Revell *et al.*
*International Journal for Educational Integrity* (2024) 20:18
https://doi.org/10.1007/s40979-024-00161-8

International Journal for Educational Integrity

ORIGINAL ARTICLE Open Access

# ChatGPT versus human essayists: an exploration of the impact of artificial intelligence for authorship and academic integrity in the humanities

T. Revell[1*†], W. Yeadon[2†], G. Cahilly-Bretzin[3], I. Clarke[1], G. Manning[1], J. Jones[1], C. Mulley[1], R. J. Pascual[1], N. Bradley[1], D. Thomas[1] and F. Leneghan[1]

[†]T. Revell and W. Yeadon contributed equally to this work.

*Correspondence: tom.revell@ell.ox.ac.uk

[1] Faculty of English Language and Literature, University of Oxford, Manor Road, Oxford OX1 3UL, UK
[2] Department of Physics, University of Durham, South Rd, Durham DH1 3LE, UK
[3] Department of Archaeology, Classics and Egyptology, University of Liverpool, 12-14 Abercromby Square, Liverpool L69 7WZ, UK

## Abstract

Generative AI has prompted educators to reevaluate traditional teaching and assessment methods. This study examines AI's ability to write essays analysing Old English poetry; human markers assessed and attempted to distinguish them from authentic analyses of poetry by first-year undergraduate students in English at the University of Oxford. Using the standard UK University grading system, AI-written essays averaged a score of 60.46, whilst human essays achieved 63.57, a margin of difference not statistically significant ($p = 0.10$). Notably, student submissions applied a nuanced understanding of cultural context and secondary criticism to their close reading, while AI essays often described rather than analysed, lacking depth in the evaluation of poetic features, and sometimes failing to properly recognise key aspects of passages. Distinguishing features of human essays included detailed and sustained analysis of poetic style, as well as spelling errors and lack of structural cohesion. AI essays, on the other hand, exhibited a more formal structure and tone but sometimes fell short in incisive critique of poetic form and effect. Human markers correctly identified the origin of essays 79.41% of the time. Additionally, we compare three purported AI detectors, finding that the best, 'Quillbot', correctly identified the origin of essays 95.59% of the time. However, given the high threshold for academic misconduct, conclusively determining origin remains challenging. The research also highlights the potential benefits of generative AI's ability to advise on structuring essays and suggesting avenues for research. We advocate for transparency regarding AI's capabilities and limitations, and this study underscores the importance of human critical engagement in teaching and learning in Higher Education. As AI's proficiency grows, educators must reevaluate what authentic assessment is, and consider implementing dynamic, holistic methods to ensure academic integrity.

**Keywords:** ChatGPT, Artificial intelligence, Assessment, Higher education, AI text detection

## Introduction

### Background

The integration of computers into Higher Education has become pervasive across disciplines. While past technological innovations were adopted gradually within education (Scherer and Teo 2019), the emergence of ChatGPT suggests a significant shift in the realm of essay writing as a mode of assessment (Yeadon et al. 2024). This freely-accessible generative AI can produce essays on any topic within seconds, potentially jeopardising the integrity of assessments - a cornerstone for Higher Education institutions, degree-awarding entities, and employers who depend on graduate outputs. Notably, two out of the five principles from the Russell Group's (24 leading UK universities) recent 'Principles on the Use of Generative AI tools in Education' concern assessment integrity and academic rigour (Russell Group et al. 2023). Authentic assessment also plays a pivotal role in nurturing student self-esteem and motivation (McArthur 2023). This paper will explore the comparative performance of ChatGPT and human writers, investigate the detectability of AI-composed essays, and also consider the potential of AI to not only pose challenges but also to bolster student learning experiences (Gupta and Chen 2022).

Since the end of 2022, publicly-accessible generative AI chatbots such as ChatGPT have demonstrated the capacity to pass (if not excel at) a number of Higher Education examinations and qualifications, including the United States medical licensing examination, the Bar, and an MBA (Ryznar 2022). However, most studies of the impact of AI in Higher Education are in STEM (Crompton and Burke 2023), while many of those in the humanities are concerned with language acquisition (Zawacki-Richter et al. 2019). Like many STEM subjects, language acquisition can be assessed more objectively, and a relatively linear progression of skills and complexity can be arranged. OpenAI's own testing of GPT-4 found that English Language and Literature was the US College Admissions Advanced Placement examination that the AI scored lowest on, by far (Achiam et al. 2023). In a small study comparing US College student essays to GPT-3 outputs, AI responses achieved a similar grade to human essays (passing, but with lower variance) in Law, Research Methods, and US History, but failed Creative Writing (Sharples 2022). Given the increasing prominence of AI in educational spheres, there remains substantial scope for exploring its implications and applications within the humanities.

While numerous organisations claim to possess the capability to detect AI-generated content, there are grounds to question the reliability of these tools. For instance, some detection tools use a segment of the text as a prompt to gauge what a Large Language Model (LLM) predicts as the subsequent sequence of words. If the predicted sequences closely align with the actual succeeding text, it is inferred that the content was likely generated by AI. However, this process depends on the predictor LLM being similar to the original LLM that created the text. Additionally, the evolution of AI systems means that they can often surpass basic tests of their output's authenticity. For instance, modern chatbots have shown prowess in passing Turing-style tests (Jannai et al. 2023). Moreover, in a comprehensive study with a large sample size ($N = 830$), researchers found that outputs from models like GPT-2, particularly in specialised domains like poetry, were hard to distinguish from human-created content (Köbis and Mossink 2021). Compounding the issue, some AI systems are being fine-tuned to produce content intentionally designed to evade human detection, making the distinction even more

challenging (Jakesch et al. 2023). Methods such as paraphrasing attacks (Sadasivan et al. 2023), where small elements of a text are paraphrased or otherwise altered, can obscure AI authorship. Additionally, detectors that emphasize the statistical characteristics of human-authored texts may inadvertently discriminate against non-native English speakers, as their phrasing might not align with native patterns (Liang et al. 2023). Beyond these challenges, other studies have highlighted textual nuances, such as the presence or absence of specific phrases and symbols, as potential markers for distinguishing between human and AI-generated content (Desaire et al. 2023). Considering the multifaceted nature of AI's text generation, there is a pressing need to assess and refine methodologies for AI text detection.

## Literature review

Before the widespread availability of AI chatbots (marked by the emergence of ChatGPT), research predominantly emphasised the manifold advantages AI could introduce into the educational sector. These advantages were often designed to reduce educator workload (and associated stress), ranging from smart classrooms (Kim et al. 2018) to tailored assessment technologies (Luckin 2017). Numerous review articles portrayed artificial intelligence as a significant asset to academic administration (Chen et al. 2020) and a pathway to more dynamic learning experiences. Particularly, adaptive learning and intelligent tutoring systems were perceived as mechanisms that could greatly alleviate the assessment responsibilities of educators. Intriguingly, challenges highlighted in AI-focused literature as recent as 2022 included the '[l]imited technical capacity of AI' and the '[i]napplicability of the AI system to multiple settings' (Celik et al. 2022), rather than the evident risks to assessment authenticity observed today. UNESCO's 2019 report on AI in education primarily emphasised sustainability, equity, improved learner outcomes, and data security, overlooking potential misuse of AI by students (Pedro et al. 2019). The swift rise of a highly adept, essay-writing AI tool freely-accessible to students was a scenario few anticipated, and one which could increase the burden on educators.

Over the past two decades, technology-driven assessments such as automated grading and essay scoring have sought to augment traditional teaching and learning practices (Shermis 2014; Vajjala 2018). Despite the promise of automation, its success has predominantly been within subjects where answers can be clearly delineated as correct or incorrect. In contrast, the humanities, a field rich in nuance and interpretation, remains underrepresented in AI assessment explorations (González-Calatayud et al. 2021). This oversight extends beyond applications, as research reveals a low presence of authors affiliated with humanities departments in the broader AI education discourse (Zawacki-Richter et al. 2019).

While numerous Higher Education entities are swiftly implementing ethics regulations and altering assessment criteria in the face of generative AI - some even considering outright bans - there is a gap in collaborative efforts between pedagogical experts and AI researchers. This collaboration is pivotal, especially in light of emerging research on students' use of generative AI in Higher Education (Lavidas et al. 2020; Smolansky et al. 2023). The urgency for comprehensive insights into AI's ramifications for both teaching and assessment has only intensified in an era where powerful generative AI tools are publicly-accessible (Zawacki-Richter et al. 2019). Ultimately, the adoption (or

prohibition) of new technologies in educational settings hinges on educators' attitudes towards these tools and their perceived utility (Scherer and Teo 2019).

### Purpose of this study

The novelty of this study is the testing of ChatGPT on an exercise central to assessment in the humanities, close reading. Not only has generative AI been uncommonly assessed by its capacity for literary critique, but the close reading task on which this study is conducted is the analysis of a passage of Old English poetry. The factors complicating the close reading task for ChatGPT, the mode of the text (verse) and the language (a medieval one, not spoken for nearly one thousand years) were chosen to further test the ability of ChatGPT in a novel way. Somewhat representative of humanities subjects, Old English scholarship is also highly-analog and philological, and a field relatively apprehensive about the application of digital methodologies and computational tools; it has yet to be the focus of any generative AI-related study. However, this study has potentially wider implications for the fidelity of examination of the humanities at large, which often employ reading comprehension and critical analysis skills in their assessments, for which ChatGPT has been less-commonly tested. As well as testing the relative performance of University of Oxford students and ChatGPT at analysing Old English poetry, the second strand to this study is human detection of authentic student and AI-generated responses. While many current approaches lean heavily on computational methods to detect AI-generated writing, our study uniquely incorporates both digital techniques and experienced human markers to evaluate essay scoring and discern authorship. This research presents the findings from an investigation where markers scored close reading stylistic commentaries on passages of Old English poetry. These commentaries were either student- or AI-generated. Additionally, markers assessed whether they believed the essays were genuine or crafted by ChatGPT. The results and potential implications for the broader impact of generative AI on Higher Education assessment fidelity are discussed below.

## Method

This study took as its focus one particular form of assessment, which presented as potentially difficult for generative AI to perform well on. In the first year of studying English Language and Literature, undergraduates at the University of Oxford are examined on a module entitled 'Early medieval literature, 650-1350' by a three-hour, in-person, closed-book examination, which includes a task whereby a short passage of poetry (20-25 lines) in either Old English or Middle English must be analysed for features of content and style. In the present work, all of the essays were the result of prompt passages in Old English. Old English is a term used to describe the language of the inhabitants of the approximate area now known as England between the fifth- and twelfth-centuries AD, a Germanic language which only somewhat resembles present-day English. The earliest literature in English - including heroic legends, saints' lives, histories, charms, law codes, and riddles - still endures. Not only are these works foundational to any English degree, but they also influence popular portrayals of medieval northern Europe across books, music, television, and film. Given the vastness of the Internet and the specialised nature of Old English literature, it's probable that such content occupies a minor portion

of a general-purpose AI model's training data. Therefore, assessing LLMs' proficiency in analysing Old English poetry, a category with a dearth of available training data, offers a valuable benchmark for their performance in the broader humanities, as it may potentially isolate the analytical and evaluative skills of humanities students.

Textual analysis and commentary is a common method of examination in the humanities, where close reading and careful consideration of written materials are of central importance to research. The additional level of complexity involved in navigating not only poetry, but a medieval language, presented an interesting opportunity to test the limits of ChatGPT. The exercise requires students to 'comment on aspects of content and style and to show that you have a good understanding of [Old] English as a literary language' (Oxford 2022). The teaching for this course includes lectures, small-group classes, and individual tutorials. Students are equipped with a basic knowledge of Old English grammar and tools for identifying and evaluating aspects of Old English poetic style. These tools include understanding alliteration, metrical scansion, and compound diction. Additionally, students develop an appreciation for how the poem's content, an understanding of the language, and a literary-critical approach to the poetic style all contribute to sophisticated rhetorical effects. The criteria assesses answers on engagement, argument, information, organisation, and presentation, and inside these categories looks for qualities including clarity, coherence, depth, accuracy, and incisiveness.

For comparison with generative AI essays, previously-submitted student commentary assignments were collected for this study. The seven Old English poetry passages used as prompts for both students and ChatGPT were sourced from three seminal works: "Beowulf" (lines 767-805a and 864-879a), "The Dream of the Rood," (lines 1-23, 57-77 and 78-94) and "The Wanderer" (lines 11b-36 and 45-69). "Beowulf" is the longest extant poem in Old English, chronicling the heroic deeds and epic battles of its titular Geatish protagonist; the portion of the poem designated for this examination focuses on Beowulf's confrontation with Grendel, spanning lines 702b-897. "The Dream of the Rood" offers a poignant vision of Christ's passion juxtaposed with the promise of celestial bliss, while "The Wanderer" resonates with the melancholy of a solitary figure, either exiled or estranged, seeking spiritual refuge. For the student assignments, the directive 'Write a critical commentary on the following passage, placing it in context and analysing significant points of content AND style:' was given before a selected passage. For the AI-generated essays, the phrasing of the directive was varied to ensure diverse responses, even when the poetry excerpt remained consistent. Prompt modifications included synonym variation and the inclusion of specific word counts, as in '... a 600-word critical. . .' - an approach inspired by prior work on AI essay generation (Yeadon et al. 2024).

A total of 48 essays were generated using GPT-4 for this study. After an initial review, 8 were excluded due to clear unsuitability. Reasons for exclusion included responses formatted as bullet-pointed lists rather than essays, essays that did not address the specified passage, and essays that closely resembled other AI-generated texts. For comparison, 28 essays penned by students were included. Student essays responded to 7 different prompt-passages of Old English, but were unevenly distributed: specifically, for the 7 passages there were 5, 5, 9, 1, 1, 3, and 4 commentaries penned by students, respectively. While approximately a quarter of these were handwritten and later digitised, the remainder were directly submitted as typed documents. Importantly, all student essays were

crafted before the release of ChatGPT in November 2022; although other generative AI technologies were available before this date, we assessed the student essays as authentic in origin. To match this distribution, 7, 7, 9, 3, 3, 5, and 6 AI-generated essays were allocated for each of the same passages for close reading, respectively. To ensure uniformity, AI-generated essays underwent minor edits: American spellings were adjusted to their British equivalents, and in one-third of the essays, the titles of the poems were italicised instead of being placed in single quotation marks. This was done to prevent essay markers from identifying a key distinction based solely on formatting. Meanwhile, student essays remained unaltered, aside from the removal of extraneous details like word counts.

Seven markers, all experienced in assessing Old English poetry commentary at the University of Oxford, evaluated the essays. To conceal their origin, each essay was assigned a unique numeric code. Markers were given between six and nine essays to assess, ensuring a mix of both student-written and AI-generated pieces with at least one from each category. They were instructed to evaluate the essays as if they were student submissions, subsequently gauging the authorship on a 4-point Likert scale ranging from 'Definitely AI' to 'Definitely human'. Informed consent was obtained for all participants and the study was performed in accordance with relevant guidelines and regulations. This study received the following ethics approval: University of Oxford, Central University Research Ethics Committee [CUREC] approval reference: R88431/RE001.

## Results

### Performance metrics: AI vs. human

After the essays were marked, we undertook a series of statistical tests. Initially, we aimed to discern any potential differences between human and AI-generated essays. Concurrently, we assessed the reliability of our research by gauging the consistency of the markers and ensuring that the chosen essay prompts did not unduly influence the scores. Subsequently, we investigated the ability of both humans and purported computational AI-detectors to distinguish between AI-generated and student-written essays.

The essays were scored based on the standard UK grading system, ranging from 0 to 100. Here, scores of 70% and above earn First-Class Honours, a mark reserved for excellent work. Scores between 60% and 69% receive Upper Second-Class Honours (2:1), which is considered a good standard and is often required for entry into graduate schemes at major UK employers. Scores from 50% to 59% qualify for Lower Second-Class Honours (2:2). Third-Class Honours (3rd) are awarded for scores between 40% and 49%. Marks below 40% are classified as a Fail. The evaluation was blinded; markers were unaware of whether they were grading an AI- or human-authored essay. The average score for AI-written essays was approximately 60.48, while the human-written counterparts scored a tad higher, averaging at 63.57.A t-test revealed a t-statistic of -1.67 and a p-value of approximately 0.10. Given that this p-value comfortably exceeds our chosen alpha threshold of 0.01, we conclude that we fail to reject the null hypothesis, indicating no statistically significant difference between the scores. Thus, AI-generated essays performed comparably well to human essays. The distribution of scores by type is illustrated in Fig. 1. Notably, there are two outliers within the AI scores (at 25 and 30), both marked

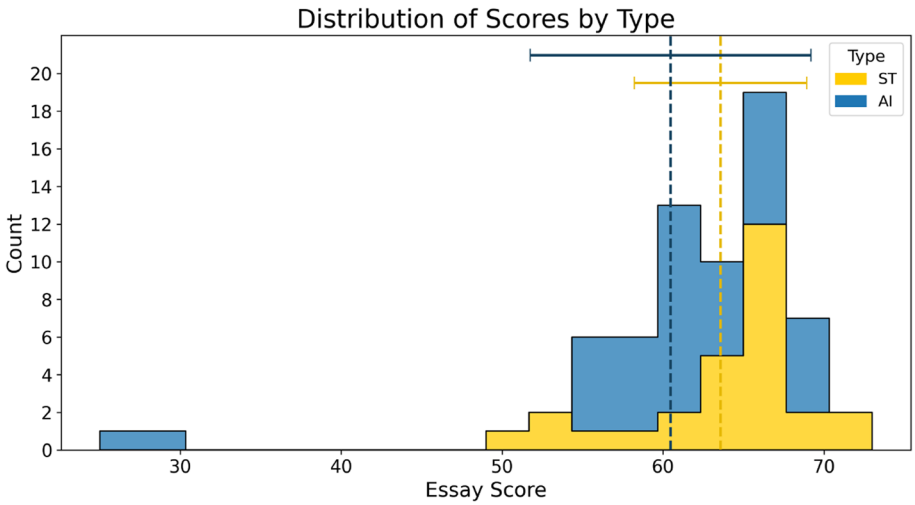


**Fig. 1** Histogram depicting the distribution of scores for AI (in blue) and students (in yellow)

by the same individual. This indicates that AI, akin to human students, can sometimes underperform.

To further understand scoring trends, we analysed consistency across the nine markers. The computed Intraclass Correlation Coefficient (ICC) was approximately 0.133, which denotes significant variability in marking. This suggests that while overarching scoring patterns can be insightful, individual marker scores demand a prudent interpretation. Prior research indicates that marker inconsistency becomes prominent in written exams without exemplars of good and subpar work (Baird et al. 2004). Our study adopted the standard assessment methodology as one marker per submission to ensure its findings were directly relevant, thereby aiding in the evaluation of AI's role in academic assessments.

Since our study used seven unique prompts, we conducted a one-way ANOVA to determine if the choice of prompt had any influence on the scores. The results showed an F-statistic of approximately 1.79 and a p-value of around 0.116. Given that this p-value surpasses our 0.01 significance threshold, we found no substantial evidence to suggest that the selection of the prompt considerably affects essay scores. The box plot in Fig. 2 clearly illustrates relatively consistent scores across prompts 1-3 and 5-7. However, the scores given to prompt 4 exhibit considerable variation, which can be attributed to the limited number of essays (only 4) for this prompt (comprising 1 human and 3 AI-generated essays). Additionally, the two distinctly low outlier scores of 25 and 30, visible in the histogram in Fig. 1, correspond to prompts 4 and 6, respectively.

#### Textual features: AI vs. human analysis

Praise for student essays, critiques of AI essays, and the criteria for achieving the top grades in the commentary assignment largely overlapped. Student essays employed knowledge of secondary criticism, early medieval English culture, and references to other parts of the poem from which the passage was taken or to other poems entirely, a level of demonstrable engagement with the text and application of relevant context that the GPT-4 essays did not feature. A nuanced consideration of minute details of poetic

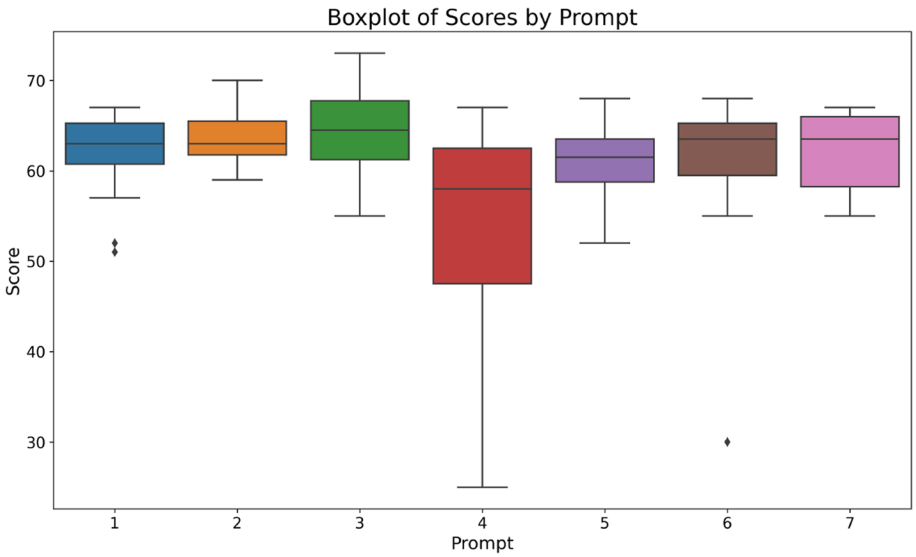


**Fig. 2** Boxplots illustrating the distribution of scores for each essay prompt, highlighting the variability and consistency among different prompts

style and effect (at the level of sounds, rhythm, words, and phrases) in the original language, combining close attention to text with a wider awareness of literature and culture of the period rather than thematic generalisations or value judgements, and original aspects of argument, observation, and interpretation were characteristics of the highest-scoring student essays that the AI essays did not match.

On the other hand, the AI-authored essays consistently scored well on structure, organisation, and academic register. GPT-4 outputs exhibited an articulate prose style, neatly-arranged using correct terminology, though perhaps with occasional odd choices of wording, or a stiff uniformity of sentence and paragraph length[1]. This aligns with the findings of Desaire et al. (2023), who noted discrepancies in the use of punctuation and the use of ambiguous, vague, or generalising language between human- and AI-writing. This is an area where otherwise strong poetic analyses can be penalised in student essays, though these aspects are not typically of central importance to submission for unseen close reading or commentary exercises. The logic and coherence of the (perceived) AI-style was often accompanied by markers' observations of non-standard or unexpected content, such as subjective value-judgments which undergraduates are taught to avoid in high-quality close reading (AI-outputs included such statements about 'the poet's prowess' or how the work belongs 'the annals of literature'), unusual mistranslations, or misapprehension of a particular rhetorical device. For example, though AI-outputs often noted ornamental alliteration as a rhetorical device in Old English poetry, occasionally the quote chosen to exemplify this device lacked alliteration, creating an odd juxtaposition of correct knowledge incorrectly applied[2].

[1] One marker specifically noted that the AI-generated answers 'break things up into small paragraphs which respond to specific parts of the prompt'.

[2] Two AI essays correctly identified *reþe ren-weardas* 'cruel [and] fearsome guards' ("Beowulf", line 770a) as an example of alliteration (on *r*); one of these essays included another correct example, *listum tolucan* 'to destroy with cunning' ("Beowulf", line 781a [where the prefix *to-* is unstressed]), but the other used *heawan þohton* 'intended to strike' ("Beowulf", line 800b), where there is no alliteration.

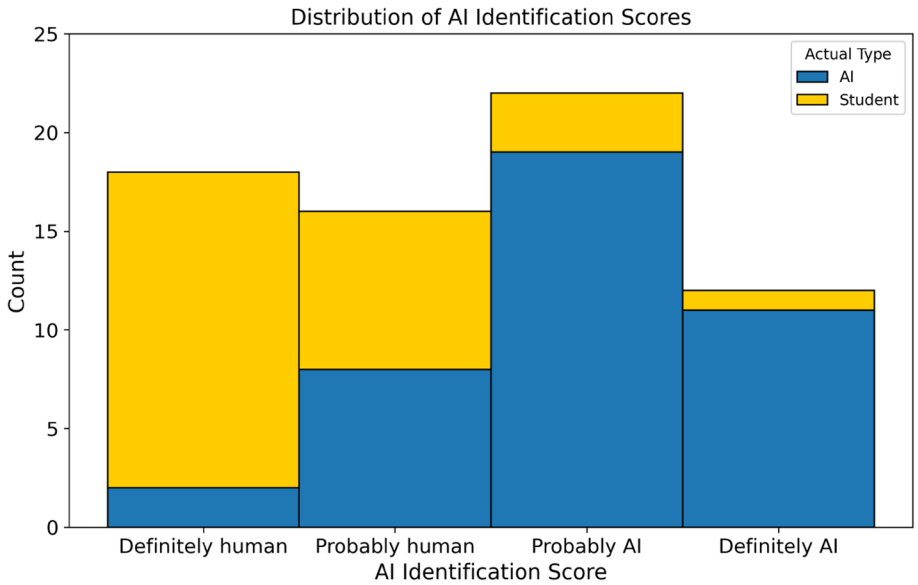


**Fig. 3** Visualisation of AI identifications distribution across the Likert scale used by the markers

The major differentiating factor between AI and authentic answers was a lack of detailed focus and critical analysis of elements of the passage. Characteristic of the ChatGPT outputs was exposition over precise consideration of individual poetic choices, "analysis" of the passage's content much more than its poetic style or effect, by describing and demonstrating rather than critically reflecting or evaluating. However, a lack of nuance and incisiveness is not uncommon in weaker students' answers, nor are generalisations of theme and context not linked explicitly to the passage at hand, so these elements cannot be regarded as a diagnostic for AI-written text.

#### AI-generated text detection: statistical measures

In our study, we also explored experienced human markers' ability to distinguish between AI-generated and student-composed essays. Markers were tasked with classifying each essay based on its perceived origin on a 4-point Likert scale ranging from 'Definitely human', 'Probably human', 'Probably AI', and 'Definitely AI'. Figure 3 shows the classifications assigned by the markers against the essays' actual authorship. Here, we can see how the 'Definitely human' and 'Probably human' categories had a higher proportion of student work compared to the 'Probably AI' and 'Definitely AI' categories, which contained mostly AI work.

The resulting confusion matrix revealed an overall identification rate of 79.41%. Breaking this down, of the 40 essays that were actually AI-authored, the human markers categorized 30 as either 'Definitely AI' or 'Probably AI'. In contrast, of the 28 essays that were student-authored, the human markers categorized 24 as either 'Definitely human' or 'Probably human'. Thus, out of the 68 essays, 54 were correctly identified, resulting in a 79.41% accuracy. These results suggest that human markers are reasonably good at differentiating between the two sources. However, it is important to note

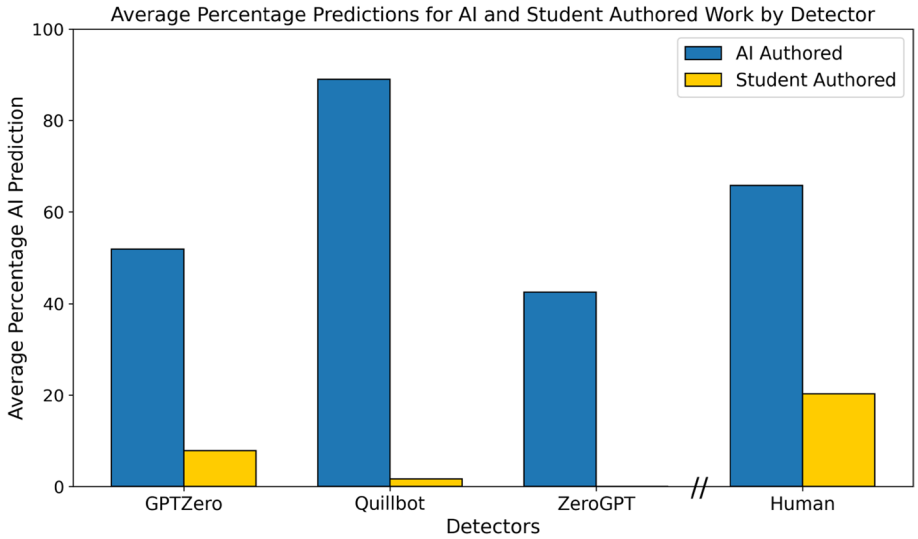


**Fig. 4** Histogram showing the performance of three purported AI-detection tools against the human markers

that 10 AI-authored essays were misclassified as human, and 2 of these were classified as 'Definitely human'.

We complemented the human authorship evaluations with a series of purported 'AI detectors'. Recent work has highlighted that this technology is proficient at identifying human-authored texts, though its accuracy in classifying AI-generated text is somewhat lower (Liu et al. 2023). Interestingly, Liu's study revealed that AI authorship was more readily discernible in Physics compared to the Humanities. As of the writing of this work, the leading methods of AI detection revolve around several approaches. These include statistical tests comparing word, character, and punctuation usage between AI-authored and human-authored texts, mimicking the AI generation process using portions of the input text and measuring how closely the subsequent generated text matches, and training a second AI on examples of both AI-authored and human-authored work to classify them. All of these methods typically yield a percentage chance of AI authorship or an otherwise estimated confidence. While some open-source detectors exist (Hu et al. 2023), many of the more popular ones, such as 'GPTZero', are proprietary. Regardless, to evaluate these detectors, we used three popular ones: 'GPTZero', 'Quillbot', and 'ZeroGPT' on the essays and compared them to the human evaluators. The results are shown in Fig. 4.

Collapsing the guesses of all detectors to binary, where $\geq 50\%$ indicates a guess of AI-authored, allows for the confusion matrix in Table 1 to be calculated. Here, we see that two out of the three detectors, 'GPTZero' and 'Quillbot', have a higher accuracy (the total of True Positives plus True Negatives over all 68 classified essays) than the aforementioned human rate of 79.41%, with 'Quillbot' having the highest at 95.59%. Thus, in this study we find that the detectors are indeed reasonable at detecting AI work. However, there are two issues to consider. First, as with plagiarism detection software, pure statistical methods are not sufficient on their own to penalize a student. Second, the test cases were either purely AI or purely human, rather than a

**Table 1** Performance metrics for various purported AI detectors compared to human evaluators. Here TP is True Positive, FP False Positive, TN True Negative and FN False Negative

| Detector | TP | FP | TN | FN | Accuracy | Precision |
|---|---|---|---|---|---|---|
| GPTZero | 33 | 2 | 26 | 7 | 86.76% | 0.94 |
| Quillbot | 37 | 0 | 28 | 3 | 95.59% | 1.0 |
| ZeroGPT | 16 | 0 | 28 | 24 | 64.71% | 1.0 |
| Human | 30 | 4 | 24 | 10 | 79.41% | 0.88 |

mixture of the two. Interestingly, while the detectors have very high precision rates - the fraction of True Positives over True Positives plus False Positives - meaning that the texts identified as AI-authored were either always or nearly always actually AI-authored, from an academic integrity perspective, a 1.0 precision should be the minimum acceptable level as falsely accusing a student of using AI is a very serious concern. Furthermore, it should be noted that the efficacy of these detectors could considerably reduce when any form of human-and-AI co-creation of text occurs Ardito (2023).

#### Qualitative evaluation of authorship

Identifying student authorship was sometimes possible by mistakes that are particularly human, such as errors of spelling or grammar. GPT-4 responses tended to make explicit reference to the wording of the questions, signpost a formalised structure of introduction and conclusion, and use primary text quotations to write whole sentences in a descriptive, story-telling style. GPT-4 output seems attuned to these as characteristics of essay-writing rather than close reading commentary, and perhaps reflective of a less mature style than expected at undergraduate-level. However, the literary-critical limitations of the AI-outputs created a dissonance between a polished academic register and a lack of attention to detail. One of the AI essays misunderstood the passage as one from Beowulf's fight *not* with Grendel, but with Grendel's mother. Though the essay would have satisfied some of the marking criteria to a reasonable degree, the fight with Grendel's mother is not one of the possible passages which could occur on the exam paper for this unseen assignment; the marker correctly identified it as an AI production, and (perhaps as a result of this) scored it a fail grade. Some of the AI essays erroneously considered as authentic student work received feedback on parts of their response which could be regarded as characteristics of generative AI style: vague or strange phrasing, lack of detailed stylistic analysis, and low attention to individual words, all contrasted with a general rigour of structure and academic tone. On the other hand, the information given to markers about the nature of the study resulted in some faulty epistemology and detection bias: a comment on one of the student essays read, 'the technical details were present, but garbled and the style suggested the 'right' information fed into a machine that doesn't quite know what to do with it'.

Despite a remarkably-high level of correct identification, markers' comments suggested a greater perceived discrepancy in performance of AI- and student-authored essays than their scoring demonstrated. One of the markers who correctly identified all of their essays and provided clear descriptions of how they did so, still awarded

their student essays an average of 65, and the AI essays an average of 62. This marker's feedback disclosed their prior experience with AI-generated outputs and clearly demonstrated an understanding of the vague, descriptive, well-structured and sophisticated style and approach of the AI in contrast to the more detailed, less scholarly tone, and looser structure of the student essays, yet the numerical discrepancy they awarded responses was ultimately not significant. The lack of consistency between markers might be a result of unequal levels of experience with generative AI: other studies have shown that markers with more awareness perform better at detection of authentic writing (Abd-Elaal et al. 2022).

## Discussion

### Overview

The findings of this study suggest that ChatGPT can effectively compete with students in generating commentary essays analysing passages of Old English poetry. While the quality measured by scoring is comparable, there is a discernible difference in the depth and nuanced understanding, particularly when it comes to human insights and cultural contexts. There might be a temptation to think that the niche nature of Old English literature could challenge a Large Language Model's capability: however, our results indicate otherwise. It appears that the LLM's training corpus encompassed enough references to Old English literary tradition, language, and poetic style, allowing ChatGPT to produce content of remarkable quality. This observation is in line with recent studies, demonstrating that even with limited but high-quality data, AI can exhibit impressive performance (Gunasekar et al. 2023).

An interesting dimension to consider is the complexity of the essay topic. This study focused on a university-level assignment, yet it's essential to recognise the potential of AI in broader educational contexts. Past research suggests that AI shows enhanced proficiency with assignments designed for younger demographics, such as those aimed at 15-16-year-olds, compared to university-level tasks (Yeadon and Hardy 2024). This suggests AI-generated work may well be superior to that of students for earlier educational levels.

As the academic community continues to grapple with the rise of generative AI, proactive engagement becomes crucial. Engaging students about the implications of AI, not just in their academic pursuits but also in the broader contexts of their future careers and societal roles, is paramount. Establishing a framework for the ethical and responsible usage of LLMs, based on diverse input from students, staff, government bodies, or NGOs, can serve as a foundation to address concerns surrounding assessment integrity and the overarching value of Higher Education qualifications. This framework should emphasise the importance of critically-evaluating AI-outputs, understanding the ethical dimensions of AI reliance, fostering self-efficacy in learning and the intrinsic value of writing as a means of both learning and reflection. Moreover, it should champion the importance of appreciating the cultural and empathetic value of delving into languages and cultures, as illustrated by this study's focus on Old English.

### Ethical implications

The advent of transformative technologies like AI in education inevitably brings with it a multitude of ethical considerations. Firstly, from a linguistic perspective, it is reassuring to note that AI-generated essays in this study, specifically those produced by GPT-4, did not exhibit any abusive or exclusionary language. In fact, GPT-4 shows an awareness of relevant, evolving issues (dependent on the cut-off date of its training data); when asked about the use of the term 'Anglo-Saxon', GPT-4 provides information on both sides of the current scholarly debate[3] surrounding this now-contentious terminology (Rambaran-Olm and Wade 2022). This is a positive step towards ensuring that AI tools are inclusive and do not perpetuate biases. However, from an academic standpoint, the results present potential dilemmas. The data suggests that AI can craft essays, without significant human intervention, that perform on par with student submissions in non-invigilated exams. This raises pressing questions about academic integrity and fairness. Beyond the immediate issue of plagiarism, there is the deeper concern that if students resort to AI assistance frequently, they risk depriving themselves of genuine learning experiences.

Further complicating the ethical landscape is the issue of equity of access. At the time of writing, GPT-4 is a premium service, while its predecessor, GPT-3.5, is freely accessible. This could inadvertently create a scenario where students with financial means are at an advantage, being able to access more advanced AI tools. While it is true that, historically, wealthier students have had the means to purchase premium educational resources, the disparity created by AI tools might be more pronounced. Moreover, students who are financially-secure benefit from peripheral advantages, such as not having to juggle part-time jobs alongside their studies, which can impact academic performance. As AI continues its trajectory in education, ongoing efforts are essential to ensure that its deployment aligns with academic standards and ethical principles.

### Impact on assessment

In the short term, alterations to assessments could prevent the exploitation of some of the biggest vulnerabilities (Susnjak and McIntosh 2024). Current limitations to LLMs could be combated by an emphasis on practical conditions of assessment (though these may cause other problems of inequity and resources); hand-written assignments; novel or unseen prompts; in-person examination; invigilation; multimodal questions (embedding video, audio, image in exam papers); and comparison with AI-generated answers to the exam questions. The nature of the task in this study with the unseen prompt seems particularly well-suited to assessment in the age of generative AI.

A reorientation of marking rubrics to focus less on organisation and presentation, and more on higher order thinking (in this case, the amount and depth of critical reflection on poetic details) may widen the gap between scoring of students and GPT-4. Such an approach would not only minimise the potential effect of generative AI, but also raise the bar on the weaker human essays that mask a lack of evaluation of poetic style and effect with sophisticated terminology and historical context (less relevant to commentary and

[3] In contemporary contexts, the term 'Anglo-Saxon' has been used more broadly to describe English-speaking peoples and their descendants. However, this usage is often criticized for oversimplifying complex cultural identities and histories. Furthermore, the term has become contentious due to its appropriation by white supremacist groups to promote racist ideologies, leading to debates about its usage in both academic and public discourse.

close reading exercise). Further alterations could involve issue spotting, problem solving, and extensive reasoning (rather than knowledge-based questions) (Ryznar 2022): although these should be considered short-term fixes, while different strategies are designed, tested, and evaluated. AI-detection tools will continue to develop and can be refined for use in specific instances, though entering an arms-race with generative AI-outputs seems futile.

However, these measures might only be short-term fixes that neither improve student or staff experience, nor equip graduates for the future world of work of which AI will be an integral part, and could soon be rendered obsolete by improvements in AI-generated outputs. Where assessments could accept, embed, or perhaps mandate the use of AI, this could encourage responsible and effective student use as a force multiplier for augmenting learning and thinking. In the longer term, innovation in assessment design is an opportunity to more effectively examine students than ever before, and could have fundamental changes to pedagogic practice too. Proactive and inclusive approaches to doing so could potentially improve the experience and outcomes for both students and educators.

The most promising methods could be more dynamic and less time-consuming. Assessing progress and contextualised knowledge-application as well as attainment could be achieved through small and regular assessments (or self-evaluations). There is potential that these could be examined in an (semi-)automated fashion, and adaptive or personalised learning algorithms could be designed to guide continued development tailored to student abilities and interests. Summative project work - constituted by multimodal artefacts, collaborative work, presentations, or creative performances - could reduce the weight of final examination, and test a range of relevant skills. The University of Oxford has the staffing and tutorial system structure in place that makes academic dishonesty a less immediate issue (unethical use of ChatGPT in writing an essay would be easy to discover in an hour-long, 1-on-1 tutorial, and where tutors have familiarity with the writing styles of their small cohorts of students); in the larger seminar setting of most UK universities, the problem is more urgent.

#### AI co-piloting

There is promising potential in using ChatGPT as a research and learning assistant (Kasneci et al. 2023). When approached as if it were an attentive undergraduate, given a brief overview of a text and a thematic focus, ChatGPT can offer a clear essay structure with prompts for each section, as illustrated in Fig. 5. However, to maximise the benefits of these suggestions, a solid understanding of the subject matter is essential. For instance, we initiated a dialogue with ChatGPT to develop an essay on the portrayal of monstrosity across Old English literature[4]. Our starting point was a mention of the monstrous figures in "Beowulf": Grendel, Grendel's mother, and the dragon, as well as Beowulf's own ambiguously-monstrous traits.

In ChatGPT's response, the thematic focus (depictions of monstrosity in Old English literature) was correctly-identified in "Beowulf", but the request for other relevant

[4] Full conversation available in supplementary materials.

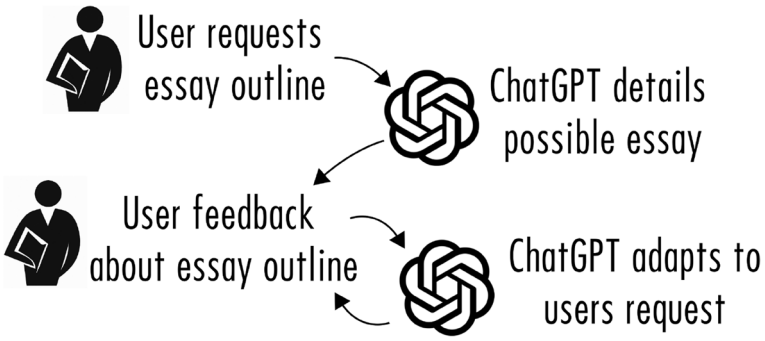


**Fig. 5** Co-piloting with ChatGPT to write an essay

texts with monstrous figures in them initially flagged three poems which do not (to any significant extent) feature monsters. Having the familiarity with these texts sufficient to know that they lack depictions of monstrosity, one can correct ChatGPT, provoking it to provide five Old English texts that do feature monsters. Each had a short description of the monstrous descriptions therein, and any of these texts could be considered in a comparative study with "Beowulf" to an interesting degree. In some instances, ChatGPT's perspective was impressively intriguing, bordering on the original. For example, it framed the phoenix in the titular poem as having monstrous elements, and indicated that "Judith" could be a compelling subject in an essay exploring monstrosity, though didn't specifically highlight the monstrous character of Holofernes.

In search of further details about where specifically one might find monstrosity in the five identified texts, ChatGPT was asked for more details about the Old English riddles, around 100 of which survive in the so-called Exeter Book manuscript. While the response identified the content of four riddles which could provide a fruitful angle on monstrosity, it mis-identified specifically which riddles these were. Furthermore, possessing prior knowledge of the poem "Andreas" revealed an error in ChatGPT's claim about a serpent featuring therein. Yet, the model aptly recognised not only St Andrew's cannibalistic foes but also underscored other elements in the poem with monstrous or supernatural undertones, some of which could be profitable starting-points for an undergraduate interested in this theme to look out for in reading this poem. Importantly, ChatGPT's responses do not reference the original language of any of the texts themselves, a fundamental part of literary study and a feature (critical and close textual evaluation) also lacking in the AI-outputs of above study.

A request for secondary criticism to consult was less successful. While recommending relevant journals, an essay by J.R.R. Tolkien, and Seamus Heaney's adaptation of "Beowulf", the other references were either mis-attributed to authors who have produced similar works, or entirely "hallucinated" (Roller et al. 2021), in each case with an inaccurate summary. A number of obvious examples of secondary works on the topic were not included.

This example is intended to provide a model of ethical use of AI for supporting teaching and learning. In generating ideas and avenues based on an initial idea, the outputs were relatively useful, providing potential angles into or perspectives on the

topic at hand which could be an inspiration to further study. These suggestions are (in varying degrees) interesting, original, and valid, and could provide a sandpit in which to experiment with ideas and approaches which might not be immediately obvious or orthodox. However, it should be clear that this exercise involved a good deal of knowledge on the primary and secondary material to navigate, correct, negotiate, and question ChatGPT's responses, to avoid its vagaries and inaccuracies and make the most out of its abilities.

Given this, educators may consider focusing any assessment on specific AI weaknesses. For example, the lack of connections in any of the AI-outputs between the focus of the essay and other relevant primary texts (such as other Old English poems like "Judith" or "The Battle of Maldon") could be leveraged to discourage students' over-reliance on AI-outputs, by making this type of analysis a required part of submissions. However, there is reason to be cautious here as these systems are rapidly evolving, and any current weaknesses may soon be resolved or otherwise mitigated or 'unhobbled' (Aschenbrenner 2024). Therefore, it is recommended to maintain transparency with both staff and students that while generative AI can create text on any given topic, the user is ultimately the judge of whether the text is useful. Our results show that detail-oriented critical analysis was the major identifiable discrepancy between AI and authentic outputs, and therefore suggest that assessors consider how to weight marking criteria more justly in light of this.

It should also be noted that the question of "Beowulf" and monstrosity is a common undergraduate essay topic, and more novel topics do not receive as detailed consideration. Not asking ChatGPT to write the essay itself is part of such a responsible approach. At this level of use, it does not (and, as in the above study, does not perform well at) analyse the texts themselves, in the original language, nor does it much consider their rhetoric, imagery, or cultural context. A student who used it in such a way would still need to go away and read the primary texts in detail and identify and analyse selected passages, contextualised by secondary material, to write a strong essay. However, the ease with which connections are made between texts and themes could, optimistically, be a force multiplier in a Higher Education setting to inspire further study.

## Conclusion

This study examines the capabilities and potential implications of generative AI within the Higher Education sphere, specifically in the humanities, represented by a task of close analysis of a passage of Old English poetry. Our results suggest that while generative AI has not yet posed an existential threat to traditional in-person, closed-book, unseen, close-reading examinations, its rapid advancements signal an imminent shift. Higher Education institutions must urgently consider designing and implementing a new generation of assessment practices. Addressing this issue now, before the next evolution of AI, could help to establish a system of assessment innovation that could be dynamically updated in line with future technological advances. These assessments should not only be adaptive to AI's capacities but also establish an effective feedback loop to ensure they both authentically and accurately test students' relevant skills, and prepare graduates for an AI world.

Quantitatively, AI-generated essays are comparable in quality to those written by undergraduate students. However, human essays still exhibit nuanced and contextualised insights that AI did not replicate. Interestingly, human markers managed to correctly discern between AI and human essays 80% of the time, yet this percentage is not robust enough for strict enforcement on any individual basis. Furthermore, while two of the AI detectors performed better than humans, one actually performed worse, including two false positives.

As the capacities of generative AI continue to increase exponentially, the education sector faces a pressing challenge: recalibrating its understanding and expectations of what constitutes authentic assessment, and determining how best to educate students for a world influenced by generative AI. An illustrative example of this was presented in the conversational essay plan between human and AI: the AI suggests ideas, while the human critically evaluates them. Such synergies may pave the way for future educational methods. The insights from this study underscore the enduring importance of human critical engagement in education, even as AI steadily makes inroads into various academic realms.

**Abbreviations**

AI Artificial Intelligence
LLM Large Language Model

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s40979-024-00161-8.

Supplementary Material 1.

**Acknowledgements**
Not applicable.

**Authors' contributions**

TR: Conceived and designed the study in collaboration with WY; led the qualitative analysis; contributed to writing and drafting half of the paper sections. WY: Conceived and designed the study in collaboration with TR; spearheaded the statistical analysis; contributed to writing and drafting half of the paper sections. The following authors were instrumental in the implementation of the study, particularly in the marking process, ensuring the blinding was maintained and upholding the integrity of the study results. They reviewed and approved the final manuscript but did not contribute to the writing of the paper: GC-B; IC; GM; JJ; CM; RJP; NB; DT; FL.

**Funding**
This research was funded in whole, or in part, by the UKRI [AHRC; Research Council Student ID: R95924E (Tom Revell)]. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

**Availability of data and materials**

The prompts used, raw essays and marks for the AI generated essays are made available as supplementary information in the published article. Due to data protection we cannot make the student essays or scores available and will destroy this data at the end of the study.

## Declarations

**Competing interests**

We know of no conflicts of interest for the work and there has been no significant financial support for this work that could have influenced the outcome.

Received: 28 January 2024 Accepted: 12 August 2024
Published online: 21 October 2024

**References**


Abd-Elaal ES, Gamage SH, Mills JE (2022) Assisting academics to identify computer generated writing. Eur J Eng Educ 47(5):725–745

Achiam J, et al (2023) Gpt-4 technical report. arXiv preprint arXiv:2303.08774. https://arxiv.org/abs/2303.08774.

Ardito CG (2023) Contra generative ai detection in higher education assessments. arXiv preprint arXiv:2312.05241

Aschenbrenner L (2024) Situational awareness: The decade ahead. https://situational-awareness.ai/. series: Situational Awareness. Accessed 22 July 2024

Baird JA, Greatorex J, Bell JF (2004) What makes marking reliable? experiments with uk examinations. Assess Educ Princ Policy Pract 11(3):331–348

Celik I, Dindar M, Muukkonen H, Järvelä S (2022) The promises and challenges of artificial intelligence for teachers: a systematic review of research. TechTrends 66(4):616–630

Chen L, Chen P, Lin Z (2020) Artificial intelligence in education: A review. IEEE Access 8:75264–75278

Crompton H, Burke D (2023) Artificial intelligence in higher education: the state of the field. Int J Educ Technol High Educ 20(1):1–22

Desaire H, Chua A, Isom M, Jarosova R, Hua D (2023) Distinguishing academic science writing from humans or chatgpt with over 99% accuracy using off-the-shelf machine learning tools. Cell Rep Phys Sci 4(6):101426

González-Calatayud V, Prendes-Espinosa P, Roig-Vila R (2021) Artificial intelligence for student assessment: A systematic review. Appl Sci 11(12):5467

Gunasekar S, Zhang Y, Aneja J, Mendes CCT, Del Giorno A, Gopi S, et al (2023) Textbooks are all you need. arXiv preprint arXiv:2306.11644

Gupta S, Chen Y (2022) Supporting inclusive learning using chatbots? a chatbot-led interview study. J Inf Syst Educ 33(1):98–108

Hu X, Chen PY, Ho TY (2023) Radar: Robust ai-text detection via adversarial learning. Adv Neural Inf Process Syst 36:15077–15095

Jakesch M, Hancock JT, Naaman M (2023) Human heuristics for ai-generated language are flawed. Proc Natl Acad Sci 120(11):e2208839120

Jannai D, Meron A, Lenz B, Levine Y, Shoham Y (2023) Human or not? a gamified approach to the turing test. arXiv preprint arXiv:2305.20010

Kasneci E, Seßler K, Küchemann S, Bannert M, Dementieva D, Fischer F, Gasser U, Groh G, Günnemann S, Hüllermeier E et al (2023) Chatgpt for good? on opportunities and challenges of large language models for education. Learn Individ Differ 103:102274

Kim Y, Soyata T, Behnagh RF (2018) Towards emotionally aware ai smart classroom: Current issues and directions for engineering and education. IEEE Access 6:5308–5331

Köbis N, Mossink LD (2021) Artificial intelligence versus maya angelou: Experimental evidence that people cannot differentiate ai-generated from human-written poetry. Comput Hum Behav 114:106553

Lavidas K, Achriani A, Athanassopoulos S, Messinis I, Kotsiantis S (2020) University students' intention to use search engines for research purposes: A structural equation modeling approach. Educ Inf Technol 25:2463–2479

Liang W, Yuksekgonul M, Mao Y, Wu E, Zou J (2023) Gpt detectors are biased against non-native english writers. arXiv preprint arXiv:2304.02819

Liu Z, Yao Z, Li F, Luo B (2023) Check me if you can: Detecting chatgpt-generated academic writing using checkgpt. arXiv preprint arXiv:2306.05524

Luckin R (2017) Towards artificial intelligence-based assessment systems. Nat Hum Behav 1(3):0028

McArthur J (2023) Rethinking authentic assessment: work, well-being, and society. High Educ 85(1):85–101

Oxford Uo (2022) English language and literature prelims handbook - university of oxford. https://oess.web.ox.ac.uk/files/ellprelimshandbook2022-2311pdf. Accessed 17 Sep 2023

Pedro F, Subosa M, Rivas A, Valverde P (2019) Artificial intelligence in education: Challenges and opportunities for sustainable development. Tech. rep, UNESCO, Paris

Rambaran-Olm M, Wade E (2022) What's in a name? the past and present racism in 'anglo-saxon' studies. Yearb Engl Stud 52(1):135–153

Roller S, Dinan E, Goyal N, Ju D, Williamson M, Liu Y, Xu J, Ott M, Smith EM, Boureau YL, Weston J (2021) Recipes for building an open-domain chatbot. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, pp 300–325. https://doi.org/10.18653/v1/2021.eacl-main.24. https://aclanthology.org/2021.eacl-main.24

Russell Group T (2023) New principles on use of ai in education. https://russellgroup.ac.uk/news/new-principles-on-use-of-ai-in-education/. Accessed 17 Sep 2023

Ryznar M (2022) Exams in the time of chatgpt. Washington and Lee Law Review Online 80:305

Sadasivan VS, Kumar A, Balasubramanian S, Wang W, Feizi S (2023) Can ai-generated text be reliably detected? arXiv preprint arXiv:2303.11156

Scherer R, Teo T (2019) Unpacking teachers' intentions to integrate technology: A meta-analysis. Educ Res Rev 27:90–109

Sharples M (2022) Automated essay writing: An aied opinion. Int J Artif Intell Educ 32(4):1119–1126

Shermis MD (2014) State-of-the-art automated essay scoring: Competition, results, and future directions from a united states demonstration. Assess Writ 20:53–76

Smolansky A, Cram A, Raduescu C, Zeivots S, Huber E, Kizilcec RF (2023) Educator and student perspectives on the impact of generative ai on assessments in higher education. In: Proceedings of the Tenth ACM Conference on Learning @ Scale. ACM: Association for Computing Machinery, New York, pp 378–382. https://doi.org/10.1145/3573051.3596191

Susnjak T, McIntosh TR (2024) Chatgpt: The end of online exam integrity? Educ Sci 14(6):656

Vajjala S (2018) Automated assessment of non-native learner essays: Investigating the role of linguistic features. Int J Artif Intell Educ 28:79–105

Yeadon W, Hardy T (2024) The impact of AI in physics education: a comprehensive review from GCSE to university levels. Phys Educ 59(2):025010

Yeadon W, Agra E, Inyang OoA, Mackay P, Mizouri A (2024) Evaluating ai and human authorship quality in academic writing through physics essays. Eur J Phys. http://iopscience.iop.org/article/10.1088/1361-6404/ad669d
Zawacki-Richter O, Marín VI, Bond M, Gouverneur F (2019) Systematic review of research on artificial intelligence applications in higher education-where are the educators? Int J Educ Technol High Educ 16(1):1–27

## Publisher's Note