RESEARCH ARTICLE

Exploring the use of AI in mathematics and statistics assessments

Siri Chongchitnan, Warwick Mathematics Institute, University of Warwick, Coventry, UK.
Email: Siri.Chongchitnan@warwick.ac.uk
Martyn Parker, Department of Statistics, University of Warwick, Coventry, UK.
Email: Martyn.Parker@warwick.ac.uk
Mani Mahal, Department of Statistics, University of Warwick, Coventry, UK.
Email: Mani.Mahal@warwick.ac.uk
Sam Petrie, Warwick Mathematics Institute, University of Warwick, Coventry, UK.
Email: Sam.Petrie@warwick.ac.uk

Abstract

The mathematical sciences and operational research (MSOR) community in higher education is still largely unprepared to adapt to the rapid rise of generative artificial intelligence (genAI) and its impact on assessment strategies. Whilst in-person exams remain an essential assessment mode for MSOR, take-home assignments are also an integral assessment tool. This work investigates concerns that current assignments are not robust against genAI and the way students use genAI. In this work, we address the following questions: 1) How well can genAI perform in current assignments? 2) To what extent do students currently use AI in take-home assignments? 3) How should assessment strategies evolve given the rapid improvement of genAI? Our research involves an investigation of genAI's performance in a range of MSOR assignments. We also conducted surveys and discussions with mathematics and statistics students and staff at the University of Warwick. We make recommendation and conclude that genAI represents a catalyst for innovation and assignments, perhaps adapted, should remain a core assessment in MSOR.

Keywords: Generative Artificial Intelligence, Mathematics, Statistics, Assessments.

1. Introduction

The mathematical sciences and operational research (MSOR) community, like all disciplines in higher education, needs to address the rapid integration of advanced AI technologies into academic environments. While in-person examinations have traditionally been the primary method of assessment in these disciplines, take-home assignments remain a critical component for evaluating student knowledge and problem-solving skills (Iannone and Simpson, 2011, 2012, 2022). The emergence of genAI presents challenges to the integrity of these assignments.

The primary purpose of this work is to explore and understand the impact of generative artificial intelligence (genAI) on mathematical assessment focusing on the (paid) large language model (LLM) GPT-4o. By examining the capabilities and limitations of GPT-4o, this article project aims to provide insights that will inform assessment strategies within the MSOR community. Initial evaluation of other genAI models demonstrated that GPT-4o provided the best responses, thus this work focuses on this model.

The research presented in this work was carried out at the University of Warwick, a large UK university where there are around 2000 taught (Undergraduate (UG) and postgraduate (PGT)) students in the mathematics and statistics departments.

This paper covers three areas.

1. How is AI's performance on current assignments? Evaluate GPT-4o's ability to solve university-level mathematics and statistics assignments.

2. Examine students' use of AI. Determine how are students using genAI to complete their assignments. What are their perceptions and understandings of these tools?

3. Assessment strategies. Discuss how assessment strategies should evolve, given the rapid improvement of genAI.

The full report for this work is available online (Chongchitnan, et al., 2024).

2. The Emergence of ChatGPT and Its Impact

ChatGPT (OpenAI, 2022) fundamentally altered the educational landscape virtually overnight. Students could suddenly, instantly, and for free, obtain answers that far exceeded the capabilities of AI task managers or search tools like Siri or Google Assistant. This shift raised concerns about the integrity of academic assessments, particularly in essay-based subjects where students could easily generate large portions - or even entire assignments - within seconds.

At the time of its release, ChatGPT was powered by a single LLM: GPT-3.5. This model quickly became synonymous with the ChatGPT brand and remains, according to our study, the most popular version used by students nearly two years later, despite being replaced by GPT-4o mini. GPT-3.5, like GPT-4o mini, was always offered for free with usage limits.

GPT-3.5 capabilities are limited by its training data, which often includes both accurate and inaccurate information (OpenAI, 2022; Huang et al., 2023). This limitation affects its performance in mathematical contexts, where rigorous logic and structured reasoning are required through multiple steps.

Since LLMs generate answers to mathematical problems through the same probabilistic mechanism used for text generation, it is not unusual to find counting or other basic mathematical errors. OpenAI provided a generic warning at the bottom of all chats that "*ChatGPT can make mistakes*". This phenomenon, where the model produces responses that seem accurate or correct but are underpinned by flawed reasoning, is known as *hallucination.* As a result, many students who initially experimented with GPT-3.5 developed a negative perception of the capabilities of LLMs broadly, but particularly in MSOR subjects (Attewell, 2024; Das and Madhusudan, 2024).

Despite these limitations, many students surveyed at Warwick use these genAI models to help with their assignments, to produce code or to act as a "study buddy", with most students relying on GPT-3.5 at the time. Some students do critically evaluate the outputs, whilst others do not, with staff reporting an increase in genAI misuse.

3. Performance of AI on university-level work

3.1 Methodology

We collected 122 assignment questions from mathematics and statistics lecturers, who submitted questions from their modules across Years 1 to 4 (FHEQ Levels 4 to 6). The questions were presented to GPT-4o with a *zero-shot* approach, i.e. the AI received no additional guidance or prompting beyond the wording in each question. We classified each question into one of two types:

Proof type. This includes questions that ask for a chain of logical reasoning, often using previous lemmas or theorems. This type of question typically requires little numerical calculations.
Examples:

- (Y1) Prove that the composition of two bijective functions is bijective.
- (Y2) Show that the partition function $p(n)$ satisfies a given recursive inequality.
- (Y3/4) Prove that a given Lie algebra is semisimple.

- ***Applied type.*** This includes questions that ask for a concept to be applied to a specific situation, requiring some symbolic manipulation or numerical calculations. The answer is typically a concrete expression, a number, a graph or code.
  Examples:
  - (Y1) Find a particular integral for a given ODE.
  - (Y2) Calculate the first three terms in the asymptotic series of a given integral.
  - (Y3/4) Suggest a proposal density for rejection sampling from a given bivariate distribution. Verify your answer by implementing it in R.

We performed the proof/applied classification to test the hypothesis that genAI is prone to making computational errors in applied-type questions, and less likely to make mistakes in proof-type questions, where the answers are more likely to be in the training data. The split between proof and applied types is shown in Table 1.

Table 1. The distribution of the 122 questions we tested by Year (1, 2, 3/4) and by type (proof, applied).

	Year 1	Year 2	Years 3/4
Proof	27	16	21
Applied	35	7	16
Total	62	23	37

We rated the correctness of GPT-4o's answers on a three-tier (traffic light) scale, where:

- **Green** (70%-100%) indicates a good solution. If produced by a student, it would demonstrate a good understanding of the topic, possibly with a few errors.
- **Yellow** (35%-69%) indicates an adequate or passable solution. If produced by a student, it would show a fair or satisfactory understanding of the topic, with some errors.
- **Red** (0%-34%) signifies a poor solution. If produced by a student, it would indicate a lack of understanding of the topic with fundamental errors.

This scale allows us to quickly analyse questions from a wide range of topics. This system also allows us to obtain an aggregate (expected score) for each year by giving each question the mean score in each category, i.e.

Expected score in each year = $\frac{85\times N_{green}+52\times N_{yellow}+17\times N_{red}}{N_{green}+N_{yellow}+N_{red}}$,

where N_i is the number of questions judged to be in category i.

In addition, the lecturers who submitted the questions were asked to re-mark a sample of 35 out of the 122 responses (approximately 30%) and rate them in terms of correctness and in three additional metrics:

- ***Similarity*** to student work (0-100%): A high score means the AI-generated solution closely resembles a typical student submission.
- ***Detectability*** as AI-generated (0-100%): A high score means the solution can be easily identified as AI-generated.
- ***Adaptability*** into student work (0-100%): A high score means the AI-generated output can easily be modified into what appears to be a genuine piece of student work.

The sample was chosen to cover a range of levels and assessment types, and the size was selected so that lecturers were not overburdened with additional work.

3.2 Results

Correctness

The performance of GPT-4o is shown in Figure 1. We see that it performed well on Year 1 assignments, achieving a first-class score. For Years 2 to 4, the performance declined. Lecturers noted that answers to proof questions were often vague, lacked detailed reasoning, or contained significant errors. The AI also struggled with complex multi-step logical arguments. The performance was not uniformly good. Overall, the performance of GPT-4o was comparable to an undergraduate at a mid 2:2 level.

	Year 1	Year 2	Year 3/4
Expected Score	74%	48%	56%

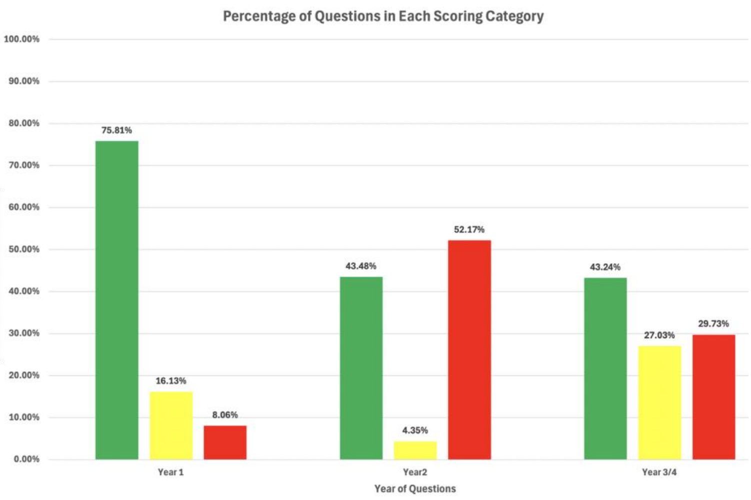

Figure 1. The average score of GPT-4o's answers across various years, and the correctness of the answers evaluated on a traffic-light scale.

Figure 2- shows a performance of proof vs. applied questions. The table shows broadly similarly performance across proof- and applied-type questions. This suggests limited evidence that GPT-4o is better at proof rather than applied questions.

Expected Score	Year 1	Year 2	Year 3/4
Proof	79%	55%	53%
Applied	74%	47%	56%

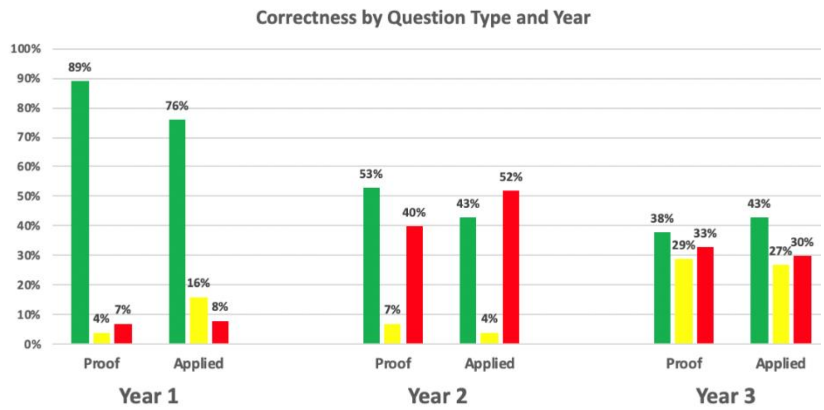

Figure 2. The correctness for proof and applied questions across all years. The scores lend weak support for the hypothesis that GPT-4o is better at proof-type than applied-type questions.

Similarity and Detectability

The similarity score (averaged across all questions in all years) is 62% (see Figure 3), although the distribution is wide. The responses indicated no significant differences between answers to proof-type and applied-type questions.

The detectability score is 53%, signifying some ambiguity in the authorship, again with negligible difference between proof-type and applied-type questions. Lecturers observed that AI-generated responses sometimes included unusual phrasing, excessive verbosity, or atypical grammar — features that could indicate AI authorship.

Adaptability

The adaptability score is 77% (see Figure 3), indicating that answers with AI characteristics could be easily modified by students to resemble their own writing style, e.g. by correcting obvious errors, adjusting the language, and removing AI tell-tale signs.

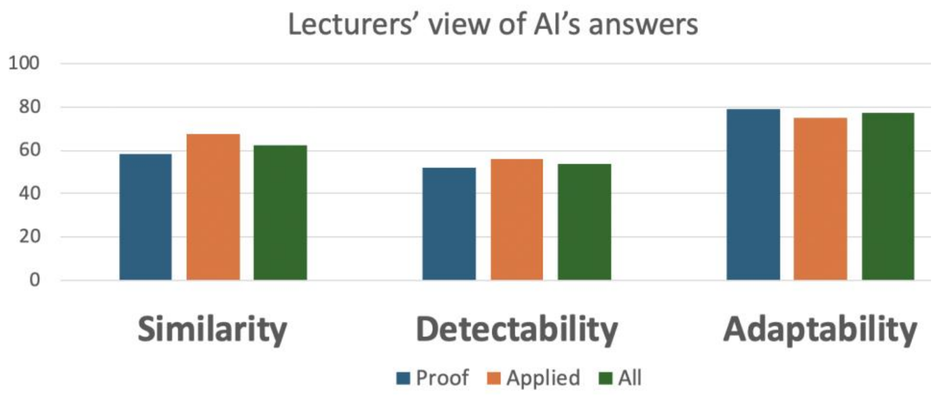

Figure 3. Lecturers' view of GPT-4o's answers, judged in terms of similarity to student work, detectability as AI, and adaptability into student work. The scores are averaged across all years.

These results highlight the nuanced capabilities of GPT-4o. While it demonstrates strong performance on simpler tasks, its limitations in complex reasoning do not necessarily prevent potential misuse if complemented by student critical evaluation of the outputs.

4. How students use genAI.

4.1 Methodology

An online survey of 145 mathematics and statistics students was conducted in June 2024 to assess their use of AI tools, ethical considerations and their attitudes towards AI. This sample represents approximately 7% of the UG and PGT population. Those completing the survey could opt in to a focus group. The respondents consisted of 86 (59%) that declared themselves AI-users (i.e. have used AI tools like ChatGPT for university work) and 59 (41%) non-AI-users.

From those that opted in, a random sample were selected for two focus groups of 6 individuals each. One group comprised AI users and the other non-AI-users.

4.2 Survey outcomes

The survey covered the following areas: ethical considerations, academic integrity, impact on degree value, student attitudes, AI assessment integration, usage patterns, and future concerns.

Three figures on the following pages summarise the survey outcomes.

Figure 4 provides a summary of questions about the students' attitude towards AI, with responses on a Likert scale (strongly disagree to strongly agree), the total number of responses in each category and their respective percentages.

Figure 5 provides a summary of responses regarding frequency of AI use. Figure 6 shows the choice of AI (if any) used by the participants, with ChatGPT being the most popular.

From these results, we made the following general observations (Chongchitnan et al., 2024).

- **Perception of cheating.** Most students regard using AI as cheating, even amongst those who have used AI in assignments.
- **Support for AI-proofing measures.** There is support for proactive measures to mitigate AI misuse, although the effectiveness of such strategies was challenged.
- **Scepticism towards AI accuracy.** Students believe that AI often provides incorrect answers to mathematics and statistics questions.
- **Apprehension about AI's role in future careers.** Students worry that AI might devalue employable skills or make them obsolete.
- **Resistance to shifting assessment methods.** Students are opposed to moving entirely to in-person exams and removing assignments altogether. This suggests a preference for maintaining a mix of assessment methods, highlighting the value students place on assignments as part of their university education.
- **Uncertainty about AI integration.** There was widespread ambivalence about the use of AI in assignments. This uncertainty was shared almost equally between AI users and non-users, suggesting that even those familiar with AI tools remain unsure about the appropriate role of AI in higher education.
- **Ethical concerns.** Some students, particularly non-AI users, refrain from using AI tools due to ethical concerns, such as the fear of cheating or undermining academic integrity. This hesitancy highlights the importance of establishing clear guidelines and educating students on the ethical use of AI in academic settings.
- **Diverse usage patterns among AI users.** While some students use AI tools regularly for assignments, the majority use them sparingly, often for specific tasks like coding assistance

or clarifying concepts. This suggests that AI is being integrated into student work more as a supplementary tool rather than a primary resource.

These findings demonstrate that students, regardless of their personal use of AI, are acutely aware of, and concerned about, the ethical implications of AI in education. It is also interesting to contrast the results in Section 3.2 (GPT-4o's performance) with student perceptions: Whilst GPT-4o can produce accurate and inaccurate responses, only those able to critical evaluate these responses can judge their value and gain educational benefit from genAI.

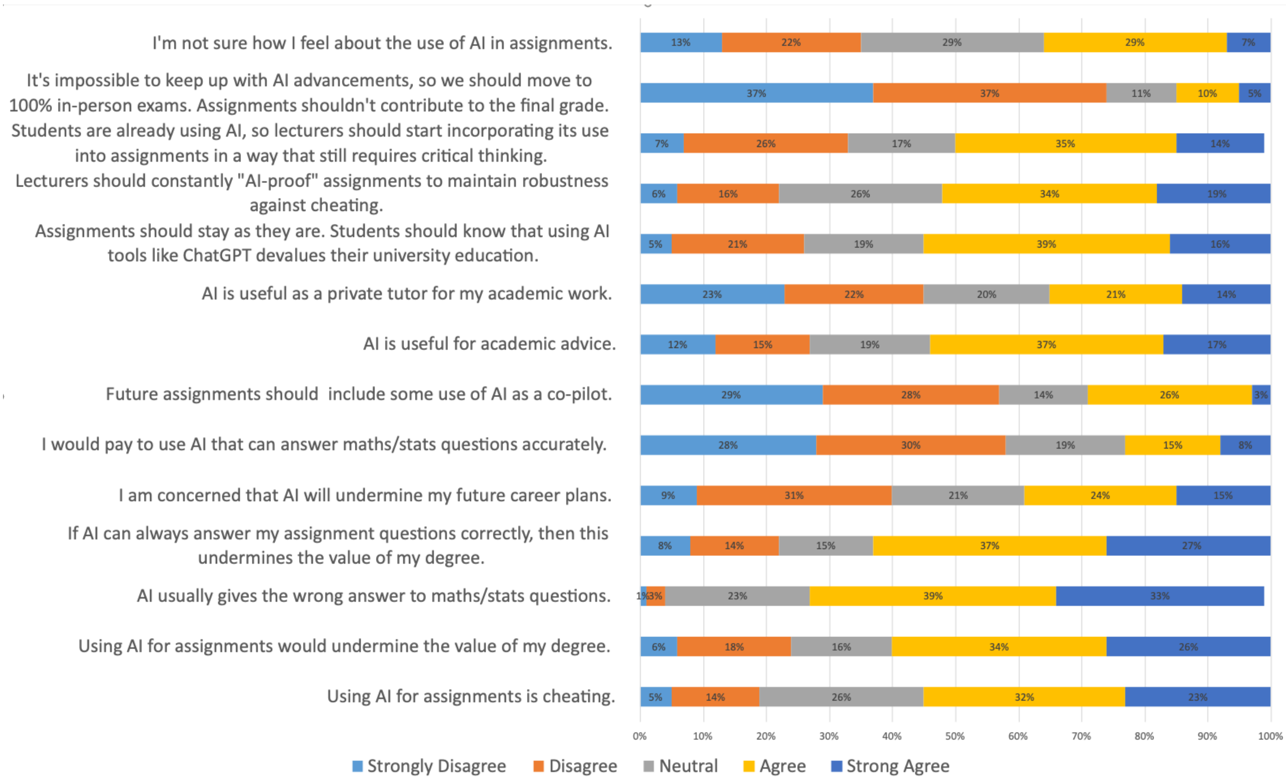

Figure 4. Survey questions with Likert-scale responses. The numbers indicate the percentages in each category.

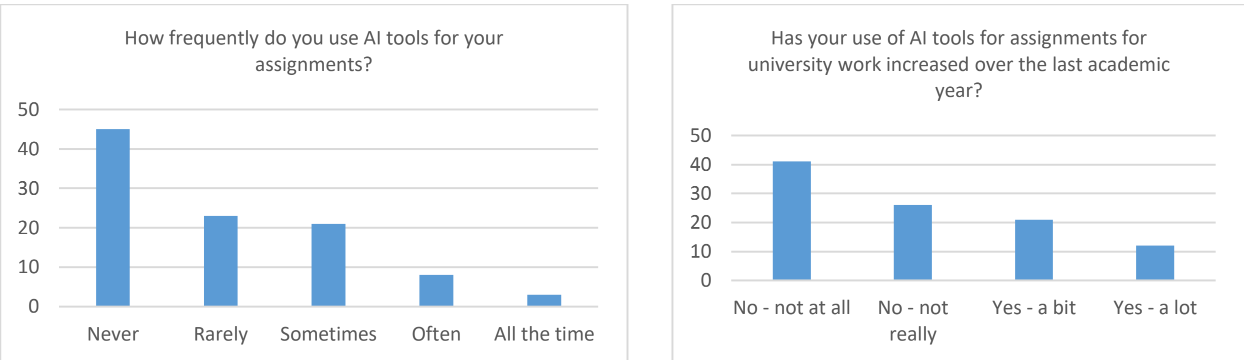

Figure 5. (Left) Percentage responses to the question "*How frequently do you use AI tools for your assignments*?" (Right) Percentage responses to the question "*Has your use of AI tools for assignments for university work increased over the last academic year*?"

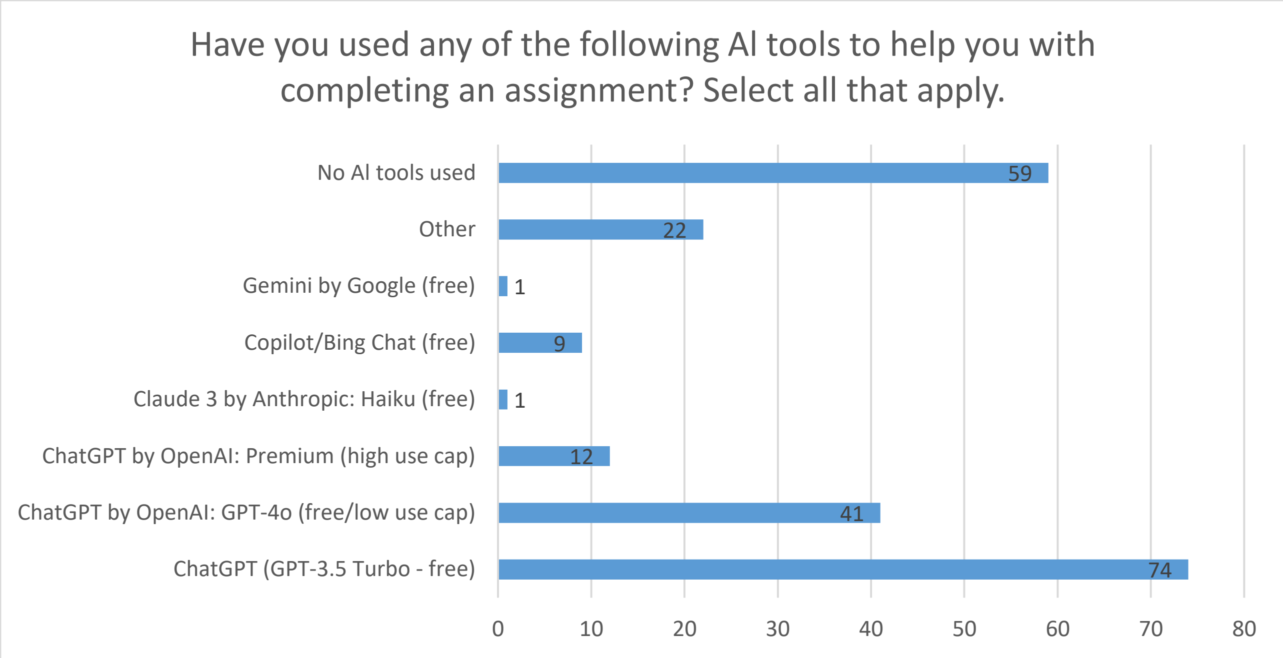

Figure 6. Summary of responses to the question "*Have you used any of the following AI tools to help you with completing an assignment? Select all that apply*."

4.4 Focus-group outcomes

This section presents key insights from focus group discussions conducted separately with AI-users and non-AI-users. The discussions aimed to capture perspectives on how genAI tools like ChatGPT are impacting learning experiences, academic integrity, and future career preparedness.

The discussion is broken down into five thematic areas: Experiences and Attitudes Towards AI, Ethical Considerations and Academic Integrity, Impact on Learning and Skills Development, The Future of AI in Education: Hopes and Fears, and Recommendations for AI Integration. Table 2 provides a summary of the key insights from these discussions. A full analysis is presented in the main report (Chongchitnan et al., 2024).

The focus groups provide individual student thoughts. For example, those who have used AI tools appreciate the support these technologies offer in studying complex concepts.

> *"One thing that* [genAI] *has an edge over asking your professors is the ability to clarify things that you don't really understand in the moment. For example, when I've been reading through my lecture notes and noticed a contradiction, I can scrutinise ChatGPT's answer line by line and ask it again, like why is it doing this?"* — Student D (AI user)

> "*I think it's served as quite a useful tool to replace Googling things... ChatGPT maybe gives you a method that helps you find* [the answers] *a little bit faster... it'll give you that little tip you need in the question to get to the next part.*" — Student C (AI user)

Some students highlighted genAI's limitations in handling advanced problems. Most AI users initially used AI models like GPT-3.5, which has shaped student views. Our work used the more advanced GPT-4o which, although better, showed inconsistent performance.

"*It's hilariously bad at maths. It's very rarely provided anything more useful than just guessing and checking.*" — Student F (AI user)

"*It's hopeless at answering any of my assignment questions.*" — Student A (AI user)

There are interesting contrasts when considering academic integrity. For example,

"*I don't really think I can consider it cheating per se because it just doesn't really give you answers*." — Student C (AI user)

"*People who are going to cheat, they're going to cheat... it's just another tool that's out there.*" — Student 4 (non-AI user)

In terms of learning experience, some provided examples of how they had used AI as a study support tool.

"*It's good for revision plans... it gave me a balance for the 11 exams that I had and helped me prepare for it*." — Student A (AI user)

Others recognised the potential for isolation.

"*I think it could be really detrimental in the fact that it cuts out that communication or that working together aspect of the degree*." — Student 5 (non-AI user)

There were interesting comments regarding potential future usage, both positive and negative.

"*I think AI will become more of a personal assistant/personal tutor that's essentially 24/7 available*." — Student C (AI user)

"*Encouraging people to use AI in their learning promotes bad habits and laziness*." — Student F (AI user)

"*If it reaches the point where it is doing our assignments... then what is the point in our degree at all?*" — Student 4 (non-AI user)

"I think it would create a situation where only students from really high wealth backgrounds are able to access that and then they'd have an extra leg up." — Student 6 (non-AI user)

Table 2. Summary of key insights from the thematic areas identified during focus group discussion.

Area	Discussion area	Key insight
Experiences and attitudes towards AI	Students' initial reactions to AI in academic settings, frequency of AI use, and overall attitudes towards incorporating AI into assignments.	AI users found AI tools helpful for coding and clarifying concepts, while non-users expressed scepticism about AI's reliability and were concerned about its potential to undermine learning.
Ethical considerations and academic integrity	Students' views on the ethical implications of using AI in assignments and whether they perceive AI use as cheating and how this perception differs between AI users and non-users.	Non-AI users largely view AI use in assignments as cheating, expressing concerns about fairness and academic integrity. AI users see it as a tool for assistance rather than a means to cheat.
Impact on learning and skill development	How AI usage affects students' learning processes and skill development, considering the benefits and potential drawbacks of AI in supporting academic growth.	AI users reported that AI helps them understand complex ideas and save time, but they also acknowledge the risk of over-reliance and encountering misinformation.
The Future of AI in Education: Hopes and Fears	Students' perspectives on the future integration of AI in education, including how AI could enhance learning and their fears about potential negative impacts on their degrees and careers.	Students are concerned that AI could devalue degrees and reduce the need for critical thinking, but they also see potential for AI to personalise learning and assist with routine tasks.
Recommendat-ons for AI Integration	Recommendations from students on how AI could be integrated into education including suggestions for guidelines, policy development, and educational practices.	Students would like clear guidelines on AI use, equitable access to AI tools, and assignments that still demand critical thinking and problem-solving skills.

5. Conclusion and discussion

The findings from this study emphasise the need for MSOR educators to develop assessment strategies and policies in response to the rapid development of genAI. We make five recommendations and suggest potential implementations.

1. **Acceptance**. AI is an integral part of the educational landscape, and entirely ‘AI-proofing’ assessments is not feasible. Although some advocate for 100% controlled-conditions assessments, this does not seem feasible. MSOR needs to create AI-ready graduates. Working with AI will involve acknowledging its capabilities and limitations, and integrating it into learning in a manner that enhances education while maintaining academic integrity, for example, when used as a study buddy or for giving additional feedback (Meyer, 2024)

2. **Assessments strategies** should be developed collaboratively with educators and students, fostering innovation and ownership to develop shared ownership of AI potential in MSOR.
3. **Demystifying AI**. Whilst most universities have drawn up generic AI policies, the MSOR discipline has unique characteristics (QAA, 2023). Departments should work with students and staff to clarify the usage policy of AI specifically in MSOR, and educate those who may feel ambivalent about using AI on its benefits and ethical usage
4. **Open dialogue and collaboration.** Encouraging conversations among students, staff and administrators could help address concerns and misconceptions about AI. Co-creation projects and collaborative work could help keep pace with technological advancements, student attitudes and evolving academic practices in the MSOR sector.
5. **Professional development.** The introduction of GPT-o1 which specialises in solving mathematical problems and the anticipated arrival of GPT-5 highlight the need for proactive approaches to maintain the quality and relevance of mathematical assessment in higher education.

GenAI provides new opportunities for innovation and to co-create initiatives where both students and lecturers engage in learning about AI tools together. The keen interest from both staff and students provides a strong opportunity to jointly critically evaluate AI in various ways.

Example 1. Students can learn to verify the accuracy and reliability of genAI in academic work. These initiatives are likely to be more formative than summative and could become part of small-group tutorial work. These sessions should encourage participants to take ownership of their learning by critically assessing AI outputs, understanding the implications of AI-generated content, and discussing the ethical responsibilities associated with AI use.

Example 2. Students create instructions on how to effectively use AI for academic tasks. This may include using AI for summarising notes, finding quotes, creating personalised learning experiences, understanding complex topics and compiling revision schedules.

Example 3. There is the opportunity to examine how assessments can be structured so that AI usage and critical evaluation is encouraged. For example, genAI can be used to generate variations of a proof of a theorem or produce a statistical analysis of a data set. Educators can use these to demonstrate and develop students' ability to critique work. This approach could provide new forms of critique-based assessments.

Example 4. Providing clear examples of acceptable and beneficial AI use. Such examples will need to be tailored to specific modules or learning contexts. For example, it may be appropriate for students to disclose their use of AI in assignments if they rely heavily on AI-generated content or include it directly in their work. In such cases, they should cite the AI tool as they would any other source. Chat logs could form part of an assessment that demonstrates critical engagement with AI. The outcomes of challenges to the outputs can be regarded as evidence of honest and transparent usage of genAI.

In conclusion, we recommend that a proactive and collaborative approach is needed to ensure that educational practices in MSOR subjects evolve in step with the rapid advancement of genAI. While some advocate for a return to fully in-person examinations, we suggest a more balanced approach that leverages the opportunities of this technology to enhance learning and to better prepare students for an AI-enhanced future.

The emergence of genAI in the higher-educational landscape gives rise to new collaboration opportunities between students and staff to engage in a meaningful dialogue. Through this dialogue, sustainable and creative strategies for AI integration might be collaboratively developed. By embracing the potential of genAI, while remaining vigilant about its challenges, we can enhance the MSOR educational experience and prepare students for a future where AI plays a significant role in professional and academic environments.

6. Acknowledgement

This work was supported by the Warwick International Higher Education Academy (WIHEA).

7. Appendix

Example year 1 question and response

We present an example question, GPT-4o input prompt and output.

The following question is a typical Year 1 statistics question asked in a Term 1 probability module.

Question. A random experiment consists of rolling three fair six-sided dice (with face values from the set $\{1,2,3,4,5,6\}$. If two or more dice show the same highest number then the three dice are rolled again. For example, if the numbers on dice are $(5,5,2)$ or $(4,4,4)$ then the three dice are rolled again. However, if the numbers shown are $(5,2,2)$ or $(6,4,4)$ then the dice show a unique highest value and the dice are not rolled again. Each roll of the three dice is called a trial.

- You may assume that all trials are mutually independent.
- You may also assume the independence of the events $\{A_i, i \in N\}$ implies the independence of the events $\{A_i^c, i \in N\}$. (You were asked to show this for three events in the lecture notes.)
- In your answers, you must make clear where you are using independence.

1 Determine the probability that the first trial results in a unique highest number.

2 Let $n \in N$ be fixed. Determine the probability each of the first n trials results in no unique highest number.

3 Let the random variable X denote the number of trials until a unique highest number is obtained. That is, if the unique highest number is obtained on the nth trial then $X = n$. Determine the probability mass function of X.

4 Let $n \in N$. Determine the probability that, if the unique highest number was obtained on the nth trial, then in at least one trial before the nth trial all the dice showed the same value.

Example input prompt. The question was split into separate parts and the question stem repeated. This approach was necessary since on occasion the question context was forgotten between parts. We illustrate the input prompt used for part 4 which is presented verbatim. GPT-4o interprets the LaTeX code correctly.

A random experiment consists of rolling three fair six-sided dice with face values from the set {1, 2, 3, 4, 5, 6}.) If two or more dice show the same highest number then the three dice are rolled again. For example, if the numbers on dice are (5, 5, 2) or (4, 4, 4), then the three dice are rolled again. However, if the numbers shown are (5, 2, 2) or (6, 4, 4), then the dice show a unique highest value and the dice are not rolled again. Each roll of the three dice is called a trial.

You may assume that all trials are mutually independent. You may also assume the independence of the events \{A_i, i \in N\} implies the independence of the events \{A_i^c, i \in N\}. In your answers, you must make clear where you are using independence.

Question: Let n \in N. Determine the probability that, if the unique highest number was obtain on the nth trial, then in at least one trial before the nth trial all the dice showed the same value.

Output. The output is a combination of Markdown and LaTeX. This output has been converted to Word and is presented below.

To determine the probability mass function (PMF) of the random variable X, which denotes the number of trials until a unique highest number is obtained, we need to utilize the fact that each trial is independent and follows a geometric distribution.

First, let's recap the key probabilities calculated earlier:

- The probability of obtaining a unique highest number in a single trial is $p = \frac{55}{72}$.
- The probability of not obtaining a unique highest number in a single trial is $1 - p = \frac{17}{72}$.

1 Geometric Distribution

The random variable X follows a geometric distribution because we are dealing with independent trials and are interested in the number of trials required to get the first success (i.e., obtaining a unique highest number).

The PMF of a geometric random variable X with success probability p is given by:

$$P(X = n) = (1 - p)^{n-1}p$$

where n is the number of trials until the first success.

2 Applying to our Specific Case

Here, the success probability p is $\frac{55}{72}$. Thus, the PMF for X is:

$$P(X = n) = \left(\frac{17}{72}\right)^{n-1}\left(\frac{55}{72}\right)$$

for $n = 1,2,3, \dots$

3 Final Probability Mass Function

The probability mass function of the random variable X is:

$$P(X = n) = \left(\frac{17}{72}\right)^{n-1}\left(\frac{55}{72}\right), n = 1,2,3, \dots$$

Thus, the PMF of X can be written as:

$$P(X = n) = \left(\frac{17}{72}\right)^{n-1}\left(\frac{55}{72}\right), n = 1,2,3, \dots$$

This result uses the fact that each trial is independent, and the geometric distribution models the number of trials until the first success.

8. References

Attewell, S. Student perceptions of generative AI. Jisc report, 2024. Available at: https://edin.ac/3Y8Y2Cd. [Accessed 10 October 2024].

Chongchitnan, S. Parker, M. Mahal, M. and Petrie, S., 2024. Exploring the use of AI in mathematics and statistics assessments, Warwick International Higher Education Academy. Available at: https://warwick.ac.uk/fac/cross_fac/academy/funding/2023-24-int-projects/ai-in-maths/ [Accessed 10 October 2024].

Das, S.R., J.V. Madhusudan, J.V., Perceptions of higher education students towards ChatGPT usage. Int. J. Technol. Educ. 7(1) (2024) 86-106.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T., 2023. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ArXiv. Available at https://doi.org/10.48550/arXiv.2311.05232

OpenAI., 2022. Introducing ChatGPT. Available at https://openai.com/index/chatgpt/ [Accessed 25 September 2024]

Iannone, P. and Simpson, A., 2022. How we assess mathematics degrees: the summative assessment diet a decade on, *Teaching Mathematics and its Applications: An International Journal of the IMA*, Volume 41, Issue 1, March 2022, Pages 22–31, https://doi.org/10.1093/teamat/hrab007

Iannone, P. and Simpson, A. 2011. The summative assessment diet: how we assess in mathematics degrees. *Teach. Math. Its Appl.*, 30, 186–196.

Iannone, P. and Simpson, A. (eds) 2012. *Mapping University Mathematics Assessment Practices*. Norwich: University of East Anglia.

Meyer, J. et al. 2024. Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions, *Computers and Education: Artificial Intelligence,* Volume 6, (2024), 100199.

QAA, 2023. Subject Benchmark Statement: Mathematics, Statistics and Operational Research. Available at https://www.qaa.ac.uk/docs/qaa/sbs/sbs-mathematics-statistics-and-operational-research-23.pdf?sfvrsn=5c71a881_12 [Accessed October 2024]