

# ChatGPT versus Human Essayists: An Exploration of the Impact of Artificial Intelligence for Authorship and Academic Integrity in the Humanities

**Tom Revell**

University of Oxford

**Will Yeadon** (✉ [will.yeadon@durham.ac.uk](mailto:will.yeadon@durham.ac.uk))

Durham University

**Glenn Cahilly-Bretzin**

University of Liverpool

**Isabella Clarke**

University of Oxford

**George Manning**

University of Oxford

**Jasmine Jones**

University of Oxford

**Clare Mulley**

University of Oxford

**Rafael Pascual**

University of Oxford

**Natasha Bradley**

University of Oxford

**Daniel Thomas**

University of Oxford

**Francis Leneghan**

University of Oxford

---

## Research Article

**Keywords:** ChatGPT, Artificial Intelligence, Assessment, Higher Education, AI text detection

**Posted Date:** November 7th, 2023

**DOI:** <https://doi.org/10.21203/rs.3.rs-3483059/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

# ChatGPT versus Human Essayists: An Exploration of the Impact of Artificial Intelligence for Authorship and Academic Integrity in the Humanities

T. Revell<sup>1†</sup>, W. Yeadon<sup>2†</sup>, G. Cahilly-Bretzin<sup>3</sup>, I. Clarke<sup>1</sup>,  
G. Manning<sup>1</sup>, J. Jones<sup>1</sup>, C. Mulley<sup>1</sup>, R. J. Pascual<sup>1</sup>, N. Bradley<sup>1</sup>,  
D. Thomas<sup>1</sup>, F. Leneghan<sup>1</sup>

<sup>1\*</sup>Faculty of English Language and Literature, University of Oxford,  
Manor Road, Oxford, OX1 3UL United Kingdom.

<sup>2</sup>Department of Physics, University of Durham, South Rd, Durham,  
DH1 3LE, United Kingdom.

<sup>3</sup>Department of Archaeology, Classics and Egyptology, University of  
Liverpool, 12-14 Abercromby Square, Liverpool, L69 7WZ, United  
Kingdom.

Contributing authors: [tom.revell@ell.ox.ac.uk](mailto:tom.revell@ell.ox.ac.uk);  
[will.yeadon@durham.ac.uk](mailto:will.yeadon@durham.ac.uk); [glenn.cahilly-retzin@liverpool.ac.uk](mailto:glenn.cahilly-retzin@liverpool.ac.uk);  
[isabella.clarke@ell.ox.ac.uk](mailto:isabella.clarke@ell.ox.ac.uk); [george.manning@ell.ox.ac.uk](mailto:george.manning@ell.ox.ac.uk);  
[jasmine.jones@ell.ox.ac.uk](mailto:jasmine.jones@ell.ox.ac.uk); [clare.mulley@ell.ox.ac.uk](mailto:clare.mulley@ell.ox.ac.uk);  
[rafael.pascual@ell.ox.ac.uk](mailto:rafael.pascual@ell.ox.ac.uk); [natasha.bradley@ell.ox.ac.uk](mailto:natasha.bradley@ell.ox.ac.uk);  
[daniel.thomas@ell.ox.ac.uk](mailto:daniel.thomas@ell.ox.ac.uk); [francis.leneghan@ell.ox.ac.uk](mailto:francis.leneghan@ell.ox.ac.uk);

<sup>†</sup>These primary authors contributed equally to this work.

## 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 significantly 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, AI-generated text detection from GPTZero revealed statistically significant ( $\alpha = 0.01$ ) differences in Perplexity and

Burstiness between AI and human essays. 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

# 1 Introduction

## 1.1 Background

The integration of computers into Higher Education has become pervasive across disciplines. While past technological innovations were adopted gradually within education (Scherer & Teo, 2019), the emergence of ChatGPT suggests a significant shift in the realm of essay writing as a mode of assessment (Yeadon, Inyang, Mizouri, Peach, & Testrow, 2023). 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, 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 & 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 & Burke, 2023), while many of those in the humanities are concerned with language acquisition (Zawacki-Richter, Marín, Bond, & Gouverneur, 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 (OpenAI, 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. Many primarily operate by using 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 align closely with the actual succeeding text, it is inferred that the content was likely generated by AI. This correlation can be quantitatively assessed using

perplexity. However, 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, Meron, Lenz, Levine, & Shoham, 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 & 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, Hancock, & Naaman, 2023). There are inherent vulnerabilities in relying solely on perplexity-based approaches. This methodology is notably susceptible to paraphrasing attacks (Sadasivan, Kumar, Balasubramanian, Wang, & Feizi, 2023), where minor text alterations can deviate enough from the AI’s expected sequence. Moreover, detectors that emphasise perplexity might inadvertently discriminate against non-native English speakers, as their phrasing might not align with native patterns (Liang, Yuksekgonul, Mao, Wu, & Zou, 2023). Beyond these challenges, other studies have highlighted textual nuances, such as the presence or absence of specific phrases and symbols, as potential markers that can distinguish between human and AI-generated content (Desaire, Chua, Isom, Jarosova, & Hua, 2023). Considering the multifaceted nature of AI’s text generation, there is a pressing need to assess and refine methodologies for AI text detection.

## 1.2 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, Soyata, & Behnagh, 2018) to tailored assessment technologies (Luckin, 2017). Numerous review articles portrayed artificial intelligence as a significant asset to academic administration (Chen, Chen, & Lin, 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, Dindar, Muukkonen, & Järvelä, 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, Subosa, Rivas, & Valverde, 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, Prendes-Espinosa, & Roig-Vila, 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, Achriani,

Athanassopoulos, Messinis, & Kotsiantis, 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 & Teo, 2019).

### 1.3 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.

## 2 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 offers a valuable benchmark for their performance in the broader humanities.

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, where students are equipped with a basic knowledge of the grammar of Old English, tools for identifying and evaluating aspects of Old English poetic style (such as alliteration, metrical scansion, and compound diction), and an appreciation for how the content of the poem, 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., 2023).

A total of 48 essays were generated using GPT-4 for this study. After an initial review, 8 were excluded due to evident poor quality or because of their close resemblance to other AI-generated essays. 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

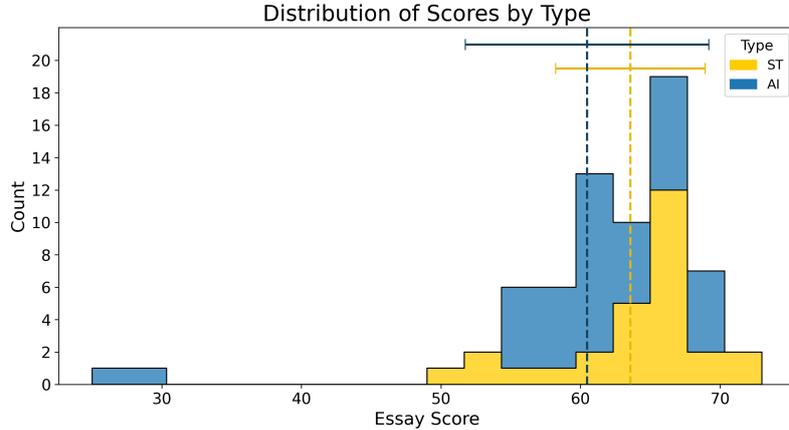
were student submissions, subsequently gauging the authorship on a 4-point Likert scale ranging from 'Definitely AI-authored' to 'Definitely human-authored'. Informed consent was obtained for all participants and the study was performed in accordance with relevant guidelines and regulations. This study received the following ethics approval: University of Oxford, Central University Research Ethics Committee [CUREC] approval reference: R88431/RE001.

### 3 Results

#### 3.1 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 to distinguish AI-generated essays. For this purpose, we employed a computational tool, GPTZero (Tian, 2023), and additionally, asked the blinded markers to ascertain whether an essay was penned by a human or an AI.

The essays were scored based on the standard UK grading system, ranging from 0–100, where a score of 65 signifies a good 2:1. 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 about 0.10. Given that this p-value comfortably exceeds our chosen alpha threshold of 0.01, we conclude that these score differences could be attributed to chance. Thus, AI-generated essays performed comparably well to human essays. The distribution of scores by type is illustrated in Figure 1. Notably, there are two outliers within the AI scores (25 and 30), both marked by the same individual. This indicates that AI, akin to human students, can sometimes underperform.

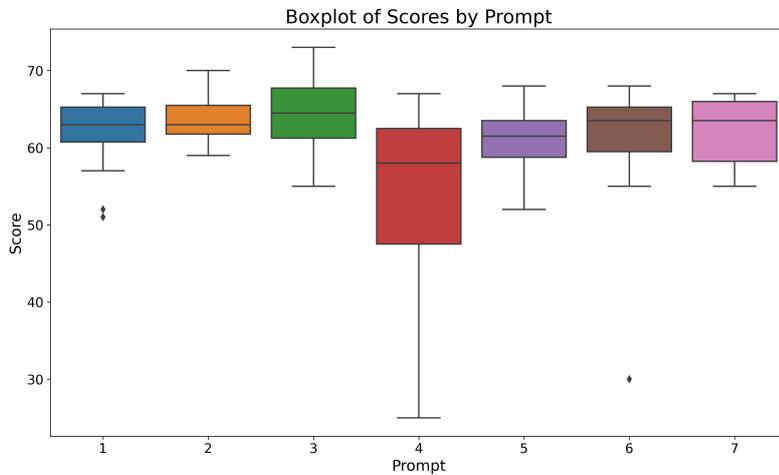


**Fig. 1** Histogram depicting the distribution of scores for AI (in blue) and students (in yellow).

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, Greatorex, &

Bell, 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 Figure 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 Figure 1, correspond to prompts 4 and 6, respectively.



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

### 3.2 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 GPT-4 was not capable of. A nuanced consideration of minute details of poetic 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<sup>1</sup>. This aligns

<sup>1</sup>One marker specifically noted that the AI-generated answers ‘break things up into small paragraphs which respond to specific parts of the prompt’.

with the findings of Desaire et al. (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’ acknowledgement of some fundamental misapprehensions of meaning of the Old English or of a particular rhetorical device, or of irrelevant value judgments of ‘the poet’s prowess’ or ‘the annals of literature’. For example, though AI-outputs often noted ornamental alliteration as a rhetorical device in Old English poetry, on occasions the quote chosen to exemplify this device lacked alliteration, creating an odd juxtaposition of correct knowledge incorrectly applied<sup>2</sup>.

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.

### 3.3 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. The resulting confusion matrix revealed that the overall identification rate stood at 79.41%, suggesting a reasonably good ability to differentiate between the two sources. However, it’s crucial to emphasise that out of the 14 misclassified essays, 10 were AI-generated but mistakenly identified as human-authored. Figure 3 visualises the classifications assigned by the markers against the essays’ actual authorship.

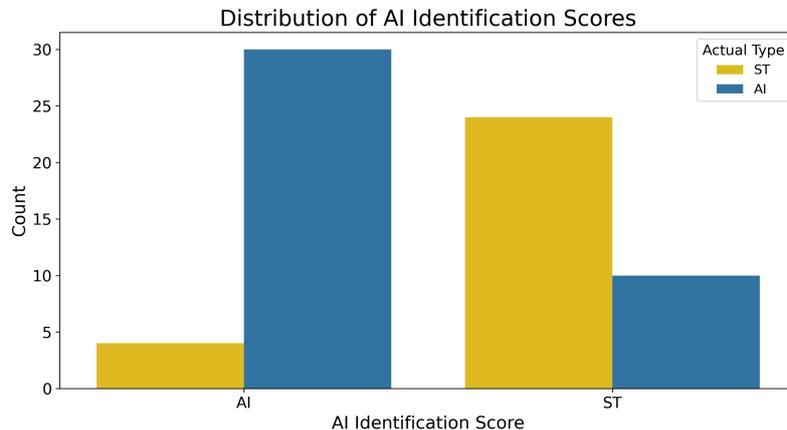
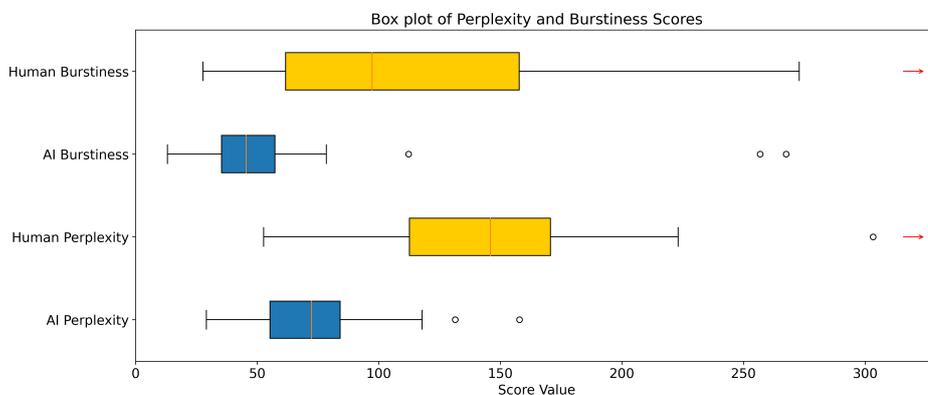


Fig. 3 Visualisation of AI identifications distribution.

We complemented human evaluations by using GPTZero to calculate Perplexity and Burstiness scores for all essays. Recent benchmarks have shown this technology to

<sup>2</sup>Two AI essays correctly identified *repe 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 pohton* ‘intended to strike’ (“Beowulf”, line 800b), where there is no alliteration.

be proficient at identifying human-authored texts, though its accuracy in classifying AI-generated text was somewhat lower (Liu, Yao, Li, & Luo, 2023). Intriguingly, Liu’s study revealed that AI-authorship was more readily discernible in Physics compared to the Humanities. Perplexity measures how well a model’s predicted probability distribution mirrors the true outcome distribution: a lower score indicates better predictability. Burstiness, on the other hand, measures the frequency of rapid event occurrences. Our findings revealed that AI essays generally had higher Perplexity scores than student essays. Conversely, while Burstiness median scores were similar for both groups, student essays exhibited a broader range. As visualised in Figure 4, a box plot of the scores for Burstiness and Perplexity found with GPTZero for human and AI essays captures the majority of the data ranges. However, it is worth noting that four human essays presented outlier scores not shown in the figure for clarity. These outliers had Burstiness and Perplexity scores considerably higher than the others, with two essays reaching Burstiness values exceeding 1600. Subsequent t-tests yielded t-statistic = -4.798, p-value =  $9.526 \times 10^{-6}$  for Perplexity, and t-statistic = -2.762, p-value = 0.007 for Burstiness. Given our 1% alpha threshold, both metrics displayed significant differences between AI and student essays.



**Fig. 4** Box plot of AI and human scores for Burstiness and Perplexity, with certain human outliers indicated but not fully depicted.

### 3.4 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, yett 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, Gamage, & Mills, 2022).

## 4 Discussion

### 4.1 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 & Hardy, 2023). 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.

## 4.2 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 evolving issues, when asked about the use of the term ‘Anglo-Saxon’ (Rambaran-Olm & 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 accessibility. 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.

## 4.3 Impact on assessment

In the short term, alterations to assessments could prevent the exploitation of some of the biggest vulnerabilities (Susnjak, 2022). 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 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 multi-modal 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.

#### 4.4 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 Figure 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<sup>3</sup>. 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 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.

---

<sup>3</sup>Full conversation available in supplementary materials.

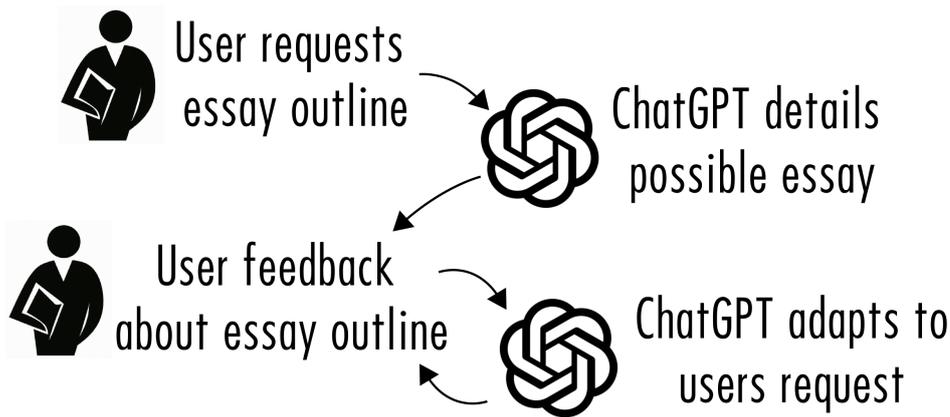


Fig. 5 Co-piloting with ChatGPT to write an essay.

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. 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.

## 5 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 cannot currently 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. Further complicating matters, computational metrics like Perplexity and Burstiness showed considerable overlap between human- and AI-produced content.

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.

## 6 Declarations

### 6.1 Ethics

This study received the following ethics approval: University of Oxford, Central University Research Ethics Committee [CUREC] approval reference: R88431/RE001. All human participants gave informed consent to be a part of this study including awareness of any outputs.

### 6.2 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.

## References

- Abd-Elaal, E.-S., Gamage, S.H., Mills, J.E. (2022). Assisting academics to identify computer generated writing. *European Journal of Engineering Education*, 47(5), 725–745,
- Baird, J.-A., Greatorex, J., Bell, J.F. (2004). What makes marking reliable? experiments with uk examinations. *Assessment in Education: Principles, Policy & Practice*, 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. *International Journal of Educational Technology in Higher Education*, 20(1), 1–22,
- Desaire, H., Chua, A.E., 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 Reports Physical Science*, ,
- González-Calatayud, V., Prendes-Espinosa, P., Roig-Vila, R. (2021). Artificial intelligence for student assessment: A systematic review. *Applied Sciences*, 11(12), 5467,
- Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C.C.T., Del Giorno, A., Gopi, S., ... others (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. *Journal of Information Systems Education*, 33(1), 98–108,
- Jakesch, M., Hancock, J.T., Naaman, M. (2023). Human heuristics for ai-generated language are flawed. *Proceedings of the National Academy of Sciences*, 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., Sekler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... others (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274,
- Kim, Y., Soyata, T., Behnagh, R.F. (2018). Towards emotionally aware ai smart classroom: Current issues and directions for engineering and education. *IEEE Access*, 6, 5308–5331,
- Köbis, N., & Mossink, L.D. (2021). Artificial intelligence versus maya angelou: Experimental evidence that people cannot differentiate ai-generated from human-written poetry. *Computers in human behavior*, 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. *Education and Information Technologies*, 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. *Nature Human Behaviour*, 1(3), 0028,
- McArthur, J. (2023). Rethinking authentic assessment: work, well-being, and society. *Higher education*, 85(1), 85–101,
- OpenAI (2023). *Gpt-4 technical report*.
- Oxford, U.o. (2022). *English language and literature prelims handbook - university of oxford*. <https://oess.web.ox.ac.uk/files/ellprelimshandbook2022-2311pdf>. (Accessed: 17 September 2023)
- Pedro, F., Subosa, M., Rivas, A., Valverde, P. (2019). *Artificial intelligence in education: Challenges and opportunities for sustainable development* (Tech. Rep.). Paris, France: UNESCO.
- Rambaran-Olm, M., & Wade, E. (2022). What’s in a name? the past and present racism in ‘anglo-saxon’ studies. *The Yearbook of English Studies*, 52(1), 135–153,
- Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., ... Weston, J. (2021, April). Recipes for building an open-domain chatbot. *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main volume* (pp. 300–325). Online: Association for Computational Linguistics. Retrieved from <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 September 2023)
- Ryznar, M. (2022). Exams in the time of chatgpt. *Washington and Lee Law Review Online*, 80, 305,
- Sadasivan, V.S., 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. *Educational Research Review*, 27, 90–109,
- Sharples, M. (2022). Automated essay writing: An aided opinion. *International Journal of Artificial Intelligence in Education*, 32(4), 1119–1126,
- Shermis, M.D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a united states demonstration. *Assessing Writing*, 20, 53–76,

- Smolansky, A., Cram, A., Radulescu, C., Zeivots, S., Huber, E., Kizilcec, R.F. (2023). Educator and student perspectives on the impact of generative ai on assessments in higher education. *Proceedings of the tenth acm conference on learning @ scale* (pp. 378–382).
- Susnjak, T. (2022). Chatgpt: The end of online exam integrity? *arXiv preprint arXiv:2212.09292*, ,
- Tian, E. (2023). Gptzero. <https://gptzero.me/>. (Accessed: 17 September 2023)
- Vajjala, S. (2018). Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education*, 28, 79–105,
- Yeadon, W., & Hardy, T. (2023). The impact of ai in physics education: A comprehensive review from gcse to university levels. *arXiv preprint arXiv:2309.05163*, ,
- Yeadon, W., Inyang, O.-O., Mizouri, A., Peach, A., Testrow, C.P. (2023). The death of the short-form physics essay in the coming ai revolution. *Physics Education*, 58(3), 035027,
- Zawacki-Richter, O., Marín, V.I., Bond, M., Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1), 1–27,