CASE STUDY

Student use of large language model artificial intelligence on a history of mathematics module

Isobel Falconer, School of Mathematics & Statistics, University of St Andrews, St Andrews, UK.
Email: ijf3@st-andrews.ac.uk

Abstract

This case study assesses experience in autumn 2023 of permitting the use of Large Language Model Artificial Intelligence (AI) in preparing essays on a module in the history of mathematics. As a check on usage and to ensure academic standards, students were required to complete two paragraphs to accompany their essays explaining their use of AI. These generated qualitative and quantitative data on student familiarity with AI, and ability to use it in a thoughtful and ethical manner, which is reported here. Findings were that over 50% of students rejected AI use, and only 9% used it extensively. There was a weak negative correlation between AI use and essay grade, for which student confidence may have been a confounding factor. The most frequent reasons for rejecting AI were ethical, personal (satisfaction and confidence), and the time needed to correct it.

Keywords: artificial intelligence, generative AI, ChatGPT, student essays, history of mathematics.

1. Introduction

This case study assesses experience in autumn 2023 of permitting the use of LLM AI (Large Language Model Artificial Intelligence, also known as Generative AI) on a module in the history of mathematics. It may help to inform the 'escalating scholarly interest in AI's role within educational contexts' (Bukar et al, 2024). However, reviews of this burgeoning literature note that the majority of studies are theoretical, and that understanding of why students adopt LLM AI, and how they engage with it, is still very limited (Schei et al 2024; Abbas et al 2024).

'Topics in the History of Mathematics', is an optional module for mathematics undergraduates in their final or penultimate year at a Scottish university, worth 15 credits (1/8 of their work for the year). Students typically enter the university as high achievers: standard entry grades are Scottish Highers: AAAAB, including A in Mathematics, or GCE A levels: A*A*A, including A* in Mathematics. The student body is roughly split into one third Scottish, one third from the rest of the UK, and one third international students. The two main motivations for taking the module are a desire to broaden their knowledge of mathematics, and a desire to acquire soft and communication skills that are less commonly fostered in core mathematics modules.

The module is assessed through two class tests (totalling 50% of grade), a preliminary essay plan (5% of grade), and an end-of-semester essay on a history of mathematics topic of the student's choice (45% of grade). This case study covers the essay component.

Our university policy is that use of LLM AI counts as academic misconduct unless a module gives explicit permission for its use. In the 2023-24 module presentation we decided to give such explicit permission for LLM AI use, while taking precautions to ensure that academic standards were maintained. After consulting the Director of Teaching and the University's LLM AI Guidance, the module team decided to require that essays be accompanied by:

- A paragraph evaluating the ways in which the student had used/not used LLM AI and explaining their decisions. The intention was to be even-handed by asking that all students justify their decisions, whether or not they decided to use LLM AI;
- A paragraph identifying the three most significant sources cited in their essay and what these had contributed to their argument. This paragraph acted as a check that they did, indeed, understand both the argument they had presented, and were familiar with at least some of their sources.

The wording of the assessment rubric relating to LLM AI was iteratively discussed between the module team and the Director of Teaching, and followed closely the University's LLM AI Guidance. The marking criteria were changed from previous presentations to put more weight on quality of argument (which AI is poor at), up to 50% from 40%, and less weight on presentation (which AI is good at), down to 10% from 20%. Students were not directed at any particular LLM AI and were not required to specify what they had used. Appendix A contains the rubric provided to the students in the Module Handbook.

We took this approach for two reasons:

1. Pragmatically, we would be unlikely to detect LLM AI use (Perkins, 2023), so hoped to avoid unknowingly awarding marks for AI-generated content.
2. Pedagogically, use of LLM AI is likely to be part of students' future employment practices so we wished to encourage critical awareness. O'Dea et al (2024, p2) report that 'globally over 43% of employees have used ChatGPT and other large language models, such as Google Gemini and Copilot to help them with their daily work'.

The module team viewed this as a relatively minor change in design of an assessment primarily aimed at evaluating students' history of mathematics skills, rather than as pedagogical research. Hence no ethics clearance was sought; this precludes quoting directly from the wealth of information on student familiarity with LLM AI, and their ability to use it in a thoughtful and ethical manner, that resulted and that is discussed in the remainder of this report.

2. Quantitative data: generation, analysis and results

Sixty-one students submitted essays. Of these, all completed the AI paragraph, and 60 completed the sources paragraph.

Essay marking was split between two markers. One marker assigned grades to the essays before reading the two AI-related paragraphs, and then occasionally adjusted the grade in the light of the AI paragraphs if:

- the paragraphs revealed significant discrepancies between declared AI use and the evidence of the essay, for example if a student identified insignificant, rather than significant, cited sources, or appeared unaware of what the presented argument was, as judged by the second of the required additional paragraphs (none did);
- credit had been given for features of the essay that turned out to be straightforwardly generated by AI as described in the first of the additional paragraphs (two scripts);
- credit had been denied for features assumed to be an over-long quotation but that turned out to be the student's own work with AI in a partner role (see below for partner role, one script).

Students' self-reported use of LLM AI was roughly grouped and labelled into four categories as shown in Table 1. Appendix B gives paraphrases of statements characteristic of each category.

Table 7. Categories of AI usage, showing category label, category name, number and percentage of students who fell within each category

Category label	Category name	No. students in category (N=61)	% students in category
0	No use	33	54
1	Limited use	13	21
2	Moderate use	10	16
3	Extensive use	5	9

Over 50% of students declared that they had not used AI at all.

With this crude categorization, there was a weak negative correlation (-0.37) between use of AI, and overall grade for the essay, i.e. students who made more use of AI tended to get weaker grades.

Grades on the essay were comparable with previous years. However, for the overall module grade a small downward scaling at the bottom end was implemented to bring the distribution in line with previous years.

Although students had not been required to use any specific LLM AI, those that specified their platform all used some form of ChatGPT: 11 specified ChatGPT but did not give the version, two specified ChatGPT-4, one ChatGPT-3.5, and one ScholarAI (a ChatGPT plugin).

3. Qualitative data: generation, analysis and results

Qualitative data came from the submitted AI paragraphs. The initial stages of a grounded theory approach were used to develop themes (Strauss & Corbin 1990), i.e. the paragraphs were all read, and coded with no pre-conceptions of what would emerge, and codes then grouped into higher level themes. Emergent themes were:

- Self confidence;
- Efficiency;
- Self-identity;
- Partner;
- Critical awareness;
- Ethics.

3.1. Self confidence

Student self-confidence appeared most frequently as a factor in students' decisions about LLM AI use and related to two main areas: 1) Language and writing skills, and 2) Existing familiarity with LLM AI. In both areas students ranged from very confident to very lacking in confidence.

Students who expressed little confidence in their *language and writing skills* used LLM AI at many levels, from help with basic vocabulary and grammar, through structuring of paragraphs, to overall structure of the essay and argument (see example comments in §7.4). Many, but far from all, of those seeking help with vocabulary and grammar were students with English as an additional language. But students across the language spectrum used AI to help with structuring at paragraph

or overall essay level, pointing out that as mathematics students they had little experience of such tasks.

Conversely, a number of students expressed absolute confidence (sometimes misplaced!) in their ability to write a high-quality essay without AI assistance (see comment in §7.1).

Some students chose not to use AI, as previous *familiarity* led them to believe that it would be of little use in this instance. More frequently, though, students claimed no previous experience and lacked confidence in their ability to instruct it effectively or to evaluate the quality of the result; they chose not to use it for these reasons (see comment in §7.1).

3.2. Efficiency

Students were split on whether using AI would save time and effort and made decisions on this basis. The main time-saving activities mentioned were:

- discovery of sources;
- summarizing sources to build up knowledge and understanding;
- summarizing sources into a literature review.

(see example comments in §7.2, §7.3 and §7.4)

However, such students were outnumbered by those who thought that fact-checking AI-generated research would take more time than it was worth (comments in §7.1). For the majority, this belief was based purely on the number of dire warnings they had read. This was particularly the case for students who had chosen fairly niche topics and were already familiar with the few extant sources; they distrusted what AI might provide if it deemed these not sufficient. Less commonly, anxiety about fact-checking effort came from experience and up-to-date knowledge, such as of a recent rise in LLM hallucinations (a false or fabricated output).

3.3. Self-identity

This theme encompasses factors clustered around students' sense of their own individuality, personal development and satisfaction. Students expressing these views fell almost entirely into the 'no use' or 'limited use' categories (example comments in §7.1).

Many students felt that the arguments they wanted to make were individual to them. They noted that AI-generated writing tended to generalize and to read as generic, whereas the students wanted to make their own precise and detailed arguments, in their own individual style, aimed at a particular audience.

Even where this was not the case, many students felt that doing all the research and writing themselves would improve their own research skills more than using AI would, and that they would derive more personal satisfaction from doing this.

3.4. Partner

Some students used AI in much the way they might use a peer, mentor or supervisor, to bounce ideas off, and check their understanding and interpretation of their sources. Examples included iteratively:

- refining from initial area(s) of interest, or a brainstorm of ideas, into a well-defined essay topic;

- checking and refining the students' translation of material from other languages. This applied not only to students with sources in their own native languages, but also to native English-speakers using sources, for example in French or German. Indeed, there seemed a slightly greater willingness among English speakers to use other-language sources than in previous years (three compared to zero previously);
- helping with understanding and presenting of proofs in unfamiliar areas of mathematics, for example, by describing the proof in simpler/more modern English, by explaining the reasoning behind the steps, or by assessing the accuracy of a student's account of the proof against the original proof.

(see example comments in §7.4)

Note that the effectiveness of these uses, judged by quality of outcome, has not been evaluated.

3.5. Critical awareness

Students developed their own critical awareness of LLM AI in two ways: through external reading, and through trial. External reading (sometimes cited) was often used to justify claims that:

- AI would not handle well topics that were very specific with few sources, rendering it more prone to hallucinating;
- AI use is unethical in a variety of ways.

Some students took the opportunity offered by the essay's rubric to trial and experiment with AI, especially if they had little or no previous experience (comment in §7.2). The majority of trials compared the AI output with something they had written independently themselves; they generally reported that AI had missed or distorted the point of their argument, and required so much correction that it was easier just to write their own text. One or two students trialed other aspects of AI, such as comparing its search effectiveness with that of Google Scholar.

3.6. Ethics

Through their reading, many students became aware of ethical issues around using AI. That most frequently raised was around the originality and authenticity of AI-generated work, as it is based on vast quantities of untraced and unacknowledged data (e.g. Chesterman 2024). The associated danger of spreading misinformation was strongly raised by some students (e.g. Xu et al 2023) (comments in §7.1).

Other ethical issues raised (by one student each) were:

- Sustainability – the environmental impact of data centres and of mining/disposing of rare earths for components (see, e.g. Henderson et al 2020);
- Racism (and many other 'isms') as LLM AI is based on historical data and hence traditional stereotypes and patterns of expression (see e.g. Bender et al 2021).

4. Discussion & Conclusions

The module's AI rubric was fairly successful in prompting students to inform themselves about LLM AI and to think critically about its use. Indeed, one or two students interpreted the rubric as meaning that they had to use AI to at least some extent, and they trialed it accordingly.

Having informed themselves, a surprisingly high number rejected its use completely, so it is not clear that any aim of enhancing skills in effective use of AI were realized. However, any such aim was

secondary to the main purpose of the assessment to develop history of mathematics skills. The most frequent reasons for rejecting LLM AI were ethical (25%), personal (satisfaction and confidence) (40%), and the time needed to correct it (40%).

A minor positive development was the increased willingness observed among English-speaking students to tackle sources in other languages. The effect was small (three students) but noticeable compared with the complete lack of such students in previous cohorts.

It seems likely that a confounding factor underlying the weak negative correlation between AI usage and grade, was student confidence and ability in written English; students whose written English was weak, as judged by the reasons they gave for using AI and their performance in class tests, were much more likely to be moderate or extensive users of AI.

It is possible that, overall, the students did better than previous cohorts who did not have access to LLM AI. The standard of the best essays seemed very comparable to the best essays of previous cohorts, but these were written by students who had rejected AI. At the lower end of the scale, the need for downward scaling of the overall module grade in order to bring grades into line with those of previous cohorts might indicate that weaker students had benefitted from AI use. However, since scaling analysis is done at module level rather than that of separate assessment components, further analysis would be required to disentangle the essay from the class test results of this and previous cohorts, before comparisons could be made. A qualitative comparison of the corpus of essays from this cohort with those of previous cohorts, could provide insights into how AI affects student writing quality and originality, but would be difficult to report on robustly given the lack of ethics clearance for any of these assessments.

Overall, this intervention proved easy to implement, taking little additional resource in class time or marking; the major time taken was in module team discussions beforehand when developing the rubric. Judging by the good correspondence between the AI statements, the sources paragraphs, and the essays themselves, it appeared that the students were honest in reporting their AI use, suggesting that the intervention was effective in its major aim of making any LLM AI use transparent to the markers; whether its success could be repeated with students who were generally less engaged and motivated, is less clear.

More forethought for the possible value of the additional paragraphs beyond the primary assessment task, might have prompted an application for ethics clearance and enabled more robust reporting of the outcomes to the wider HE mathematics community; seeking ethics clearance would seem advisable if there is even a remote likelihood that outcomes may form the basis for research, however the impact on what they write of asking students for the necessary informed consent has also to be considered.

Although the approach taken was, on balance, a success, we cannot assume that it can be repeated on the next presentation in 2025-26, as AI and student skills with it, will have moved on considerably by then.

5. Acknowledgements

I would like to thank Dr Deborah Kent, who co-lectured this module, and the students on MT4501 who were a joy to work with.

6. Appendix A: Project requirements and marking criteria as stated in the module handbook

Project requirements:

- Free choice of topic – provided it is about history of mathematics;
- An essay, normally of 2500-3000 words (depending a bit on how many equations or tables you use), containing:
  - First page with Title, your student ID, an abstract of 3-5 sentences describing the content of the essay;
  - Introduction, including your research questions and thesis statement;
  - Body of the Essay (may be divided into sections but does not have to be);
  - Conclusion;
  - Citations and references in a standard format (see below);
- PLUS;
  - a compulsory paragraph of up to 200 words evaluating the ways in which you have used/not used AI. If you have used AI say how, and what it contributed to your essay; if you have not, explain your decision not to;
  - a compulsory paragraph of around 200 words that identifies the three most important sources you have used and analyses the ways in which those were important to your argument.

Use of LLM/AI (e.g. ChatGPT)

On MT4501 we recognize the benefits of learning to use LLM/AI effectively and intelligently.

You may use LLM/AI for your project, but we want to know how and why you have used it. If you have not used it, tell us why you decided not to. Either answer is equally acceptable. You **must** submit a paragraph accounting for this along with your project essay. At a minimum, we expect you to have verified all the references and "facts" contained in your essay, and to have chosen an essay structure that provides the most effective support for your argument.

LLMs may be useful for:

- writing your essay for you (!);
- revising your drafts to improve the quality of your English, especially if you are a non-native English speaker;
- structuring your essay (in a common way).

But we expect you to demonstrate awareness of limitations of LLMs such as:

- LLMs may generate misinformation, as they prioritize coherence and plausibility over factual accuracy;
- LLMs generate text that is coherent, contextually relevant, and plausible, but they do not "think" and cannot generate an argument;
- LLMs may generate an essay structure that is common and presents the content in a coherent manner, but this may not be the best structure to support your argument.

Note also the University's guidance and policy on Good Academic Practice has sections on unauthorised use of AI and how to avoid it.

We will discuss use of LLM/AI in a tutorial.

Project marking criteria

Essays will be marked according to the following criteria:

- Quality of argument/analysis (weighted approx. 50%), including:
  - Originality/independence of approach;
  - Difficulty/ambition of project;
  - Critical writing, analysis and interpretation;
  - Understanding of concepts;
  - Were appropriate assumptions made & appropriate conclusions/inferences drawn?;
  - Were appropriate tools/methods used?;
  - Was the argument well-supported by the evidence?;
- Quality of content (weighted approx. 40%), including:
  - Choice of appropriate sources and examples;
  - Amount of work undertaken;
  - Appropriate use of diagrams, tables, images;
  - Factual accuracy;
  - Understanding of detail;
- Presentation and exposition (weighted approx. 10%), including:
  - Statement of aims and objectives;
  - Structure and organization of material;
  - Clarity and readability;
  - Literacy and grammar;
  - Citation and referencing.

The two compulsory additional paragraphs will be used to assess your essays against the standard School grade descriptors.

The paragraph on your use/non-use of LLM/AI will be assessed according to:

- Depth of reflection on effective ways to use LLMs, and their limitations;
- Correspondence between your use/non-use of LLMs and the evidence of your essay.

The paragraph on your most important three sources will be assessed according to:

- Appropriateness of your selection of sources to discuss;
- Quality of your argument about what they have contributed to your essay.

7. Appendix B: Characteristics of AI use categories

Paraphrases (not direct quotations) of AI statements characteristic of each category.

7.1. No use

'I chose not to make use of AI.' Such statements were often followed by reasons such as:

> 'I have not used AI before and did not want to spend time learning to use it effectively, rather than researching for my essay.'
>
> 'There are very few sources on my specific topic and I was worried that AI would generate false sources to fill the gaps.'
>
> 'Fact-checking every AI output would take more time than writing the essay myself.'
>
> 'I felt confident in my ability to write clearly and concisely, and doing so would give me more personal satisfaction.'
>
> 'I could not justify using AI for an academic essay, due to its environmental costs.'
>
> 'AI relies on the creative work of individuals who are not acknowledged or paid; I could not use it in good conscience.'

7.2. Limited use

'I did not use LLM/AI for very much.' Such statements were usually followed by an account of trials they had performed with AI on their own initiative to assess its usefulness, finding it not useful, for example:

> 'I tested ChatGPT by asking it to summarise this source, but the summary was over-simplified and omitted key points, and the writing felt impersonal.'

'I used AI to provide an initial structure to help me get started, but abandoned the structure as my research progressed.'

'I used AI to suggest potential avenues of research, but then checked them out and decided whether to pursue them, researching them on my own. I did not use AI for any of the writing.'

'Once I had written my essay, I used AI minimally to suggest improvements to spelling, grammar and clarity.'

7.3. Moderate use

'I have used AI to help structure my essay, and to improve the quality of my English by fixing the grammar and suggesting more varied word choices.'

'I chose to use AI to help express my points more concisely and reduce my word count. I also used it to help generate a title and abstract.

7.4 . Extensive Use

'I used AI to organise my thoughts and refine my essay plan, and then to help break the plan down to actionable subsections. I used it extensively to check my writing style. Finally, I used it to find extra sources from the internet.'

‘I have used AI throughout my essay. During the research I would ask it questions, and its responses would provide me with the key features of a subject – which I could then check whether I wanted to include. I did not use it for help with writing text, although I did use it to highlight grammatical errors. I also used it to search out answers to technical questions about use of LaTeX.’

‘I used AI extensively, especially to help me understand the unfamiliar style of proofs. I asked ChatGPT to explain the proof, then wrote the proof in my own words and asked ChatGPT to check my proof against the original to make sure I had not omitted key steps or information.’

‘Using AI tools improves quality and efficiency in the essay writing process. It enhances the precision of language, and streamlines discovery and review of the literature.’

8. References

Abbas, M. (2024). Is It Harmful or Helpful? Examining the Causes and Consequences of Generative AI Usage among University Students. *International Journal of Educational Technology in Higher Education* 21(1). https://doi.org/10.1186/s41239-024-00444-7.

Bender, E.M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York: Association for Computing Machinery. pp.610-23. https://doi.org/10.1145/3442188.3445922.

Bukar, U.A., Sayeed, M.S., Razak, S.F.A., Yogarayan, S., and Sneesl, R. (2024). Decision-Making Framework for the Utilization of Generative Artificial Intelligence in Education: A Case Study of ChatGPT.” *IEEE Access* 12 pp.95368–89. https://doi.org/10.1109/ACCESS.2024.3425172.

Chesterman, S. (2024). Good Models Borrow, Great Models Steal: Intellectual Property Rights and Generative AI. *Policy and Society*, 00(00), pp.1–15. https://doi.org/10.1093/polsoc/puae006.

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D. and Pineau, J. (2020). Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. *Journal of Machine Learning Research* 21(248) pp.1–43. http://jmlr.org/papers/v21/20-312.html [accessed 23 December 2024].

O’Dea, X., Tsz Kit Ng, D., O’Dea, M., & Shkuratskyy, V. (2024). Factors affecting university students’ generative AI literacy: Evidence and evaluation in the UK and Hong Kong contexts. *Policy Futures in Education*, *0*(0). https://doi.org/10.1177/14782103241287401.

Perkins, M. (2023). Academic Integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching & Learning Practice*, 20(2). https://doi.org/10.53761/1.20.02.07.

Schei, O.M., Møgelvang, A., and Ludvigsen, K. (2024). Perceptions and Use of AI Chatbots among Students in Higher Education: A Scoping Review of Empirical Studies. *Education Sciences* 14(8): 922. https://doi.org/10.3390/educsci14080922.

Strauss, A. and Corbin, J.M. (1990). *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. Thousand Oaks, CA, US: Sage.

Xu, D., Fan, S. and Kankanhalli, M. (2023). Combating Misinformation in the Era of Generative AI Models. *Proceedings of the 31st ACM International Conference on Multimedia*. New York: Association for Computing Machinery. pp.9291–98. https://doi.org/10.1145/3581783.3612704.