Research Matters / 37
A Cambridge University Press & Assessment publication

ISSN: 1755-6031

Journal homepage: https://www.cambridgeassessment.org.uk/our-research/all-published-resources/research-matters/

Does ChatGPT make the grade?

Jude Brady (International Education Research Hub), **Martina Kuvalja** (Digital Assessment and Evaluation), **Alison Rodrigues** (International Education Research Hub) **and Sarah Hughes** (Digital Assessment and Evaluation)

To cite this article: Brady, J., Kuvalja, M., Rodrigues, A., & Hughes, S. (2024). Does ChatGPT make the grade? *Research Matters: A Cambridge University Press & Assessment publication, 37*, 24–39. https://doi.org/10.17863/CAM.106034

To link this article: https://www.cambridgeassessment.org.uk/Images/research-matters-37-does-chatgpt-make-the-grade.pdf

Abstract:
This study explores undergraduate students' use of ChatGPT when writing essays. Three students were tasked with writing two essays each for a coursework component for a Cambridge qualification facilitated by access to ChatGPT. After writing the essays, they participated in semi-structured interviews about their experiences of using the technology. Researchers compared the transcript of the chatlog between the students and ChatGPT with the submitted essays. Analysis showed that the students relied on ChatGPT outputs to different extents, although they followed a similar process of engagement. The students shared their misgivings and points of appreciation for the technology.

Does ChatGPT make the grade?

Jude Brady (International Education Research Hub), **Martina Kuvalja** (Digital Assessment and Evaluation), **Alison Rodrigues** (International Education Research Hub) **and Sarah Hughes** (Digital Assessment and Evaluation)

This study took place in March 2023, just four months after OpenAI launched ChatGPT as a free service into the public domain. The generative artificial intelligence (AI) platform attracted over 100 million users in two months (Milmo, 2023). Its rapid growth and potential uses sparked a great deal of discussion about the implications for teaching, learning and assessment.

Our research is an early attempt from within Cambridge University Press & Assessment to explore the ways in which ChatGPT might be used in an assessment context. For this research, we engaged three undergraduate students to complete coursework essays with the assistance of ChatGPT. We selected a coursework task from the Cambridge IGCSE™ Global Perspectives syllabus because the syllabus encourages learners to think about and explore solutions to significant global issues. Students need to consider different perspectives and contexts. They also develop transferable skills to complement learning in other curricular areas.

We chose Component 2, the Individual Report (IR), which is a coursework component requiring candidates to write an essay. The skills assessed are:

- researching, analysing and evaluating information
- developing and justifying a line of reasoning
- reflecting on processes and outcomes
- communicating information and reasoning.

Component 2 requires that learners respond to tasks relating to the following topical issues of global importance:

- Belief systems
- Biodiversity and ecosystem loss
- Changing communities
- Digital world
- Family
- Humans and other species
- Sustainable living
- Trade and aid.

Given the broad range of skills required and breadth of topics available for the Global Perspectives Component 2, we considered that it lent itself well to our research task which required students to write essays with the assistance of ChatGPT. Through a comparison of the students' final essays with their ChatGPT outputs, analysis of ChatGPT chat logs, and interviews with students, the research explores the different extents to which students might use generative AI to help with essay-based coursework assessments. Findings from the work map the process that students adopted to navigate the ChatGPT platform and its outputs to arrive at a complete essay.

Findings from our work cannot be generalised to all possible uses of ChatGPT in essay writing due to the specific context and small sample size, but the broad remit of the work allows us to draw useful initial conclusions and identify areas for further research. The study highlights areas for consideration from a compliance and policy perspective; it invites discussion around the strengths and weaknesses of ChatGPT as an assessment aid and the limits of acceptable use. Additionally, when considering the future of assessment, questions can be raised about what skills are being measured and assessed when students engage with this type of technology in comparison to traditional methods. We hope that this early work will provide some ideas to help construct a holistic portfolio of research which examines and further explores the strengths and weaknesses of generative AI in essay writing and assessment.

Literature review

What is ChatGPT?

ChatGPT is a chatbot driven by a generative AI program. ChatGPT, like other generative AI chatbots such as Bing Chat, Microsoft Copilot and Bard, can generate novel outputs in response to prompts from the user, just like having a conversation. In practice, this means that a user can type in a question or instruction and the chatbot will generate a new response every time. Its responses are based on training data comprised of millions of websites, media outputs, journals, and books. The outputs are human-like and content is generated quickly. The chatbot is easily accessible provided the user has a device with an internet connection.

How does it work?

The ChatGPT program is based on a Large Language Model (LLM). Yosifova (2023) indicates that the model is "large" because it is informed by masses of data, and the model itself has many dimensions, layers, and connections or pathways between its different parts. In its training phase, the GPT-3.5 model developed and improved its ability to predict based on 175 billion parameters (*ibid*). The training sources were dated up to September 2021. Sometimes LLMs can be trained for a specific purpose, such as translation; however, ChatGPT uses a general model that aims to produce human-like language. As a result of the general training, the LLM is very powerful because it can be used for a huge range of tasks from chat and summarising materials to solving mathematics problems, and rewriting code.

GPT-3.5 training involved receiving human feedback in the form of rankings and accuracy ratings of its outputs. These human inputs helped to further the LLM's improvement.

Organisational approaches to generative AI

Cambridge University Press & Assessment is exploring a variety of LLM research. Areas which are being investigated include:

- production of texts for students at specified CEFR levels[1]
- content creation capability, including multiple-choice questions and quizzes (Galaczi, 2023)
- learning and assessment-focused applications and automarking capability.

Our work sits alongside these other investigations to offer some insight into the use of generative AI by students in an assessment context. The following review, drawn from academic literature and grey literature including blogs and opinion pieces, explores the perceived risks and opportunities of generative AI in this area.

Risks for assessment

Academic integrity

The most pertinent risk of generative AI to assessment is the potential for misuse. There are concerns that students could pass off AI-generated work as their own (Eke, 2023). Currently, AI detection tools are not reliable enough to be used to determine whether responses or answers to an assessment are partially or completely AI-generated, which means that students could be falsely accused of academic malpractice or indeed get away with cheating (Dalalah & Dalalah, 2023).

Reliability of information

There is evidence of ChatGPT generating false information and "hallucinations" which take the form of very plausible sounding references that do not exist (Dale, 2021; Perkins, 2023). If students are not trained to verify their sources, they will encounter challenges in distinguishing between facts and fabrications and possibly develop a knowledge base founded on fiction. Further to this, research has shown that GPT-3.5 is susceptible to different types of faulty reasoning (Marcus & Davis, 2020).

Inbuilt bias

Dwivedi et al. (2023) infer a risk when they describe how the information generated by AI could exhibit bias and privilege. The GPT model is largely trained on English language materials meaning these sources are not reflective of the diversity of views, perspectives and cultural truths that are prolific across the world (Lebovitz et al., 2023). Politically this is important because if AI is used to provide students with information and answers, to suggest ways of phrasing, mark assessments, author assessments, screen university or teaching applicants, or even to inform decision-making in education, it may well privilege and perpetuate one kind of perspective (e.g., a global north and white-centric viewpoint).

1 The Common European Framework of Reference for Languages.

Legal concerns

There are unresolved complications around the use of generative AI in relation to copyright infringement and intellectual property rights (Dalalah & Dalalah, 2023; Lee, 2023). For example, if an assessment is authored mainly with the aid of generative AI, who owns the content? (Dippenaar, 2023). There are some instances of writers listing ChatGPT as a co-author on work (King & ChatGPT, 2023). However, high-profile journals such as *Nature* and *Science* have rejected this practice and will not accept chatbots as authors (Stokel-Walker, 2023).

Ethical concerns

Wider ethical concerns lie alongside these legal questions. Such concerns are not yet satisfactorily resolved because the extent to which generative AI has the capacity to cause harm is not fully understood and ethical frameworks for ChatGPT use are in development (CMS/W, 2023; Dwivedi et al., 2023).

Opportunities for learners, teachers, and assessors to exploit AI

Despite the risks, generative AI provides a set of unique assessment opportunities. Not only could such chatbots be used to help students learn, but they could also be used to author assessments and mark them.

Efficiency for teachers and students

Given the range of potential uses, the popularity of generative AI with educators is unsurprising. A survey of teachers in November 2023 suggests that 42 per cent of teachers are using AI to help with their work. Generative AI has the potential to improve efficiency for teachers if they use it with an awareness of its limitations. Kasneci et al. (2023) suggest that teachers could use chatbots to produce a text or model answer or to generate discussion prompts or lesson ideas, although these outputs would need reviewing by a human expert in the discipline. However, this could potentially save teachers time because the time spent authoring materials from scratch or searching online for appropriate teaching and learning supports could be reduced.

For students, ChatGPT and its equivalents could be used to provide a starting point for their research into a given and unfamiliar topic. The production of AI-generated content also opens the possibility of teachers and students reviewing these outputs together with a critical lens that invites discussion around the reliability, ethics, and efficacy of AI in education. It should be noted, however, that this kind of meta-reflection on the uses of generative AI introduces a new component into the teaching and learning arena. Educators would need to consider if or how courses and assessments should be adapted to accommodate and recognise learning which has taken place using generative AI.

Improved personalisation

Chatbots such as ChatGPT offer a unique opportunity for highly personalised learning. School students can learn through chatbot generated quizzes, summaries, and step-by-step explanations of how to solve specific problems (Kasneci et al., 2023). ChatGPT's "Socratic mode" allows students to be guided

towards understanding through the Socratic questioning method.[2] Despite this potential, recent research into the Socratic mode's teaching of physics concepts found that the bot is unreliable at correcting misconceptions (Gregorcic & Pendrill, 2023). The authors suggest that the chatbot may prove more useful in producing erroneous explanations which students can then correct.

Availability and accessibility

One of the perceived advantages of generative AI is that there is a variety of models available (some free, some paid for), which can be accessed anytime and anywhere with an internet connection and suitable device. Students have the possibility of a learning dialogue with a chatbot without having to wait for peer or teacher assistance. For example, students could use ChatGPT as a revision aid and chatbots could assist learners who have additional needs (Kasneci et al., 2023). Finally, students who are learning in a language other than their first language may benefit from the instant translation, paraphrasing and clarification possibilities offered by ChatGPT.

Methodology

Research question

The research answers the following question:

- How do students use ChatGPT in essay writing?

The question was addressed from two angles. Firstly, we quantified the extent to which the students relied on ChatGPT-generated outputs to form the content of their essays. Secondly, we analysed interview data to gain insight into *how* students interacted with ChatGPT and their process of engagement with the technology for the set task. Students also shared their perceptions of the strengths, weaknesses, and purpose of ChatGPT in assessment.

Task

The focus of the study was to qualitatively explore how undergraduate students used ChatGPT technology to support them in a written assessment. Three undergraduate university students were invited to write two essays each for the Cambridge IGCSE Global Perspectives Individual Report (IR). Convenience sampling was used to select the students, and all students were reimbursed for expenses and paid for their time. An IGCSE Global Perspectives assessment task was selected because of the wide range of skills demanded. Furthermore, the assessment task requires students to gain (through research) a broad understanding of a topic of which they were expected to have limited prior knowledge. As the assessment is intended for IGCSE candidates, it was considered that the undergraduates would already have a good command of the skills required to engage effectively and meaningfully with the task, but that they would not have an in-depth knowledge of the topic areas. For these reasons, we expected that the undergraduates would be able to engage with the assessment task with relative ease and we could retain the research focus on their uses of ChatGPT to aid with essay writing.

2 Socratic questioning is a shared dialogue by teachers, or in this case the chatbot, posing thought-provoking questions. The students then engage by asking their own questions. The discussion continues back and forth.

Although we chose to conduct the study with undergraduate-level students, the typical IGCSE student is 15–16 years old. It is likely that students of different age groups and/or with different levels of education would engage with both the technology and the process of essay writing in a different way to the undergraduates included in this research. Undergraduate students are also likely to be more skilled and experienced in essay composition and research-related tasks than their IGCSE counterparts. With these comments in mind, it should be noted that the findings from this qualitative and explorative study are not intended to be transferable or suggestive of the behaviours of wider populations. It intends to present a qualitative analysis of the practices of three undergraduate students who were provided with access to ChatGPT in an artificial assessment set-up.

The students had access to the ChatGPT "premium plan" to enable reliable access. They also had the choice of using either the version based on GPT-3.5 or the more recent GPT-4.0 version. Syllabus familiarisation training was provided. In this training, students gained an overview of the syllabus and its requirements, the Assessment Objectives and marking criteria, and they looked at an exemplar essay. After the familiarisation training, the students were invited to select two essay titles from 13 options that had been randomly selected in advance by the researcher. The titles were selected from genuine IR assessment titles submitted in November 2019.

Students were provided with the instructions shown in Figure 1.

Use ChatGPT to write a 1500–2000-word essay on your topic:

- Aim to make the essay look like it was written by a student
- Aim to make the essay high scoring
- Present the essay in Word
- Try to cite the sources used in the essay
- Keep your ChatGPT history for this task

You can use

- ✓ Example essay
- ✓ Suggested essay structure
- ✓ Wider internet access

Figure 1: Instructions provided to research participants

To avoid a scenario where the students deliberately authored poor essays, believing that these were reflective of the level of the typical (IGCSE) student, we asked them to try and make the essays "high scoring". As per the syllabus requirements, the essays needed to be presented in Microsoft Word and sources cited in a consistent way. As Figure 1 shows, the students were invited to use the training materials and internet as well as their prior knowledge. They were told that their approach was their decision and the researcher explained that the task was purposefully not over-prescriptive because we were interested in *how* they used the generative AI technology. It was also for this reason that the students were not provided with ChatGPT familiarisation training, although they each reported prior awareness of the technology.

Document comparison

Students retained the transcripts of their interactions with ChatGPT. These were submitted to the research team alongside the essays. An "overall plagiarism" percentage was calculated by comparing the ChatGPT transcripts to the students' final essays using Copyleaks' document comparison tool: "text compare". The two text types (the essay and chatlog) were input into Copyleaks' "text compare" for the tool to output a plagiarism percentage score. The score denotes how much of the essay had been copied and pasted or adapted from the ChatGPT outputs.

Copyleaks' developers state that the "text compare" tool works by using "advanced algorithms" which "[look] for matches within the submitted text" (Jacob, n.d.). The explanation continues to outline how it uses:

> "lexical analysis, semantic analysis, and machine learning [...to...] uncover even subtle instances of plagiarism [...]. The algorithm then does a deep-dive, using fuzzy matching to uncover patterns and stylometry to check for differences in writing style." (Jacob, n.d.)

The output documents highlight which sections of the student's essay have been flagged as which type of plagiarism. The identification of paraphrasing in particular could be more sensitive than a human evaluator, although to the best of our knowledge there are no publicly available studies comparing the "text compare" identification of paraphrasing with that of humans.

The term "plagiarism" is used throughout this article, although the extent to which each of the students engaged in academic malpractice is debatable. Further to the overall plagiarism percentage score, the analyses output a percentage score for each of three levels of plagiarism: "identical" where sentences or phrases were lifted word for word from ChatGPT's outputs; "minor amendments" where students made small changes to the ChatGPT content, and "paraphrased" content. The researcher compared the overall plagiarism score for each essay and student and made a qualitative and relative judgement between students as to whether each undergraduate had relied on ChatGPT to a "high", "medium" or "low" extent when constructing the essays.

Interviews

Upon the submission of their essays, the students were interviewed by one of three researchers about their experience of using ChatGPT in writing their essays. The purpose of the interviews was to gain insights into the process of using ChatGPT in producing essays in contexts where students had little previous knowledge of the topic.

The one-to-one semi-structured interviews took between 40 minutes and 1 hour. Overall, the students were asked (i) how they had used ChatGPT to write essays, (ii) how they had integrated ChatGPT-generated content into their essays, and (iii) how well they thought ChatGPT had helped with their essay writing. The interview protocol was followed and some follow-up clarifying questions were

asked by researchers, as appropriate. The interviews were audio recorded and transcribed using Microsoft Teams. The transcripts were then anonymised, edited, and analysed by applying thematic analysis. This approach included reviewing the transcripts, indexing segments into categories and, finally, identifying common themes across the interviews.

Ethical considerations

This research followed the British Educational Research Association's guidelines for conducting educational research. The students gave their written and verbal consent to participate in the research study and they were provided with opportunities to ask questions and to withdraw from the study. The identities of the students are obscured throughout the article by the use of pseudonyms and gender-neutral pronouns (e.g., "they" to refer to an individual).

Findings

To what extent did students rely on ChatGPT-generated outputs to form the content of their essays?

Following the Copyleaks analysis, it appeared that the students had taken three different approaches to the use of ChatGPT in their essay-writing tasks. When looking at the overall plagiarism percentage scores, Kim had the highest level of plagiarised content (with plagiarism scores of 70 per cent and 64 per cent for their essays). This finding indicates that Kim had interpreted the task in such a way that they relied heavily on the chatbot for essay content. Relative to Kim and Ronnie, Charly engaged in a "medium" level of overall plagiarism (44 per cent and 41 per cent), suggesting a relative mid-level of reliance on ChatGPT in constructing the essays. With Ronnie there was no evidence that they had copied and pasted (0 per cent) or minorly amended (0 per cent) text from the ChatGPT chatlog. Furthermore, there was very little evidence that ChatGPT generated content had been paraphrased (7 per cent and 4 per cent).

Table 1: Copyleaks' plagiarism scores for the ChatGPT-assisted essays

Essay	Student	Plagiarism overall	Identical	Minor changes	Paraphrase
Religion and conflict	Kim	70%	21%	17%	32%
Animal rights*	Kim	64%	13%	29%	22%
Legalising abortion*	Charly	44%	17%	12%	15%
Celebrities as role models	Charly	41%	2%	7%	31%
Capital punishment	Ronnie	7%	0%	0%	7%
Sweatshops	Ronnie	4%	0%	0%	4%

** Essay written with ChatGPT-4 instead of ChatGPT-3.5*

We interpret these different levels of overall plagiarism scores as indicating different levels of reliance on ChatGPT for essay writing in this task. In the findings below these approaches are referred to as "high", "medium", and "low" dependence, and this is a relative judgement made by the researchers by comparing the students' essays, plagiarism scores, and interviews.

Low dependence

Ronnie, who adopted the low plagiarism approach, was sceptical about the idea that ChatGPT alone, without human editing and input, could achieve good grades at university level. This scepticism is perhaps reflected in their approach to the task. As Table 1 illustrates, Ronnie did not simply plagiarise from ChatGPT. Instead, they developed the essay argument, and then used ChatGPT to elicit sources of information to vindicate their ideas:

> Ronnie: "[A] lot of the [essay] ideas just [depended] on the way that I wanted to steer it, and where I wanted to take it. Obviously, I was going to argue that [capital punishment] is not ethical because that's what I believe. So, I'd steer [ChatGPT] down that road …"

Here Ronnie explains that the essay argumentation and ideas were their own. Before engaging with the technology, they had decided that they would argue that capital punishment was unethical. Accordingly, they "steered" ChatGPT to provide information that was useful to the construction of this stance. To elicit the desired information, Ronnie needed to alter their prompts. They explained that ChatGPT's inbuilt safeguards initially mean that the system refused to answer their questions:

> Ronnie: "First of all it [ChatGPT] didn't like really want to register the word 'death' […] I switched to 'capital punishment' and then it went and started giving me a more detailed answer …"

Ronnie rapidly found a way around ChatGPT's safeguarding mechanisms by avoiding the word "death", and this allowed them to gain information relevant to the case that they wanted to argue. Although Ronnie later claimed that ChatGPT "doesn't really have an opinion", they found ways to force it to mimic an opinion, for example by asking the system to give a perspective from a particular standpoint (such as "from the perspective of someone that lived in Bangladesh" for the essay question about the ethics of "sweatshops").

Medium dependence

Charly adopted a middle approach with 41–44 per cent of their essays plagiarised in some way. Where they had adopted a "copy and paste" approach, they were concerned that this was at odds with the rest of their essay's style:

> Charly: "There were very rare moments where I just copied and pasted straight from GPT […] at the points where I did, it didn't feel real. It felt like a fact file rather than an essay."

Charly may have objected to the "fact file" style for aesthetic reasons and considered that the ChatGPT voice was ill-suited to an essay of this type. Whatever Charly's reasons, they indicated a preference for expressing content in their own words. This preference is partly borne out in the findings displayed in Table 1. For the essay about celebrities, Charly's use of "copy and paste" was limited to just 2 per cent, but they paraphrased a further 31 per cent of their essay material from ChatGPT's outputs. In the essay about legalising abortion, however, 17 per cent of the overall essay was copied directly from the ChatGPT chatlog

and 15 per cent was paraphrased. It is unclear why Charly adopted these slightly different approaches to ChatGPT use across the two essays. It could be that by the time Charly started work on their second essay (the legalising abortion essay), they were running short on time and so resorted to a higher degree of direct dependence on ChatGPT's outputs.

High dependence

Kim's essays were constructed more heavily from ChatGPT outputs compared to both Ronnie and Charly. Despite plagiarising between 64–70 per cent of the ChatGPT log, Kim noted that a complete "copy and paste" approach would be unlikely to be successful:

> Kim: "Overall ... [ChatGPT is] not great [for essay writing]. [...] you can't just [...] ask it to write it and then copy and paste it. It just wouldn't work. So I think, it definitely takes more work than you think it would initially do. I didn't think it was going to take 4 hours to write an essay with GPT, whereas it does."

Kim was unimpressed with the idea of writing an essay with only ChatGPT. In line with the other two students, they felt that this "wouldn't work". Even though Kim's essays included the greatest proportion of content plagiarised from the ChatGPT transcript, they undertook "more work" than they initially had anticipated would be necessary. Much of this work was to do with the selection and synthesis of information, which was also remarked on by Charly.

Kim's interview indicated that they found the process of using ChatGPT unfulfilling. When asked about their perception of the quality of their essays, Kim responded:

> Kim: "I literally have done no work for this assignment. But somehow, I managed to do OK. [...] it was like copying and pasting and just more like trying to find the sources where they got the information from rather than just like you know, researching it yourself. I don't know — I think it [that] it didn't feel as momentous, you know?"

They felt dissatisfied with the process of essay writing with ChatGPT because they perceived that they had "done no work for this assignment", and they viewed their input largely in terms of locating sources and "copy and pasting". Kim may have perceived the skills required to elicit information from ChatGPT, verify sources, and select and synthesise ChatGPT-generated content as less valuable than those needed to "[research] it yourself". As a result, Kim felt underwhelmed rather than pleased with the "momentous" achievement of having researched, processed, synthesised, structured, and authored high-quality work. Kim's comments open a further question about the way in which engagement with ChatGPT affects the constructs measured in a coursework assessment and the motivation and satisfaction of learners.

How did students interact with ChatGPT in the process of essay writing?

Each of the students used ChatGPT-generated content in different ways and to different extents. Nonetheless, the analysis found similarities in their process of writing essays while interacting with this technology. This process (shown in Figure 2) comprised of:

1. orientation
2. specific enquiries
3. occasional verification
4. structuring; and
5. writing up.

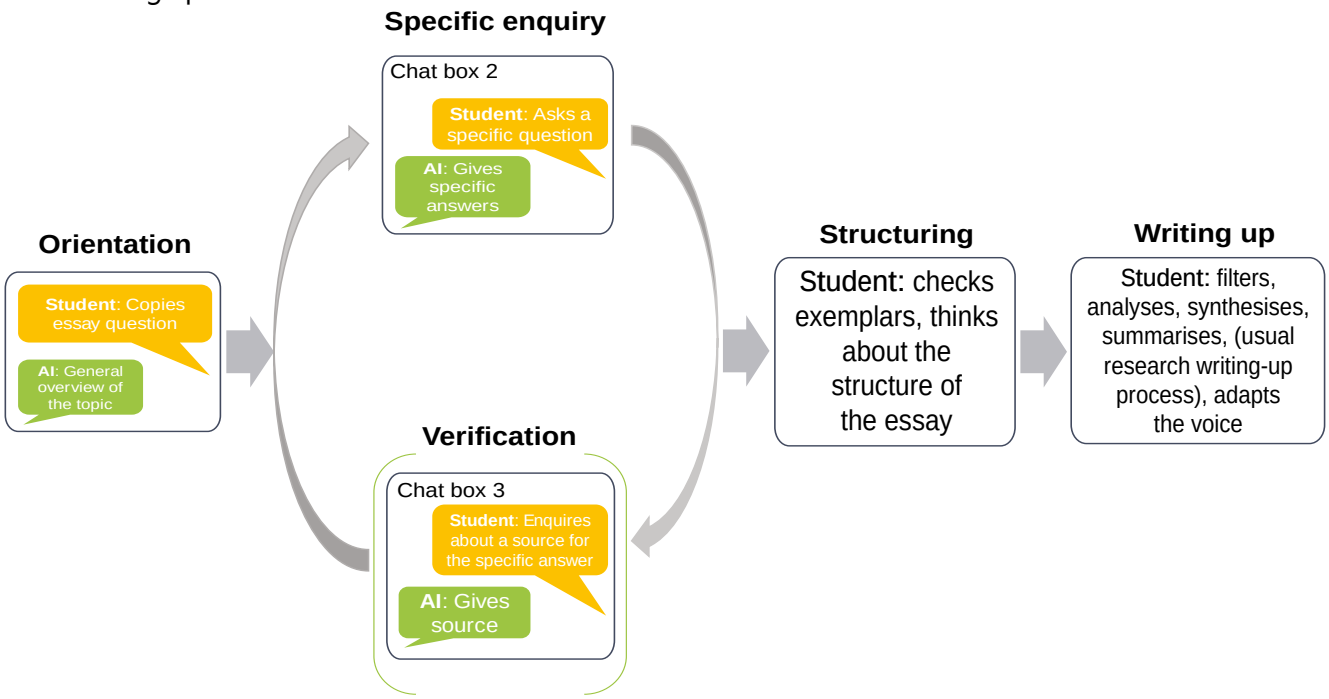

Figure 2: Students' essay-writing process using ChatGPT. The verification step is presented in brackets to show that students did not consistently verify information.

Orientation and specific enquiries

Students reported starting off by entering the essay question into the ChatGPT chat box or asking a general question about the topic ("orientation"). It helped students, who did not know much about a topic, to get a general overview of it. This was followed by multiple "specific enquiries"; students would start multiple chats or enter new questions into the same chat to enquire about specific aspects of the topic (Figure 1):

> Charly: "So, I'd start new chats [...] the pattern I followed was: I'd ask it a general question like [what is the] Christian debate on abortion, for example, and it would give maybe four or five points. And then after that, I'd ask more with details, facts and statistics."

Students were, overall, quite impressed with the speed of access to information through ChatGPT. They felt that the purpose of ChatGPT was very similar to the purpose of any other internet search engine – to provide information, just much quicker and more user-friendly:

> Ronnie: "It's just [...] a better Wikipedia. It's like times... infinity better [than] Wikipedia."

Verification

Occasionally, students attempted to verify the ChatGPT-generated content ("verification") by either asking ChatGPT to provide the source or by searching for the sources without the help of ChatGPT by using internet search engines. The students had a strong sense of ethics around verification of the AI-generated content and referencing, but when asked about verification practices and referencing used in their essay writing, they reported they often skipped this part of the process. This was due to the difficulty of verifying the content, time constraints and a lack of understanding around the expected standards of verification and referencing:

> Kim: "[...] we didn't have time to [...] verify it. I probably should do that because it's probably [...] I don't know how reliable it is [...] it doesn't specifically give you sources when it gives you the information. You can't do that for every piece of information that you're gonna get, because otherwise, it's just gonna go on."

Tracking down the source of ChatGPT's information was frustrating. Kim felt that it "would have been better" if they had found and referenced the sources in the first place, rather than relying on ChatGPT and later trying to verify its information, which is now possible (e.g., see Microsoft Copilot).

Ronnie, on the other hand, used ChatGPT output and "searched on Google Scholar" to find "text that link[ed]" to the ChatGPT output. They would then reference the Google Scholar source. With this method of searching for and then citing likely-sounding sources in the essay, Ronnie found that all the output they wished to include was "quite supportable" mainly because there was "nothing outrageous" to justify.

Structuring

At the next "structuring" stage, the students gathered all the content that ChatGPT generated and started thinking about the form of their essays. They reported checking the exemplar essay to gain an idea of an appropriate format and content type. Ronnie and Kim started to structure their essays after the orientation stage, whereas Charly first gathered 22 pages of information from ChatGPT before starting to structure the essay. The structuring process was not contained or linear for any of the students and all showed evidence of returning to the stage at different points in the essay-writing process. In the structuring stage, the students reported difficulties in obtaining appropriate introductions and conclusions from ChatGPT:

> Charly: "I messed around with trying introductions and conclusions. And they were very poor because [ChatGPT] doesn't come to a conclusion."

Writing up

In this stage, students needed to decide how to write up the essay so that it had a coherent argument and followed the component's suggested essay structure.

This process was complex and required higher-level critical thinking skills, including analysis, synthesis, and evaluation of the ChatGPT-generated content (and students were aware of that):

> Charly: "[Using ChatGPT is] not that bad because you can then apply your research skills and select and synthesise. You will get a very low mark just [by] using ChatGPT. [...] You need to have [...] the skills developed [...] but then I feel like those skills are developed from not using a source like ChatGPT [...] it's [...] a paradox."

Integrating ChatGPT-generated content into an essay required an adaptation of style and voice. The ChatGPT outputs were often deemed unsuitable for copying verbatim into the essays because they lacked "style" and were "too logical":

> Kim: "You make it more so that it sounds like it's more appropriate for that essay, as your own voice [...] because [...] it was very directive... it's more third person [...] It's more of a telling rather than like you're explaining."

As a further observation, the students did not use the entire set of functionalities offered by ChatGPT (e.g., restructuring, copyediting, proofreading, and providing feedback). This could be due to the lack of training and time pressures. Finally, at the time of this study, ChatGPT had been widely available for only four months, which means that students may not have had sufficient time to familiarise themselves with all its features. It is reasonable to presume that students' general familiarity with ChatGPT will be significantly more advanced by the time this journal article is published.

Conclusion

As suggested throughout the article, there are several limitations that must be recognised in the interpretation of the data and presentation of findings. Unlike a naturalistic setting, the research participants were explicitly asked to use ChatGPT to write their essays. As such, they may have engaged with generative AI to a greater extent than if they had been genuinely studying towards this qualification. Secondly, the research participants were undergraduates aged 18–22, and it should not be assumed that students in this category reflect the behaviours of a younger cohort who may have a different approach and skill level when it comes to engaging with generative AI. Thirdly, the undergraduate students had little previous knowledge of the essay topics, and had only two days to produce the essays. These factors may have affected their approach to the process of essay writing compared to a naturalistic setting.

Despite the limitations, findings from this research provide an indication of how the selected students engaged with ChatGPT and offer insight into their perceptions of the utility and ethics of using such a tool to assist with essay writing in an assessment context. Notably, despite different levels of reliance on ChatGPT, the students used the technology in a similar way: primarily as an information gathering and producing tool. There was limited evidence of them exploiting ChatGPT to its full potential, as an editor, proofreader, or to provide formative

feedback – for example. As previously noted, this seemingly limited awareness of ChatGPT's potential may be because the students had not explored it in depth prior to the research task and did not have time to test it or be creative with it during the essay-writing task.

The students understood that ChatGPT generated both accurate and false (and outdated) information, and they did not always verify the information provided to them. They recognised this as a problem and suggested that ChatGPT's overall capabilities and outputs were not of a high enough standard to facilitate top marks in an essay at IGCSE level or above. As such, they may not have deployed ChatGPT's full suite of uses because they did not believe that these would add value to their essays. Had the students been provided with ChatGPT familiarisation training or with the addition of exploration time prior to starting the essay-writing tasks, they might have uncovered the technology's capabilities, used it in more varied ways or been more impressed by its functions and applications. Similarly, had the LLM output included the sources behind its content (for example, see Microsoft Copilot), the students might have interacted with it in a different manner and had more confidence in using the AI-generated content.

As the technology evolves and as users become more accustomed to its potential applications and uses, students such as those in our study sample may develop into more skilled users of generative AI and they may perceive that it can, in fact, outperform humans in tasks such as essay writing at IGCSE level.

As well as the students' perceptions, this research has highlighted the importance of higher-order thinking skills for AI-assisted essay writing, and that is unlikely to change any time soon. The students reported challenges in using ChatGPT for content generation because they could not easily verify the chatbot's outputs, nor did they find ChatGPT's default voice to be appropriate to their task. These findings accord with current wider claims that ChatGPT does not necessarily excel in these areas (University of Cambridge, 2023). Given that students with lower academic performance often also display poor critical thinking (Behrens, 1996; Fong et al., 2017) and poor metacognitive skills (Pintrich & De Groot, 1990; Young & Fry, 2008), it would be fair to assume that low-performing students in particular could be potentially negatively affected by using ChatGPT for content generation.

With a view to the future, other research in the area of AI and assessments has mapped out what is possible and what is desirable (Abu Sitta et al., 2023). Such research explores how to capitalise on AI in such a way that it enhances rather than diminishes human capabilities. Future-oriented research combined with research about current practice and engagement in AI can help to inform institutional policies and guidelines which are under continual review, given the fast-developing nature of the area. It could be useful to include undergraduate students' voices in the design of such policies and guidelines because, as the students in this study have shown, they may have valuable perspectives on what ethical and legitimate use of generative AI could look like in an assessment context.

References

Abu Sitta, F., Maddox, B., Casebourne, I., Hughes, S., Kuvalja, M., Hannam, J., & Oates, T. (2023). *The futures of assessment: Navigating uncertainties through the lenses of anticipatory thinking*. DEFI & Cambridge University Press & Assessment.

Behrens, P. J. (1996). The Watson-Glaser critical thinking appraisal and academic performance of diploma school students. *Journal of Nursing Education, 35*(1), 34–36.

CMS/W. (2023, January 13). *Advice and responses from faculty on ChatGPT and A.I.-assisted writing*. MIT Comparative Media Studies/Writing.

Dalalah, D., & Dalalah, O. M. A. (2023). The false positives and false negatives of generative AI detection tools in education and academic research: The case of ChatGPT. *The International Journal of Management Education, 21*(2), 100822.

Dale, R. (2021). GPT-3: What's it good for? *Natural Language Engineering, 27*(1), 113-118.

Dippenaar, B. (2023, June 12). *ChatGPT and AI: Navigating uncharted copyright territory*. Lexology.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., ... Wright, R. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71*, 102642.

Eke, D. O. (2023). ChatGPT and the rise of generative AI: Threat to academic integrity? *Journal of Responsible Technology, 13*, 100060.

Fong, C. J., Kim, Y., Davis, C. W., Hoang, T., & Kim, Y. W. (2017). A meta-analysis on critical thinking and community college student achievement. *Thinking Skills and Creativity, 26*, 71–83.

Galaczi, E. (2023). *English language education in the era of generative AI: Our perspective*. Cambridge University Press & Assessment.

Gregorcic, B., & Pendrill, A.-M. (2023). ChatGPT and the frustrated Socrates. *Physics Education, 58*(3), 035021.

Jacob, S. (n.d.). Copyleaks Plagiarism Review – Originality.AI. Copyleaks. Retrieved 8 January 2024, from https://originality.ai

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274.

King, M. R. & ChatGPT. (2023). A conversation on artificial intelligence, chatbots, and plagiarism in higher education. *Cellular and Molecular Bioengineering, 16*(1), 1–2.

Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. (2023). The No. 1 question to ask when evaluating AI tools. *MIT Sloan Management Review, 64*(3).

Lee, J. Y. (2023). Can an artificial intelligence chatbot be the author of a scholarly article? *Journal of Educational Evaluation for Health Professions, 20*(6).

Marcus, G., & Davis, E. (2020, August 22). *GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about.* MIT Technology Review.

Milmo, D. (2023, February 2). ChatGPT reaches 100 million users two months after launch. *The Guardian*.

Perkins, M. (2023). Academic integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching and Learning Practice, 20*(2).

Pintrich, P. R., & De Groot, V. E. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*, 33–40.

Stokel-Walker, C. (2023). ChatGPT listed as author on research papers: Many scientists disapprove. *Nature, 613*(7945), 620–621.

University of Cambridge. (2023). *ChatGPT (We need to talk)*.

Yosifova, A. (2023, June 28). *ChatGPT: How to understand and compete with the AI bot*. 365 Data Science.

Young, A., & Fry, J. D. (2008). Metacognitive awareness and academic achievement in college students. *Journal of the Scholarship of Teaching and Learning, 8*(2), 1–10.