

CURRICULUM DEVELOPMENT AND ASSESSMENT

ChatGPT and academic integrity: a case study of assessment vulnerability in biological and environmental sciences

 Danielle Hinchcliffe,  Chrysanthi Fergani, and  Susanne R. K. Zajitschek

School of Biological and Environmental Sciences, Liverpool John Moores University, Liverpool, United Kingdom

Abstract

The rise of generative artificial intelligence (AI) has accelerated content creation across sectors, including education and academia. While these tools can streamline workflows and support learning, their use by students to complete assignments poses risks to academic integrity. This case study quantitatively evaluated the vulnerability of existing coursework assessments and exams to cheating using generative AI. Specifically, we compared ChatGPT-4's performance to student grades on 131 assessments from 40 modules in the School of Biological and Environmental Sciences at a UK University. ChatGPT excelled in "exam-like" assessments, including coursework-associated tests, exams (particularly those involving multiple choice), and, at some levels of study, in essay-style assessments. Conversely, AI performance and hence vulnerability were low on assessments requiring in-class data collection, collaboration, or the creation of reports, posters, or presentations. Overall, ChatGPT-produced work was often superficial and vague but better than expected, posing risks to assessing learning in all fields within the biological sciences, including physiology. While generative AI holds promise as a tool to help students understand, organize, and structure, it raises concerns around misuse. Designing creative, individualized tasks that require genuine intellectual engagement is essential, with greater emphasis on ethical conduct and the importance of integrity. Synthesizing information across multiple biological scales and applying clinical or experimental scenarios is recommended to develop higher-order physiological reasoning in students. Continuous, collaborative monitoring of generative AI performance on assessments should become part of routine assessment design and evaluation, given the continuous development and improvement of these tools. In cases where basic knowledge testing remains necessary, reversion to in-person examinations is recommended to replace online testing.

NEW & NOTEWORTHY This study reveals that current assessments in biological and environmental sciences are highly vulnerable to generative artificial intelligence (AI) misuse, which threatens assessment validity. ChatGPT-4 either outperformed or approximated students in exam-style testing and on some essay assignments but struggled with tasks involving data collection, collaboration, and presentations. The authors recommend redesigning assessments to require deeper reasoning, ethical engagement, and clinical or experimental application, alongside ongoing monitoring of AI capabilities and increased use of in-person exams.

artificial intelligence; assessment; pedagogy; quality education

INTRODUCTION

The recent rise of freely available and easily accessible generative artificial intelligence (AI) tools has prompted opportunities and challenges for various sectors, including education. AI tools have the power to lead to more efficiency and reduce time-consuming workloads, as they offer significant potential to assist in improving work through the generation of drafts, outlines, and analysis pathways (1, 2). However, due to the ability of generative AI to create unique, human-quality text, images, and code (e.g., Ref. 3), its potential impact on assessments in higher education is a matter of growing concern. Some of the implications AI poses for academics, as researchers and educators, are 1) a loss of rigor and compromised authenticity of academic work; 2) work

intensification, despite intentions to reduce workload, as AI tools lead to higher expectations for output and having to evaluate and verify AI usage; 3) a reluctance to work collaboratively if AI tools can serve as an alternative; and 4) digital disparities due to differential access to and use of AI tools. In the context of physiology education, unique challenges for students involve the integration of concepts from molecular to whole body scales and understanding often complex data and dynamic systems. AI tools can assist students in breaking down complexities, but the vulnerability of assessments that can test this complex, integrated understanding is a largely untested concern.

The implications of AI tool mis- or overuse carry predictable consequences, including a loss of academic autonomy due to overdependence on AI for content creation, which



Correspondence: S. R. K. Zajitschek (s.r.zajitschek@ljmu.ac.uk).
Submitted 30 October 2025 / Revised 4 February 2026 / Accepted 2 July 2026



risks reducing human intellectual contribution and leading to a homogenization of thought and research. This reliance also risks the devaluation of the academic, as AI becomes capable of performing increasingly complex academic tasks. Additionally, there is the risk of AI eroding the unique role universities traditionally have played in nurturing human ingenuity, creativity, and critical thinking (4). AI is impacting the way in which universities assess students. Now more than ever, we are challenged to discourage malpractice involving the use of generative AI in our assessments (5, 6). The ability of AI to produce high-quality content raises questions about the authenticity of student work, undermining the integrity and validity of academic evaluations. If generative AI can be used to generate essays, research papers, or code without detection, it could undermine the purpose of assessments as measures of student learning and understanding (7).

Evaluating AI performance on a wide range of assignments can help develop strategies for credible assessments and to mitigate the risks of academic dishonesty. This project, encompassing assessments within the Biological and Environmental Sciences at a UK University, including core modules in Physiology and related disciplines, aimed to evaluate the performance of ChatGPT-4 on 131 distinct assessments. The goal was to quantitatively identify assessment types most vulnerable to AI misuse and, critically, to determine if assessment resilience correlates with tasks requiring higher-order physiological reasoning, data interpretation, and experimental application, to inform the development of robust, credible assessment strategies and mitigate academic dishonesty risks for wider adoption by module leaders.

METHODS

At the start of the academic year 2023/2024, module leaders (MLs) were contacted about the opportunity to investigate the performance of ChatGPT on available assessments. Interested MLs reached out to the corresponding author, and a research assistant (RA) enrolled in the 40 participating modules. The RA was able to access necessary course materials and assessment information for each module. Our initial aim was to grade AI submissions mixed in anonymously with student submissions to minimize bias for or against AI-produced content by graders. However, due to data protection issues, the preferred option of within-module anonymous participation was not possible for the RA. We recognize that awareness that a submission was AI generated rather than student created may have introduced some bias in assessments that were not automatically graded in Canvas. However, we asked MLs for an objective assessment of the Chatbot-produced work, in line with student submissions, and to provide a short explanation for why a particular grade was given to AI-produced work [except multiple-choice tests or exams, which were automatically graded by our virtual learning environment (VLE) based on correct and incorrect answers]. To avoid accidental interference with the course content, we exported the identified assessment items into a sandbox module, where deadlines and access restrictions were lifted to allow the RA to asynchronously produce work using ChatGPT. Prompts for ChatGPT adopted a “minimal interaction” approach. Prompts were based solely on the official assessment briefs and

submission requirements, with no supplementary information, interpretation, or iterative refinement applied to the initial AI output. This standardized approach controlled for human variation, providing a direct baseline measure of the AI's inherent capability to respond to typical assessment instructions. While this may not represent how most people interact with chatbots, it allowed insights into AI performance without adding too much variation based on human individuality. This approach also intentionally excluded any clarifications that students might normally seek during classes. By not adjusting prompts to account for such interactions, we ensured that all AI inputs reflected only the information available to a student who relies exclusively on the published assessment documentation. The premium version of ChatGPT-4 (OpenAI, model GPT-4, released March 2023; AI answers generated between February and June 2024) was used for all assessments to allow completion of the full range of tasks, particularly those involving advanced functionalities, including data analysis, which were not available in the free versions at the time of the study. Settings were not changed from default mode for the version obtained [no tools: browsing/plugins off; memory/custom instructions: off/empty, temperature/randomness: fixed (no user control available)]. Answers provided by ChatGPT were entered back into Canvas manually and submitted for grading. Multiple-choice questions (MCQ) exams and Canvas quizzes were automatically graded; reports, essays, talk slides, and posters were sent to MLs for marking. The ML grade justifications were double-checked by the RA and the corresponding author to ensure consistency with the established marking rubrics and to confirm that feedback strictly adhered to the stated criteria for student work, thereby minimizing potential bias. All ChatGPT-scored grades were collated and combined with assessment grades achieved by students at the end of the academic year.

Liverpool John Moores University (LJMU) Research Governance Approval was obtained and registered as 251211LJMUREC201. The project did not require full ethical approval, as no personal data were collected and no information related to individual students was obtained. All data used in this project were fully anonymous.

This study included 40 modules and 131 different assessments covering a range of formats: exams, essays, reports, posters, talks, journal clubs, online tests, practical skills assessments, blog articles, and portfolio components. All assignments from participating modules were included (see Supplemental Table S2 for the list of participating modules) and grouped within the assessment types provided in Table 1. We identified ~20% of assessment as AI-proof “by design”: all in-person practical skills assessments ($n = 2$) as well as in-person handwritten exams/tests ($n = 13$), individual live presentations where AI reliance would not have helped preparation ($n = 6$), and reports requiring substantial amounts of individual data collection ($n = 9$) were assigned AI scores of “0.” Retaining rather than removing these assessments allowed us to include valuable information on student performance in the absence of AI misuse. For in-person, PC-based exams, the AI submission was included for grading as a baseline to investigate the hypothetical effects of AI misuse under exam conditions, but we acknowledge the low probability of genuine AI reliance in a proctored environment. We

Table 1. Definition of assessment types and approach to marking of artificial intelligence-generated content

Assessment	Definition	Grading Approach for AI Submissions
Skills Test	Practical assessment, hands-on skills, or lab/field performance. Typically online, may have been invigilated or taken in own time, incorporating MCQ, short answer questions, or calculations. While heavily based on practical experience, the online delivery makes it vulnerable to external resource use, including AI.	Automatic zero MCQ: automatic grading in VLE Short-answer questions: graded by ML and treated like student
Poster	Assessment of scientific poster (grading may include presentation skills or assess visual and scientific content only).	General/visual/scientific content: treated like student Individual subcomponents (oral skills): zero
Seminar	Mostly group-based presentation, where individual contribution, presentation skills, and teamwork components are assessed alongside scientific content.	General/visual/scientific content: treated like student Individual subcomponents (oral skills): zero
Report	Individual reports, typically based on data collected or provided during in-class practicals. Success often requires analysis and interpretation of specific, nonpublic data sets. See Supplemental Table S1 for typical structure.	Graded by ML and treated like student
Essay	May be part of coursework or final exam but involves a structured written reply to a prompt, typically requiring synthesis and critical evaluation. When posed as exam often prepared under in-person exam conditions.	PC based: graded by ML and treated like student Exam hall/handwritten: automatic zero
Exam	Formal, high-stakes assessment containing a majority of MCQ elements, offered online or in person.	PC based: MCQ: automatic grading in VLE Short-answer questions: graded by ML and treated like student Exam hall/handwritten: automatic zero

AI, artificial intelligence; MCQ, multiple-choice question; ML, module leader; VLE, virtual learning environment.

also tested AI performance on the components of in-person presentations (regardless of individual or group work) that required individual preparation and internet use. For these presentation assessments, the AI's presentation skills and teamwork components, which would have to be performed in-person by students, were scored as "0" on marking rubrics.

Grades from AI results and anonymized individual student grades were compiled for each assessment and module and together with assessment type information (see Table 1) and study level imported into R (RStudio 2025.050, Posit Software, R version 4.5.0) for visual analysis and significance testing. A linear mixed-effects model (LMM) was employed to analyze the relationship between AI and student performance depending on assessment types and level of study, using packages lme4 (8) and lmerTest (9), with degrees of freedom and *P* values calculated via Satterthwaite's approximation. The model was specified with "grade" as the dependent variable and a three-way interaction between "taker" (student vs. AI), "type" (type of assessment; Table 1), and "level" (level of study, L3–L7) as fixed effects (grade ~ taker * type * level). To account for the nested structure of the data and variability in grading across different subjects, "Module" was included as a random intercept (1|Module), controlling for nonindependence of assessments within the same module. Due to the imbalance in sample sizes of student versus AI grades for assessments within levels (students: *n* range = 7–1,447; AI: *n* range = 1–33; see Table 2), follow-up pairwise comparisons were conducted using Games-Howell post hoc tests using the package rstatix (10). For filtered post hoc comparisons, zero scores were excluded when they reflected nonattempts rather than meaningful assessment performance; this applied to both student and AI observations.

Plots were produced using ggplot2 (11), exploring differences between student and AI performance on different assessment types and across levels. We confirm that AI was

exclusively used for generating assessment outcomes and minor code cleaning but not for any stage of the manuscript preparation.

RESULTS

We report both descriptive statistics and model-based estimates; where relevant, we indicate whether results refer to filtered or full datasets. The significant interaction ($F_{10,7867} = 3.924$; $P < 0.0001$) in the overall model (which contained "Module" as random effect; for the full list of modules, see Supplemental Table S2) indicates that the difference in performance between students and ChatGPT is not uniform but is contingent on both the type of assessment (see Tables 1 and 2 for outcomes) and the level of study (L3: foundation year; L4–L6: undergraduate study, years 1–3; L7: postgraduate study, MSc). When looking at all assessment types across all levels (L3 to L7), students outperformed ChatGPT in most assessment categories (skills, poster, seminar, report, and essay; Fig. 1). In written PC-based tests or exams, AI and students received more similar grades (Fig. 1 and Fig. 2, bottom right; Figs. 3 and 4). While the post hoc results reported below and in Figs. 2, 3, and 4 are based on a filtered data set excluding all zero scores, these exclusions primarily reflect nonattempts rather than true indicators of performance. This includes AI zero scores assigned where an assessment was effectively nonattemptable under the study design, as well as student zero scores arising from nonattempts. Interpretations and overall results were similar in the full data set, for which post hoc results are reported separately in Supplemental Table S3. Unless otherwise stated, differences reported from descriptive statistics refer to the difference in mean grades (student vs. AI), whereas model estimates and associated confidence intervals reflect AI minus student values, with negative values indicating higher AI performance (Supplemental Table S3).

Table 2. Summary statistics across assessments, split by level of study and taker

Taker	Level	Type	<i>n</i>	Median	Mean	SD	95% CI
Student	L3	Exam	174	68	67.59	15.97	[65.21, 69.96]
AI	L3	Exam	2	79	79	24.04	[45.68, 112.32]
Student	L3	Report	173	77	73.41	17.27	[70.83, 75.98]
AI	L3	Report	2	61.5	61.5	33.23	[15.44, 107.56]
Student	L4	Exam	557	70	68.13	11.62	[67.17, 69.1]
AI	L4	Exam	7	82	80.29	12.05	[71.36, 89.21]
Student	L4	Report	423	63	65.56	18.71	[63.77, 67.34]
AI	L4	Report	4	44	38.25	16.09	[22.48, 54.02]
Student	L4	Seminar	29	68.26	68.27	10.46	[64.47, 72.08]
AI	L4	Seminar	1	65	65	NA	[NA, NA]
Student	L4	Skills	75	51	51.65	11.16	[49.13, 54.18]
AI	L4	Skills	NA	NA	NA	NA	[NA, NA]
Student	L4	Test	1,222	63.64	63.88	21.59	[62.67, 65.09]
AI	L4	Test	15	63	61.13	16.12	[52.98, 69.29]
Student	L5	Essay	81	58	56.68	9.83	[54.54, 58.82]
AI	L5	Essay	3	50	46.67	5.77	[40.13, 53.2]
Student	L5	Exam	271	54	53.38	13.34	[51.8, 54.97]
AI	L5	Exam	5	60	57.4	11.08	[47.69, 67.11]
Student	L5	Poster	220	60	59.85	11.07	[58.39, 61.32]
AI	L5	Poster	4	31	35.5	12.79	[22.96, 48.04]
Student	L5	Report	1,039	62	60.6	14.86	[59.69, 61.5]
AI	L5	Report	11	42	42.91	13.79	[34.76, 51.06]
Student	L5	Seminar	52	60	62.17	7.65	[60.09, 64.25]
AI	L5	Seminar	1	43	43	NA	[NA, NA]
Student	L5	Test	1,447	60	61.04	22.01	[59.91, 62.18]
AI	L5	Test	30	60.5	60.83	22.22	[52.88, 68.78]
Student	L6	Essay	347	60	60.07	12.19	[58.79, 61.35]
AI	L6	Essay	11	30	31.64	10.81	[25.25, 38.02]
Student	L6	Exam	98	62	60.07	12.31	[57.63, 62.51]
AI	L6	Exam	3	55	55.33	6.51	[47.97, 62.7]
Student	L6	Report	908	66	65.41	14.3	[64.48, 66.34]
AI	L6	Report	11	33	27.64	21.07	[15.18, 40.09]
Student	L6	Seminar	507	67	66.18	10.2	[65.29, 67.07]
AI	L6	Seminar	3	20	23.67	14.84	[6.87, 40.46]
Student	L6	Test	102	72	69.3	20.24	[65.38, 73.23]
AI	L6	Test	NA	NA	NA	NA	[NA, NA]
Student	L7	Essay	38	68	66.68	6.54	[64.61, 68.76]
AI	L7	Essay	3	67	61.67	10.12	[50.22, 73.11]
Student	L7	Report	7	71	72.14	8.99	[65.48, 78.8]
AI	L7	Report	1	15	15	NA	[NA, NA]
Student	L7	Skills	7	88	86.14	9.23	[79.31, 92.98]
AI	L7	Skills	NA	NA	NA	NA	[NA, NA]

Number (*n*), median, mean, SD, confidence interval (CI), and artificial intelligence (AI) versus student are shown. "Automatic zero" for AI was excluded. NA, not applicable.

AI versus student performance changes depending on the study levels, particularly for scientific reports (Fig. 2). Overall, this holds true for most coursework assessments (Fig. 2), where the discrepancy in grades increases with higher study levels, and students tend to perform better than AI. In reports, students performed significantly better than AI {Games-Howell post hoc tests: L4: student outperforming AI with a 27-point difference (95% CI [1.90, 52.71]), $P = 0.042$; L5: 18-point difference (95% CI [8.81, 27.37]), $P < 0.0001$; L6: 28-point difference (95% CI [16.27, 38.98]), $P < 0.0001$; L7: 56-point difference (too few AI data points for formal analysis); L3: 12 higher student grades than AI but no formal significance testing (too few datapoints in AI condition)}. The pattern of generally higher student than AI performance was also seen in seminars {L4: students 3 points more than AI on average; L5: 19-point difference with students outperforming AI; both excluded from formal post hoc comparisons; L6: students outperforming AI by 42 points (95% CI [5.81, 79.21]), $P < 0.0001$ } and in posters, which were only examined in few modules {L5 only: students outperformed AI with 24-point difference (95% CI

[4.17, 44.54]), $P = 0.031$ }. Skills assessments precluded AI involvement, leading to much better student than AI performance (L4: 51-point difference; L7: 86-point difference, not analyzed; only shown across levels of study; Fig. 1). In contrast, coursework-associated tests showed no significant difference between AI and student performance {L4: students scoring 3 points more than AI (95% CI [-5.60, 12.36]), $P = 0.434$; L5: 0.4-point advantage for AI (95% CI [-7.99, 8.74]), $P = 0.927$; L6: 63-point difference, excluded from formal post hoc comparisons, AI scored automatic zero}.

Overall, the gap between AI and student performance increases with the level of study in assessment types with low vulnerability for AI misuse (reports, skills, and seminars; Fig. 2; majority of data via reports). The relative resilience of the "report" assessment is particularly noteworthy for physiology education, as such tasks often require students to interpret unique datasets from core physiological laboratories (e.g., analyzing blood pressure responses or muscle contraction data), which are highly context specific and unavailable in ChatGPT's training data.

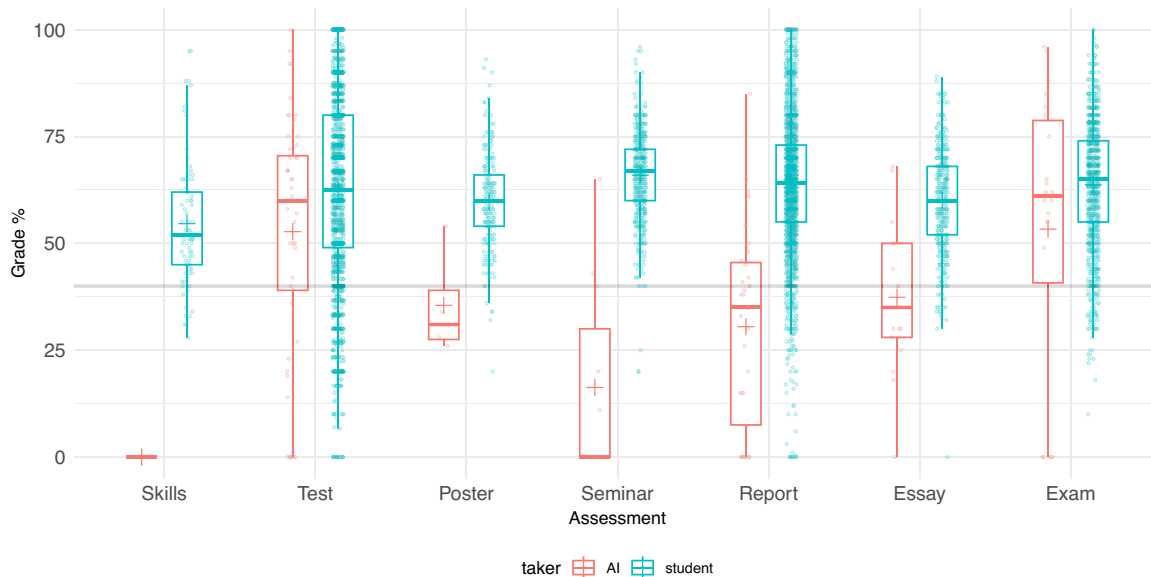


Figure 1. General overview of student and ChatGPT performance across all levels (L3–L7). Boxplots show medians and interquartile ranges; crosses indicate mean grades. Gray line: 40% pass grade. Included are all assessments, including in-person exams, on which artificial intelligence (AI) was assigned an automatic zero (see Table 1). Blue/green: student grades; orange/red: AI-scored grades.

Essays and exams, which generally test general knowledge rather than individual skills, were expected to reveal relatively high vulnerability, with AI either outperforming or not being different from student grades. Indeed, performance differences between students and AI appear to be small on most levels of study (Fig. 3, but see L6). Assessed essays on L5 and L7 indicate no significant difference in grades between the students and AI submissions {L5: students performing better than AI by a 10-point difference (95% CI $[-2.70, 22.73]$), $P = 0.082$; L7: 5-point advantage of student vs. AI (95% CI $[-19.03, 29.07]$), $P = 0.482$; but see L6: students

outperforming AI with 28 points difference (95% CI $[21.28, 35.93]$); $P < 0.0001$.

In exam-based assessments, AI performance was often similar to student performance, with some descriptive comparisons suggesting higher AI mean scores in lower-level exams (see also Fig. 4); however, these differences were not consistently statistically significant. Patterns were similar across all levels of study {Fig. 3, right, and Fig. 4; L3: AI outperforms students with a 11-point difference, no post hoc test performed; exam L4: 12 points more for AI than students (95% CI $[-23.30, -1.00]$), $P = 0.037$; L5: 4 points higher AI



Figure 2. Different coursework types across levels 3 to 7 (L3–L7) and comparison of artificial intelligence (AI) and student performance. Boxplots show medians and interquartile ranges, crosses indicate mean grades, and violin plots illustrate the distribution of grades. Instances where AI submissions received an automatic zero were excluded. Zero scores not representing meaningful assessment performance were excluded from this filtered dataset.

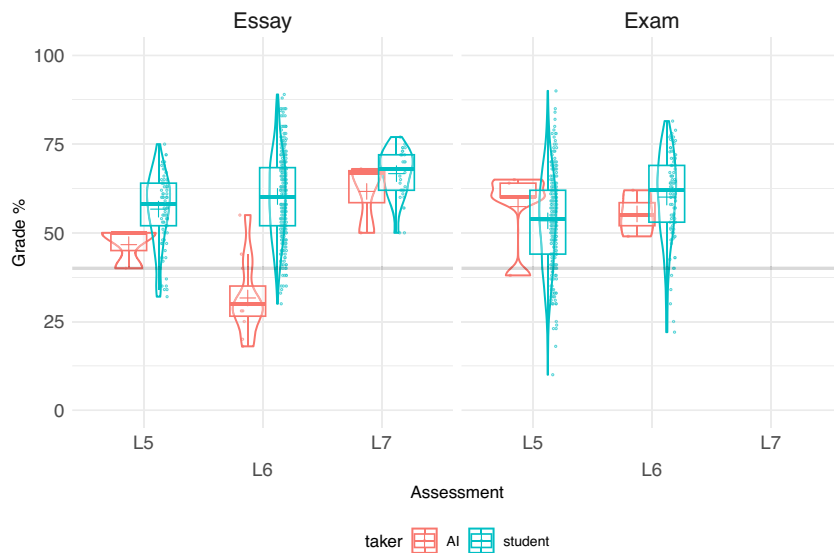


Figure 3. Essay- and exam-based assessments undertaken by students from levels 5 to 7 (L5–L7), comparing student performance with artificial intelligence (AI)-generated responses. Exams are usually taken in person on invigilated PCs. This dataset includes PC-based exams only; assessments where AI responses received an automatic zero due to manual submission were excluded. Boxplots show medians and interquartile ranges, with mean values indicated by crosses; violin plots illustrate the distribution of grades.

than student performance (95% CI $[-17.68, 9.65]$), $P = 0.466$; L6: students outperforming AI with a 5-point difference (95% CI $[-9.56, 19.34]$), $P = 0.334$).

Our results reflect the differences in targeting higher-order cognitive skills (analysis and evaluation, e.g., reports), compared to lower-order skills (remembering and understanding, e.g., L3 MCQs), explaining the AI's higher success rate in the latter.

DISCUSSION

We have evidenced that AI performance on university assessments varies substantially depending on different assessment types and levels of study. AI performs particularly well on essay- and exam-based assessments (which is particularly concerning when these are PC based or taken

remotely). However, its performance declines drastically in coursework and scientific report-based assessments that often involve individual experiments, data collection, or collaborative learning processes. The relatively robust performance of ChatGPT on essays and exams indicates that when assessing general knowledge, exam-type assessments and essays need to be reconsidered. While we implemented checks to mitigate marking bias, the inherent knowledge that a submission was AI generated may have subconsciously led markers to scrutinize the work for superficiality or lack of original insight, potentially leading to lower mean scores compared to if the work had been submitted by a student. This suggests that our reported vulnerability levels might be a conservative estimate. In addition, while our minimal interaction prompt strategy established a reliable baseline of AI capability, it does not incorporate iterative prompting and refinement techniques, which more closely reflect how users are likely to engage with these tools. It is expected that these would achieve significantly higher scores, particularly on essay-based assessments. Future research should investigate the efficacy of “prompt-aware” assessment design.

With the rise of AI use in university settings, the need for authentic assessment has never been greater. By this, we are referring to assessments that focus on applying knowledge and skills in practical, meaningful ways. This does not necessarily indicate the need for “AI-proofing”: AI may be used to enhance the assessment and supplement the students' existing skills. However, assessments must still test the student for core skills that are highly desirable by potential graduate employers in a specific sector. Key features of authentic assessments include real-world relevance, application over memorization, and process-oriented skills such as critical thinking, problem-solving, and creativity. Collaboration and teamwork are important factors valued by employers and expected in many real-world situations (12). There are two approaches we could take to implement AI into university education: efficiency (educator centered) and effectiveness (learner centered). For the latter approach, authentic assessment focuses on personalized learning, adaptive content, and using AI as an assistant (13). Our results provide quantitative evidence to support the learner-centered approach, as

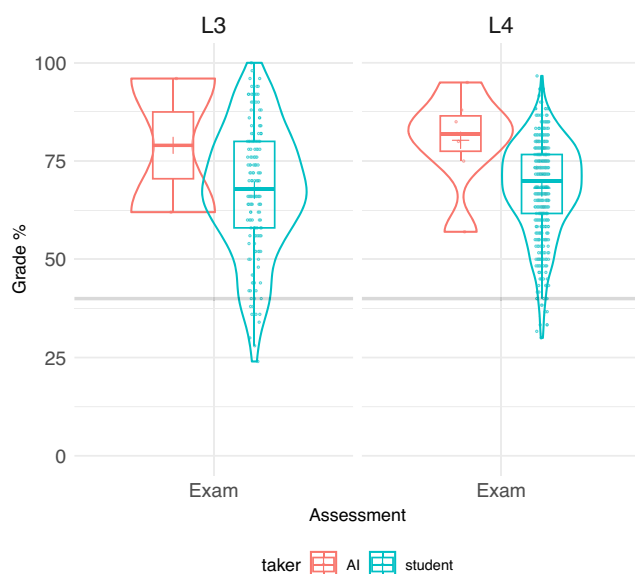


Figure 4. Exam-based assessments for level 3 (L3) and level 4 (L4) modules, comparing student and artificial intelligence (AI) performance in PC-based exams. Boxplots show the median and interquartile range, crosses indicate mean grades, and violin plots illustrate the distribution of grades.

coursework and report-based assessments that involved individual or group-generated data consistently challenged the AI performance.

Scientific reports proved to be a highly resilient assessment format across all undergraduate levels. This resilience stems from the context-specific and iterative nature of scientific inquiry, which demands human critical thinking and original interpretation. These are skills that ChatGPT-4 did not reproduce well under the conditions tested here. This finding is highly relevant for physiology education: physiological concepts are best learned through active, hands-on data collection (in virtual or wet laboratories) followed by analysis. The necessity to integrate unique, real-world physiological data into the report forces students to move beyond pattern recognition to true physiological interpretation, a critical skill that generalist AI models cannot currently replicate. Furthermore, these reports require students to apply context-specific knowledge referenced from internal teaching materials. Ultimately, while AI can support the process of scientific writing (e.g., clarifying language or organizing ideas), it does not replicate the critical, contextual, and reflective nature of student-generated work inherent in a quality scientific report based on practical work.

Aside from scientific reports, teamwork-based and/or visual and/or oral presentations are relatively unsuitable for AI overreliance. Importantly, these types of assessments aid in developing critical human competencies and fostering peer learning (14–16). Teamwork requires good communication, conflict resolution, and division of labor, developing interpersonal skills (17, 18).

Unexpectedly, marks in online tests based on practical content and experiential material did not differ overall between AI and students. The design of truly authentic assessments requires an iterative process and the integration of multiple dimensions (19–21). While the integration of practical work constitutes an important cornerstone for authentic assessment in the biological and environmental sciences, our findings indicate that practical knowledge is often tested through general logic at lower levels. Therefore, approaches that involve self-assessment, peer interaction, and mandatory reflection on personal experiential learning may provide additional dimensions that support deeper learning and learner autonomy, if applied appropriately (22, 23).

Academic essay-style-based assessments (whether part of a timed final exam or coursework) are a traditional method of evaluating learners and require students to demonstrate knowledge and critical evaluation of content while also assessing language and writing skills. Traditionally, formulating insightful, cohesive, well-justified, and meticulously researched arguments to convey a focused message is associated with higher grades, whereas lower grades are indicative of surface-level cognitive products such as knowledge “listing.” Generative AI has been shown to enhance essay writing when used appropriately but constitutes poor academic conduct if students over-rely on its use for text and content generation.

On the one hand, AI-derived writing tools can strengthen the writing process via proofreading student-written text and making corrections to language and grammar (24, 25). For many students, this can undoubtedly lead to increased clarity and coherence in their written essay work. For example, students with learning disabilities, such as attention-

deficit/hyperactivity disorder, dyslexia, dyspraxia, and autism, which can affect writing skills (26), as well as nonnative English speakers (27), benefit greatly from AI tools.

On the other hand, using AI technology for writing academic essays poses serious challenges. A recent report by Freeman (28) states that 53% of UK students use generative AI to help them with assessments and 13% have used it to generate text, which is sometimes not further refined or edited. An overreliance on text-generating tools creates lazy and superficial learners (29), impeding skills in critical evaluation and communication. Our data indicate that broad, literature-review-type essays provide ideal circumstances for AI to score comparably well.

Because AI models are trained on biased datasets relating to political views, sexism, racism, and cultural stereotypes (30–32) and may lack accuracy or specificity (33), rigorous human critical evaluation is necessary to verify answers to prompts, linking assessment integrity back to curriculum design. For example, AI-generated text is expected to perform poorly if the prompt required 1) opinion-based argumentation (i.e., students are asked to support their opinions and perspectives on a controversial topic using published literature); 2) expansion on concepts, activities and/or data specifically highlighted in class or collected during a practical; 3) discussion and analysis of particular case studies or scenarios; and 4) text generation aimed at educating the public on a particular topic in a specific format (blog/vlog/podcast).

There is no denying that these powerful technologies will significantly impact the future of education. As educators and researchers, it is our responsibility to guide students in building the skills and knowledge necessary for their success. Banning AI tools altogether is not an option, as this risks hindering student progress and limiting their opportunities beyond the classroom. Instead, we need to learn how to harness the power of AI in ways that benefit our students' futures (34). Whether for formal assessments or independent study, students will inevitably and increasingly use AI to assist their learning. Integrating AI for research and organization is beneficial, so long as the student remains able to prove their own understanding of the subject matter. At the same time, a shift away from high-stakes, knowledge-recall exams to data-driven reports and collaborative projects necessitates a shift in faculty development, requiring module leaders to become experts in assessment design rather than simply content delivery. This requires institutional support for staff training in authentic, AI-resilient pedagogy, where continuous, collaborative monitoring of AI performance (as demonstrated in this case study) becomes a routine, institutional requirement. In this age of generative AI, assessment integrity has become a constantly shifting target, demanding ongoing adaptation and validation.

Conclusions

This study provides detailed evidence on how generative AI, specifically ChatGPT-4, performs across a wide range of traditional assessment types in biological and environmental sciences in higher education. Our analysis reveals that AI performs well in knowledge-based exams and traditional essay-style assessments, particularly at lower program levels and when assessments are delivered nonproctored (PC based or remote).

In contrast, AI consistently underperforms in tasks that require practical-based skills, individual or group-based data collection, collaboration, and context-specific critical analysis. These findings have important implications for academic integrity, assessment design, and pedagogical practice and necessitate a dual strategy for assessment redesign. First, assessments highly susceptible to AI, such as unsupervised online tests and broad essay assignments, must be replaced with proctored in-person examinations or radically redesigned to incorporate local, unique data and context-specific requirements. In physiology, this mandates a shift away from high-stakes, knowledge-recall exams focusing on well-documented pathways toward problem-solving scenarios that demand the synthesis of physiological variables and the application of integrated, whole-body concepts. Second, we must evolve towards AI-aware assessment formats that value authentic human competencies. Tasks that engage students in collaboration, complex problem-solving, and reflective processes (such as scientific reports, group seminars, and portfolios) are areas where AI reliance remains limited and where students can leverage AI tools for organizational or grammatical support while retaining their critical intellectual contribution. Thoughtful integration, supported by institutional-level pedagogical training, is essential for maintaining academic integrity while preparing students for an AI-augmented professional future.

DATA AVAILABILITY

The data that support the findings of this study are openly available on the Liverpool John Moores University (LJMU) data repository (<https://opendata.ljmu.ac.uk/id/eprint/243>; S. Zajitschek, Liverpool John Moores University, 2025).

SUPPLEMENTAL MATERIAL

Supplemental Tables S1–S3: <https://doi.org/10.6084/m9.figshare.31249807>.

ACKNOWLEDGMENTS

We thank research assistant Anthony Imhenkuomon for contribution to the project, LJMU's Teaching and Learning Academy for funding and technological support, and all participating module leaders for providing access to module assessment information, student outcomes, and additional grading efforts.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the authors.

AUTHOR CONTRIBUTIONS

S.R.K.Z. conceived and designed research; S.R.K.Z. analyzed data; D.H., C.F., and S.R.K.Z. interpreted results of experiments; S.R.K.Z. prepared figures; D.H. and C.F. drafted manuscript; D.H., C.F., and S.R.K.Z. edited and revised manuscript; D.H., C.F., and S.R.K.Z. approved final version of the manuscript.

REFERENCES

- Crawford J, Cowling M, Allen KA. Leadership is needed for ethical ChatGPT: character, assessment, and learning using artificial intelligence (AI). *J Univ Teach Learn Pract* 20: 3, 2023. doi:10.53761/1.20.3.02.
- Ghimire A, Pather J, Edwards J. Generative AI in education: a study of educators' awareness, sentiments, and influencing factors. In: *2024 IEEE Frontiers in Education Conference (FIE)*. Institute of Electrical and Electronics Engineers, 2024, p. 1–9. doi:10.1109/FIE61694.2024.10892891.
- King MR, chatGPT. A conversation on artificial intelligence, chatbots, and plagiarism in higher education. *Cell Mol Bioeng* 16: 1–2, 2023. doi:10.1007/s12195-022-00754-8.
- Watermeyer R, Lanclos D, Phipps L. Does generative AI help academics to do more or less? *Nature* 625: 450, 2024. doi:10.1038/d41586-024-00115-7.
- Department of Education (DfE). Generative Artificial Intelligence in Education (Online). GOV.UK, 2023. <https://www.gov.uk/government/publications/generative-artificial-intelligence-in-education> [2023 Aug 21].
- Farrelly T, Baker N. Generative artificial intelligence: implications and considerations for higher education practice. *Educ Sci (Basel)* 13: 1109, 2023. doi:10.3390/educsci13111109.
- Dabis A, Csáki C. AI and ethics: investigating the first policy responses of higher education institutions to the challenge of generative AI. *Hum Soc Sci Commun* 11: 1006, 2024. doi:10.1057/s41599-024-03526-z.
- Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Soft* 67: 1–48, 2015. doi:10.18637/jss.v067.i01.
- Kuznetsova A, Brockhoff P, Christensen R. lmerTest package: tests in linear mixed effects models. *J Stat Soft* 82: 1–26, 2017. doi:10.18637/jss.v082.i13.
- Kassambara A. rstatix: Pipe-Friendly Framework for Basic Statistical Tests (Online). 2023. <https://rpkgs.datanovia.com/rstatix/> [2026 Jan 30].
- Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, 2016. doi:10.1007/978-3-319-24277-4_9.
- Moore T, Morton J. The myth of job readiness? Written communication, employability, and the 'skills gap' in higher education. *Stud High Educ* 42: 591–609, 2017. doi:10.1080/03075079.2015.1067602.
- Pratschke BM. *Generative AI and Education: Digital Pedagogies, Teaching Innovation and Learning Design*. Springer, 2024.
- Boud D, Cohen R, Sampson J. *Peer Learning in Higher Education: Learning from & with Each Other*. Psychology Press, 2001. doi:10.4324/9781315042565.
- Boud D, Lee A. Peer learning' as pedagogic discourse for research education. *Stud High Educ* 30: 501–516, 2005. doi:10.1080/03075070500249138.
- Dochy F, Segers M, Sluijsmans D. The use of self-, peer and co-assessment in higher education: a review. *Stud High Educ* 24: 331–350, 1999. doi:10.1080/03075079912331379935.
- Järvelä S, Kirschner PA, Panadero E, Malmberg J, Phielix C, Jaspers J, Koivuniemi M, Järvenoja H. Enhancing socially shared regulation in collaborative learning groups: designing for CSCL regulation tools. *Educ Tech Res Dev* 63: 125–142, 2015. doi:10.1007/s11423-014-9358-1.
- Kagwesa AM. Peer interaction and learning: a study of higher education students initiated group work activity. *Int J Know Learn* 9: 179–193, 2014. doi:10.1504/IJKL.2014.068916.
- Ashford-Rowe K, Herrington J, Brown C. Establishing the critical elements that determine authentic assessment. *Assess Eval High Educ* 39: 205–222, 2014. doi:10.1080/02602938.2013.819566.
- Gulikers JT, Bastiaens TJ, Kirschner PA. A five-dimensional framework for authentic assessment. *Educ Tech Res Dev* 52: 67–86, 2004. doi:10.1007/BF02504676.
- Schultz M, Young K, K. Gunning T, Harvey ML. Defining and measuring authentic assessment: a case study in the context of tertiary science. *Assess Eval High Educ* 47: 77–94, 2022. doi:10.1080/02602938.2021.1887811.
- Fawns T, Bearman M, Dawson P, Nieminen JH, Ashford-Rowe K, Willey K, Jensen LX, Damşa C, Press N. Authentic assessment: from panacea to criticality. *Assess Eval High Educ* 50: 396–408, 2025. doi:10.1080/02602938.2024.2404634.
- Villarroel V, Bloxham S, Bruna D, Bruna C, Herrera-Seda C. Authentic assessment: creating a blueprint for course design. *Assess Eval High Educ* 43: 840–854, 2018. doi:10.1080/02602938.2017.1412396.

24. **Alharbi W.** AI in the foreign language classroom: a pedagogical overview of automated writing assistance tools. *Educ Res Int* 2023: 4253331, 2023. doi:10.1155/2023/4253331.
25. **Dale R, Viethen J.** The automated writing assistance landscape in 2021. *Nat Lang Eng* 27: 511–518, 2021. doi:10.1017/S1351324921000164.
26. **Zhao X, Cox A, Chen X.** The use of generative AI by students with disabilities in higher education. *Internet High Educ* 66: 101014, 2025. doi:10.1016/j.iheduc.2025.101014.
27. **Gayed JM, Carlon MK, Oriola AM, Cross JS.** Exploring an AI-based writing assistant's impact on English language learners. *Comp Educ Art Intell* 3: 100055, 2022. doi:10.1016/j.caeai.2022.100055.
28. **Freeman J.** Provide or Punish? Students' Views on Generative AI in Higher Education. Higher Education Policy Institute, Policy Note 51 (Online). 2024. <https://www.hepi.ac.uk/reports/provide-or-punish-students-views-on-generative-ai-in-higher-education/> [2025 Jul 10].
29. **Zhao X, Cox A, Cai L.** ChatGPT and the digitisation of writing. *Humanit Soc Sci Commun* 11: 482, 2024. doi:10.1057/s41599-024-02904-x.
30. **Deshpande A, Murahari V, Rajpurohit T, Kalyan A, Narasimhan K.** Toxicity in ChatGPT: analyzing persona-assigned language models. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, 2023. doi:10.18653/v1/2023.findings-emnlp.88.
31. **Motoki F, Pinho Neto V, Rodrigues V.** More human than human: measuring ChatGPT political bias. *Public Choice* 198: 3–23, 2024. doi:10.1007/s11127-023-01097-2.
32. **Rettberg JW.** How generative AI endangers cultural narratives. *Issues Sci Tech* 40: 77–79, 2024. doi:10.58875/RQJD7538.
33. **Williams A.** Integrating artificial intelligence into higher education assessment. *Intersect J Intersect Assess Learn* 6: 128–154, 2025. doi:10.61669/001c.131915.
34. **Tseng W, Warschauer M.** AI-writing tools in education: if you can't beat them, join them. *J China Comp Assist Lang Learn* 3: 258–262, 2023. doi:10.1515/jccall-2023-0008.