Physics Assessment in the age of AI

Will YEADON

Department of Physics, Durham University, Lower Mountjoy, DH1 2LE, UK

Abstract. This contribution explores the profound effects of generative AI, particularly ChatGPT models, on physics education. By melding performance analyses of GPT-3.5 and GPT-4 across diverse assessments - including 593 physics exam questions, 300 coding submissions, and 300 essay submissions - we unveil nuanced insights into AI's effect on assessment. Our findings reveal that AI rivals human performance in essay writing and approaches it in coding tasks, yet falls short in physics written exams. This comprehensive evaluation not only highlights AI's potential and limitations in academic contexts but also sets the stage for discussing its pedagogical implications and future integration.

Introduction:

As we stepped into 2023, the landscape of education encountered an unprecedented shift with the advent of advanced AI technologies, such as ChatGPT. The ease with which students might leverage these AI tools for completing assignments ignited concerns over academic integrity, prompting a re-evaluation of traditional educational practices. This talk seeks to quantify how easily AI can complete various forms of physics assessment.

Research questions:

This talk addresses the following key research questions (RQs), aimed at uncovering the impact of generative AI on physics education and assessment:

- RQ1: How do generative AI models, specifically GPT-3.5 and GPT-4, perform in comparison to students when answering physics exam questions? This question probes the extent to which AI can replicate student success in traditional exam formats [1].
- RQ2: Can generative AI submissions, specifically those from GPT-3.5 and GPT-4 (both with and without prompt engineering), achieve or approach parity with student submissions in university-level coding assignments? Furthermore, we ask whether the independent markers, blinded to the source of submissions, can accurately distinguish between AI-generated and student-generated work.
- RQ3: How does the performance of the latest AI model, GPT-4, compare to student performances in writing physics essays? This aspect investigates not only the comparative quality of AI-generated and human essays but also explores the ability of markers, unaware of each essay's origin, to discern whether essays were generated by AI or by human students [2].

Methods and findings:

For the physics exams, we evaluated AI's proficiency on 42 physics exam papers with 593 questions, administered over several years at Durham University. Despite AI's performance improvements over time, it still trailed behind, particularly on pre-COVID exams, with GPT-4 scoring 50.8% and GPT-3.5 at 41.6%, while post-COVID scores showed a slight decrease. This suggests that even with adaptive exam formats, AI has not surpassed the level of weaker students.

Turning to coding, 100 university-level coding assignment submissions - 50 from students and 50 AI-generated - were blindly assessed by three independent markers, providing 300 data points. The accompanying histogram (Figure 1) clearly shows that students outperformed AI, with students averaging a score of 91.1% against 79.6% for GPT-4 submissions with prompt engineering. While prompt engineering notably enhanced AI performance, the submissions were still reliably distinguishable from student work, with an 85.3% accuracy rate in binary identification.

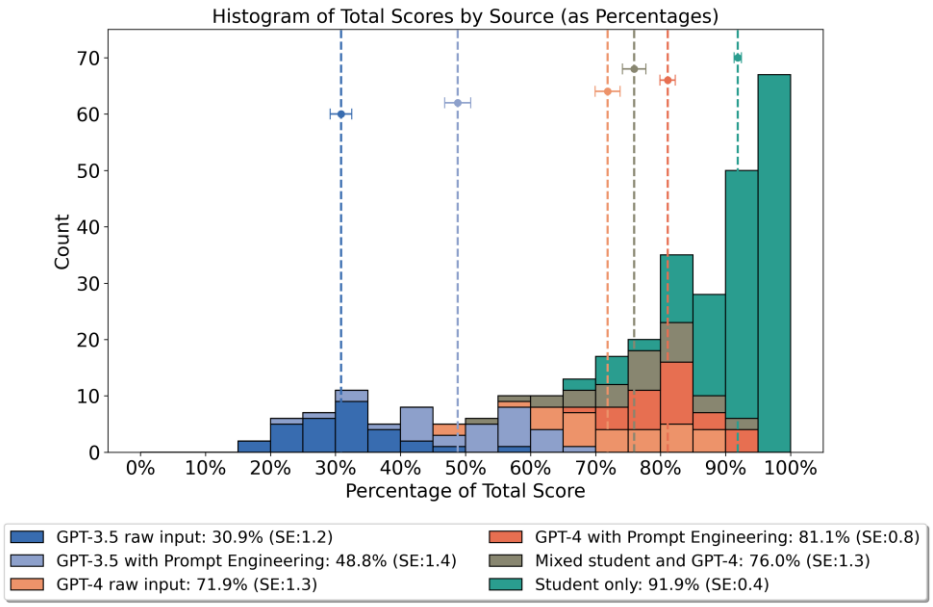

Figure. 1. Histogram of scores achieved by students and various AI models on a series of coding assignments.

Lastly, our analysis of 300 short-form physics essays - half written by students and half generated by AI - revealed no significant difference in scoring, as adjudicated by five independent markers. However, these markers performed only marginally better than chance when trying to identify AI-authored content. The evaluation of commercial authorship identification tools found ZeroGPT to be highly accurate, suggesting such tools may help maintain integrity in academic assessments amidst the rise of AI assistance.

Conclusion:

Our studies collectively demonstrate that while AI can emulate certain aspects of student performance, particularly in essay writing, it still falls short in complex physics exams, indicating the need for careful integration of AI within educational frameworks.

References

[1] W. Yeadon, D. P. Halliday, *Exploring Durham University physics exams with large language models*. arXiv preprint arXiv:2306.15609. 2023.

[2] W. Yeadon, E. Agra, O. Inyang, P. Mackay, A. Mizouri, *Evaluating AI and Human Authorship Quality in Academic Writing through Physics Essays*. arXiv preprint arXiv:2403.05458. 2024.