task for an accredited physics course at Durham University, UK. Employing a blinded assessment methodology, markers evaluated essays without knowledge of their origin to eliminate bias and focus on content quality and adherence to academic standards. Essays were selected from submissions prior to the widespread adoption of modern LLMs, ensuring a fair comparison by minimizing potential AI-generated content in the control group. Authorship was assigned post-evaluation to maintain the integrity of the assessment process. This approach allows for a critical examination of AI's capabilities in producing academically sound essays and an investigation into the broader implications for educational practices and assessment integrity. As LLMs evolve, understanding their benefits and limitations is crucial for educators, students, and academic institutions.

1.2. State of the art

There are many powerful LLMs available, and benchmarking them is a comprehensive field of study aimed at evaluating the capabilities of these models across a broad spectrum of cognitive and linguistic tasks [6]. As of this writing, OpenAI's GPT-4 consistently ranks at or near the top of most LLM benchmark leaderboards, scoring highly on prominent benchmarks such as the Massive Multitask Language Understanding (MMLU), according to its technical report [7].

The MMLU benchmark, known for its extensive range of subjects from humanities to hard sciences, provides a rigorous test of a model's ability to understand and generate responses across diverse knowledge areas. Similarly, the SuperGLUE benchmark [8], which focuses on tasks that require a deep understanding of language, such as question answering, inference, and reasoning, challenges models to demonstrate advanced levels of comprehension and the ability to handle nuanced linguistic constructs. GPT-4's performance, independently supported by scoring over 90% on SuperGLUE [9], underscores its capabilities.

Given this, our study uses GPT-4 as a representation of current LLM essay writing abilities. While in principle other LLMs could perform better, GPT-4's proficiency ensures that our test is a valid measure of the state of the art (SoTA) in AI-driven essay composition.

1.3. Detection of AI-authored text

Following the introduction of ChatGPT, researchers soon investigated the detectability of the latest generation of sophisticated AI-authored texts [10, 11]. Early on, it seemed that merely paraphrasing or editing AI-generated content could bypass existing detection technologies [12] and some initial detection tools were criticized for their bias against non-native English speakers [13]. In response to these challenges, one of the leading plagiarism detection software companies, Turnitin, introduced an AI content detection feature. This move, however, was met with skepticism by numerous UK universities [14] leading several institutions to forgo its use citing concerns around accuracy amongst other issues.

Despite the challenges with early implementations, advancements in AI content detection technology are steadily progressing, enhancing both accuracy and reliability. There are a few principal methods for detection. One method involves statistical-based techniques, such as evaluating the intrinsic dimensionality, which measures the complexity and structure inherent in the text, like the variety and similarity of words used. This can help identify statistical differences between human and AI-generated content [15]. Another method is looking for prior added watermarks within text. These watermarks take the form of subtle, detectable patterns or sequences of words and punctuation intentionally inserted by the LLMs' creators. Finally, neural-based detectors use other deep learning models trained on large datasets to differentiate between human and AI-generated texts to find patterns and features that distinguish the two [16].

While some of these methods show promise, the stakes for a plagiarism accusation within Higher Education are high, especially for work such as theses that constitute a significant portion of a student's final grade. Being found guilty of plagiarism can lead to expulsion from an institution. Therefore, purely computational methods, even with $>99\%$ accuracy, would likely not be sufficient on their own to find a student guilty. Human judgment would undoubtedly still play a crucial role in such determinations.

1.4. Aims

There is significant apprehension regarding the potential of LLMs to automatically complete Physics assessments, and they have already demonstrated some success in this area [17, 18]. Writing ability is an important skill to develop during a Physics degree, and written essays are a common form of assessment. Therefore, this work has three aims: to evaluate the quality of SoTA LLMs in answering essay-like questions in the context of Physics Education, to determine whether human and AI-authored work can be distinguished, and to provide guidance for the Physics Education community in how to adapt written assignments to generative AI. To ensure a fair comparison, essays will be evaluated blindly. To detect AI-authored work, we will first conduct a type of Turing test by categorizing the essays on a Likert scale. Subsequently, we will apply various purported computational detection methods.

2. Method

2.1. Overview

This investigation entailed the blind assessment of 60 PDF documents, each comprising of five short-form essays, by five independent markers. This structure resulted in a collective evaluation of 1500 separate short-form essays with an average length of 285 words, combined into $n = 300$ graded submissions (60 PDFs $\times$ five markers) for evaluation purposes. The documents were equally divided between human-authored texts and those produced by

Table 1. Evaluation criteria for the essays, taken exactly from the 'Physics is Society' module at Durham University.

Element	Evaluation criteria
1	Is there a high academic content, at a suitably advanced level, indicating familiarity with 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 material presented logically, coherently, and concisely, supported by appropriate factual information?

OpenAI's GPT-4 LLM. These essays are components of the 'Physics in Society' module facilitated by the Department of Physics at Durham University. The module's curriculum focuses on exploring the historical and philosophical dimensions of physics, including its evolution and the ethical implications surrounding the societal integration of its technological advancements. The module features a take-home essay assignment where students have 48 hours to respond to five questions in essays not exceeding 300 words each. Questions such as '*Is physics based on facts that follow from observations?*' and '*Is there a satisfactory interpretation of quantum mechanics?*' explore the history, philosophy, communication, and ethics of physics. The module's autumn semester provides formatively assessed questions to prepare students before they tackle a summatively assessed assignment on a new set of questions in the spring.

The study used three sets of student essays from a total of 20 unique authors. Specifically, 10 students from the 2021/22 cohort submitted two sets of five short-form essays each: one for the 2021/22 Formative assignment and one for the 2021/22 Summative assignment. Additionally, 10 different students from the 2022/23 cohort each submitted five short-form essays for the 2022/23 Formative assignment. This totals 30 submissions, each containing five short-form essays, from 20 student authors. In a typical cohort of Physics students at Durham University, approximately 85% are home students. Thus, we expect the vast majority of these submissions to have been written by native English speakers. Table A1 shows the 2021/22 formative assignment questions, Table A2 the 2021/22 summative questions, and Table A3 the 2022/23 formative questions. Although students submitted these essays as their own work, the presence of language models like GPT-2 since 2019 means authorship cannot be guaranteed. However, since ChatGPT was released on 30 November 2022, and mainstream adoption of AI chatbots began in 2023, essays submitted before this date are assumed to be student-authored. Essays from the 2022/23 Summative and 2023/24 Formative assignments were excluded from this study due to potential authorship ambiguity.

The evaluators assessed the essays based on their effectiveness in addressing five key elements, as outlined in table 1. Each submission, consisting of five essays, was marked holistically as a group, allowing students to strategically emphasize the aspects most relevant to the questions posed. For example, the question '*How did natural philosophers understanding of electricity change during the 18th and 19th centuries?*' encourages detailed discussions of specific physics concepts. This method mirrors the exact assessment process employed in the actual 'Physics in Society' module at Durham University; all evaluators in this study had prior experience working within the module.

Grading followed the standard UK university criteria on a scale out of 100, with scores of 70% and above qualifying for First-Class Honours, reflecting exceptional comprehension and skill. Next is Upper Second-Class Honours (2:1) for scores between 60% and 69%, considered a very good standard and often the minimum requirement for graduate positions and postgraduate study in the UK job market. This is followed by Lower Second-Class Honours (2:2) for 50% to 59%, Third-Class Honours (3rd) for 40% to 49%, and a Fail for marks below 40%. Approximately 30% of UK students achieve a First-Class Honours degree, a figure that varies by institution and subject but underscores the high standard of achievement these grades represent [19]. To ensure the integrity of degree classifications, UK universities typically require that marks do not vary significantly from year to year, resulting in overall averages that often hover around 65%.

Each element of the essays was graded on a scale from 0 to 100, in 5-point increments. Performance across these elements was categorized into seven distinct levels, and the average of these scores was calculated to give a final score out of 100. Scores from 80 to 100 indicated 'Exemplary' performance, showcasing exceptional insight and mastery beyond standard expectations. 'Excellent' scores ranged from 70 to 75, reflecting superior understanding and application, albeit not at the exemplary level. 'Good' (60–65) denoted solid competence and satisfactory execution, while 'Sound' (50–55) represented basic adequacy with some weaknesses. 'Acceptable' scores (40–45) signified marginal performance that just met minimum criteria, and 'Insufficient' (30–35) indicated notable deficiencies and a lack of basic understanding. The lowest category, 'Unacceptable' (0–25), marked failure to meet fundamental requirements, showing profound inadequacies in knowledge or execution. This grading framework provided a structured approach to evaluating essays, clearly delineating between varying levels of academic achievement.

In addition to marking the submissions for content quality and relevance, evaluators were tasked with assigning authorship to each essay using a four-point Likert scale. The options on the scale were 'Definitely AI', 'Probably AI', 'Probably Human', and 'Definitely Human'. This step of authorship assignment was deliberately conducted after the essays had been marked to prevent any potential bias in scoring based on the presumed origin of the text. The outcomes of these human evaluations are intended for comparison with results from various computational techniques designed for AI text detection. This comparative analysis aims to assess the efficacy of human judgement against automated methods in distinguishing between AI-generated and human-written texts.

2.2. AI essay generation

Given the performance of the LLM branded 'GPT-4' within the ChatGPT web app has been shown to change over time [20], AI-generated essays were produced use OpenAI's API with the gpt-4-1106-preview model. This approach ensured replicability by utilizing a specific version of the GPT-4 family and allowed for efficient essay generation through a Python script rather than via the ChatGPT web app. Each question from the assignments underwent paraphrasing 10 times, incorporating a push for novelty. For instance, the question '*Was there a scientific revolution in 17th-century Europe?*' was transformed into '*In 250 words, analyze whether the 17th-century European developments, such as the Copernican model and Galileo's telescope observations, truly signify a scientific revolution*'. Similarly, '*Is there a satisfactory interpretation of quantum mechanics?*' was reworded to '*Examine in 242 words the de Broglie/Bohm theory as an alternative to mainstream quantum mechanics interpretations and its approach to wave-particle duality*'. The specific suggestions for novelty in the prompts were sourced from the module syllabus. For example, the de Broglie/Bohm

theory, the Copernican model, and Galileo's telescope observations are all covered as part of the course. This approach yielded 50 unique paraphrased prompts for each of the three assignments, resulting in a comprehensive collection of 150 prompts, detailed in the supplementary material[1]. The word count specified in the prompts often varied around 245, as this typically yielded essays close to 300 words in length.

The 150 AI-authored essays, generated from prompts, had a mean word count of 286.68 (SD = 26.44), comparable to the human-authored essays, which had a mean of 283.69 (SD = 13.51). A t-test indicated no significant difference in essay length between AI-generated and human-written essays (p-value = 0.219), suggesting that distinctions in content quality and complexity are the primary differentiators. Upon submitting the AI essays to Turnitin, the average plagiarism levels detected were 6% for the 2021/22 Formative assessment, 0% for the 2021/22 Summative assessment, and 1% for the 2022/23 Formative assessment. This result shows that the AI-authored essays are novel.

3. Analysis and results

3.1. Overview

The combined scores from five markers for sixty submissions resulted in a dataset of $n = 300$. Human-authored essays had an average score of 66.9 with a standard error of 0.5, while AI-authored essays averaged 65.7, with a standard error of 0.5. Here, the standard error is found by dividing the respective standard deviation by the square root of their sample sizes of $n = 150$ for both AI and human essays. A t-test revealed a p-value of 0.107, indicating no statistically significant difference between the scores at an α level of 0.05. A histogram of the scores, as illustrated in figure 1, shows similar distributions for both sets of essays. These findings suggest that AI and human authors are reaching parity in writing short-form physics essays, where assignments can be completed by ChatGPT within seconds and achieve scores comparable to those of human authors. For a detailed breakdown of the scores awarded by each marker and an analysis of the grading consistency among markers, please refer to appendix B.

The decision to submit AI-generated work without knowing its potential score poses a risk, particularly for stronger students expecting high grades. In contrast, weaker students might find using AI advantageous, as the score distribution for AI-authored essays suggests that choosing a random essay from those created by the prompts detailed in section 2.2 could result in a first or upper second-class grade 81.43% of the time, according to the cumulative probability from a simple z-score ($\mu = 65.73$, $\sigma = 6.41$) calculation with a value of 60—the boundary for a 2:1. These results echo those of Ghassemi *et al* [21], who found that below-average performers gained more from using AI than above-average performers. The widespread availability of powerful language models is thus leveling the playing field in physics essay writing, offering students the opportunity to secure solid grades with minimal effort.

After evaluating the essays, markers were asked to classify the authorship on a 4-point Likert scale, with 1 representing 'Definitely human' and 4 as 'Definitely AI'. The results, depicted in figure 2, showed a relatively constant amount of AI-authored texts across each category, while the proportion of human-authored text decreased from 'Definitely human' to 'Definitely AI'. This suggests a slight bias towards markers identifying essays as human-authored. Moreover, their confidence in the authorship was generally proportional to their accuracy. A CochranArmitage test for trend revealed a negative trend statistic (−0.093) for

[1] Also available at https://github.com/WillYeadon/AI-Exam-Completion.

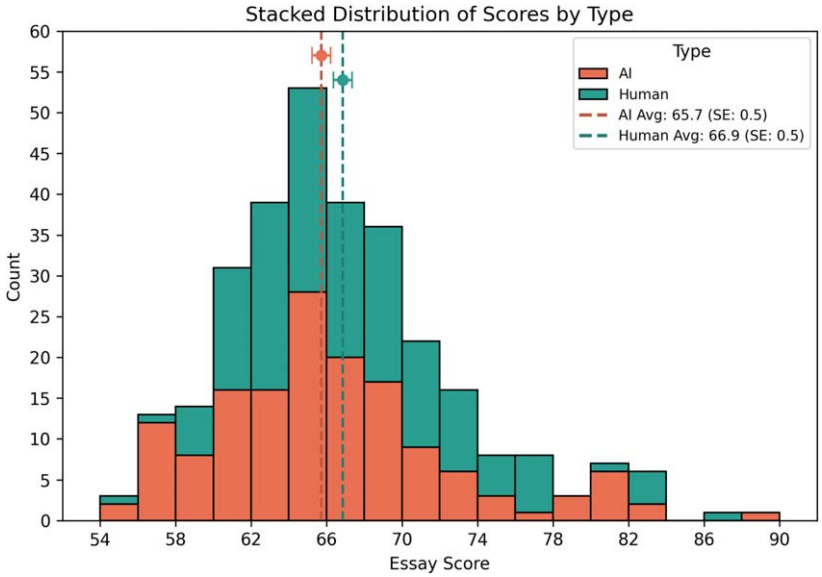

Figure 1. Histograms of scores for AI and Human authored essays for all markers combined totalling $n = 300$ data points. The distributions look visually similar and a t-test ($SD_{AI} = 6.41$, $SD_{H} = 5.71$) reveals they are statistically indistinguishable.

human-authored texts with a p-value of 0.027. This p-value is just below our $\alpha = 0.05$ indicating a mild at-best trend in markers being more confident in AI authorship as the proportion of human-authored texts decreases.

To better understand how markers can distinguish between essays written by AI and humans, we simplified their responses on the Likert scale to a straightforward yes or no decision, combining the 'Definitely' and 'Probably' categories. The accuracy rates recorded for the five markers—68%, 67%, 63%, 57% and 57% - were only marginally better than a default 50% success rate[2]. Interestingly, markers were more likely to flag essays as AI-generated if they contained unusual features, such as the presence of em dashes (—), or if the essays were structured as a list of numbered points. Despite these indicators, the relatively low discernment success underscores the limitations of relying solely on human judgment to detect AI-generated work. Given the serious consequences of wrongly accusing a student of academic dishonesty, there is a pressing need for more reliable methods of authorship verification, such as employing lexical analysis or specialized AI detection tools.

Looking further at two examples of AI-authored work, we focus on the one most often correctly classified when combining all five markers and the one most often misclassified. Figure 3 shows a revealing snippet of the AI-authored work most accurately classified as AI —a submission to the 2021/22 Summative assignment. Here, the AI answers the essay question using both lists and an apparent attempt at HTML formatting with bolded item heads. This is an example of an AI chatbot failure mode, whereby a response more akin to an essay plan was produced rather than a contiguous essay. In practice, a human intent on using AI to write an essay would likely notice this and ask for another iteration. However, complete AI-only work or careless copy-pasting without checking can clearly feature these kinds of

[2] Assuming a simplistic strategy where assigning all essays to a single category would result in 50% accuracy, given an even split between AI and human-authored essays.

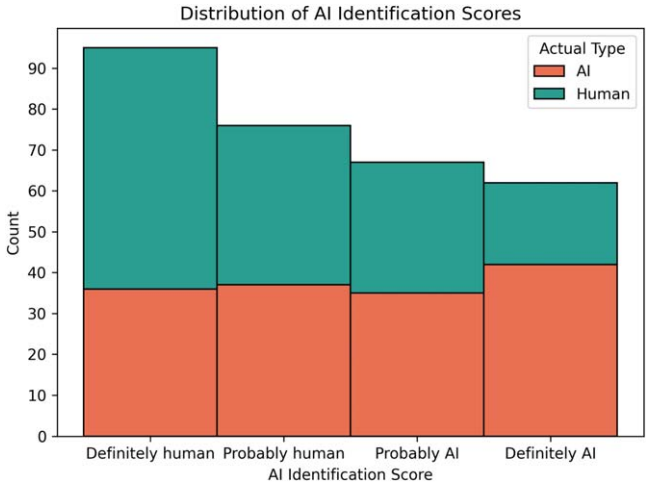

Figure 2. Histogram showing the markers' assigned authorship versus actual authorship of the 300 assessed submissions. The amount of AI-authored texts is roughly constant in each category whereas the percentage of human-authored texts goes from 62% 'Definitely human', 51% in 'Probably human' then 48% in 'Probably AI' and 33% in 'Definitely AI'.

Question: If most theories have been shown to be false, do we have any reason to have confidence in our theories?

Answer snippet: Falsification contributes to the survival of the fittest theories in several ways: 1. **Rigorous Testing**: Falsification encourages rigorous testing of theories. By continually subjecting theories to potential refutation, only those that withstand repeated attempts at falsification survive. This process eliminates weaker theories that cannot account for observed phenomena or that make incorrect predictions. 2. **Progressive Refinement**: As theories are tested and falsified, the surviving theories often undergo refinement. This leads to more precise and robust theories that have a greater explanatory power. Over time, science thus advances by building on these increasingly fit theories.

Figure 3. A revealing snippet of the AI-authored submission most often correctly classified as AI-authored by the markers. Note the list style response and attempt at HTML formatting with the ** to make sections bold.

mistakes. Within the submission, this answer was the only one featuring the list-style response, yet it was sufficient to alert the markers.

Comparing this to a snippet of the AI-authored document most often misclassified as human, from the 2021/22 Formative assignment, shown in figure 4. An interesting feature here is the frequent use of 'we' within the passage, which may inadvertently cause readers to anthropomorphize the AI, assuming it to be human. This effect is even observed when humans know they are interacting with an AI, as highlighted by Thelot: 'People can and will assign sentience to things whether we prove or disprove their sentience scientifically' [22].

Question: Is physics based on facts that follow from observations?

Answer snippet: It would be naive to strictly say that physics is only based on facts that've followed from observations, otherwise, we would limit ourselves to the possibilities to which only what we see. The term "observed" is itself ambiguous, this implies we have only "seen" such phenomena. There have been multitudes of discoveries found from either: inferring data, understanding behaviours of specific systems or by exploring the links of known science in hopes to understand others. An example of that being gravitational waves. Gravitational waves always made theoretical sense, yet couldn't be measured.

Figure 4. A snippet from the AI-authored essay most often classified as Human-authored by the markers. Note the anthropomorphic phrases such 'we see'.

Beyond the use of 'we', the passage is clear, cogent, and scientific enough to be classified as human by professional university staff. All AI-authored essays used in this study are available at[3].

Nominally, the marking and authorship assignments were independent tasks; however, the markers were aware that they were possibly evaluating AI-generated work. This awareness introduces a risk of subconscious or conscious bias either for or against AI-generated work. To test for this bias, we converted the Likert scale into a numerical one, with 'Definitely human' =1 and 'Definitely AI' =4, and then calculated the Pearson correlation coefficient between each marker's assigned score and their assigned authorship. Marker #1 had a correlation of 0.1298, indicating a slight tendency to mark essays higher if they believed them to be AI-authored. Conversely, Marker #2 had a correlation of −0.3840, suggesting they scored essays lower if they believed them to be AI-authored. Similarly, Marker #3 showed a correlation of −0.4308, and Marker #4 showed a moderate negative correlation of −0.6763, the strongest among all markers. Marker #5 had a negligible correlation of 0.0961. Combining the results from all markers yielded an overall correlation of −0.3980. This indicates a weak-to-moderate negative correlation, suggesting a potential bias against AI-authored work on average. Although the average correlation was not particularly strong, these results highlight the possibility of bias in the marking process. This finding suggests an avenue for further study, particularly focusing on the psychological impact of AI on evaluators.

3.2. Lexical analysis

An examination of the essays' lexical characteristics revealed that human-authored essays had an average of 154.96 unique words (SD = 12.77), while AI-authored essays featured 159.47 unique words on average (SD = 15.10). Additionally, the average word length was 5.31 (SD = 0.30) for human essays and 5.78 (SD = 0.24) for AI essays. Statistically significant differences were observed for both metrics, with p-values of 5.594×10^{-3} for unique words and 3.454×10^{-38} for average word length, indicating a slightly richer vocabulary in the AI-generated texts and a preference for longer words. These findings suggest that AI essays not only employ a broader lexicon but also engage with complex language structures more frequently than their human counterparts. However, this lexical diversity and sophistication failed to translate into better scores; the mere presence of lexical richness does not guarantee

[3] https://github.com/WillYeadon/AI-Exam-Completion

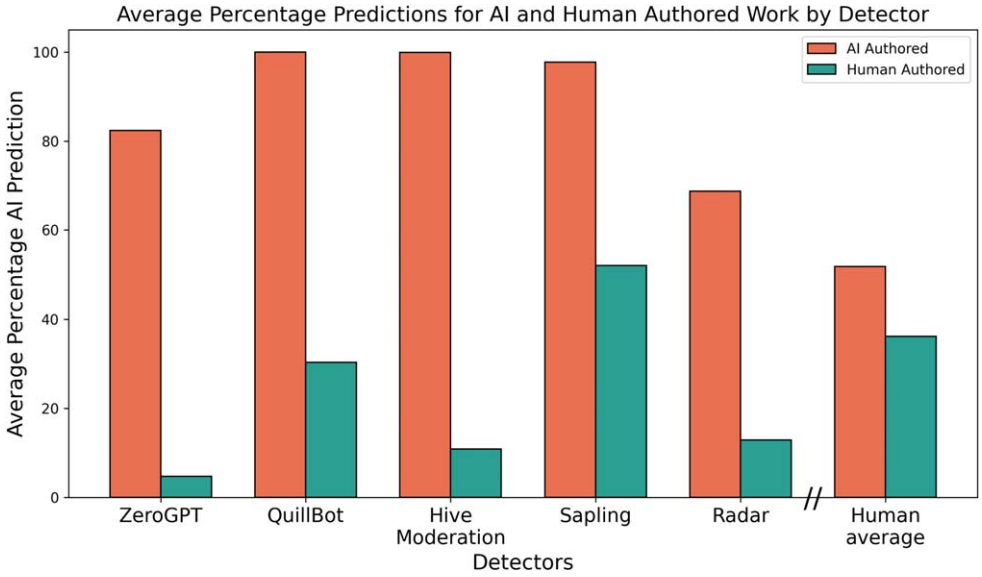

Figure 5. Histogram showcasing the performance of five AI detection tools in differentiating between AI-authored and human-authored text. On average, all detectors rated the AI-authored content as more likely to be AI-generated than the human-authored text thou there was considerable variation in the percentages assigned to each submission. On the far right, the human marker's Likert scores for each document are averaged whereby Definitely human = 0%, Probably human = 33.$\dot{3}$%, Probably AI = 66.$\dot{6}$%, and Definitely AI = 100%. Here, human performance is worse than all detectors at discerning AI from human text.

comprehension or analytical insight. The true effectiveness of these essays is their ability to articulate a clear understanding of the underlying physics concepts and to address the posed questions with precision and depth.

3.3. AI authorship computational detection

To comprehensively examine the landscape of AI text detection, we evaluated all the essays in our study using five different AI detector tools: 'ZeroGPT', 'QuillBot', 'Hive Moderation', 'Sapling', and 'Radar [23]'. These AI detection platforms employ varied methodologies to assess texts, yielding metrics such as 'Burstiness', 'Simplicity', 'Readability', and 'Perplexity'. The output from these tools can range from categorical assessments (e.g., 'Mostly AI-written' or 'Partly AI-written') to a straightforward binary indication of AI content presence. Typically, these applications also quantify the likelihood of AI authorship in terms of a percentage confidence level. Figure 5 offers a side-by-side comparison of the average percentages of the 'AI authorship' metric[4] assigned to AI-authored and human-authored submissions by each of the five detectors utilized in our study.

Applying a confusion matrix to the five AI text detectors, as shown in table 2, reveals that 'ZeroGPT' exhibits the best precision with no false positives (FP) and only one false negative (FN), achieving an accuracy of 98%. Next, 'QuillBot' shows a higher rate of false positives, mistaking human-written text for AI-authored content 10 times, which reduces its precision to 75% and accuracy to 83%. 'Hive Moderation' records 4 FPs and no FNs, while 'Sapling' has

[4] The exact wording used varied depending on the detector.

Table 2. Confusion matrix components, accuracy, and precision for each detector. Of the detectors tested, 'ZeroGPT' performs the best although the one open source detector, 'Radar', shows good performance also. The final row shows the average human marker's scores for each document, where the previously used Likert scale is converted to numerical values: Definitely human = 1, Probably human = 2, Probably AI = 3, and Definitely AI = 4. Scores above 2.5 are categorized as 'AI' and those below 2.5 as 'human'.

Detector	TP	FP	TN	FN	Accuracy	Precision
ZeroGPT	29	0	30	1	0.98	1.00
QuillBot	30	10	20	0	0.83	0.75
Hive moderation	30	4	26	0	0.93	0.88
Sapling	30	18	12	0	0.70	0.62
Radar	25	2	28	5	0.88	0.93
Human	15	6	24	15	0.65	0.71

the highest number of FPs at 18. Compared to these proprietary models, 'Radar' demonstrates moderate performance with a precision rate of 93% and an accuracy of 88%, although it has 5 FNs, suggesting that while it is quite reliable when it detects AI-generated content, it occasionally misses such content. Just as it is apt to measure generative AI quality with GPT-4 rather than GPT-3, the apparent state-of-the-art model, 'ZeroGPT', shows very strong accuracy and precision in detecting AI text.

This evaluation indicates that while AI-generated text may be detectable, the effectiveness of detection tools against content modified by paraphrasing remains uncertain. Our experiments with various paraphrasing tools suggest minimal impact on detection rates, which could reflect the limitations of these specific paraphrasing tools rather than an inherent robustness of detection algorithms. The possibility of human-assisted paraphrasing introduces an additional layer of complexity, effectively adding an 'editorial phase' where AI-generated text is reviewed and modified by humans to evade detection at which point it may no longer be considered AI-text. Further, if paraphrasing significantly degrades the text's quality or coherence, its utility might be questioned. Thus, any integration of human judgment with AI-generated content ventures into a realm of ethical and practical ambiguity.

4. Discussion

4.1. Impact on higher education

The findings from this study indicate that short-form physics essays, when not invigilated, are highly susceptible to being completed by AI, rendering this assessment method ineffective. If universities wish to retain this form of assessment, they must either implement stringent measures to verify authorship, such as oral examinations about the essays, or put their trust in AI detection tools like ZeroGPT. Our analysis, conducted without specific prompt engineering or iterative AI improvement, showed no discernible difference in quality between AI and human-authored essays, as determined by five independent markers. In reference to the paper's aims, we can see that human and AI-authored work cannot be distinguished by humans alone, though computational techniques may show more promise. The imperative is clear: adapt take-home essay assignments or transition to in-person assessments.

This said, there may not be as much cause for immediate concern in other areas of Physics Education. For instance, Polverini and Gregorcic [24] found GPT-4's performance in

interpreting kinematics graphs comparable to high school students, though with some unusual errors due to the novelty of the responses. If staff were trained to identify such novelty, like the curious text formatting in figure 3, it might be possible to eliminate egregious AI use. Similar results were found in [18], where GPT-4 approached physics problems using methods not covered in the syllabus. Additionally, both studies noted that GPT-4 refused to select one of the available answers around 5% of the time in multiple-choice questions. This said, AI excels in certain areas; for instance, in the case of Physics coding assignments, the best AI results actually match those of humans [25]. Furthermore, LLMs could aid physics education by helping correct misconceptions and enhancing educators' explanatory skills [26]. Given this, while we conclude that written short-form essay assignments are no longer an apt form of assessment, this is not universally true among physics assignments.

4.2. Recommendations

With AI tools becoming increasingly prevalent and integrated into software packages, it is crucial for the Physics Education community to recognize this reality and establish clear guidelines on acceptable use. We propose that any work with $\leqslant$50% AI-generated content be classified as human-authored. This threshold is future-proof, maintains a strong human authorship contribution, and simplifies enforcement, as any detector would need to show AI usage far in excess of 50%.

We do not yet know whether progress in AI writing capabilities will follow an exponential trajectory, surpassing human capabilities significantly, or a more sigmoidal trajectory, leveling off as training data is exhausted and model size increases yield diminishing returns. Regardless, models are becoming faster and more accessible. An increasing trend is 'unhobbling,' where LLMs overcome previous limitations by using more effective methods, such as running code to solve mathematical problems instead of relying solely on language processing [27]. Consequently, the specific failure modes of AI-authored text, such as idiosyncratic styles and approaches reported in previous studies [18, 24, 25], may not remain characteristic features forever. Ultimately, the physics knowledge typically required to answer undergraduate or master's level assessment questions is likely to be contained within the training distribution of foundation LLMs. In principle, you could be strategic with the question format to exploit areas where Transformers, a key element in modern LLMs, struggle, such as chaining various pieces of information to infer a final answer [28]; however, this too may be 'unhobbled' in the future as AI models continue to advance.

Our findings indicate that humans often cannot discern AI-authored text. However, several AI detectors, as shown in figure 5, displayed promise. By setting the limit at 50%, any significant AI use detected or identified idiosyncratic features would need to be egregious to surpass this threshold thus protecting academic integrity. It is crucial to remember that Physics education does not exist in isolation. If AI technology is widely used in the professional world, it should be allowed to some extent in academic settings. Yet, this integration should not undermine the value of human effort in obtaining a university degree. Thus, setting a threshold of $\leqslant$50% AI-generated content ensures that students still engage meaningfully with their work while adapting to technological advancements.

4.3. Limitations and future work

The primary limitation of this study stems from the utilization of raw, unedited output from GPT-4 for the essays. Two significant considerations emerge from this approach. First, allowing the AI to iteratively refine its answers by identifying and amending its own mistakes could potentially enhance its knowledge in physics [29]. Second, regarding assessment integrity, our

analysis was limited to essays authored entirely by AI, rather than examining a blend of human and AI contributions. The use of 100% AI-generated content could be seen as analogous to outsourcing essay writing to a paid service, a practice widely regarded as academic misconduct. However, the ethical landscape becomes murkier with minor AI contributions, such as using ChatGPT to rewrite a few sentences for improved clarity. The ambiguity increases further when considering tools like Microsoft Copilot, which are integrated directly into word processing software, presenting an evolving challenge without clear-cut solutions.

5. Conclusion

In the last 18 months, the rise of AI across various fields has been remarkable. Its impact necessitates considerable adjustments within physics education. This paper has shown that traditional non-invigilated short-form physics essays are losing relevance as humans and AI show similar writing abilities. This said, the outright dismissal of all non-invigilated exams may be premature; the burgeoning capabilities of AI detection technologies and idiosyncratic AI writing styles, like the inclusion of em dashes (—), suggest that it may be possible to limit egregious academic misconduct where AI work is copy/pasted and passed off as a student's own. Given this, we suggest $\leqslant$50% AI-generated content to be the boundary for human-authored text as it is inclusive of a future with ubiquitous AI assistance whilst also respecting human-authorship.

This evolution is akin to the 'hype cycle' model [30], where initial fervour for new technologies peaks and then wanes, only to stabilize as improvements in the underlying technology continue steadily. Consequently, educators must not panic but must evolve, incorporating AI into teaching and assessment strategies. This might involve assignments encouraging the exploratory use of tools like ChatGPT, allowing students to form their own opinions as to how they might utilize these powerful tools. While AI promises to augment the educational process, enriching rather than replacing the human touch, the authors would be reluctant to endorse a wholesale substitution of physics students with deep learning models [31]!

Data availability statement

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

Appendix A. Original questions from 'Physics in Society'

Table A1. Questions used for the 2021/22 formative assignment.

Question number	Question
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	Was there a scientific revolution in 17th-century Europe?
4	How did natural philosophers' understanding of electricity change during the 18th and 19th centuries?
5	Does Kuhn or Popper give a more accurate description of physics?

Table A2. Questions used for the 2021/22 summative assignment.

Question number	Question
1	If most theories have been shown to be false, do we have any reason to have confidence in our theories?
2	How has scientists understanding of charge changed during the 19th and 20th centuries?
3	Is there a satisfactory interpretation of quantum mechanics?
4	Was the project to build an atomic bomb typical of science in the twentieth century?
5	Why should the public believe scientists claims?

Table A3. Questions used for the 2022/23 formative assignment.

Question number	Question
1	In your experience, is physics based on facts that follow from observations?
2	What, in your opinion, was the most important advance in natural philosophy between 1100 and 1400?
3	Was there a scientific revolution in 17th-century Europe?
4	How did natural philosophers' understanding of electricity change during the 18th and 19th centuries?
5	Would you judge that Kuhn or Popper gives a more accurate description of physics?

Appendix B. Breakdown of marks by marker

A stacked histogram displaying the marks awarded by each marker is depicted in figure B1. To assess the similarity of the marks assigned by the five markers, an Analysis of Variance (ANOVA) was performed, followed by Levene's test for equality of variances. The ANOVA revealed a highly significant p-value of 5.106×10^{-29}, indicating strong evidence against the null hypothesis that all group means are equal, suggesting that at least one marker's scores significantly differ from the others. Similarly, Levene's test yielded a significant p-value of 8.437×10^{-10}, indicating unequal variances among the groups, violating the assumption necessary for ANOVA.

To evaluate the consistency of grading among multiple markers for the 60 submissions, we employed the Intraclass Correlation Coefficient (ICC) as a statistical tool to examine grading reliability across markers. The ICC1 model was utilized to measure the absolute agreement among markers, with each submission being evaluated by a different set of markers selected randomly. This model specifically assesses the consistency of scores given to each submission. Conversely, the ICC2 model was applied to assess the extent of agreement among markers on the relative ranking of the submissions, rather than the exact scores. An ICC value of 1 indicates perfect agreement among markers, whereas a value of -1 signifies complete disagreement. The results from these models are presented in table B1.

The results reveal a notable variability in grading consistency among markers in both ICC1 and ICC2 analyses, with ICC values nearing zero. This suggests that the grades for essays might have been significantly influenced by the markers' individual interpretations of the elements outlined in table 1. In the actual 'Physics in Society' module at Durham University, only one

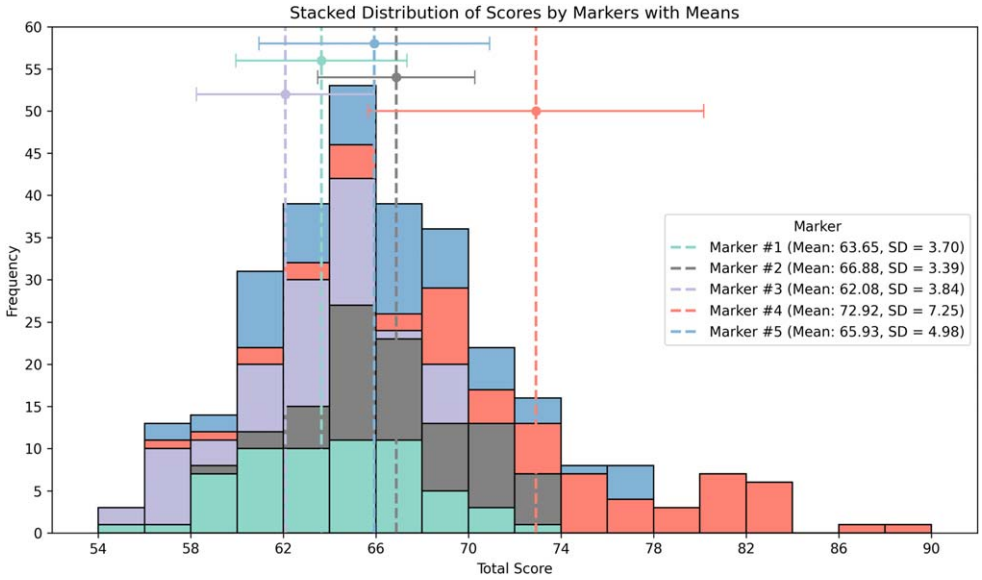

Figure B1. Stacked histogram of the scores awarded by the five independent markers. Both the ANOVA and ICC models used find that the markers were not consistent in their evaluations.

Table B1. Intraclass correlation coefficient (ICC) analysis results.

Type	ICC	*F*	df1	df2	*p*-value	CI 95%
ICC1	−0.053	0.749	59	240	0.907	[−0.11, 0.03]
ICC2	0.035	1.323	59	236	0.076	[−0.01, 0.11]

Table B2. Statistical results for t-tests for each marker individually showing that each marker's results also do not show a statistically significant difference in their means for AI-authored and human-authored essays.

Marker	Author	Mean	Std	*p*-value
Marker #1	AI	63.07	4.15	
Marker #1	Human	64.23	3.15	0.225
Marker #2	AI	66.33	3.36	
Marker #2	Human	67.43	3.39	0.212
Marker #3	AI	61.37	4.47	
Marker #3	Human	62.80	3.00	0.150
Marker #4	AI	72.53	8.16	
Marker #4	Human	73.30	6.34	0.686
Marker #5	AI	65.33	4.53	
Marker #5	Human	66.53	5.41	0.355

marker is assigned per submission. However, given the observed variability among markers previously involved in the module, there is a potential risk of impact on students' academic outcomes unless efforts are made to align standards and minimize subjective variability.

Despite the observed variability in grading among markers, this study treats the evaluation as comprising $n = 300$ separate assessments, evenly divided between essays authored by humans and AI. This approach yielded an average score of 66.86 (SD = 5.70) for human-authored essays and 65.73 (SD = 6.41) for AI-authored essays. A t-test analysis revealed a p-value of 0.107, suggesting that there is no statistically significant difference between the scores of human and AI-authored essays at a significance level of $\alpha = 0.05$. This finding holds consistent across a more granular analysis, where the dataset is considered as five separate sets of $n = 60$ submissions, one for each marker. As detailed in table B2, this comparative analysis between human and AI-authored essays was replicated across all five markers, reinforcing the initial conclusion that there are no significant differences in scoring between the two groups.

Appendix C. Ethics statement

This project received ethical approval from the Durham University Physics Ethics Committee ref: EDU-2023-03-14T14_02_18-hvxg44. All methods were performed in accordance with Durham University's Research Integrity Policy and Code of Good Practice. All students who participated in this study completed a statement of informed consent.

ORCID iDs

Will Yeadon https://orcid.org/0000-0002-9444-108X
Oto-Obong Inyang https://orcid.org/0000-0002-9001-0418

References

[1] Woolcock N 2023 ChatGPT marks end of homework at Alleyn's school The Times
[2] Grimaldi G *et al* 2023 Machines are about to change scientific publishing forever *ACS Energy Lett.* **8** 878–80
[3] Dowling M and Lucey B 2023 ChatGPT for (finance) research: the bananarama conjecture *Finance Res. Lett.* **53** 103662
[4] Hargreaves S 2023 Words are flowing out like endless rain into a paper cup ChatGPT & law school assessments *Legal Education Review* **23** 69
[5] Revell T *et al* 2023 *Research Square* ChatGPT versus human essayists: an exploration of the impact of artificial intelligence for authorship and academic integrity in the humanities (https://doi.org/10.21203/rs.3.rs-3483059/v1)
[6] Laskar M T R, Bari M S, Rahman M, Bhuiyan M A H, Joty S and Huang J X 2023 A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets (https://doi.org/10.48550/arXiv.2305.18486)
[7] Achiam J *et al* 2023 GPT-4 technical report (https://doi.org/10.48550/arXiv.2303.0877)
[8] Wang A, Pruksachatkun Y, Nangia N, Singh A, Michael J, Hill F, Levy O and Bowman S 2019 Superglue: a stickier benchmark for general-purpose language understanding systems *Adv. Neural Inf. Process. Syst.* **32**
[9] Singh M *et al* 2023 Mind meets machine: unravelling GPT-4's cognitive psychology arXiv:2303.11436 (https://doi.org/10.1016/j.tbench.2023.100139)
[10] Pu J, Sarwar Z, Abdullah S M, Rehman A, Kim Y, Bhattacharya P, Javed M and Viswanath B 2023 Deepfake text detection: limitations and opportunities *2023 IEEE Symposium on Security and Privacy (SP)* (IEEE) pp 1613–30
[11] Mitchell E, Lee Y, Khazatsky A, Manning C D and Finn C 2023 Detectgpt: zero-shot machine-generated text detection using probability curvature *International Conference on Machine Learning* (*ICML'23: Proceedings of the 40th International Conference on Machine Learning*)

[12] Sadasivan V S, Kumar A, Balasubramanian S, Wang W and Feizi S 2023 Can AI-generated text be reliably detected? (https://doi.org/10.48550/arXiv.2303.11156)
[13] Liang W, Yuksekgonul M, Mao Y, Wu E and Zou J 2023 GPT detectors are biased sgainst non-native english writers *Patterns* **4** 100779
[14] Staton B 2023 Universities express doubt over tool to detect AI-powered plagiarism Financ. Times
[15] Tulchinskii E, Kuznetsov K, Kushnareva L, Cherniavskii D, Nikolenko S, Burnaev E, Barannikov S and Piontkovskaya I 2024 Intrinsic dimension estimation for robust detection of AI-generated texts *Adv. Neural Inf. Process. Syst.* **36** 39257–76
[16] Wu J, Yang S, Zhan R, Yuan Y, Wong D F and Chao L S 2023 A survey on llm-gernerated text detection: necessity, methods, and future directions (https://doi.org/10.48550/arXiv.2310.14724)
[17] Gregorcic B and Pendrill A-M 2023 ChatGPT and the frustrated socrates *Phys. Educ.* **58** 035021
[18] Yeadon W and Halliday D P 2023 Exploring durham university physics exams with large language models (https://doi.org/10.48550/arXiv.2306.15609)
[19] Office for Students (OfS) 2022 *Analysis of degree classifications over time: changes in graduate attainment from 2010-11 to 2020-21* Office for Students (OfS)
[20] Chen L, Zaharia M and Zou J 2023 How is ChatGPT's behavior changing over time? *Harvard Data Science Review* **6** 1–26
[21] Dell'Acqua F, McFowland E, Mollick E R, Lifshitz-Assaf H, Kellogg K, Rajendran S, Krayer L, Candelon F and Lakhani K R 2023 Navigating the jagged technological frontier: field experimental evidence of the effects of AI on knowledge worker productivity and quality *Harvard Business School Technol. Oper. Mgt. Unit Working Paper (24-013)* **1** 1–58
[22] Thelot R 2023 Searching for sentience *AI Soc.* **1** 1–3
[23] Hu X, Chen P-Y and Ho T-Y 2024 Radar: Robust AI-text detection via adversarial learning *Adv. Neural Inf. Process. Syst.* **36** 15077–95
[24] Polverini G and Gregorcic B 2024 Performance of ChatGPT on the test of understanding graphs in kinematics *Phys. Rev. Phys. Educ. Res.* **20** 010109
[25] Yeadon W, Peach A and Testrow C P 2024 A comparison of human, GPT-3.5, and GPT-4 performance in a university-level coding course (https://doi.org/10.48550/arXiv.2403.16977)
[26] Gregorcic B and Polverini G 2024 ChatGPT as a tool for honing teachers' Socratic dialogue skills *Physics Education* **59** 045005
[27] Aschenbrenner L 2024 Situational Awareness: The Decade Ahead. Series: Situational Awareness
[28] Wang B, Yue X, Su Y and Sun H 2024 Grokked transformers are implicit reasoners: a mechanistic journey to the edge of generalization (https://doi.org/10.48550/arXiv.2403.16977)
[29] Polverini G and Gregorcic B 2024 How understanding large language models can inform the use of ChatGPT in physics education *Eur. J. Phys.* **45** 025701
[30] Dedehayir O and Steinert M 2016 The hype cycle model: a review and future directions *Technol. Forecast. Soc. Change* **108** 28–41
[31] Davis J P and Price W A 2017 Deep learning for teaching university physics to computers *Am. J. Phys.* **85** 311–2