The Integrity of Psychology Assessments in the AI Age: A Critical Examination

Matthew Ivory[1]* (0000-0002-5296-5897), Samuel Finnerty[1](0000-0002-6617-3866), Kirsty Dunn[1](0000-0003-0863-8757), Richard Philpot[1] (0000-0002-0359-2123), and Lara Warmelink[1] (0000-0003-1218-9448)

[1]Lancaster University, United Kingdom

Statements and Declarations: The authors declare they have no relevant or financial interests that relate to the research described in this paper.

Funding: The Lancaster University Department of Psychology supported the research with a small, internal grant.

Data Availability: the datasets generated by this research are available at https://doi.org/10.17605/OSF.IO/ERYTS. They are embargoed until April 2026.

Author Contributions: **Matthew Ivory**: Conceptualization, Methodology, Investigation, Formal analysis, Writing - Original Draft, Writing - Review & Editing, Data Curation. **Samuel Finnerty**: Conceptualization, Methodology, Investigation, Writing - Original Draft, Writing - Review & Editing. **Kirsty Dunn**: Conceptualization, Writing - Original Draft, Writing - Review & Editing. **Richard Philpot**: Funding Acquisition, Conceptualization, Writing - Original Draft, Writing - Review & Editing. **Lara Warmelink**: Conceptualization, Writing - Original Draft, Writing - Review & Editing.

*Correspondence concerning this article should be addressed to Matthew Ivory, Lancaster University, Department of Psychology, Lancaster University, UK, LA1 4YD, UK, Email: matthew.ivory[at]lancaster.ac.uk

Abstract

Generative artificial intelligence's (Gen-AI) increasing prevalence within higher education raises concerns over its usage by students, particularly over their development in skills like creativity and critical thinking. One important component of the standard undergraduate programme are assignments, where skill development can be assessed and constructive feedback offered to enable students to continue to build upon their skills. In this case study, we submitted the assignments from a three-year Psychology undergraduate programme to ChatGPT using three levels of prompting – basic, intermediate, and advanced – to answer the question, can Gen-AI obtain a psychology bachelor's degree? Responses were reviewed by researchers experienced in assessment marking and deemed as pass or fail. We report on the finding that Gen-AI was capable of successfully passing almost all assessment types. Of 40 individual assessments, 36 were of passing quality. We discuss the implications of this susceptibility, where Gen-AI can be used positively as a learning tool but is simultaneously dangerous when it generates incorrect, but convincing, information. The positive marking approach enables incorrect values or basic formatting to be condoned reducing the desire or need for students to make significant effort to alter the Gen-AI content to evade detection. We recommend that traditional written assessments, where not invigilated, should be adapted either in terms of assignment or marking criteria to adapt to the aforementioned challenges.

Keywords: Gen-AI; assessment; higher education; generative AI; artificial intelligence

The Integrity of Psychology Assessments in the AI Age: A Critical Examination

Introduction

Generative Artificial Intelligence (Gen-AI) models, such as ChatGPT, Bard, Claude, and Gemini (amongst others) have gained worldwide attention for their ability to perform complex language and content generation tasks. Since Gen-AI's inception, its user base has grown exponentially (Vázquez-Cano et al., 2023). Following this trend, higher education (HE) students have also increasingly leveraged these tools to support them in their academic work (Ibrahim et al., 2023). Although Gen-AI offers many positive opportunities for teaching and learning in HE (see Farrokhnia et al., 2024; Lo, 2023), it also raises concerns (Lingard, 2023). For example, a reliance on Gen-AI may hinder the development of a student's own critical thinking, creativity, and writing skills (Farrokhnia et al., 2024; Stokel-Walker, 2022). Gen-AI may also reduce a student's motivation to complete their own work, and whether intentional or not, undermine academic integrity should work be submitted for assessments that is AI-generated (Farrokhnia et al., 2024; Lee et al., 2024). This threat to academic integrity is amplified by a growing market of targeted ads, social media posts, and resources teaching students how to use AI for assessments, as well as to 'humanise' AI responses, evade detection, and even directly 'cheat' with AI (Rudolph et al., 2023).

Increasingly, AI tools are influencing social and cultural development (Phillips et al., 2024; Watson & Romic, 2025) and being viewed as a positive part of a collaboration (Luther et al., 2024). As this perception changes, students are increasingly more likely to integrate Gen-AI into their work flows (Namatovu & and Kyambade, 2025), with varying levels in how they perceive it as academic integrity infractions (Lund et al., 2025).

In a recent study that tested how students engaged with Gen-AI when answering Multiple Choice Questions (MCQ) assignments, it was seen that students interacted differently depending on their general academic performance. Russell et al. (2025) reported that high performing students used AI assistance selectively, often following their own failed attempts, whereas lower performing students would default to use AI before trying to answer the questions themselves. Another group emerged, of those who did not use AI assistance despite multiple failed attempts on the assignments, again highlighting potential issues around successful navigation and use of Gen-AI appropriately. Others report that academic dishonesty around Gen-AI is more likely to occur in those who hold a positive perception of ChatGPT (Ofem et al., 2025), an effect consistent across sex, yet stronger for younger Nigerian students. In an Australian sample, Gruenhagen et al. (2024) found those who reported use of AI in assessments were less likely to view this as cheating, yet found no relationship between a wide range of demographic factors and likelihood of using AI on assessments. Across cultures, those in countries categorized as high uncertainty avoidant (UAI), were less likely to see the benefit of use of AI and over 3.5 times more likely to categorise use of tools as cheating than those in low UAI cultures (Yusuf et al., 2024). At this early emerging stage, large heterogeneity in factors and samples examined makes the relationship between student perceptions of AI, and misuse in assessments difficult to disentangle from cultural indices, level of experience with AI, or variance in higher education systems in clarity in institutional guidance. However, it is clear that there are both within- and between-culture differences that must be considered when integrating AI into educational policies.

Given the risk Gen-AI poses to the integrity of degree programs, and the technology's growing use by students – a trend expected to continue with growing social influence and with the tools becoming more ubiquitous, user-friendly, and effective (Acosta-Enriquez et al., 2024) – research has turned its attention to the extent to which Gen-AI can complete HE assessments. In an early example, Choi et al. (2022) examined how well ChatGPT could perform on university law exams, comprising essay-style and multiple-choice questions. By blind scoring ChatGPT's responses alongside actual student submissions it was found that the Gen-AI could achieve a below-average passing grade of C+. Here, ChatGPT typically performed better on essay questions than multiple-choice questions, but its performance was highly variable within both assessment types.

In the intervening years, similar evaluations of Gen-AI's capabilities to complete HE assessments have been conducted across many disciplines including Medicine (Kung et al., 2023), Physics (Yeadon et al., 2023), Psychology (Scarfe et al., 2024), Management Science (Terwiesch, 2023), Economics (Geerling et al., 2023), and Computer Science (Jalil et al., 2023). Although variations are found in the methodologies and study results, the consensus is that Gen-AI is more than competent at completing HE assessments and poses a significant threat to the academic integrity of the education sector (Cotton et al., 2024). Supporting these assertions, a recent and comprehensive evaluation of ChatGPT's performance across 32 university courses and a plethora of assessment types found that the Gen-AI scored comparably or better than students in 38% of occasions (Ibrahim et al., 2023). Once again there was variation in the success of the Gen-AI depending on the task at hand – showing strengths in structured and factual tasks but underperforming in tasks requiring higher levels of creativity and complex analytical thinking.

With the demonstrated and evolving proficiency of Gen-AI, research into the educational advantages and disadvantages of its use in assessment are naturally left behind. This has resulted in many universities and academic departments scrambling to 'AI proof' their assessments (Cassidy, 2023; Rudolph et al., 2023), to ensure that students are graded on their own contributions and that degree programmes are not devalued. In this move, many universities have turned to AI detection tools to vet assessment submissions and extract those with a high likelihood of being generated or enhanced by AI. While the evidence for the effectiveness of detection tools remains mixed (Chaka, 2024; Ibrahim et al., 2023; Walters, 2023), these are often used in combination with other more qualitative indicators of AI and conversations with students meaning that time associated with academic integrity assurance has soared. Proven cases of Gen-AI misconduct rose from 1.6 cases per 1,000 students in 2022-23 to 5.1 per 1,000 in 2023-24 (Goodier, 2025), highlighting the increased workload placed upon these integrity assurance teams. Further, this offers a single snapshot of the increased workload associated with AI investigations compared to 'traditional' pre-AI forms of plagiarism where the text can be directly found in existing sources.

There remains a lack of understanding for many in academia regarding the extent to which *specific* assessments are vulnerable to AI use and, given the typical broad range of assessments across a degree programme, the extent to which *a programme itself* overall is vulnerable to student Gen-AI use (see Pimbblet & Morrell, 2025 for a recent exception in physics). In particular, there is currently a lack of understanding about the susceptibility of the entire psychology undergraduate programme covering a broad range and style of assessments. The current study is the first to provide a comprehensive review of such a programme through a systematic application of differing levels of AI complexity (basic, intermediate, and advanced). In doing so, this novel study contributes to the emerging literature on AI-assessment interactions by providing Psychology-specific evidence at a programme-level, offering a unique insight into how collectively vulnerable assessments can potential create a vulnerable and weak degree. At this juncture, rapid improvements in Gen-AI continue to outpace institutional policy increasing assessment vulnerability and educator uncertainty. Without an understanding of how a degree programme is vulnerable, individual and institutional responses risk being reactive, inconsistent, or misaligned with current Gen-AI capability.

The current study

The current study adopted an instrumental case study design (Priya, 2021; Stake, 1995 Yin, 2017), employing a pass/fail evaluation to examine the extent to which Generative AI can produce passable submissions across an undergraduate psychology degree programme (Priya, 2021). An instrumental case study approach was appropriate as this represented a bounded case to examine a broader phenomenon (AI vulnerability and risks to academic integrity). The case comprised a single British Psychological Society (BPS) accredited Psychology degree at a UK university, bounded to all coursework-based assessments available to students during the 2023/2024 academic year. Each assessment served as an embedded unit of analysis, allowing evaluation of both assignment-level vulnerability and cumulative programme-level vulnerability.

The Bologna agreement ensured mutual recognition of degrees across the European area (EC, 2022) and several non-European countries also use the Bachelor/Master/PhD degree system (e.g. the US). However, no international agreement on outcome classification and grading systems has been reached. This means that the pass/fail borderline is the only internationally recognized definition of a higher education degree. However, UK universities, including the one in the case study, traditionally classify students' outcomes into five categories, known as degree classifications: First class, Upper second, Lower second, Third and Fails. The first four classification represent passes, yet only firsts and upper second are considered "good" degrees (Bachan, 2018). How these classifications are derived from individual assessment grades differs across universities and can lead to student inequalities (Smith et al. 2024). This study focused on the crucial and more clearly defined pass/fail borderline. This borderline is far more salient outside of the UK undergraduate system, where degree classifications are not used as thresholds for job requirements, and so a third class degree may be treated equal to a first class degree.

Building on prior work demonstrating that most AI-generated psychology submissions went undetected (94%) and were on average a half grade boundary higher than student-written submissions (Scarfe et al., 2024), the present study examines whether students using ChatGPT in a low-effort manner, could generate assignment content likely to meet passing standards, regardless of formatting. The resulting outputs were reviewed to determine their adequacy—whether they were likely to achieve a passing grade under typical departmental standards. These judgements were made by two researchers with extensive experience marking undergraduate psychology assessments across all years of the programme. In cases where submissions were initially insufficient to pass, the researchers explored more advanced prompting strategies, such as rephrasing or adding specificity to the task instructions, to assess whether performance could be improved.

The study was guided by two main research questions:

1. Can ChatGPT generate passable submissions for undergraduate psychology assessments?
2. Can advanced prompt engineering improve the pass rate of AI-generated submissions within a fixed time frame (one hour)?

Methods

Sample

The research team all work for a psychology department in the UK. The coding and assessment of AI-generated outputs were conducted by authors MI (worked in the department since 2020) and SF (worked in the department since 2019). Both have extensive experience marking undergraduate psychology assessments across all three years of the programme, including essays, reports, web-based assessments, and class tests, and are experienced in applying in departmental grading standards and typical student performance. Their expertise ensures that binary pass/fail judgements reflect realistic academic expectations. This department has slightly below average student numbers, a median staff to student ratio, but above average entry tariffs and research quality for a UK university (Guardian, 2023; Complete University Guide, 2025). The assessments included in this case study were all used for the three-year BSc Psychology degree in the academic year 2023/2024. All open answer assessments were marked with letter grades (passing grades of A+, A, A-, B+, B, B-, C+, C, C-, D+, D, D-, and failing grades F1, F2, F3, F4). University guidelines stipulate that across a cohort assessments should normally have a mean grade between a B+ and a C+. Closed answer questions are marked on a percentage scale. Like many psychology degrees in the UK, the curriculum and assessment strategy for this degree followed the guidelines for the British Psychological Society (https://www.bps.org.uk/). All coursework-based assessments were included. Exams, which are taken under time pressure and invigilation, and dissertations which are highly unique to each student, were excluded from the sample.

There were 16 types of undergraduate assessment, spread across the three years and covering 40 unique assessments (not counting individual web-based assignments per module). See Table 1 for a summary of how the assessments are spread over each year.

Table 1. Types of assessment as used in each academic year

Assessment Type	First Year	Second Year	Third Year
essay	✓	✓	✓
web-based assessment	✓	✓	✓
report	✓	✓	
critical report		✓	✓
class test		✓	✓
presentation		✓	✓
individual brief summary		✓	
research proposal		✓	✓
course assignment (creative application)			✓
paper review			✓
grant proposal			✓
course assignment (media analysis)			✓
poster			✓
abstract			✓

Assessment Type	First Year	Second Year	Third Year
case study			✓
reflective portfolio			✓

Assessments were collected using the student-facing side of the learning management platform, Moodle. This ensured no staff-restricted content was leaked into the conversations that could have influenced the AI responses. Assessment details were then populated using all the material available to students. This included core and supplementary documents (e.g. templates and further details). All documents were stored in a shared repository and were used as the reference source during the analysis. The transcripts can be accessed at: https://osf.io/eryts/?view_only=79f1bcabd97d4cac9ff9109ba159b3e0.

Long-form assessments comprised 65% of assessments tested (78% inclusive of exams and dissertation). These tend to require students to submit written content, such as essays or research reports. Reports, in the context of a psychology degree, typically refer to a research report that details a qualitative or quantitative study in a format like a journal article. These assignments are often multi-faceted requiring data analysis, followed by the interpretation and discussion of the results.

The remaining two assessment types were the web-based assessments and class tests which predominantly deploy multiple-choice questions, or fill-in-the-blank style responses, enabling automated scoring systems to be used. Class tests are invigilated and typically conducted in-person in computer labs minimising the risk of students being able to use Gen-AI, however the WBAs were unsupervised assessments completed outside of contact time or invigilation.

Procedure

To collect the data from ChatGPT, the first two authors independently started new conversations with ChatGPT for each module assessment. Personalisation settings were switched off to minimise contamination across conversations. The first two authors requested the Gen-AI to provide responses to the assignment brief. Most responses were generated using ChatGPT 3.5, with some using 4o when document uploads were required. At the time of writing, 4o was available to use freely for a limited number of uses per day and so is available to all undergraduate students. Conversations were generated in June and July 2024.

To evaluate the Gen-AI output, the assessments were evaluated by two experienced markers—the first and second authors—each with over four years of experience grading undergraduate psychology assignments. The focus was on quick, minimal-effort assessments of AI-generated outputs, reflecting the study's goal of determining whether ChatGPT could meet basic standards with limited intervention. The assessors classified each submission as either passable or not, without applying in-depth grading or providing detailed feedback, replicating the kind of time-constrained, practical, and binary decision-making often employed in initial marking (Bloxham et al., 2011).

Three levels of prompting—basic, intermediate, and advanced—were used to generate responses:

- Basic: Each assignment was entered as a single prompt containing the task's title, word count, assignment type, and academic level. If no reference list was generated, a second prompt was used to request one. This was usually completed within 5 minutes.
- Intermediate: Submissions that failed at the basic level were prompted more extensively, with each component (e.g., abstract, introduction, methods) requested separately. Prompts were unrestricted, and responses were refined until deemed passable. This was usually completed within 5 to 20 minutes.

- Advanced: Submissions that failed at the intermediate level, as well as one example of each assessment type, underwent advanced prompting. This included document uploads (e.g., PDFs), use of prompt engineering to improve the performance of ChatGPT to generate the intended output.
    - Prompt engineering was performed in a separate chat, where a tailored and structured prompt was designed to guide ChatGPT toward producing section-specific, detailed, and well-organised outputs. These prompts typically outlined the task structure (e.g., sections and their word limits), provided style requirements (e.g., lay summary or academic proposal), and included specific instructions, such as incorporating citations, addressing core readings, or responding in a particular tone or voice. This process allowed the researchers to maximise the Gen-AI's potential for generating outputs that were coherent, contextually relevant, and aligned with assessment requirements.
    - To maintain consistency with the study's focus on minimal-effort scenarios, the total time for the advanced prompting process—including the preparation of tailored prompts—was limited to one hour per assessment. This constraint was implemented to reflect practical conditions under which students might reasonably use ChatGPT to complete complex tasks within time limits typical of academic workloads or deadlines.

For MCQ assessments, only basic-level prompting was used, with questions and answer options input into ChatGPT without further refinement. ChatGPT's performance was measured as the percentage of correct answers.

Table 2. Action types available for each level of prompting

Action	Basic	Intermediate	Advanced
Assignment type	✓	✓	✓
Request references	✓	✓	✓
Specify academic year	✓	✓	✓
Word count	✓	✓	✓
Multiple prompts		✓	✓
Word counts per section		✓	✓
File upload		✓	✓
Prompt engineering			✓

Action	Basic	Intermediate	Advanced
Query content			✓
Request output revisions			✓

All assignments underwent the basic procedure. The intermediate prompting method was used for all assessments that were deemed a fail at the basic level. The advanced process was used for all those that failed the intermediate stage, as well as ensuring that at least one example of every assessment type was subject to the advanced prompting. See Table 2 for the different actions available in each level of prompting.

The separate responses from a conversation were then combined into a single document ready for reviewer assessment. No content was added or removed that was not provided by the Gen-AI. The output was evaluated with a binary pass/fail based upon the expert judgment of two researchers with experience in constructing, delivering, and assessing assignments at the undergraduate level. MCQs were scored as a percentage correct using scoring systems available to staff.

Results

This section is separated by assignment types where we report on the observations that arose from the procurement and assessment of the Gen-AI outputs.

It was observed that the Gen-AI was unable to adhere to word counts particularly well. Most basic level prompts resulted in a mean average of 75% of the total wordcount being used (min = 38%, max = 107%), highlighting the inability for the output to be consistently writing to the word limit. In the intermediate and advanced prompts, word counts could be requested for each section, which even when given word counts that amounted to the total limit, would exceed these limits with average wordcounts being 81%, with a minimum of 20% and up to a maximum of 200%.

The creative and applied nature of assessments that required students to creatively engage with novel tasks/information and apply psychological theory often posed as a limiter for the Gen-AI. They typically required more advanced prompts that guided the Gen-AI towards specific ideas or applications. In doing so, a core aspect of the assessment (the application of psychology in a creative manner) was removed from the Gen-AI and passed to the student. An example of this was one module (PSYC303) where students were required to describe a novel religion grounded in psychological theory. Multiple attempts (five) resulted in the same religion (Harmonia) being described across different prompting levels, and even when using different user accounts (the Harmonia religion was also observed in a real-world student submission by one of the authors). By providing Gen-AI with novel scenarios (such as with case studies, e.g. PSYC370) it was more than capable of building upon these and generating passable assignment content. In this example (see full prompt and response in the online materials), the assignment provided a rich description of a novel scenario, "What we know: Initial reports of road traffic collision (RTC) involving a white van and 2 smaller cars in middle section of tunnel. Traffic standstill – no vehicles allowed in tunnel at either side. Smoke coming from tunnel. Police colleague has set up a Rendezvous Point (RVP) outside mouth of Liverpool side of tunnel. No emergency services available immediately (2 police making their way on foot)… at the scene, there are a growing number of people late for meetings and irritated by the lack of information being provided". This description was followed by structured questions that provided obvious direction for answers, "Using the cylinder model as your framework, what are the different messages the on-the-scene police officer could communicate to the crowd?"

In assessments that required the critical analysis of existing work (such as ‘critical reports’ or ‘paper reviews’), the Gen-AI was incapable of generating passable content at the basic level by nature of our prompt levels. This was because without being supplied the file to be scrutinised, the Gen-AI generated false information. In a paper review assignment (PSYC305), the basic prompting reviewed a nonexistent paper, "The Role of Social Interaction in Early Language Acquisition: Evidence from Neuroimaging Studies" by Johnson, Smith, and Taylor (2022) including an unresolvable DOI in its reference. Only with more advanced techniques, such as file uploads could the Gen-AI create passable outputs. This may be the result of the original research papers not being included within the Gen-AI training data and so it cannot access the relevant information it was trained upon. Even when given the paper to use, the Gen-AI still generated basic mistakes about the original article but not at a level dissimilar to those mistakes that students would make, meaning that they are not distinct markers of Gen-AI use, and are often condoned in marking criteria. For example, in the advanced prompting for PSYC305, a paper was chosen and uploaded, and the Gen-AI response intimated that R-squared values should have been reported (which they were), and fabricated information such as the sample size, stating that 378 participants were involved whereas the paper reported a total of 329 participants. These "errors" would likely be picked up by an attentive marker but would not be considered close to AI misconduct.

Out of the five assessments that were referred to as ‘essays’, only one passed at the basic level (a first-year cognitive psychology topic), but all passed at the intermediate level. During the researcher's reading of the answers, it was noted that many of the arguments presented within the Gen-AI essays were very basic and lacked references, or any complexity and criticality that an assessor would expect from higher grade submissions. Despite this, the arguments presented were typically sound and based on current psychological thinking. One aspect of apparent criticality can come from leveraging previous research and appropriately backing up claims using citations. By default, the Gen-AI would produce a relatively small reference list (between three and nine items for basic prompts, and between seven and 14 for intermediate/advanced prompts). Again, when reviewing the responses, it was noted that arguments typically lacked sufficient references to indicate a thorough attempt at the assignment. The claims within the AI-generated output were sometimes considered to be unsubstantiated, which is often an indicator of a lower-performing student. For example in the basic level of an essay on social processes in the Black Lives Matter movement (PSYC203), the output contained the following non-referenced claim, " Realistic conflict theory (RCT), another important concept in social psychology…" which would be noted by a marker as requiring a citation.

For assignments that required visual media output, such as posters and presentation slides, the Gen-AI was incapable of producing sufficient content as the model could not produce complete presentation slides nor poster content. Adhering to our criteria for assessing the success of the AI output, these could not pass as they did not generate the slides needed but the content that it offered was sufficient that a student would simply need to copy and paste the content into slides. More recent versions of ChatGPT will not only produce the content, but they can also generate downloadable PDF and presentation slide files. This is not a feature limited to ChatGPT as there are numerous other software tools that could produce visual media suitable for poster assignments.

Some assignments require students to conduct data analysis before reporting and interpreting the results in their writeup. Where the Gen-AI does not have access to the dataset, it would often note that it may be more accurate if it did have access, but it would then proceed to simulate data and analyses resulting in incorrect information (see the PSYC123 report at both basic and intermediate level for examples). This behaviour poses a risk for students, as the presentation of values may seem ‘good enough’ for those who are misusing the Gen-AI and do not necessarily have the time or the understanding to check the output. The susceptibility of the assignment is also increased where markers are encouraged to condone incorrect values and positively mark, focusing on student efforts to use the correct analyses and interpretation. Further Gen-AI is prone to the generation of incorrect information and the hallucination of sources and references that do not exist. This leaves the student at risk of not only submitting content that lacks credibility but also learning misinformation that may then be used in later examination. These inaccuracies can be spotted but require a marker who is expert on the given topic, and/or a thorough and time-consuming review of the accuracy of reference section where the current marking criteria call for a more efficient review concerned more with the formatting of such things. The Higher Education Policy Institute (HEPI) Student Generative AI Survey 2025 shows a jump from 53% to 88% of students making use of AI in assessments in one year. At the most extreme category of use (without editing), there is an increase on 2024 from 5% to 8%. With 23.900 students starting a psychology degree in the UK each year (Palmer et al., 2021), a more thorough review of hallucinated or incorrect referencing takes significant time resource. Without provision of this resource, in both student learning and course validity are at risk.

Short Form Assessments

Multiple Choice Questions (MCQs) were typically deployed within the department in class tests or as part of web-based assessments that could be scored automatically. Class tests are invigilated and conducted in-person in computer labs minimising the risk of students being able to use Gen-AI, however the WBAs are often weekly, unsupervised assessments taken in the student's own time.

Across all individual WBAs, the mean average score was 70% (min = 35, max = 100, SD = 17.30), highlighting their susceptibility to being answered by Gen-AI. When grouped by module, average accuracy increased to 74% (min = 58, max = 93, SD = 14.90).

As would be expected from Gen-AI built upon the vast quantity of declarative information available on the internet, the questions most easily answered were declarative questions. However, the Gen-AI was also capable of making choices based on *likely* answers rather than interpreting the question. For example, some modules presented images of linear model outputs and required students to identify key values from this. Despite not providing the image to the Gen-AI, it was often able to identify the correct answer based on the question text and provided answers by selecting the most *likely* value. For example, if asked to identify the *p*-value of a statistical test, it would choose the answer that was within the bounds of possible values (between 0 and 1). Similarly, providing images can be enough for the Gen-AI to choose the correct answer, such as with statistics-based content where students may be presented with equations and asked to explain specific information, but by naming the equation, the Gen-AI does not need access to the image to determine the correct answer. For example, in the prompt shown in table 3, figure 1 was displayed on-screen to students which was not uploaded to ChatGPT.

Table 3. A prompt and response with the Gen-AI highlighting that it was capable of answering an MCQ correctly without a provided image.

Source	Text
Prompt	Please examine the following formula: We know that this is a between-participants F statistic ratio because: a. Of the 'F' in italics, on left of equals sign b. It includes individual differences on both the numerator and denominator, whereas a within-participants ANOVA does not. c. It includes random (residual) errors on both the numerator and denominator, whereas a within-participants ANOVA does not. d. No. We cannot tell whether this is a within or between participants F statistic ratio e. There are degrees of freedom association with it
Response	This is a between-participants F statistic ratio because of the following reason: **b. It includes individual differences on both the numerator and denominator, whereas a within-participants ANOVA does not.**

$$F = \frac{\text{treatment effects + individual differences + random (residual) errors}}{\text{individual differences + random (residual) errors}}$$

Figure 1. The equation presented to students but not included in a ChatGPT prompt. Despite its omission, ChatGPT was successful in answering the question correctly.

Some class tests used a ‘higher MCQ’ format, where scenarios were described (such as patients presenting symptoms in PSYC317) before a series of questions were asked. The Gen-AI was adept at breaking down the scenarios and typically choosing the correct answers, scoring as highly as 87%, highlighting that the application of declarative knowledge to novel scenarios is a task very easily managed by Gen-AI when given the same information as is given to students.

When presented with multiple answers, the Gen-AI would sometimes arrive at a decision in a manner that presented as process of elimination. Gen-AI does not possess actual thought, and it is therefore not the case that AI determines the answer by true process of elimination (indeed, it does not determine the answer at all, rather it puts words in a string based on statistical probability). Despite this, where it imitates human though processes, it can make the answer look more convincing to a human reader. Interestingly, in one interaction (PYSC317 question eight) the response identified that option B was correct as it went through each option, but when it reported the final decision at the end of the response, it chose an incorrect answer, highlighting that despite correctly identifying the right answer, it 'hallucinated' and perhaps got waylaid by its own content. Table 4 provides this interaction.

Table 4. A prompt and response example of where the Gen-AI provides the wrong answer after identifying the correct one initially. Whether a user would read the entire response or skip to the emboldened decision would determine whether they answered this correctly.

Source	Text
Prompt	Cocaine was assessed as less harmful either to users or to others, or both, than alcohol and heroin. Which of the following one or more statements about cocaine are correct? a. In the CNS, the effects of cocaine are mediated by its ability to bind to dopaminergic postsynaptic receptors. b. In the CNS, the effects of cocaine are primarily mediated by its ability to inhibit the reuptake of monoamines. c. Cocaine is not very addictive as there are very few obvious withdrawal symptoms once a binge is terminated. d. As cocaine enters the brain slowly its reinforcing effects are small. e. Both a) and c).
Response	Let's evaluate each statement about cocaine to determine which one(s) are correct: **a.** This statement is partially correct. Cocaine primarily exerts its effects by inhibiting the reuptake of dopamine, serotonin, and norepinephrine, leading to increased levels of these neurotransmitters in the synaptic cleft. This action contributes significantly to its reinforcing and euphoric effects. **b.** This statement is correct. Cocaine's main mechanism of action is to block the reuptake transporters for dopamine, serotonin, and norepinephrine, thereby prolonging their action in the brain and enhancing their effects. **c.** This statement is incorrect. Cocaine is highly addictive due to its rapid onset of euphoria and its ability to quickly lead to tolerance and dependence. Withdrawal symptoms from cocaine can be intense and include cravings, depression, fatigue, and disturbed sleep. **d.** This statement is incorrect. Cocaine actually enters the brain very rapidly, which contributes to its intense reinforcing effects and addictive potential. The rapid onset of euphoria after using cocaine is a key factor in its abuse liability. **e.** Since statement a) is partially correct and statement c) is incorrect, the correct answer is:

a. In the CNS, the effects of cocaine are mediated by its ability to bind to dopaminergic postsynaptic receptors.

Statistics modules in the Psychology department use the programming language R to apply statistics content and MCQs in these modules can include questions that require students to run R code and answer questions based on the outputs. The Gen-AI struggled to answer many of these questions simply because it was not given access to the datafile and could not execute code. It was able to produce code for the user to run themselves but often used object names that were not present in the actual datafiles, resulting in an inability of the user to execute the code without modification. This being said, when taking the code provided by the Gen-AI and running it (with the addition of having the dataset named correctly), the accuracy was 60% for these types of questions, and it did not score higher mostly because it guessed at column names. This would not be a complex fix for a student to provide the correct column names (or even provide the Gen-AI with the dataset column names), further driving up accuracy rates.

A final observation made was that when the Gen-AI provided incorrect answers, the associated explanations were often convincing, particularly on content outside the expertise of the researchers. An example of this can be seen in the first question of the PSYC205 WBA 6 transcript (shown in Table 5). To a marker unfamiliar with developmental psychology, the response is convincing enough to be believed. Such explanations would be detrimental to students who may try using Gen-AI as learning aids or revision tools (outside the context of using them to answer assignments) where the information given is incorrect but wholly convincing.

Table 5. Question one from the 6th WBA in the 205 module. The correct answer is (a).

Source	Text
Researcher	Remember the green sloth from our previous weeks? It’s back for the final time! Sam is 10 months of age when he visits the zoo in a foreign tropical country during a family trip. There he encounters an unusual animal he has never seen before – a Pygmy Three-Toed Sloth. The sloth has started to slowly and deliberately groom itself. Sam has never seen such grooming behaviours before. What is Sam most likely to do as he observes the sloth? a. Look at his caregiver b. Over-imitate its actions c. Use ostensive cues d. All of the above
ChatGPT	Given Sam’s age and the scenario described, the most likely option is “d. All of the above.” At 10 months old, Sam is likely to exhibit a combination of behaviors as he observes the sloth grooming itself: - **Look at his caregiver:** Infants often look to their caregivers for cues on how to react to new or unfamiliar situations. Seeing a new animal like the Pygmy Three-Toed Sloth, Sam might glance at his caregiver for reassurance or to gauge their reaction. - **Over-imitate its actions:** Over-imitation is a common behavior in young children where they mimic even unnecessary or irrelevant actions they observe in others. If Sam is intrigued by the sloth’s grooming behavior, he might attempt to imitate it, even if he doesn’t fully understand the purpose or mechanics behind it. **Use ostensive cues:** Ostensive cues are social cues that draw attention to specific objects or actions. Sam might use these cues, such as pointing or vocalizations, to draw attention to the sloth or to communicate his interest or confusion about its behavior. So, in summary, Sam is likely to exhibit all of the behaviors mentioned in the options provided.

Discussion

In this case study, we explored a set of undergraduate psychology assignments with the intention to identify their susceptibility to large language models, specifically OpenAI’s ChatGPT. Researchers tasked ChatGPT to generate responses at three levels of prompt complexity, before assessing the output as being of sufficient quality to achieve a passing grade. Our results show that all but four assessments were vulnerable to AI, replicating findings from

Scarfe et al. (2024) and answering the question as to whether Gen-AI can produce passable submissions. With 90% of assessments being passable in the present study, and Scarfe et al. (2024) reporting a rate of 94%, this suggests that psychology might be more vulnerable than other disciplines. In other disciplines, such as engineering, Gen-AI successfully answered 65% of examination questions (Borges et al., 2024). Our work contributes to the literature in strengthening the position that Gen-AI is a genuine risk to higher education (Lingard, 2023; Rudolph et al., 2023). Pithily, ChatGPT could obtain an undergraduate degree in Psychology.

Importantly, we found that MCQs ranged in quality of output, with scores as low as 35% and as high as 100%, an indication that Gen-AI should not be wholly trusted to provide the correct information. A reliance on Gen-AI as a learning tool in this context is misplaced. While it often performs at a passing level, at times it confidently produces inconsistent and often incorrect answers. Gen-AI may be best used when students possess an adequate background in the subject at hand (Shoufan, 2023). Without this prior knowledge, students may become misinformed. It was also noticeable that where images form part of the question (often screenshots of statistical test output tables), but the image was not provided to the Gen-AI, the Gen-AI could identify the correct answer based on the question text and the possible answers. This suggest that the answer options in these MCQs are not wholly dependent on the data provided and AI can extract the *most likely* option, even without the extra information.

Assignments requiring novelty and originality were more resilient to the Gen-AI output. Gen-AI is, in general, good at incremental creativity (Lee & Chung, 2024). However, creativity in a manner significantly departing from previous content was more difficult. This echoes previous findings by Ibrahim et al. (2023) where Gen-AI underperformed when tasks required higher levels of creativity or analytical thinking.

Historical Context of ChatGPT

Due to the development and speed in which Gen-AI is made available to the public, this section briefly provides a description of the capabilities of ChatGPT at the time of data generation, as well as how it has changed since then. In June of 2024, the free tier of ChatGPT enabled access to GPT-3.5 with limited usage of the advanced GPT-4o model which expanded capabilities allowing users to upload files/images and ask questions about them with the model. Once limits of 4o were reached, users were restricted to 3.5 until a time period had elapsed during which continuing conversations that involved file uploads were restricted. Paid users had access to a faster version, GPT-4. Since mid-2024, OpenAI have released several updates with the current offering (at time of writing) being GPT-5.2. 3.5 and 4o possess some basic "reasoning" features but it was easily derailed, which resulted in hallucinations. It was only in later versions that reasoning became a feature that offered multi-step processing and the ability to search the web for results. As the capabilities of Gen-AI develops, it is important that institutions are aware of the changes and how these may impact assignments. It is also worth noting that ChatGPT is not the only Gen-AI tool on the market, with others such as Claude, Gemini, DeepSeek to name a few. There are even specific tools for generating essays targeted at undergraduate students.

Future Developments for AI and Social Impact of AI

AI capabilities are growing rapidly and are likely to continue doing so in the near future (DSIT, 2025), as they access more training data and AI models become more sophisticated.

Research has shown that positive perceptions of AI, such as ease of use (Al Shamsi et al., 2022) or usefulness (Kashive et al., 2020), increase student's likelihood to use it for academic dishonesty (Ofem et al., 2025). This means that using AI for malpractice purposes is likely to become more attractive to students and harder to combat for educators, as ease of use and reliability increase. This leaves educators with a difficult dilemma. They must balance the value of allowing students to gain experience with using AI with the risks to academic integrity these tools present. Conversely, educators' decisions will affect the acceptance of AI in wider society. Many students will be first introduced to the ethics of AI in the classroom. Educators will also be crucial in facilitating students' ability to use AI to support their learning (Gado et al., 2022; Namaziandost & Rezai, 2024).

Limitations

One major limitation of this work is its likely underestimation of how students would truly utilise Gen-AI content for assessments. We placed multiple constraints on the quality of the final output, prompting with basic, intermediate, and advanced levels. We made no edits to the content, beyond joining sections of content produced in multiple exchanges. In practice, it is not unreasonable to expect students may rewrite or incorporate additions from their writings and elsewhere. Additionally, we did not use ChatGPT to its greatest capabilities. For example, it may have been more effective if we requested analysis code and then updated the code manually with correct variable names or offered it snippets of data to supply more accurate code. Our findings assume a naive student rather than a more sophisticated AI user. Despite limiting the possible interactions, the fact that 90% of the assignments were considered a passing grade indicates that the use of Gen-AI presents a serious and very credible threat to the suitability of the assignments.

As a result, the risks and susceptibilities that we identified may be underestimated or even misplaced in comparison to situations where students leverage Gen-AI in ways not sanctioned by the department.

Another limitation is that the assessors knew that all the content was AI-generated. Previous work has used a blind approach by assessing AI content and student content together (such as Choi et al., 2022). In the present study most long-form assignment responses were clearly atypical 'submissions' due to incorrect formatting styles and submission guidelines (for example, the absence of title sheets with a student's unique identifier). Our primary interest in this study, however, was to assess whether the content, irrespective of its presentation, was passable, not whether it could fool a human assessor. Our intent was not to compare AI and student submissions (which would have required human intervention and editing to fit formatting requirements), but rather our efforts were a representation of the absolute minimum effort, intended to determine the quality of Gen-AI responses to simple prompts. A further and important implication is that Gen-AI responses may be more likely to achieve a passable submission where marking criteria and routine marking practices over-reward academic writing precision over evidence of deeper comprehension and critical evaluation. Our study did not directly test this. However, the current findings do stress the importance of assessment designs and making criteria that are tightly aligned with intended learning outcomes, and which require demonstrable evidence of understanding, rather than allowing fluent academic writing to be sufficient for a pass.

Recommendations for Educators

1) Require version control. One method to discourage Gen-AI malpractice could be through requiring student to create an auditable “evidence trail”. This could be achieved in various ways, but one option includes encouraging version history tools to be adopted, which is included in many different software tools. Being able to inspect the timeline in which an assignment developed and how it changed would allow educators a greater insight into cases of suspected Gen-AI misconduct. It may also deter students from using AI as the amount of information that needs to be generated increases exponentially. This would also provide students with workflow tools that can be used in other modules and projects to protect against file corruption or lost data.
2) Require reproducible assessments. Reproducibility, and the reproducibility crisis (Reed, 2018), are often covered in an undergraduate psychology degree. For assignments that contain data analysis, educators can assess students on their ability to generate reproducible documents. Like recommendation 1, this provides an auditable trail that can be examined in suspected Gen-AI situations. Incorporating this additional, but minimal requirement (those who have completed the assignment properly would have analysis files to hand) might also deter students from using AI, as the amount of information that needs to be generated increases. Additional benefits might also include strengthening the students focus on process rather than outcome.
3) Creative assessment should ask for more than incremental creativity. With AI’s strength in incremental creativity (Lee & Chung, 2024), greater levels of creativity should be focused on more as this would require students to provide novel ideas.

4) Consider integrity reviews for students who submit work with contextually irrelevant information. Some of the Gen-AI responses produced in the current dataset referred to concepts and methods that were not delivered as part of the taught content. Whilst excellent students are expected to go beyond the taught material and incorporate their own learning, where the assignments include incongruent ideas and lack conceptual clarity (perhaps they refer to confirmatory factor analysis in a module that focuses on linear regression) this can be a sign of Gen-AI use (Busch & Hausvik, 2023), and, thus, a relevant factor in reducing susceptibility of AI in assessments.
5) Include referencing in marking criteria. One common issue with much of the Gen-AI output in our case study was the quality of the referencing. The output was inconsistent as to whether it would produce the reference list within the first prompt (easily mitigated by asking for the reference list), and often the number of references included was deemed insufficient to be awarded higher grades. Previous work with ChatGPT has highlighted that the actual quality of the references can be very low, with one study finding that only 7% of supplied references were both real and relevant to the prompt (Bhattacharyya et al., 2023). Noticing these errors and incorporating them into the mark will encourage students to ensure they are correct.
6) Align assessment criteria and marking guidance with the intended learning outcomes. Given Gen-AI's capability to produce high quality academic writing, it is important to ensure that assessment passes are awarded for evidenced attainment of the intended learning outcomes, rather than on writing fluency alone.

Recommendations for Institutional Leaders and Policy Makers

1) Resource Management for academic integrity and assessment reform. In a context of increasing academic workload and workforce contraction, time is a diminishing resource. Education institutions must prioritise protected time for markers to review integrity in proportion with student numbers, and reform assessment in the context of AI, where required. The identification of Gen-AI influenced work, and subsequent meeting with student, is skilled, time-intensive work that cannot be sustainably absorbed into already compressed academic roles. Without adequate resourcing, courses risk lowered quality and standards, resulting in graduates who do not reliably possess the skills and competencies that employers and professional bodies expect. Over time, this risks discrediting degrees at programme and sector level, likely to translate directly into reduced student recruitment and retention (Hemsley-Brown et al., 2015). Adequate resourcing for academic integrity is thus, likely to uphold educational quality, reputational stability, and long-term institutional sustainability.
2) Clear AI policies. Recent surveys show variance in AI policies in institutions across countries with just 19% overall indicating their institution had a formal AI policy in place and 42% with policy in development. The recent HEPI (2025) survey reports 80% of students in the UK felt that their institution had a clear AI policy. These policies though, have been found to have significant variance in comprehensiveness, educational approach, enforcement and approach (Atkinson-Toal et al., 2024). Long-term commitment of higher education regulators,

professional bodies, and institutions to collaborate in offering norms and standardisations of effective strategies for modernising programmes whilst maintaining integrity would significantly support decision making in teaching and assessment.

3) Ensure assessors are trained well and using updated marking criteria. HE is often subject to strong resource pressures (Collins, 2025; Universities UK, 2025). Many markers are also under substantial time pressure. Ensuring that all assessing staff are aware of new assessments and marking guidelines and implementing them will be crucial.

Reflexivity Statement

All the authors involved in this manuscript are researchers and educators who are invested in the maintenance of the undergraduate curriculum delivered at our institution. During this case study, various conversations between the research team often circled back to an expression of concern for the future of psychology degrees and education in general, and specifically over how students interpret the role of Gen-AI. We also recognise that Gen-AI can be greatly beneficial to students and educators alike, when used appropriately.

Conclusion

In this exploration of the risks of AI for undergraduate assessments within a Psychology programme, we report that even with a relatively basic understanding of leveraging AI, it is possible to obtain content that would likely result in passing grades. With a little more effort, and with no resources unavailable to the average undergraduate student, nearly all the assessment types are susceptible to being passed solely using AI. Our approach scored assignments with a binary pass/fail response reflecting the manner in which assessors typically first view assignments. This approach meant the generated content was untouched or edited prior to being assessed, with a limitation to the generalizability of the findings as the content can only be considered through the pass/fail metric.

This AI susceptibility leaves the degree programme vulnerable to devaluation and risks the many hours that are spent on student feedback being rendered unlikely to serve the purpose of developing their own skills. The assessment types that were less vulnerable were those that required an in-person aspect, such as presentations or visual media (e.g. posters), or those that required more input than was possible during the hour time limit, such as the specifics of datasets for analysis. In the discussion of this work, we highlight the issue of how Gen-AI can mislead students with convincing responses despite being incorrect, and that the assessment vulnerability cannot be blamed solely upon student usage, but that the pressures of the HE domain mean that failings of the AI output may be condoned resulting in greater likelihood of AI content being given passing grades.

In answer to our research questions of ‘can ChatGPT generate passable submissions for undergraduate assessment’, the answer is yes (with the caveat this was assessed for minimum passing criteria), and that advanced engineering does indeed improve the pass rate of the AI-generated submissions.

References

Acosta-Enriquez, B. G., Ramos Farroñan, E. V., Villena Zapata, L. I., Mogollon Garcia, F. S., Rabanal-León, H. C., Angaspilco, J. E. M., & Bocanegra, J. C. S. (2024). Acceptance of artificial intelligence in university contexts: A conceptual analysis based on UTAUT2 theory. *Heliyon*, *10*(19), e38315. https://doi.org/10.1016/j.heliyon.2024.e38315

Al Shamsi, J. H., Al-Emran, M., & Shaalan, K. (2022). Understanding key drivers affecting students' use of artificial intelligence-based voice assistants. *Education and Information Technologies*, *27*(6), 8071–8091. https://doi.org/10.1007/s10639-022-10947-3

Atkinson-Toal, A. (2024). Student Value of a Transdisciplinary Approach to Curriculum Development. *Journal of Marketing Education*, *0*(0). https://doi.org/10.1177/02734753241288182

Bachan, R. (2018). *The drivers of Degree Classifications* (pp. 1–63). UK Standing Committee for Quality Assessment. https://www.universitiesuk.ac.uk/sites/default/files/uploads/Reports/drivers-of-degree-classifications.pdf

Bhattacharyya, M., Miller, V. M., Bhattacharyya, D., & Miller, L. E. (2023). High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content. *Cureus*, *15*(5), e39238. https://doi.org/10.7759/cureus.39238

Bloxham, S., Boyd, P., & Orr, S. (2011). Mark my words: The role of assessment criteria in UK higher education grading practices. *Studies in Higher Education*, *36*(6), 655–670. https://doi.org/10.1080/03075071003777716

Borges, B., Foroutan, N., Bayazit, D., Sotnikova, A., Montariol, S., Nazaretzky, T., Banaei, M., Sakhaeirad, A., Servant, P., Neshaei, S. P., Frej, J., Romanou, A., Weiss, G., Mamooler, S., Chen, Z., Fan, S., Gao, S., Ismayilzada, M., Paul, D., … EPFL Data Consortium. (2024). Could ChatGPT get an engineering degree? Evaluating higher education vulnerability to AI assistants.

Proceedings of the National Academy of Sciences, *121*(49), e2414955121. https://doi.org/10.1073/pnas.2414955121

Busch, P. A., & Hausvik, G. I. (2023). Too Good to Be True? An Empirical Study of ChatGPT Capabilities for Academic Writing and Implications for Academic Misconduct. *AMCIS 2023 Proceedings, 21*. Americas Conference on Information Systems. https://aisel.aisnet.org/amcis2023/sig_odis/sig_odis/21

Cassidy, C. (2023). Australian universities to return to 'pen and paper' exams after students caught using AI to write essays. *The Guardian*. https://www.theguardian.com/australia-news/2023/jan/10/universities-to-return-to-pen-and-paper-exams-after-students-caught-using-ai-to-write-essays

Chaka, C. (2024). Reviewing the performance of AI detection tools in differentiating between AI-generated and human-written texts: A literature and integrative hybrid review. *Journal of Applied Learning & Teaching*, *7*(1), 115–126. https://doi.org/10.37074/jalt.2024.7.1.14

Choi, J. H., Hickman, K. E., Monahan, A. B., & Schwarcz, D. (2022). ChatGPT Goes to Law School. *Journal of Legal Education*, *71*(3), 387–400. https://jle.aals.org/home/vol71/iss3/2/

Collins, D. E. (2025, April 24). *The great collapse of US higher education has begun*. Al Jazeera. https://www.aljazeera.com/opinions/2025/4/24/the-great-collapse-of-us-higher-education-has-begun

Complete University Guide. (2025, June 10). *Psychology Subject League Table 2026*. https://www.thecompleteuniversityguide.co.uk/league-tables/rankings/psychology

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, *61*(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148

DSIT. (2025, July 17). *Engines of AI primed to accelerate new breakthroughs, economic growth, and transform the UK into an AI maker*. https://www.gov.uk/government/news/engines-of-ai-primed-to-accelerate-new-breakthroughs-economic-growth-and-transform-the-uk-into-an-ai-maker

Erickson, R. A., Archer, A. A., & Fienen, M. N. (2024). A Reproducible Manuscript Workflow With a Quarto Template. *Journal of Fish and Wildlife Management*, *15*(1), 251–258. https://doi.org/10.3996/JFWM-24-003

European Commision. (2022). *The Bologna Process and the European Higher Education Area—European Education Area*. https://education.ec.europa.eu/education-levels/higher-education/inclusive-and-connected-higher-education/bologna-process

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2024). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International*, *61*(3), 460–474. https://doi.org/10.1080/14703297.2023.2195846

Fourali, C. (1997). Using Fuzzy Logic in Educational Measurement: The Case of Portfolio Assessment. *Evaluation & Research in Education*, *11*(3), 129–148. https://doi.org/10.1080/09500799708666923

Gado, S., Kempen, R., Lingelbach, K., & Bipp, T. (2022). Artificial intelligence in psychology: How can we enable psychology students to accept and use artificial intelligence? *Psychology Learning & Teaching*, *21*(1), 37–56. https://doi.org/10.1177/14757257211037149

Geerling, W., Mateer, G. D., Wooten, J., & Damodaran, N. (2023). ChatGPT has Aced the Test of Understanding in College Economics: Now What? *The American Economist*, *68*(2), 233–245. https://doi.org/10.1177/05694345231169654

Goodier, M. (2025). Revealed: Thousands of UK university students caught cheating using AI. *The Guardian*. https://www.theguardian.com/education/2025/jun/15/thousands-of-uk-university-students-caught-cheating-using-ai-artificial-intelligence-survey

Gruenhagen, J. H., Sinclair, P. M., Carroll, J.-A., Baker, P. R. A., Wilson, A., & Demant, D. (2024). The rapid rise of generative AI and its implications for academic integrity: Students' perceptions and use of chatbots for assistance with assessments. *Computers and Education: Artificial Intelligence*, *7*, 100273. https://doi.org/10.1016/j.caeai.2024.100273

Hemsley-Brown, J., & Oplatka, I. (2015). University choice: What do we know, what don't we know and what do we still need to find out? *International Journal of Educational Management*, *29*(3), 254–274. https://doi.org/10.1108/IJEM-10-2013-0150

Higher Education Policy Institute. (2025). *Student Generative AI Survey 2025*. https://www.hepi.ac.uk/reports/student-generative-ai-survey-2025/

Ibrahim, H., Liu, F., Asim, R., Battu, B., Benabderrahmane, S., Alhafni, B., Adnan, W., Alhanai, T., AlShebli, B., Baghdadi, R., Bélanger, J. J., Beretta, E., Celik, K., Chaqfeh, M., Daqaq, M. F., Bernoussi, Z. E., Fougnie, D., Garcia de Soto, B., Gandolfi, A., … Zaki, Y. (2023). Perception, performance, and detectability of conversational artificial intelligence across 32 university courses. *Scientific Reports*, *13*(1), 12187. https://doi.org/10.1038/s41598-023-38964-3

Jalil, S., Rafi, S., LaToza, T. D., Moran, K., & Lam, W. (2023). ChatGPT and Software Testing Education: Promises & Perils. *2023 IEEE International Conference on Software Testing,* *Verification and Validation Workshops (ICSTW)*, 4130–4137. https://doi.org/10.1109/ICSTW58534.2023.00078

Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies*, *28*(12), 15873–15892. https://doi.org/10.1007/s10639-023-11834-1

Kashive, N., Powale, L., & Kashive, K. (2020). Understanding user perception toward artificial intelligence (AI) enabled e-learning. *The International Journal of Information and Learning Technology*, *38*(1), 1–19. https://doi.org/10.1108/IJILT-05-2020-0090

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., Leon, L. D., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, *2*(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198

Lee, B. C., & Chung, J. (2024). An empirical investigation of the impact of ChatGPT on creativity. *Nature Human Behaviour*, *8*(10), 1906–1914. https://doi.org/10.1038/s41562-024-01953-1

Lee, V. R., Pope, D., Miles, S., & Zárate, R. C. (2024). Cheating in the age of generative AI: A high school survey study of cheating behaviors before and after the release of ChatGPT. *Computers and Education: Artificial Intelligence*, *7*, 100253. https://doi.org/10.1016/j.caeai.2024.100253

Lingard, L. (2023). Writing with ChatGPT: An Illustration of its Capacity, Limitations & Implications for Academic Writers. *Perspectives on Medical Education*, *12*(1), 261–270. https://doi.org/10.5334/pme.1072

Lo, C. K. (2023). What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. *Education Sciences*, *13*(4), 410. https://doi.org/10.3390/educsci13040410

Lund, B. D., Lee, T. H., Mannuru, N. R., & Arutla, N. (2025). AI and Academic Integrity: Exploring Student Perceptions and Implications for Higher Education. *Journal of Academic Ethics*, *23*, 1545–1565. https://doi.org/10.1007/s10805-025-09613-3

Luther, T., Kimmerle, J., & Cress, U. (2024). Teaming Up with an AI: Exploring Human–AI Collaboration in a Writing Scenario with ChatGPT. *AI*, *5*(3), 1357–1376. https://doi.org/10.3390/ai5030065

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, *6*(4), 462–472. https://doi.org/10.3758/BF03197480

Moseley, S. (2025, April 15). *Automating Deception: AI's Evolving Role in Romance Fraud* (Centre for Emerging Technology and Security Briefing Papers). https://cetas.turing.ac.uk/publications/automating-deception-ais-evolving-role-romance-fraud

Namatovu, A., & Kyambade, M. (2025). Leveraging AI in academia: University students' adoption of ChatGPT for writing coursework (take home) assignments through the lens of UTAUT2. *Cogent Education*, *12*(1), 2485522. https://doi.org/10.1080/2331186X.2025.2485522

Namaziandost, E., & Rezai, A. (2024). Special Issue: Artificial Intelligence in Open and Distributed Learning: Does It Facilitate or Hinder Teaching and Learning? *The International Review of Research in Open and Distributed Learning*, *25*(3), i–vii. https://doi.org/10.19173/irrodl.v25i3.8070

Ofem, U. J., Owan, V. J., Iyam, M. A., Udeh, M. I., Anake, P. M., & Ovat, S. V. (2025). Students' perceptions, attitudes and utilisation of ChatGPT for academic dishonesty: Multigroup analyses via PLS–SEM. *Education and Information Technologies*, *30*(1), 159–187. https://doi.org/10.1007/s10639-024-12850-5

Palmer, W., Schlepper, L., Hemmings, N., & Crellin, N. (2021). *The right track: Participation and progression in psychology career paths*. The Nuffield Trust. https://www.nuffieldtrust.org.uk/sites/default/files/2021-07/1625671007_nuffield-trust-the-right-track-pipeline-of-psychologists-web2.pdf

Phillips, N., Kalvapalle, S., & Kennedy, M. (2024). Beyond the Turing Test: Exploring the implications of generative AI for category construction. *Organization Theory*, *5*(3), 1–20. https://doi.org/10.1177/26317877241275113

Pimbblet, K. A., & Morrell, L. J. (2025). Can ChatGPT pass a physics degree? Making a case for reformation of assessment of undergraduate degrees. *European Journal of Physics*, *46*(1), 015702. https://doi.org/10.1088/1361-6404/ad9874

Priya, A. (2021). Case Study Methodology of Qualitative Research: Key Attributes and Navigating the Conundrums in Its Application. *Sociological Bulletin*, *70*(1), 94–110. https://doi.org/10.1177/0038022920970318

Reed, W. R. (2018). A Primer on the 'Reproducibility Crisis' and Ways to Fix It. *Australian Economic Review*, *51*(2), 286–300. https://doi.org/10.1111/1467-8462.12262

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, *6*(1), 342-363. https://doi.org/10.37074/jalt.2023.6.1.9

Russell, J.-E., Smith, A. M., George, S., Pratt, J., Fodale, B., Monk, C., & Brummett, A. (2025). Unlocking Insights: Investigating Student AI Tutor Interactions in a Large Introductory STEM Course. *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 451–461. https://doi.org/10.1145/3706468.3706524

Scarfe, P., Watcham, K., Clarke, A., & Roesch, E. (2024). A real-world test of artificial intelligence infiltration of a university examinations system: A "Turing Test" case study. *PLOS ONE*, *19*(6), e0305354. https://doi.org/10.1371/journal.pone.0305354

Shoufan, A. (2023). Exploring Students' Perceptions of ChatGPT: Thematic Analysis and Follow-Up Survey. *IEEE Access*, *11*, 38805–38818. https://doi.org/10.1109/ACCESS.2023.3268224

Smith, S., Sutherland, N., & Allen, D. (2024). Opening the Black Box of Degree Classification Algorithms: Towards a Research Agenda. *Teaching in Higher Education*, *29*(5), 1398–1408. https://doi.org/10.1080/13562517.2024.2302555

Stake, R. E. (1995). *The Art of Case Study Research* (1st edn). SAGE.

Universities UK. (2025, October 10). *Universities grip financial crisis – but at what cost to the nation?* Universities UK. https://www.universitiesuk.ac.uk/what-we-do/creating-voice-our-members/media-releases/universities-grip-financial-crisis-what

Stokel-Walker, C. (2022). AI bot ChatGPT writes smart essays—Should professors worry. *Nature*. https://doi.org/10.1038/d41586-022-04397-7

Terwiesch, C. (2023). *Would Chat GPT3 Get a Wharton MBA?* [White paper]. Mack Institute for Innovation Management at the Wharton School. https://mackinstitute.wharton.upenn.edu/wp-content/uploads/2023/01/Christian-Terwiesch-Chat-GTP-1.24.pdf

The Guardian. (2023). The Guardian University Guide 2024 – the rankings. *The Guardian*. https://www.theguardian.com/education/ng-interactive/2023/sep/09/the-guardian-university-guide-2024-the-rankings

UNESCO. (2025, September 2). *UNESCO survey: Two-thirds of higher education institutions have or are developing guidance on AI use*. https://www.unesco.org/en/articles/unesco-survey-two-thirds-higher-education-institutions-have-or-are-developing-guidance-ai-use

Vázquez-Cano, E., Ramírez-Hurtado, J. M., Sáez-López, J. M., & López-Meneses, E. (2023). ChatGPT: The brightest student in the class. *Thinking Skills and Creativity*, *49*, 101380. https://doi.org/10.1016/j.tsc.2023.101380

Walters, W. H. (2023). The Effectiveness of Software Designed to Detect AI-Generated Writing: A Comparison of 16 AI Text Detectors. *Open Information Science*, *7*(1), 20220158. https://doi.org/10.1515/opis-2022-0158

Watson, S., & Romic, J. (2025). ChatGPT and the entangled evolution of society, education, and technology: A systems theory perspective. *European Educational Research Journal*, *24*(2), 205–224. https://doi.org/10.1177/14749041231221266

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. https://doi.org/10.1088/1361-6552/acc5cf

Yin, R. K. (2017). *Case Study Research and Applications: Design and Methods* (6th edn). SAGE Publications.

Yusuf, A., Pervin, N., & Román-González, M. (2024). Generative AI and the future of higher education: A threat to academic integrity or reformation? Evidence from multicultural perspectives. *International Journal of Educational Technology in Higher Education*, *21*(1), 21. https://doi.org/10.1186/s41239-024-00453-6