International Journal of Artificial Intelligence in Education
https://doi.org/10.1007/s40593-024-00452-8

ARTICLE

ChatGPT in the Classroom: Evaluating its Role in Fostering Critical Evaluation Skills

Angela Oates[1] · **Donna Johnson**[1]

Accepted: 15 December 2024

Abstract

The increasing prevalence of artificial intelligence in educational domains raises both opportunities and challenges in the context of academic integrity and pedagogical efficacy. This study outlines an innovative project that investigates the use of ChatGPT as a tool for enhancing the critical evaluation skills of master's students in biomedical science. Using a dual approach combining AI-generated essay writing with subsequent student-led critical evaluation, this project sought to foster deeper critical evaluation skills in learners. By having participants critically assess AI-generated essays, supported critical evaluation based on peer-reviewed literature, the project aimed to deepen their evaluative skills. Outputs from the tasks were compared against academic benchmarks considering factors such as marks, writing, and overall quality. Participant perceptions were collected through a combination of a focus group session and an evaluation questionnaire. The key finding of this project was that while ChatGPT demonstrated proficiency in structural coherence and grammatical accuracy, it did not augment academic performance– participant marks for the AI-generated essays aligned closely with their overall module marks, showing no overall improvement. However, this study did see an increase in marks for participants' critical evaluations. This suggests that ChatGPT was more effective as an assessment tool when used for critical evaluation tasks, aligning with pedagogical emphasis on nurturing critical evaluation skills. User interaction with AI emerged as a significant variable that influenced the tool's efficacy, highlighting the need for a nuanced approach to its integration into educational settings. The study concludes that while ChatGPT offers promising avenues for both drafting and assessment, and demonstrated a high level of factual accuracy, it is not a substitute for human-led academic enquiry, and students preferred writing their own essays.

Keywords ChatGPT · Artificial intelligence · Assessment · Critical evaluation

Extended author information available on the last page of the article

Published online: 24 January 2025

Introduction

The emergence and subsequent surge in adoption of artificial intelligence (AI) across diverse sectors has left an indelible mark on contemporary society. Nowhere is this more evident than in education, which stands at the confluence of tradition and innovation (Gill et al., 2024; Tahiru, 2021; Zhai et al., 2021). When we speak of the contemporary digital age, we refer to an era dominated by rapid technological advancements and increased connectivity. This age has ushered in profound shifts in societal paradigms, and AI has been at the forefront of these transformations, carving out a distinctive niche. The traditional classroom, with its whiteboards and printed textbooks, has largely been replaced by digital interfaces, adaptive learning platforms, and AI-driven tools (Mallik & Mallik, 2017). This integration is not just for the sake of modernisation but has brought tangible benefit s(Ranasinghe & Leisher, 2009). Students can now access resources from any corner of the globe, teachers can tailor learning experiences to individual student needs, and educational administrators can streamline operations, all thanks to the capabilities provided by educational technology(Haleem et al., 2022).

In education, AI applications span multiple functions that address traditional educational challenges through intelligent automation, adaptive learning and personalised tutoring systems (Wang et al., 2024). Such applications extend from intelligent assessment and management tools, which provide real-time feedback to learners, to predictive profiling systems that enable educators to understand learners' strengths and areas for improvement before these emerge as issues. The application of conversational agents in particular, showcase a move towards emotionally intelligent interfaces capable of assessing and responding to learner's emotional states. These agents have the potential to mitigate stress in high-stakes assessments through interactions that mimic human empathy and support while still rigorously evaluating knowledge and understanding (Alaswad et al., 2023).

The recent development of ChatGPT and similar generative AI tools highlights the potential of these technologies to transform creation of teaching materials, engagement and understanding. These tools have quickly become popular tools in education due to their ability to provide accessible responses to a wide array of topics (Gill et al., 2024).Generative AI platforms are not just coded programmes; they are designed to mimic human-like interactions. Such a design allows for more than just information dissemination; learners and educators can engage in meaningful dialogue with these agents by challenging AI-systems with queries, seeking clarifications, and even brainstorming ideas. In essence, they provide a semblance of the tutor-student dynamic but within a digital framework. ChatGPT, and similar platforms, are not just passive repositories of knowledge. Their intrinsic value lies in their adaptability and responsiveness. They can modify their responses based on the user's needs, making the learning experience truly dynamic. For example, a learner struggling with a complex concept might receive a more detailed explanation, while another looking for a summary might receive a concise overview. This tailored approach to teaching and learning sets these conversational agents apart (Alaswad et al., 2023; González-Castro et al., 2021).

Personalised learning is, at its heart, a recognition of the diversity of the student body (Li & Wong, 2021). No two learners share an identical academic journey. Each individual gains an educational experience informed by their past experiences, cultural background, cognitive strengths, and areas of challenge. Traditional educational models, while effective for many, often face challenges in fully supporting the vast spectrum of learning styles and paces. Herein lies the transformative potential of AI. Imagine a classroom where every query, no matter how complex or simple, is met with patience and precision. Where feedback isn't just a standardised mark, but a comprehensive breakdown tailored to an individual's strengths and weaknesses. This is the environment conversational AI can foster. By analysing a student's inputs AI can potentially offer feedback that addresses specific misconceptions and suggests further resources tailored to their interests and needs.

It is not hyperbolic, then, to equate the capabilities of systems like ChatGPT to those of a personal tutor (Conati et al., 2021). Traditional tutors, while invaluable, are bound by constraints of time and geography. In contrast, AI-driven tutors are always available, ready to assist at any hour. This constant availability is particularly advantageous for adult learners or those in different time zones. What we witness, then, is a bridging of the age-old chasm between the one-size-fits-all approach of standardised education and the tailored guidance of individualised instruction (Belda-Medina & Calvo-Ferrer, 2022).

Scientific writing, an integral component of university education, is more than just an exercise in stringing words together; it's an activity that demands a fusion of critical thinking, comprehensive research, and coherent articulation. For many students, especially in the rigorous academic climate of a masters-level science programme, grasping this can be daunting. The stakes are high, with these written pieces often serving as the bedrock of their academic assessment and intellectual growth.

Unlike some academic tasks, writing is iterative. Rarely does a learner produce a perfect piece in a single attempt. The process entails drafting, reviewing, revising, and perhaps even starting from scratch. Conventional approaches to soliciting feedback, such as waiting for tutor comments or peer reviews, can be time-consuming. Here, AI's instantaneous nature shines. Students can submit a draft and promptly receive feedback, allowing them to immediately address any potential issues. This fluidity not only makes the writing process more efficient but also makes it more dynamic and responsive. The scope of feedback provided by platforms like ChatGPT is another area where their transformative potential becomes evident. Itgoes beyond grammar checks or vocabulary suggestions by addressing intricate complexities of academic writing. Feedback can extend into the depth of content, identifying areas where arguments lack clarity or where evidence is weak. Structural anomalies can be highlighted, ensuring that the narrative flow of the essay or paper remains unbroken. The learner's writing style can be evaluated too, ensuring a consistent and appropriate tone for the intended audience. But perhaps the most significant aspect of this feedback is its potential to elevate the depth and originality of a learner's thought processes. Effective scientific writing goes beyond presenting facts; it involves weaving these facts into a detailed, compelling argument. AI can assist a learner in recognising gaps in their reasoning or introducing perspectives they may not have considered.

This not only results in improved writing, but also cultivates a deeper understanding of the subject matter.

However, integrating AI into education brings not only positives but also important considerations (Sok & Heng, 2023). Given the potential of this technology, addressing its appropriate use is a matter of utmost importance. The allure of sophisticated platforms can sometimes overshadow the foundational principles of academic pursuits. While these tools possess the capacity to transform the way learners engage with tasks, there exists a key concern: the potential risk of over-reliance without a critical consideration of the output.

Effective scientific writing demands the integration of various cognitive skills. Learners must assimilate knowledge from diverse sources, apply critical thinking to dissect arguments and positions, and use their analytical skills to present coherent and compelling narratives. These are not tasks that can be outsourced entirely, even to advanced AI, without compromising the core educational experience. While platforms like ChatGPT can offer guidance, they cannot and should not replace the intellectual effort that learners must exert. If learners heavily rely on AI for writing tasks, there is a potential risk of curtailing their own capacity for original thought. Prolonged dependence on AI could contribute to a homogenisation of thought processes, as learners might inadvertently align their thinking too closely with AI-generated content, suppressing their unique perspectives and voices. This underscores the crucial role of educators. The introduction of AI tools in academic settings extends beyond offering learners with an additional resource; it involves integrating this resource into the broader pedagogical framework. Educators play a pivotal role in guiding learners on the judicious use of AI, not as a dependency, but as a supplement. Workshops, guidelines, and assessment criteria can be developed to ensure students use AI responsibly. Striking a balance is key, leveraging the advantages of AI to enhance writing without letting it overshadow the learner's authentic voice and development of their skills. A vital part of this education process is helping learners discern the boundary between assistance and over-reliance, just as we wouldn't use a calculator to perform every basic arithmetic operation, students shouldn't turn to AI for every aspect of their writing. They must understand where AI's capabilities can be beneficial and where human cognition should take precedence.

We can see two potential pathways to handling AI in education: one involves adopting a stringent stance, monitoring and penalising (mis)use, while the other embeds it into pedagogical methods as an educational asset. This project explored the latter approach. Initially, learners were introduced to both the capabilities and limitations of ChatGPT in an academic context. Aiming to evaluate the efficacy of ChatGPT as a pedagogical tool in enhancing critical evaluation skills, learners were tasked with creating an essay using ChatGPT, followed by critically evaluating the output using peer-reviewed sources. Critical thinking, often perceived as an elusive skill among learners, remains a cornerstone of academic success, and by anchoring their critiques on AI-generated essays, learners engaged in a focused evaluative task. Each statement warrants verification, driving learners to delve more deeply into the literature. This not only solidifies their skills in reading and assessing academic literature but also refines their analytical skills as they dissect the essay's strengths,

flaws, and coherence. Such a method emphasises critical evaluation's significance, laying down markers for the learner's own writing.

This study builds on existing AI research in education by moving on from more traditional uses of AI as a an assessment or feedback tool (Ali & Abdel-Haq, 2021; Zawacki-Richter et al., 2019). While much of the current literature focuses on using AI for knowledge acquisition or as an aid to streamlining educational processes, this study positions AI as a catalyst for critical thinking– a move from seeing AI as a solution to educational challenges to seeing it as a tool for developing essential academic skills. By directly embedding ChatGPT into the pedagogical framework, the study takes advantage of the potential of Gen-AI to enhance rather than bypass learner effort, underscoring the idea that AI's most valuable role in education may lie in developing active rather than passive engagement with material.

This study also contributes a unique perspective to AI research by showing how Gen-AI can support personalised learning goals in critical thinking. Unlike traditional educational models that might present critical thinking as a stand-alone skill, here, critical evaluation becomes an integrated part of learning through structured, task-oriented engagement with AI. Each statement generated by ChatGPT requires verification, which drives learners into an iterative process of assessment and validation that strengthens their research skills, comprehension of the literature and confidence in evaluating scientific content.

Much of the Gen-AI discourse centres on its threat to academic integrity (Benke & Szőke, 2024; Gupta, 2024; Meça & Shkëlzeni, 2024), in this study instead it is framed as a tool that can support it when used purposefully. Rather than AI being a risk that could lead to thought homogenisation or misconduct, this study demonstrates how AI can be used responsibly to enhance originality and criticality in learner work. Participants were guided on how to use AI critically, recognising it as an aid to deeper thinking instead of a shortcut to task completion. This contrasts with approaches that rely on surveillance or penalties to address AI misuse, suggesting that the right pedagogical frameworks can integrate AI as a constructive learning tool.

Methods

This project involved the participation of ten postgraduate students enrolled in Master's programmes at Leeds Beckett University, specifically in Medical Microbiology, Medical Biochemistry, or Biomedical Science. Recruitment was conducted via an email campaign targeting students enrolled in these specific courses. The email contained a hyperlink to detailed information outlining the study's objectives and emphasising the voluntary nature of participation.

Following recruitment, participants undertook an orientation session where a comprehensive overview of the research activities was provided. This session encompassed open discussions about ChatGPT, scrutinising its merits and drawbacks with a particular focus on issues relating to accuracy and academic integrity. A systematic explanation was delivered concerning the procedure for essay generation and this included the significance of formulating an accurate prompt and the editorial adjustments necessary for producing an essay of suitable level and quality. The session also

involved practical exercises such as creating a question matrix for a sample question and a step-by-step demonstration of essay generation.

The main activity of the project involved participants' use of ChatGPT to generate an essay for the questions, "Discuss the impact of the COVID-19 pandemic on the field of biomedical science. How has the crisis shaped research priorities, funding, and global collaboration? What lessons can be learned from the pandemic response to better prepare for future health emergencies?" Participants then critically evaluated the essay, following guidelines taught during their master's course, locating and using peer-reviewed literature to support their assessment. Students were also instructed to document their prompts and submit these along with their essays and evaluations.

The timeframe allocated for essay completion and its critical evaluation spanned four weeks, with an estimated workload ranging between 10 and 15 h. Evaluation metrics for the outputs were assessed by the authors using predefined rubrics (Tables 1 and 2), which participants were provided with during the introduction session. These metrics were contrasted with the participants' cumulative academic performance and benchmarked against similar types of assessments within their respective courses. Essays were rigorously assessed on multiple dimensions, including structural integrity, content, factual accuracy, and adherence to the essay question. Critical evaluations were examined for their analytical depth, clarity of argumentation, and the use of evidential sources.

For the analysis of the output marks, SPSS v28 was used to identify any statistically significant disparities between the participants' academic averages and their project marks, using a T-test to determine these differences. The T-test is particularly appropriate here as it is designed to test the means of two groups, in this case the participants overall academic averages and their performance in this project. SPSS is particularly well-suited for this type of analysis due to its comprehensive range of statistical tests, allowing calculation of both descriptive and inferential statistics.

Participant perspectives were collected through a focus group that lasted approximately one hour. A semi-structured methodology was used to discuss participants' opinions on their engagement with the project activities. The session also allowed for broader discussions concerning perceptions of AI's role in educational settings. An experienced moderator guided the discussion, and audio recordings were made with participants' consent. These recordings were subsequently processed in Adobe Audition and transcribed in Microsoft Word. Any necessary amendments were executed by the authors. Thematic analysis was used to identify recurring themes and patterns from the focus group's feedback.

Additional evaluation of the project was gathered through a structured questionnaire, designed in Microsoft Forms. It incorporated multiple-choice questions for structured feedback and free-text queries to capture more nuanced responses. The multiple-choice section gauged participants' attitudes towards their use of ChatGPT in academic activities, whilst the free-text section examined participant engagement, the perceived strengths and weaknesses of the assessment framework, and participant views on requisite training for optimising AI in educational contexts. Thematic analysis was again employed for interpreting the free-text responses.

The study strictly adhered to the institutional ethical protocols for human subject research and received formal approval from the Local Ethics Review Coordinator.

Table 1 Essay rubric

Weight	Section	0	1	2	3	4	5	6	7	8	9	10
10	Introduction	No introduction present.	The introduction exists but lacks any clarity or focus. No identifiable thesis statement or purpose.	The introduction is disorganised, with vague hints at what the essay might discuss but no concrete thesis statement or objectives.	The introduction provides minimal context and has a poorly articulated thesis statement that is difficult to identify.	Some context is provided, and a thesis statement is present but lacks clarity or specificity.	The introduction sets the context and contains a thesis statement, though either one or both could be more clearly articulated.	Adequate context and a clear thesis statement are present, providing a general direction for the essay.	Good introduction with clear context and a well-defined thesis statement that could be further refined for specificity or insight.	Very strong introduction with well-set context and a clear, insightful thesis statement.	Excellent introduction that is well-focused, provides comprehensive context, and contains a sharply articulated thesis statement.	Exceptional introduction that excels in clarity, focus, and specificity. The thesis statement is not only clear but also insightful and thought-provoking.
10	Research priorities	No mention of research priorities.	Mentions research priorities but lacks detail.	Limited discussion of research priorities without examples.	Some discussion of research priorities but lacks depth.	Fairly detailed discussion but misses key points.	Adequate coverage of research priorities with some examples.	Good discussion of research priorities but could include more examples.	Very good discussion with relevant examples.	Nearly exhaustive discussion of research priorities with pertinent examples.	Excellent discussion of research priorities, well-supported by examples.	Exceptional discussion, thoroughly supported by relevant examples.

Table 1 (continued)

Weight	Section	0	1	2	3	4	5	6	7	8	9	10
20	Funding	No mention of funding.	Mentions funding but lacks detail.	Limited discussion of funding without examples.	Some discussion of funding but lacks depth.	Fairly detailed discussion but misses key points.	Adequate coverage of funding with some examples.	Good discussion of funding but could include more examples.	Very good discussion with relevant examples.	Nearly exhaustive discussion of funding with pertinent examples.	Excellent discussion of funding, well-supported by examples.	Exceptional discussion, thoroughly supported by relevant examples.
20	Global Collaboration	No mention of global collaboration.	Mentions global collaboration but lacks detail.	Limited discussion of global collaboration without examples.	Some discussion of global collaboration but lacks depth.	Fairly detailed discussion but misses key points.	Adequate coverage of global collaboration with some examples.	Good discussion of global collaboration but could include more examples.	Very good discussion with relevant examples.	Nearly exhaustive discussion of global collaboration with pertinent examples.	Excellent discussion of global collaboration, well-supported by examples.	Exceptional discussion, thoroughly supported by relevant examples.
20	Lessons Learned	No mention of lessons learned.	Mentions lessons learned but lacks detail.	Limited discussion of lessons learned without examples.	Some discussion of lessons learned but lacks depth.	Fairly detailed discussion but misses key points.	Adequate coverage of lessons learned with some examples.	Good discussion of lessons learned but could include more examples.	Very good discussion with relevant examples.	Nearly exhaustive discussion of lessons learned with pertinent examples.	Excellent discussion of lessons learned, well-supported by examples.	Exceptional discussion, thoroughly supported by relevant examples.

Table 1 (continued)

Weight	Section	0	1	2	3	4	5	6	7	8	9	10
5	Conclusion	No conclusion present.	Conclusion exists but lacks any clarity or focus.	Conclusion is disorganised, with vague hints at what the essay discussed.	The conclusion provides minimal context and is poorly articulated.	Some context is provided, but the conclusion lacks clarity.	The conclusion sets the context but could be more clearly articulated.	Adequate context and a clear conclusion providing a general direction for further thought.	Good conclusion with clear context that could be further refined for specificity or insight.	Very strong conclusion with well-set context.	Excellent conclusion that is well-focused, provides comprehensive context.	Exceptional conclusion that excels in clarity, focus, and specificity.
5	Language and Style	Incoherent and unreadable.	Barely readable with numerous errors.	Readable but lacks clarity and has frequent errors.	Somewhat clear but style is inconsistent.	Moderately clear with some inconsistencies in style.	Clear but could benefit from more varied sentence structures.	Generally clear and readable with minor style issues.	Clear, concise, and mostly well-structured.	Very clear and well-structured.	Excellent clarity and style, almost flawless.	Exceptionally clear, well-structured, and engaging.
5	Structure and Organisation	Completely disorganised and incoherent.	Essay is mostly incoherent and disorganised.	Lacks logical flow; ideas are disconnected.	Some semblance of structure but often veers off track.	Mostly organised but lacks transitions and flow.	Adequately organised with some logical flow.	Well-organised but could benefit from better transitions.	Very well-organised with good logical flow.	Nearly flawless organisation and logical flow.	Excellent organisation and logical flow of ideas.	Exceptional organisation and seamless logical flow.

Table 1 (continued)

Weight	Section	0	1	2	3	4	5	6	7	8	9	10
5	References and Citations	No references or citations.	Minimal or incorrect citations.	Some citations but many are incorrect or irrelevant.	Moderate number of citations but lacks key references.	Adequate citations but some are irrelevant or incorrect.	Good number of citations but could be formatted better.	Very good citations but lacks some key references.	Excellent citations with minor formatting issues.	Nearly flawless citations and excellent choice of references.	Excellent citations, well-formatted and relevant.	Exceptional citations, perfectly formatted and highly relevant.

Table 2 Critical evaluation rubric

Weight	Section	0	1	2	3	4	5	6	7	8	9	10
15	Understanding of Subject Matter	Does not address the subject matter.	Demonstrates no understanding but makes an attempt.	Misunderstands the subject matter almost entirely.	Shows minimal understanding; mostly incorrect.	Limited understanding with several misconceptions.	Basic understanding but lacks details and depth.	Fair understanding but with noticeable gaps.	Above average understanding; some misconceptions.	Good understanding but lacks some depth.	Shows comprehensive understanding, missing only minor details.	Demonstrates an exceptional understanding of both the capabilities and limitations of ChatGPT, highlighting nuances.
50	Critical Analysis	Does not engage in critique.	Makes an attempt but misses the point entirely.	Fails to critique effectively; mostly incorrect.	Minimal critique; major flaws in reasoning.	Limited critique; mostly descriptive.	Basic critique; lacks depth.	Fair critique but not well-supported.	Above average critique; leans towards either strengths or weaknesses.	Good critique but lacks some nuance.	Strong critique with minor gaps in reasoning.	Provides an exceptional, nuanced critique with multiple examples.
15	Clarity and Coherence of Argument	Incoherent and unclear.	Makes an attempt but largely incoherent.	Barely understandable; severe coherence issues.	Significant issues in both clarity and coherence.	Mostly unclear and incoherent.	Lacks both clarity and coherence.	Fair but with noticeable issues in clarity and coherence.	Above average; some issues with clarity or coherence.	Clear but could be more coherent.	Mostly clear and coherent with minor lapses.	Exceptionally clear and coherent argument.

Table 2 (continued)

Weight	Section	0	1	2	3	4	5	6	7	8	9	10
15	Use of Evidence and Sources	No use of sources.	Makes an attempt but fails to use sources effectively.	Almost no use of sources.	Minimal use of sources; not integrated.	Few sources; poorly integrated.	Limited use and variety of sources.	Fair; noticeable lack of sources or citation issues.	Above average; some minor issues in sourcing or citation.	Good use of sources but lacks variety or integration.	Uses multiple good sources; minor integration or citation issues.	Employs a wide range of high-quality, well-integrated sources, properly cited.
5	Presentation and Writing Style	Unreadable or not presented.	Makes an attempt but largely unreadable.	Barely readable; significant errors throughout.	Numerous errors; severely impacts readability.	Poorly written; hard to understand.	Frequent errors; affects readability.	Fair; multiple errors but readable.	Above average; some noticeable errors.	Good writing but with some room for improvement.	Minor grammatical or stylistic errors; well-written.	Flawless presentation; high level of academic writing.

Results

Impact on Assessment Quality

There was no significant difference between the study essay marks and the overall course mark averages for the participants. There was, however, a significant improvement in the outcomes for the critical evaluations when compared to the overall average course marks (paired t test, $p=0.04$, mean difference of 9%). When compared to similar assessment types within the course, there were improvements (mean difference of 7%) within this study for the critical evaluation marks, but decreases for the essay marks (mean difference of 2%) however, neither difference was significant (Table 3).

Appraisal of Key Essay Characteristics and Critical Evaluation

In marking the essays and evaluations, we followed the provided rubrics, considering key characteristics such as structural coherence and content. This process provided a foundation for understanding the assessment outcomes, illustrating the specific qualities that contributed to the marking process.

Written English: The standard of written English across the essays was high, and generally of a higher standard compared to other examples of participant work for both native (four participants) and non-native (six participants) speakers, suggesting that ChatGPT performs well in this regard for all students.

Structure: Submitted essays had a high standard of structure which closely followed the PEEL framework—Point, Evidence, Explanation, and Link (Costello, n.d. [2000]). This structural choice offers several advantages. First, it contributes to the cohesion of the essay. By adhering to the PEEL format, each paragraph becomes a self-contained unit of thought, which enhances the essay's overall unity. Second, employing this structured approach inherently leads the reader through the essay, facilitating a logical flow from one point to the next. This enhances the ease of comprehending and engaging with the presented argument. Third, the use of the PEEL structure promotes a level of critical engagement, a quality frequently expected in academic/scientific writing. It prompts the writer to substantiate assertations with

Table 3 Marks for the essay and critical evaluation

Participant	Essay	Critical Evaluation
1	75	35
2	85	92
3	47	80
4	70	84
5	48	67
6	74	72
7	81	88
8	60	84
9	64	71
10	56	59
Average	66	73

relevant evidence and well-reasoned explanations. This not only strengthens the argument but also imbues it with the academic rigour that is expected at the master's level.

Content: The essays demonstrated a high level of comprehensiveness, encompassing all essential points and thereby ensuring a holistic exploration of the subject matter. This level of coverage ensured that the essays met all the requirements of the question. This approach strengthened the arguments within the essays, rendering them more compelling and robust, as well as contributing to a thorough and clear understanding by the reader.

Critical evaluation: The participants' critical evaluation of their essays maintained a high standard; however there was a tendency to emphasise the reliability of references rather than the factual accuracy of the content itself. Participant insights presented balanced perspectives on the use of ChatGPT for academic purposes, highlighting both its strengths and weaknesses. One of the key issues raised was the simplicity of the generated content. Whilst simplicity holds merit in specific contexts, particularly in academic writing at the master level, it can signify a lack of depth or sophistication, potentially undermining the complexity required for success at this level. Another limitation identified was the repetition of the content. Repetitive arguments or statements can significantly undermine the impact of an essay, as they can suggest a lack of comprehensive research or insufficient engagement with the subject matter. This is a concern as repetitiveness not only diminishes the overall quality of the written assessment but also could raise questions about its originality.

The main concern raised in all the critical evaluations pertained to the veracity of the references provided by the AI software. ChatGPT can generate citations that appear accurate, attributing to authors actively publishing in relevant areas, however, closer examination often reveals that the suggested publications and citations do not exist. In scientific writing, the quality and accuracy of references are paramount. The evaluations indicate that ChatGPT falls short in this regard, posing significant drawbacks in its use. The inclusion of inaccurate or unreliable references can compromise the integrity of the entire piece, leading to a loss of credibility.

On the positive side, speed was highlighted as a distinct advantage of using ChatGPT. The ability to generate content quickly was deemed to be incredibly valuable, especially in time-sensitive scenarios such as meeting tight assessment deadlines. Nevertheless, it is crucial to consider whether this efficiency comes at the expense of depth and accuracy, as indicated by the identified limitations. While ChatGPT is recognised for its ability to provide quick information, there are reservations about the depth and quality of the information produced. The inherent trade-off between efficiency and depth becomes a central consideration, implying that while ChatGPT can be a useful tool for rapid content generation, it may lack rigour.

The marks assigned to the critical evaluations were generally higher than for the essays. This variation could be attributed to the emphasis placed on critical evaluations in the assessment criteria, along with the formal instruction learners received in crafting such evaluations during their master's programme. In contrast, it may have been sometime since they had similar guidance and experience for essay writing. This suggests that while ChatGPT might be proficient in generating content that

aligns with formal academic structures, its utility might depend on the specific type of academic work and the level of expertise needed.

Prompt Usage

In order to generate the essays, participants were required to input appropriate prompts into ChatGPT. While they approached this in different ways, prompts generally fell into two categories: (1) thematic prompts, that aimed to extract g broader subject specific information and (2) Information prompts that focused on expanding the details for specific aspects of the essay.

1. Thematic Alignment:

- COVID-19 and Biomedical Research: These prompts were directly aligned with the essay's central theme.
- Research Methodologies: These prompts helped students identify research priorities and changing methodologies.
- Future Outlook: These prompts align with the essay's final question about lessons for future health emergencies.
- Specificity: These prompts were generally used for addressing specific considerations such as global collaborations and funding.

2. Information detail:

- Depth: These prompts were used for expanding the detail in the information provided in the more general prompts.
- Focus: These prompts were used to address the level and complexity of the information so that it met the required level.

Focus Group

While the focus group revealed that participants' preference was for creating their own essays over using ChatGPT, it also highlighted the perception that the tool could provide a degree of flexibility by accommodating diverse personal preferences and learning styles. Overall, the participants expressed the view that ChatGPT served as an initial resource for grasping the fundamentals of a topic or by sparking ideas rather than a more comprehensive aid in essay preparation.

Five key themes emerged from analysis of participant comments: the use of technology, ethical considerations, academic quality, skills development and personal preferences.

Theme 1: Use of the Technology

The participants discussed how technology, particularly generative AI like ChatGPT, can be instrumental in various aspects of academic work. They felt it could be useful beyond simple essay writing - it could summarise complex text and even adapt to

intricate prompt structures to overcome limitations. Participants suggested that we could think of AI more like an assistant, one that can 'think' for itself to some extent. In academic work, this means it can help with more complex tasks, like suggesting research methods or helping to plan out a project. Leading on from this, they thought that AI worked best with their oversight. ChatGPT could handle a lot of information and input quickly, but it still needed human oversight to get the most suitable outcomes.

Theme 2: Ethical Considerations

Participants expressed concerns regarding the employment of AI in generating academic content, highlighting specific concerns that AI-generated material could potentially circumvent traditional checks for academic integrity and perceived this as a form of 'cheating'. This raises a noteworthy concern: if AI can generate unique but not genuinely original content, what implications does this have on the principles of academic integrity? This is a complicated question. On the one hand, AI tools like ChatGPT are able to assist and facilitate academic work, but on the other, their capabilities could be misused to produce work that is not genuinely the product of the student's effort.

Another ethical question raised was the issue of fairness, particularly in the equitable distribution of educational resources. Participants raised concerns regarding unfair advantages, prompting us to consider whether having access to advanced AI tools, such as those behind a paywall, confers an advantage to those who can afford to pay for these tools over others who may not have the same level of access. This underscores the significance of ensuring equitable access to resources in educational settings where variations in resources accessibility among learners can be significant.

Theme 3: Academic Quality

The focus group highlighted the limitations of relying solely on AI for academic work. Participants noted that ChatGPT lacks the capability to access journal databases such as PubMed or Google Scholar, which significantly impacts its use for scientific work. This limitation is crucial because databases like these are repositories of peer-reviewed articles that serve as the cornerstone for scientific research and preparation of assessments at scientific master's level. The absence of such access diminishes the depth and credibility of research conducted solely using AI platforms.

The participants agreed that while AI can be a valuable asset, it should not replace human skills and judgement. While the technology can generate a scaffold for written coursework, the critical aspects such as evaluation, argument development, and fine-tuning must be driven by the individual. AI use at the initial stages of writing can be crucial for overcoming writer's block or organising thoughts. However, it is essential to recognise that this scaffold is just a starting point. Generative AI can provide the foundational structure of an essay, but it falls short of the nuanced tasks that render academic writing compelling and credible.

Theme 4: Skills Development

Consensus among participants was that the essential aspects of critical evaluation, argument development, and fine-tuning of an essay should come from the individual. These tasks require skills that AI cannot currently replicate, including critical thinking, research acumen, and ethical judgement. For example, while AI can summarise a complex text, it cannot evaluate its credibility or relevance in the context of a broader academic debate. Similarly, AI can generate text based on input prompts, but it cannot develop an argument that requires nuance, counter-arguments, and a deep understanding of the subject matter. Participants did, however, feel that using AI as a tool in the writing process could offer a valuable opportunity for skills development. By starting with an AI-generated scaffold they could focus on honing their skills in critical evaluation and argument development. Participants thought this iterative process would help students improve their academic writing skills over time, but that it was crucial that they understood the limitations of AI and the irreplaceable value of human skills in this process.

Participants were also concerned that relying too heavily on AI tools for academic writing could stunt the development of essential skills. They emphasised the need to develop these skills for their intrinsic value and for effectively interpreting and using the output generated by AI. While the utility of AI tools offers the promise of a beneficial learning experience, it's also important to balance this with the development of individual skills. The participants mostly used the AI as a supplementary tool, which suggests an awareness of the importance of honing their own skills in research, critical thinking, and writing.

Theme 5: Personal Preferences

One of the more subtle themes that emerged from the focus group was the adaptability of technology to different learning styles and preferences. Some participants used the AI as a springboard for ideas, while others used it for more detailed tasks. This adaptability underscores the potential for tailoring the technology to individual needs, allowing for a more personalised approach to learning and academic work. One speaker even referred to ChatGPT as "a personalised Wikipedia," highlighting the tool's ability to cater to specific user requirements.

While AI can adapt to various tasks, its efficacy is often determined by the extent of user involvement. A participant noted that enhanced personalised input could elevate the quality of the output. This speaks to the collaborative nature of AI in academic work; capable of accommodating diverse needs and styles yet requiring active user engagement for optimal results.

Opinions on the Assessment

The usefulness of the assessment was assessed via a questionnaire. It contained a range of multiple-choice questions about participant's thoughts on aspects of the assessment and their skills development. The responses showed that students had a much higher awareness of the ethics of using AI in education than before the project

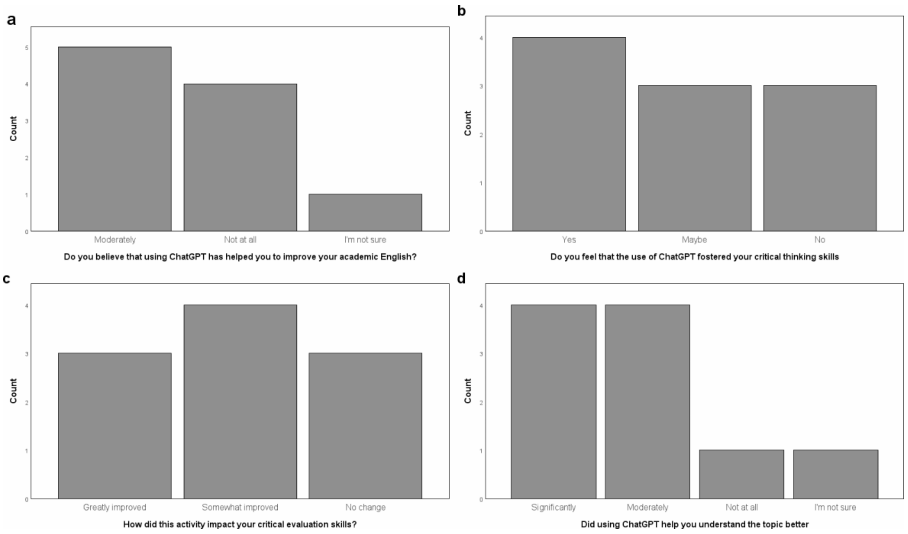

Fig. 1 Q1

and that there was some improvement in academic skills (Fig. 1). Most participants also responded that they had to always correct the information provided by ChatGPT but that it was somewhat effective at generating a coherent essay.

There were tangible benefits identified, such as saving time, making the learning process more interesting, and helping to generate ideas for the work (Fig. 2). Participants also found the process of writing an essay easier when using ChatGPT. Overall, however, participants were largely neutral about integrating ChatGPT into future assessments, their general experience of using ChatGPT and the potential impact on AI in education.

Free text questions were also used to assess participant engagement in the assessment and their thoughts on improving it.

If Your Engagement Level was Different, Why?

The engagement levels in response to the task varied considerably among the participants, reflecting a spectrum of experiences with differing causes. Some participants found no significant change in their engagement, attributing it to a consistent level of effort required for task completion. In contrast, others noticed a decrease in overall workload, which inadvertently led to reduced engagement levels.

A subset of participants noted that their engagement was affected by the time they had to invest in cross-verification, particularly in scrutinising references. This seems to indicate that despite potential efficiency gains in certain aspects of the process, the overall effort remained relatively constant when compared to traditional approaches. A contrasting perspective emerged from individuals who reported heightened engagement due to the novelty of employing a new method for essay writing. However, it is important to note that motivational challenges were also prevalent among some

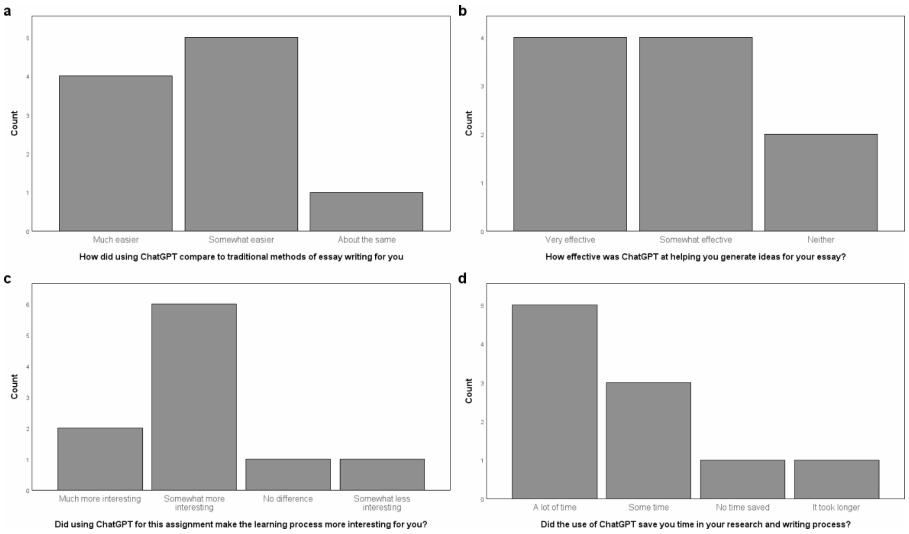

Fig. 2 Tangible benefits were identified. **a**: How did ChatGPT compare to traditional methods of essay writing for you; participants felt using ChatGPT made the essay writing process easier. **b**: How effective was ChatGPT in generating a coherent, factually accurate essay; most felt ChatGPT was somewhat effective for essay generation. **c**: Did using ChatGPT for this assignment make the learning process more interesting for you; most participants experience an increased interest. **d**: Did the use of ChatGPT save you time in your research and writing process; most participants reported a lot of time or some time saved

participants with difficulties in interacting with the chatbot being identified as a contributing factor.

Overall, the range of responses highlights the multifaceted nature of engagement, influenced by factors such as workload, novelty, and user interface experience.

What are the Main Strengths and Weaknesses you Found with this Assessment and How Would You Improve the Assessment in the Future?

Positively, the assessment was lauded for enabling a deeper understanding of the subject, promoting critical thinking, and aiding grammatical refinement. The use of ChatGPT in rapidly generating essays and providing initial ideas was frequently cited as a significant advantage.

On the negative side, a recurring concern centred on the reliability of the information provided by ChatGPT, particularly in relation to referencing the source material and citation. Several participants pointed out the need for extensive cross-verification, which offset time saved in other aspects of the project. Some also noted that the AI's output was at times overly simplistic, repetitive, or even erroneous, hindering the in-depth analysis required for more scientific topics.

Concerns were also raised about the assessment's novel approach, which made it difficult for participants to compare their performance with traditional methods. Some participants felt that the AI led to reduced engagement, as it performed tasks that would usually demand more intensive thought and research.

For future improvements, suggestions varied from opting for simpler topics that align better with the AI's capabilities, to enriching assessments by integrating more evidence-based information. Additionally, participants proposed adjustments in the word count and a greater emphasis on critical evaluation to more effectively challenge students.

Overall, the responses suggest that while the assessment holds promise in integrating AI into assessments, careful adjustments are needed to address its current limitations.

To Make the Most of AI in Education, What Support or Training Would You Like as Part of Your Course?

The responses indicate a clear need for support and training to harness the potential of AI in educational settings. One predominant theme was the need for explicit guidelines detailing how students should engage with and use AI for assessments. Training appears to be another area of focus. Participants expressed a keen interest in courses designed to equip them with the skills to effectively leverage AI in various aspects of their degree programmes.

There was also a demand for an introduction to AI that includes guidance on avoiding academic misconduct, which is crucial for maintaining academic standards. Some respondents suggested that AI could be particularly useful for introducing new terms or topics, serving as a supplementary educational tool. There was also an emphasis on the need for students to be taught the limitations of AI, particularly its inability to produce large, factually accurate essays without human oversight. The significance of verifying factual accuracy of any sources used was also highlighted.

Overall, the feedback suggests a need for a well-rounded educational framework that equips students with the knowledge and skills to use AI responsibly and effectively.

Discussion

The results of this project offer a nuanced and layeredaceted insight into the role and constraints of ChatGPT in academic writing and assessment within a master's level science programme. While capable of generating essays that meet certain academic standards, particularly in the areas of grammar and structural coherence, ChatGPT it is not without limitations. A key issue is the absence of a significant difference between the marks for the essays and the overall course averages. This may suggest that while ChatGPT can assist in generating academically acceptable content, it does not necessarily contribute to exceptional performance. In other words, while ChatGPT may serve as a useful tool for generating drafts or initial ideas, the data imply that it is not a substitute for human-led academic enquiry.

Interestingly, the study revealed that employing critical evaluation of ChatGPT outputs proved to be a more effective assessment tool. When students were asked to critically evaluate the essays generated by ChatGPT, these were marked higher than the essays themselves. This raises important questions about the pedagogical impli-

cations of integrating AI into academic settings. Perhaps the true educational value lies not in using AI to replace human effort, but in analysing and critiquing its output as a means to support knowledge acquisition and foster critical evaluation skills.

Petrovska et al., (2024) saw similar outcomes when integrating ChatGPT into software development education. Learners were asked to examine AI-generated code alongside their own, leading to a higher level of engagement and a deeper understanding of programming concepts (Petrovska et al., 2024). This process encouraged learners to recognize errors, analyse stylistic choices and improve the code by critically reflecting on the AI's suggestions. This approach enabled learners to not only refine their programming skills but also develop their capacity for critical thinking.

Dickey et al. (2023) saw similar outcomes with their AI-Lab Framework, which was designed to balance structured instruction and self-reflection while using gen-AI in programming courses (Dickey et al., 2023). Again, learners were tasked with critically evaluating AI output and this guided interaction promoted a healthy scepticism, equipping learners with the evaluative skills needed to use AI as a supplementary tool rather than as a replacement.

This approach has also been used in the context of authentic assessments in economics education (Nguyen Thanh et al., 2023). They evaluated the performance of Gen-AI across different levels of Bloom's Taxonomy (Bloom, 1956) and found that ChatGPT handled basic recall and comprehension tasks effectively but often struggled with more complex tasks such as critical evaluation. When learners were asked to critically evaluate AI-generated responses to complex economics questions, they strengthen their skills in evaluation by identifying logical weaknesses and considering the coherence and evidence within AI arguments. Such exercises expose the limitations of AI while developing learners' skills to critically evaluate content, turning Gen-AI into a tool for deeper engagement rather than passive reliance.

Impact on Assessment Quality

The data reveal an interesting pattern regarding assessment performance within the study's cohort. While the marks for essays align closely with the overall course marks, this was not the case for the critical evaluations, where a comparative increase was seen. This disparity can be interpreted in various ways, but one explanation centres on the curriculum's particular emphasis on critical evaluations. It's plausible that the pedagogical strategies employed have equipped students to excel in this form of assessment. This may stem from targeted teaching methods, specific course materials, or even a combination of both, which have collectively enhanced students' proficiency in critical evaluations over essay writing. Beyond the immediate academic context, these findings have broader implications, especially when considering the value of critical evaluation skills for employment and future academic pursuits (Demaria et al., 2018). Many professions and postgraduate courses demand the ability to critically evaluate information, therefore, if a curriculum can effectively teach these skills, it not only serves the academic aims but also better equips students for future professional endeavours.

For course designers, these findings serve as a catalyst to re-examine the current balance of assessment types. If the course aims to prepare students for real-world

challenges and further studies, and critical evaluation skills are highly valued in those contexts, then it might be prudent to allocate greater weight to these types of assessments. Conversely, if essay writing is also deemed a crucial skill—either for the course's academic objectives or for future employability—then additional pedagogical interventions may be needed to bring essay performance up to the level of critical evaluations.

The study findings propose a shift in perspective, indicating that generative AI might serve educators more effectively as a content-creation tool rather than a student resource. In this role, AI could generate a diverse array of materials—ranging from essays and articles to case studies and data sets. These generated resources can then be presented to students for critical evaluation fostering a dynamic and interactive learning experience. This approach aligns with the observed strengths in students' critical evaluation skills, capitalising on an existing area of proficiency. By focusing on the critical evaluation of AI-generated content, educators can promote a more active form of learning. This requires students to delve deeply into the material, applying their analytical skills and making reasoned judgments. These are key competencies that are highly valued in higher education and the professional world alike (Rakowska & de Juana-Espinosa, 2021). This approach not only offers a versatile framework for assessment but also provides a direct link to real-world applicability, particularly given the high value placed on critical evaluation skills in various professional fields, including STEM and healthcare. From a logistical standpoint, the use of AI to generate assessment content could offer significant time savings for educators. This would free them to focus on other crucial aspects of teaching, such as personalised instruction, curriculum development, and even their own research activities. Of course, the quality of the AI-generated content would need to meet certain academic standards, which educators could control by setting appropriate parameters for the AI. Nevertheless, a shift towards AI-generated content for critical evaluation also raises important ethical considerations. For instance, to enhance the transparency of the assessment process, it becomes essential to inform students that the content they are assessing is machine-generated. The introduction of AI-generated content could then serve as a catalyst for broader discussions about the ethical implications of using artificial intelligence in both academic and professional settings (Gill et al., 2024).

The study reveals that ChatGPT performs exceptionally well in certain aspects of academic writing, particularly in the quality of written English and its adherence to the PEEL (Point, Evidence, Explanation, Link) framework (Costello, n.d. [2000]). These strengths suggest that generative AI can serve a valuable function in facilitating high-calibre academic writing, particularly when it comes to the mechanics of sentence construction and the overarching structure of the text. Such capabilities could be especially beneficial for students who struggle with these foundational elements of writing, offering a form of automated assistance that brings their work up to an academically acceptable standard. However, it also uncovers limitations in ChatGPT's output, particularly when evaluated against the high standards expected at master's level. Specifically, the content generated by ChatGPT lacks the depth of analysis and complexity of thought that are considered hallmarks of advanced academic work. While the AI can construct sentences that are grammatically correct and structure an argument according to the PEEL framework, it falls short in deliv-

ering the insights and original contributions to knowledge that are expected at this advanced stage of study.

This dichotomy between form and substance has significant implications for both students and educators. For students, particularly those at the master's level, the use of generative AI like ChatGPT could serve as a double-edged sword (Hisan & Amri, 2023). On one hand, it can assist in generating drafts that are structurally sound, thereby saving time and effort that can then be devoted to refining the content. On the other hand, there's a risk of overreliance on the tool, which could result in work that is polished on the surface but lacking in intellectual rigour. For educators, these findings could inform decisions about the integration of AI tools into the educational process. ChatGPT and similar technologies could be employed as supplementary aids for teaching the basics of academic writing and structuring arguments, but they should be accompanied by clear guidelines and limitations on their use, particularly for tasks that require a higher level of expertise (Aiken & Epstein, 2000; Kumar, 2019).

Limitations of ChatGPT

The study highlights a crucial limitation in the capabilities of ChatGPT, concerning its inability to access repositories of peer-reviewed sources such as PubMed or Google Scholar. This shortcoming is especially significant when considering the tool's utility for supporting work at the master's level, where access to peer-reviewed, academic sources is indispensable for generating high-quality work. This inability to tap into these databases essentially restricts ChatGPT's usefulness to the surface layers of academic writing and research. For instance, without the ability to source and cite authoritative academic publications, any content generated by ChatGPT would likely lack the depth of research and breadth of perspectives that are expected in master's-level work.

This limitation has several implications for both students and educators. For students engaged in advanced academic work, it serves as a caution against relying too heavily on AI tools for support (Sok & Heng, 2023). While ChatGPT may provide a useful starting point for framing a research question or generating an initial draft, it cannot replace the extensive review of the literature and in-depth analysis of scholarly sources that are central to master's-level work; learners would still need to engage intensively with the literature to meet the standards expected at this level. For educators, this limitation of ChatGPT raises questions about its appropriate role in the educational ecosystem. While it might serve effectively as a tool for teaching the basics of academic writing and structuring (Schmohl et al., 2020), its use in more advanced courses, particularly at the master's level, would likely need to be limited and clearly defined. Educators may consider using it as a supplementary tool for specific tasks, such as brainstorming or initial draft writing, while also emphasising the importance of primary research and direct engagement with scholarly sources.

The Role of Prompts

The study's findings highlight the critical role that the type and quality of user-generated prompts play in determining the utility of ChatGPT (Bozkurt & Sharma, 2023; Zamfirescu-Pereira et al., 2023). Fundamentally, this highlights a symbiotic relationship between ChatGPT and the user. The quality of the generated content is not solely a function of the AI's capabilities; it is also significantly influenced by the user's adeptness in posing focused and relevant questions. This interdependence has several implications. It suggests that ChatGPT's utility is not fixed but rather can be optimised through effective user interaction. For students who are adept at asking well-formulated, specific questions, ChatGPT could prove to be a highly valuable resource for generating initial drafts, brainstorming ideas, or even performing basic data analyses. For these users, the AI tool becomes a more potent asset, capable of producing output that is closer in quality to what might be expected in an academic context. Alternatively, this relationship also reveals a potential pitfall: if the user lacks the ability to ask the right questions, the AI's output may be general, unfocused, and of limited academic value. This is a particularly crucial consideration for educators who might be contemplating the integration of ChatGPT into their teaching methods. While the tool has the potential to facilitate certain aspects of academic work, its effectiveness is, to some extent, contingent on the user's proficiency in posing questions—a skill that frequently requires training and experience.

This interplay between the user's question-framing skills and the AI's output quality could have an impact on assessments. For example, if a student uses ChatGPT to assist with an assignment, the mark they receive may not only reflect their understanding of the subject matter but also their ability to effectively interact with AI tools. This introduces an additional layer of complexity to the evaluation process and may require educators to consider new assessment criteria that take into account the learner's interaction with AI.

Student Opinions on Integration of AI into Assessment

The participant opinions offer a range of perspectives on the integration of ChatGPT into assessments. There was a clear recognition of the transformative potential of AI in academic endeavours; from its capacity to swiftly outline essays or offer guidance on appropriate research methods, AI is perceived as a valuable asset capable of significantly streamlining academic workflow. This enthusiasm is however, tempered by a prevailing sentiment that places AI as a supplemental tool rather than a full substitute for human intellect and effort.

The consensus seems to be that while AI can act as a powerful assistant in academic work, its role should largely be confined to that of a facilitator. For instance, while ChatGPT can quickly generate essay scaffolds, these are viewed not as end products but as starting points requiring further refinement, a task that is inherently human. The participants are unequivocal in their view that human oversight is not just desirable but essential for achieving the level of quality and rigour expected in academic work. This is especially pertinent in a landscape where the stakes are high, as in master's level or research-intensive studies.

This balanced perspective serves to lend a level of pragmatic realism to the broader discussion on AI's role in academia. While it's tempting to view AI through a utopian lens as a solution to various academic challenges, the participants' viewpoints serve as a grounding mechanism. They underscore that AI, for all its capabilities, still has limitations—whether it's the inability to access peer-reviewed content or the ethical concerns surrounding plagiarism and equitable access. These limitations aren't just technical challenges to be solved but are issues that require thoughtful discussion, ethical considerations, and perhaps even institutional policy changes.

Ethical considerations surfaced as a prominent theme in the participant opinions, indicating a depth of thought about the broader implications of integrating AI into the academic arena. One of the most striking concerns was the capability of AI-generated content to evade traditional cheating detection mechanisms. This possibility raises far-reaching ethical questions about the nature of originality and academic integrity in the era of advanced AI technologies. The participants' apprehension signals a pressing issue that extends beyond academic misconduct; it calls into question the frameworks and systems that educational institutions have long relied upon to maintain integrity. This suggests an urgent imperative for educational institutions not just to adapt but to radically rethink and revise existing policies and guidelines concerning academic integrity.

The ethical concerns intensified with the introduction of the fairness principle, particularly concerning equitable access to AI tools. The participants raised concerns that students with access to more sophisticated AI tools may gain an unfair advantage over those who don't. This issue introduces an ethical dimension that extends beyond the academic context; delving into the broader societal issues of inequality and access to educational resources (Kacperski et al., 2023; Trucano, 2023). In a system where some students can afford state-of-the-art AI assistance while others cannot, the playing field is inherently biased, and the academic outcomes may not serve as a reliable measure of individual capability or effort. This fairness issue has implications for how educational institutions might choose to integrate AI tools into their curricula. Will these tools be provided as a common resource to all students, or will students be required to procure them individually? If the latter, how will institutions ensure that all students have fair access? These are immediate questions that require careful consideration, not just from an operational standpoint but from an ethical one.

Academic integrity was also raised in the context of ethical use of AI, with participants concerned about what constitutes originality and plagiarism when we consider use of AI. Traditional models of academic integrity focus on concepts such as plagiarism, originality and unauthorised assistance, which were simpler to assess and enforce before the advent of sophisticated Gen-AI tools like ChatGPT. However, as participants noted, AI tools challenge these conventional definitions by making it possible to produce seemingly original work that evades standard plagiarism detection software. This raises the question of what 'original work' truly means when AI has contributed substantially to content generation, potentially prompting educators and institutions to reconsider their frameworks.

The introduction of AI tools also brings about a shift in what constitutes unauthorised assistance. While AI usage may be viewed as an extension of study aids, the line between acceptable support and academic misconduct becomes increasingly blurred.

If institutions permit AI as a learning tool, they must also establish clear guidelines on acceptable use, making clear the boundaries between what constitutes AI support that enhances learning and where it risks undermining academic integrity. This may require the development of new ethical guidelines and educational polices that both acknowledge the potential benefits of AI and set boundaries to preserve individual accountability and learning outcomes.

To address these challenges, educators and institutions may need to adopt a dual approach. First, by implementing robust, AI-aware policies that clearly define acceptable usage and secondly, through proactive education around AI ethics, ensuring learners have the understanding needed to navigate the use of AI responsibly. Academic integrity in the age of AI will depend as much on policy as it will on developing a culture of ethical engagement.

The idea of academic quality emerged as a theme in the participant opinions as well. While participants acknowledged the proficiency of ChatGPT in generating well-structured and grammatically sound content, they expressed significant reservations about its applicability to more research-intensive tasks. This concern about these limitations aligns closely with the participants' broader focus on the indispensability of human skills in academic work. For instance, while AI can provide a scaffold, the analysis and critical evaluation are competencies uniquely human and vital for academic rigour. Participants emphasised that the details– the depth of understanding needed to evaluate claims, contextualise findings and engage in reflective analysis, as well as the counter-arguments, the weighing of evidence, and the ethical considerations that are crucial to scholarly work, especially at a master's level, are beyond the purview of current AI capabilities.

While the participants maintained this stance, this wasn't entirely reflected in their critical evaluations. Though there was an improvement in their evaluations compared to previous efforts, the ability to develop robust-counter arguments remained comparatively undeveloped. This gap suggest a further potential are where AI could be used to improve skills. Despite AI's current limitations in mimicking the complete range of human cognitive abilities, it has proven effective in structuring arguments and identifying logical fallacies or gaps in reasoning, which can serve as a foundation for strengthening argumentative skills.

To build on this, a further stage could be added to this activity, where students interact directly with AI to refine their evaluations. Specifically, they could be encouraged to present their arguments and counter-arguments to the AI, which can then offer feedback and suggest additional points that may have been overlooked. Such an interaction would not only help the students to see their arguments through a different lens but also improve their ability to construct them.

The participants' views also seem to underscore that academic quality is not merely a function of informational accuracy or structural integrity. It encompasses a wide array of skills that include not just the ability to gather and present information but also to critique and to generate new knowledge through synthesis and analysis. These skills are critical in postgraduate studies, where students are expected not just to be consumers of existing knowledge but also contributors to their field. In this context, ChatGPT's limitations in delivering the expected level of depth and academic rigor become even more pronounced.

Skills development emerges as a potential issue in the participant feedback. On one hand, they see a distinct advantage in leveraging AI for scaffolding essays. This use of AI in handling the more mechanical aspects of academic work—structure, grammar, and basic data collection—can free up students to invest more time and cognitive resources in tasks that demand higher-order thinking skills. By providing a solid foundation upon which to build, AI tools like ChatGPT can serve as catalysts for skills development, allowing students to focus on refining their analytical abilities and enhancing the depth and breadth of their arguments. On the other hand, however, this optimistic view is tempered by concerns about the potential downsides of AI dependency. Participants worry that an over-reliance on AI could inadvertently lead to a form of skills atrophy, particularly in critical areas. If students become accustomed to relying on AI performing a significant portion of the research and drafting work, they may find themselves less equipped to handle these tasks independently. This is especially concerning in an academic context where mastery of research skills and the ability to navigate ethical dilemmas are not just useful competencies but essential skills.

These viewpoints suggest the need for a moderated and balanced approach to integrating AI into the academic workflow. They advocate for a model where AI serves as a supplementary tool that can handle specific tasks, freeing students to focus on the more complex aspects of their work. However, they also underscore the importance of not letting AI take over functions that are critical for the development of essential academic skills.

In terms of personal preferences, participants indicate that the effectiveness of AI as an educational tool is highly dependent on the level of user engagement. They suggest that the effectiveness of AI tools like ChatGPT is not solely predicated on the technology's capabilities but is linked with the level of user engagement. This perspective reframes our understanding of AI in academia, transforming it from a passive service provider to an interactive platform that thrives on active user participation. Participants appreciate the adaptability of AI, acknowledging its capacity to be tailored to diverse academic needs, however, this adaptability reaches its full potential only when met with a high level of engagement from the user. For example, while AI can generate a broad array of content based on general prompts, the quality of this content can be significantly elevated through more personalised, specific input from the user. It's a symbiotic relationship; the AI can offer a range of services, but the depth and nuance of these services are enhanced when the user actively engages with the tool.

This emphasis on user engagement also implies that the technology is not merely a tool to be used but a collaborative partner in the academic process. The participants seem to suggest that for optimal results, users must not just operate the AI but engage with it—questioning its outputs, refining its inputs, and tailoring its functions to better align with specific academic objectives and personal learning styles. In this way, the AI becomes more than just a machine that executes commands; it becomes a dynamic educational asset that can evolve and improve through ongoing interaction with the user. The participants' opinions point towards a future of academic work where AI tools serve not as passive, one-size-fits-all solutions but as dynamic, interactive platforms that require active human engagement for optimal effectiveness. This suggests

a collaborative model where both the AI and the user adapt to each other's capabilities and limitations. In this model, the AI serves as a highly adaptable tool that can cater to a broad array of academic tasks and individual preferences, while the human user serves as the curator, customising the AI's functions to suit specific needs and ensuring that the output meets academic standards.

Limitations and Future Work

The limitations of the project present some important caveats that should be considered when interpreting its outcomes. First and foremost, the small sample size of 10 participants raises questions about the generalisability of the findings. A sample of this size might not adequately reflect the diversity of opinions or needs of a broader student population, potentially restricting the external validity of the results.

This limitation is particularly important when considering the application of these findings to educational policy or curriculum design, as the strategies that proved effective with this small, relatively homogeneous group may not yield the same outcomes in a larger or more diverse cohort. For example, this group included motivated and tech-savvy students who mostly had positive opinions about AI and as such the finding may not be directly applicable to those who are less comfortable with it or that have greater reservations about using AI in their academic activities. This may lead to skewed interpretations, potentially overlooking the needs or challenges of other learner populations not represented in this group. Expanding the sample size in future work would provide a more robust basis for drawing conclusions, enabling more reliable insights that could guide broader educational reforms.

The lack of diversity in terms of the educational skills and attitudes of the participants also raises potential concerns about how well the results capture the academic needs present in a typical learner population. Learners may approach AI tools with different expectations and skill sets, influencing how they interact with and benefit from these tools. In a larger, more diverse sample the study might reveal nuances in how AI affects learner's critical thinking across disciplines, cultural backgrounds or prior technological exposure. To ensure the findings are applicable to a wider educational context, future studies would benefit from including a larger, more heterogeneous group of participants, allowing educators to understand how different learner profiles respond to AI in educational settings, making the results more relevant to an expansive and varied student body.

The voluntary nature of participation could also introduce a self-selection bias into the study. It's plausible to assume that volunteers for a project involving AI and academic assessment might be more technologically adept or intrinsically motivated than the average learner. Such a bias could skew the results, making them less applicable to students with varying degrees of technological proficiency or motivation. This is especially pertinent given the participants' generally positive views on the adaptability and utility of AI, views that might not be shared by less tech-savvy or less motivated students.

Another potential limitation is the possibility of a ceiling effect among the participants. The above-average marks of the participants suggest that they were already performing near their academic best, limiting the scope for any significant improve-

ment via the use of AI. This has important implications for interpreting the study's findings; the lack of significant improvements in essay scores might not reflect limitations of the AI tool itself but rather the already high-performance levels of the participants. In other words, the AI's impact might be more pronounced among students who have greater room for academic improvement. The academic strength of the participants presents another layer of complexity. Given their expertise and high performance in coursework, these students might be better equipped to navigate or compensate for the limitations of the AI tool in understanding complex academic topics. In a more diverse academic setting, where the range of expertise and performance levels is broader, the limitations of the AI tool might be more glaringly exposed.

A promising area for future work would be to investigate the role of instructional support in AI-based tasks, examining the extent to which guidance and scaffolding enhance learners' ability to use AI tools effectively for critical evaluation. Instructional support could take various forms, such as initial training as used here, or use of structured prompts or frameworks for assessing AI-generated content. By assessing how learners perform with different levels of instructional support, educators could gain valuable knowledge about how guidance influences learners' critical engagement and overall learning outcomes.

Understanding the role of support is important because it would enable educators to develop targeted pedagogical frameworks that effectively and ethically incorporate AI. Using this information, educators could design AI-based tasks that progressively reduce instructional scaffolding, helping learners to gradually develop independence in their critical thinking skills.

Conclusion

Generative AI technologies like ChatGPT have the potential to disrupt how we think about academic writing and assessment. While ChatGPT manifests a commendable proficiency in generating structurally coherent and grammatically accurate essays, it does not necessarily elevate the academic performance of students to an exceptional level. This is highlighted by the lack of significant disparities between essay marks and overall course averages, suggesting that ChatGPT's use, in its current form, is confined largely to drafting and idea generation rather than acting as a surrogate for in-depth, human-led scholarly enquiry.

Interestingly, the pedagogical utility of ChatGPT seems to be inverted; it appears to serve more effectively as a tool for assessment rather than as an aid for students. When the focus shifts from generating essays to critically evaluating AI-produced material, the learner's performance remarkably improves. This finding dovetails with the broader pedagogical emphasis on critical evaluations, casting a spotlight on the current assessment structures within academia. For course designers, this presents an opportunity to recalibrate assessment types, leaning more heavily into critical evaluation tasks that not only align with students' demonstrated proficiencies and offer real-world applicability.

The ethical dimensions of integrating AI in academic settings are also far from trivial. From questions of academic integrity and originality to concerns about equi-

table access to AI tools, there is a labyrinth of ethical considerations that educational institutions must navigate. This is compounded by ChatGPT's inability to access scholarly databases, which limits its use for in-depth research— a cornerstone of advanced academic work. The ethics of AI integration also present significant considerations for policy. Issues of academic integrity, originality and equitable access highlight the complexities institutions must address to maintain fairness and uphold academic standards. Policymakers may consider guidelines that differentiate between acceptable use of AI for initial research scaffolding and inappropriate uses that bypass authentic academic engagement.

The interaction between ChatGPT and the user emerges as a significant variable influencing the tool efficacy, pointing to the necessity embedding the developing AI literacy amongst learners within the curriculum. This symbiotic relationship implies that the use of the AI is dynamic, contingent on the user's ability to ask the right questions. This adds a layer of complexity to its integration into educational settings, as it necessitates a level of proficiency in interacting with AI tools—a skill set that itself requires pedagogical attention. Student opinions corroborate the complexity of this AI-human interface, emphasising the supplemental role of AI. While they acknowledge the potential efficiencies brought about by AI in tasks like essay scaffolding or basic data collection, they are unequivocal in their stance that these efficiencies cannot supplant the need for human intellect and effort, especially when the academic stakes are high.

The integration of ChatGPT and similar AI technologies into academic settings is a double-edged sword (Hisan & Amri, 2023). While promising as a tool for initial drafts and as a unique assessment mechanism, it falls short of being a panacea for academic writing and research challenges. What becomes clear is that the future of academic assessment and writing is likely to be a blended one, combining the computational power of AI with the nuance, ethical considerations, and critical faculties that are uniquely human. This fusion, if ethically managed and pedagogically sound, has the potential not only to reshape the contours of academic practice but also to equip students with a more rounded skill set, better preparing them for the complexities of the professional world.

Author Contributions DJ conceptualised the research, all authors were involved in the collection and analysis of the data and of preparing and reviewing the manuscript.

Funding The funding for this project was provided by the Centre for Learning and Teaching at Leeds Beckett University as part of their Teaching Excellence Project scheme.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Competing Interests The authors declare no competing interests.

References

Aiken, R. M., & Epstein, R. G. (2000). Ethical guidelines for AI in education: Starting a conversation. *International Journal of Artificial Intelligence in Education*, *11*(2), 163–176.

Alaswad, S., Kalganova, T., & Awad, W. (2023). Investigating the Value of Using Emotionally Intelligent Artificial Conversational Agents to Carry out Assessments in Higher Education. *2023 International Conference on IT Innovation and Knowledge Discovery (ITIKD)*, 1–5.

Ali, M., & Abdel-Haq, M. K. (2021). Bibliographical analysis of artificial intelligence learning in higher education: Is the role of the human educator and educated a thing of the past? *Fostering communication and learning with Underutilized technologies in Higher Education* (pp. 36–52). IGI Global.

Belda-Medina, J., & Calvo-Ferrer, J. R. (2022). Using chatbots as AI Conversational partners in Language Learning. *NATO Advanced Science Institutes Series E: Applied Sciences*, *12*(17), 8427.

Benke, E., & Szőke, A. (2024). Academic integrity in the time of artificial intelligence: Exploring student attitudes. *Italian Journal of Sociology of Education 16(Italian Journal of Sociology of Education*, *16*(2), 91–108.

Bloom, B. S. (1956). *Taxonomy of Educational objectives: The classification of Educational Goals*. Longmans.

Bozkurt, A., & Sharma, R. C. (2023). Generative AI and prompt Engineering: The art of whispering to let the Genie out of the Algorithmic World. *Asian Journal of Distance Education*. http://asianjde.com/ojs/index.php/AsianJDE/article/view/749

ChatGPT. (n.d.). Retrieved September 22 (2023). from https://chat.openai.com

Conati, C., Barral, O., Putnam, V., & Rieger, L. (2021). Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence*, *298*(103503), 103503.

Costello, C. (2020). (n.d.). *PEEL Paragraph Writing*. Virtual Library. Retrieved May 20, from https://www.virtuallibrary.info/peel-paragraph-writing.html

Demaria, M. C., Hodgson, Y., & Czech, D. P. (2018). Perceptions of transferable skills among Biomedical Science Students in the final-year of their degree: What are the implications for Graduate Employability? *International Journal of Innovation in Science and Mathematics Education*, *26*(7). https://openjournals.library.sydney.edu.au/index.php/CAL/article/view/12651

Dickey, E., Bejarano, A., & Garg, C. (2023). Innovating computer programming pedagogy: The AI-Lab framework for Generative AI adoption. In *arXiv [cs.CY]*. arXiv. https://doi.org/10.1007/s42979-024-03074-y

Gill, S. S., Xu, M., Patros, P., Wu, H., Kaur, R., Kaur, K., Fuller, S., Singh, M., Arora, P., Parlikad, A. K., Stankovski, V., Abraham, A., Ghosh, S. K., Lutfiyya, H., Kanhere, S. S., Bahsoon, R., Rana, O., Dustdar, S., Sakellariou, R., & Buyya, R. (2024). Transformative effects of ChatGPT on modern education: Emerging era of AI chatbots. *Internet of Things and Cyber-Physical Systems*, *4*, 19–23.

González-Castro, N., Muñoz-Merino, P. J., Alario-Hoyos, C., & Kloos, C. D. (2021). Adaptive learning module for a conversational agent to support MOOC learners. *Australasian Journal of Educational Technology*, *37*(2), 24–44.

Gupta, A. (2024). *When generative artificial intelligence meets academic integrity: educational opportunities & challenges in a digital age*. 14.

Haleem, A., Javaid, M., Qadri, M. A., & Suman, R. (2022). Understanding the role of digital technologies in education: A review. *Sustainable Operations and Computers*, *3*, 275–285.

Hisan, U. K., & Amri, M. M. (2023). ChatGPT and Medical Education: A double-edged Sword. *Journal of Pedagogy and Education Science*, *2*(01), 71–89.

Kacperski, C., Ulloa, R., Bonnay, D., Kulshrestha, J., Selb, P., & Spitz, A. (2023). Who are the users of ChatGPT? Implications for the digital divide from web tracking data. In *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/2309.02142

Kumar, D. N. M. (2019). Implementation of artificial intelligence in imparting education and evaluating student performance. *Journal of Artificial Intelligence and Capsule Networks*, *1*(1), 1–9.

Li, K. C., & Wong, B. T. M. (2021). Features and trends of personalised learning: A review of journal publications from 2001 to 2018. *Interactive Learning Environments*, *29*(2), 182–195.

Mallik, A., & Mallik, L. (2017). A review of Education Technology in Digital Age: Classroom Learning for Future and Beyond. *Psycho-Educational Research Reviews*, *6*(3), 80–92.

Meça, A., & Shkëlzeni, N. (2024). Academic integrity in the face of generative language models. *Lecture notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering* (pp. 58–70). Springer Nature Switzerland.

Nguyen Thanh, B., Vo, D. T. H., Nguyen Nhat, M., Pham, T. T. T., Trung, T., H., & Xuan, H., S (2023). Race with the machines: Assessing the capability of generative AI in solving authentic assessments. *Australasian Journal of Educational Technology*, *39*(5), 59–81.

Petrovska, O., Clift, L., Moller, F., & Pearsall, R. (2024, January 5). Incorporating generative AI into software development education. *Proceedings of the 8th Conference on Computing Education Practice*. CEP '24: Computing Education Practice, Durham United Kingdom. https://doi.org/10.1145/3633053.3633057

Rakowska, A., & de Juana-Espinosa, S. (2021). Ready for the future? Employability skills and competencies in the twenty-first century: The view of international experts. *Human Systems Management*, *40*(5), 669–684.

Ranasinghe, A. I., & Leisher, D. (2009). *The benefit of integrating technology into the classroom*. m-hikari.com. http://www.m-hikari.com/imf-password2009/37-40-2009/ranasingheIMF37-40-2009.pdf?utm_source=Buncee%26utm_campaign=aa3b6ee8c0-EMAIL_CAMPAIGN_2020_07_30_10_40_COPY_01%26utm_medium=email%26utm_term=0_2a223d00c6-aa3b6ee8c0-410994845

Schmohl, T., Watanabe, A., Fröhlich, N., & Herzberg, D. (2020). How artificial intelligence can improve the Academic Writing of students. In *Conference Proceedings. The Future of Education 2020*. conference.pixel-online.net

Sok, S., & Heng, K. (2023). ChatGPT for Education and Research: A Review of Benefits and Risks. In *Available at SSRN 4378735*. https://doi.org/10.2139/ssrn.4378735

Tahiru, F. (2021). AI in education: A systematic literature review. *Journal of Cases on Information Technology (JCIT)*, *23*(1), 1–20.

Trucano, M. (2023). *AI and the next digital divide in education*. https://policycommons.net/artifacts/4514806/ai-and-the-next-digital-divide-in-education/5324561/

Wang, S., Wang, F., Zhu, Z., Wang, J., Tran, T., & Du, Z. (2024). Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications*, *252*(124167), 124167.

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023, April 19). Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23: CHI Conference on Human Factors in Computing Systems, Hamburg Germany. https://doi.org/10.1145/3544548.3581388

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education– where are the educators? *International Journal of Educational Technology in Higher Education*, *16*(1), 1–27.

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., Liu, J. B., Yuan, J., & Li, Y. (2021). A review of Artificial Intelligence (AI) in education from 2010 to 2020. *Complexity*, *2021*. https://doi.org/10.1155/2021/8812542

Authors and Affiliations

Angela Oates[1] · **Donna Johnson**[1]

✉ Donna Johnson
Donna.johnson@leedsbeckett.ac.uk

Angela Oates
a.oates@leedsbeckett.ac.uk

[1] Biomedical Science, School of Health, Leeds Beckett University, Leeds, UK