Using AI for question generation in mathematics education: what are the advantages and disadvantages?

Lucy Rycroft Smith & Darren Macey

University of Cambridge

Writing mathematical questions, tasks or activities - along with criteria for 'correctly' answering them - is specialised and important work which has been previously done by teachers and designers with expertise and experience in their use. In this study, we asked three different large language models (Copilot, Claude, and ChatGPT3.5) to design questions to assess students' understanding in the topic of the area of a rectangle, using ten prompts with nuanced differences. We found variety in the length, mathematical correctness, and ways to surface student understanding across the three models, suggesting that not all large language models give the same kinds of response when question generating, and teachers and designers may wish to experiment with different LLMs when considering which might be best for their needs. We suggest that future research can build on this study by exploring more dialogic co-design between LLMs and human agents.

Keywords: artificial intelligence; large language model; task design; neural networks; area; teacher professional learning; design

Introduction

The design of effective tasks, questions and problems in mathematics education is an important field, with various theoretical frameworks associated with it. One way in which to consider these is the differentiation between design that uses theories of mathematics learning in general, and that which focuses on the development of students' learning in specific mathematical areas (Kieran et al., 2015). We question this separation, wondering whether all frameworks that theorise regarding learning in a specific topic also must, however obliquely, embed theories of learning about mathematics in general, and also assumptions about the purpose and practice of mathematics itself.

These questions have become particularly urgent since the rapid take-up of artificial intelligence large language models (LLMs) for everyday use. It has been suggested that these models could support teachers in a variety of ways, including the design of tasks. In this paper, we consider the advantages and disadvantages of using LLMs to design mathematical tasks, using examples from three different natural language interfaces (Copilot, Claude, and ChatGPT) on the topic of the area of a rectangle to exemplify some of these ideas. We ask, in particular, what next for research on mathematics education question design which aims to use artificial and human intelligence collaboration (AHIC)?

What makes an effective task in mathematics education?

Several schools of thought exist regarding the criteria of an effective mathematics task. For example, Prusak et al.'s (2013) five principles of task design suggest tasks should, in order to foster a culture of problem solving and conceptual learning: encourage the production of multiple solutions; create collaborative situations; support engagement in socio-cognitive conflicts; provide tools for checking hypotheses; and invite students to reflect on solutions. One area of research that is still in its infancy is whether and how to use such principles when working with artificial intelligence agents, either to prompt better quality tasks, evaluate tasks that have been written, or indeed both. We note also the use of 'task' when referring to human design, and more often 'question' when connected to the 'generation' activity performed by artificial intelligence agents. Regardless, it seems clear to us that defining meta-goals for task design/question generation could also be fruitful when interacting with artificial intelligence interfaces, beyond just giving content-specific prompts.

Using large language models in mathematics education

There are many suggested ways of using large language models in mathematics education, including:

- **Provocation:** to generate questions and prompts that encourage students to think critically, and to analyse and interpret the information presented to them
- **Inclusivity:** in combination with speech-to-text or text-to-speech solutions to help people with visual impairment; can be used to develop inclusive learning strategies by adding support in tasks such as adaptive writing, translating, and highlighting of important content in various formats
- **Planning:** to assist teachers in the creation of (inclusive) lesson plans and activities
- **Material generation:** to generate (large amounts of) ideas, tasks, solutions, questions and answers, often in response to particular needs or adapting to particular contexts
- **Dialoguing:** to chat with students as a pedagogical agent; to stimulate the curiosity of children and enhance question-asking skills to chat with teachers to provide professional learning
- **Assessment**: to generate prompts for formative assessment activities that provide ongoing feedback to inform teaching and learning
- **Personalisation**: to create individually-tailored, adaptive learning experiences for students
- **Elucidation:** to provide explanations, step-by-step solutions, and detailed justifications, which can help students (and teachers) develop analytical and creative thinking

(Abdelghani et al., 2023; Baidoo-Anu & Ansah, 2023.; Dijkstra et al., 2022; Kasneci et al., 2023).

Given the ongoing development of these technologies, it is likely that the effectiveness of any given LLM will vary across uses and continue to evolve and change over time. With this in mind, users must be aware that updated versions of LLMs may change the assumptions that can be made about which use types can be

most and least effectively deployed for any given model. Similarly, there are many known pitfalls when using AI in mathematics education:

- **Lack of interpretability**: it can be difficult to understand the reasoning behind the model's predictions
- **Ethical considerations:** models are biased by the content they learn from, and are likely to reproduce inequity; models impact employment, may be misused or deployed unethically, and are currently built on exploitative/extractive structures
- **Lack of contextual understanding**: models do not have awareness of students, schools, or environments
- **Access:** not everyone has fair access to large language models (for example languages, internet access, cost)
- **Overextrapolation**: generative models are based on statistical patterns in the data they are trained on, and they do not have a true understanding of the concepts they are 'explaining'
- **Lack of creativity:** models can only use data that already exists; they cannot invent 'new' data
- **Embedded bias**: models use data that is biased to create new data, replicating and perpetuating biases
- **Need for human oversight**: models can be incorrect, harmful, or inappropriate
- **Data privacy/security:** data that is inputted to the model effectively becomes public
- **Sustainability**: models have a high energy consumption
- **Fair use**: copyright issues may emerge
- **Accuracy**:models may confidently 'hallucinate' incorrect information
- **Data cannibalism**: the more data that models produce, the more they are likely to encounter/ingest their own data, leading to closed loops

(Abdelghani et al., 2023; Baidoo-Anu & Ansah, 2023.; Dijkstra et al., 2022; Kasneci et al., 2023).

While we do not focus on the ethical implications of LLM usage in this study, it is important to foreground that choosing to deploy an LLM to support educational purposes involves an (often invisible) ethical statement by the user. Issues with the development of LLMs including copyright infringement, exploitative business practices, and negative environmental impact have been identified (Mökander et al., 2023).The LLM user therefore must make an informed choice as to how the balance any potential benefits to using LLMs against this backdrop. Below, we address a few further disadvantages as seen in the responses we received from our prompts.

Methodology

We gave three different large language models (Copilot, Claude, and ChatGPT3.5) a series of prompts to write questions focused on the topic of the area of a rectangle, reasoning that this topic was straightforward, covered both primary and secondary mathematics, and was narrow enough to reveal variation where it occurred. We varied the prompts systematically. The baseline prompt was: W*rite a question with a markscheme to test student understanding of the area of a rectangle*. The variations were:

> Variation A: Write a question with a markscheme to test student understanding of the concept of the area of a rectangle.
>
> Variation B1: Imagine you are an excellent secondary mathematics teacher. Write a question with a markscheme to test student understanding of the area of a rectangle.
>
> Variation B2: Imagine you are an excellent primary mathematics teacher. Write a question with a markscheme to test student understanding of the area of a rectangle.
>
> Variation C: Write a question with a markscheme and a diagram to test student understanding of the area of a rectangle.

We tried all the combinations for each of the three large language models (i.e. A, B1, B2, C, AB1, AB2, AC, B1C, B2C).

Results

All of the questions generated by the LLMs fell within our expectations, and were also approximately in line with what we might expect from human agents of varying expertise and experience in mathematics education design. From the perspective of practical use, perhaps the most important element of the models' output to consider is whether the questions and markschemes were mathematically correct, which we defined to mean that all parts of the solution provided were correct (including correct units). Two of the responses were borderline – they gave correct solutions but omitted parts of the question, which we encoded as 'correct'. Of the three models, Claude generated the highest proportion of mathematically correct questions (Fig. 1) with just a single incorrect solution. It is noteworthy that in relation to the prompts that requested a diagram (Variation C), only the questions generated by Copilot included diagrams, and these were responsible for all four of the mathematical errors in the model's output (the model gave 'diagrams' which were generic images of cuboids). Further, two of these were the result of prompts that did not request a diagram at all (A and B1). Given that we did not penalise the models which failed to produce a diagram at all, a case could be made that the Copilot model was the most 'mathematically reliable' model if all diagrams are removed from consideration.

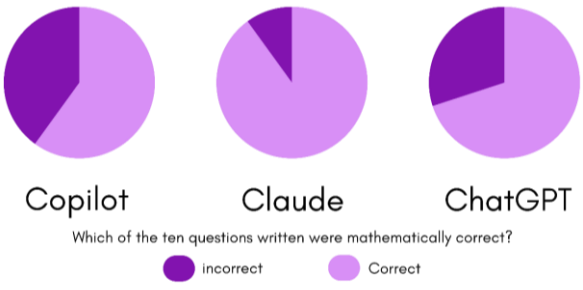
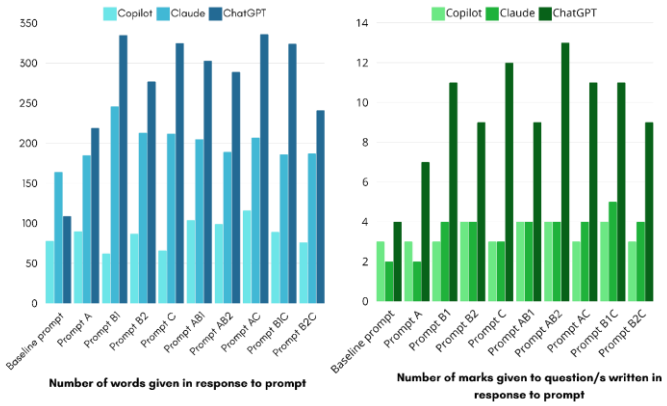


Figure 1: proportion of the ten questions written by each LLM in response to the ten prompts that were categorised as mathematically correct

Further analysis of the outputs addressed key aspects of the structure of the questions that affect their utility in the classroom. Firstly, there was some consistency in the number of words within each of the models (Fig. 2), although this could be a feature of the similarity of the prompt variations. There did not seem to be any clear difference in word count between questions aimed at primary students and secondary students, which is perhaps surprising. The three models each produced sets of

questions with broadly similar word counts, although across the models there was a marked difference, with Copilot consistently producing fewer words, and ChatGPT producing the most. There was a clear correlation between number of words produced and number of marks allocated. It is also noteworthy that Claude, in addition to giving questions and solutions, gave potentially useful meta-commentary such as:

> "The question tests if students conceptually understand: What area represents as the space inside a 2D shape; Why the length x width calculation gives the total area inside a rectangle. The markscheme awards marks for the visual representation and logical explanation of the area concept, without requiring the precise formula wording. This allows assessment of conceptual understanding." (Claude's response to prompt variation A)



Figures 2 and 3: The number of words, and the number of marks, given by each LLM in response to each prompt variation

ChatGPT consistently generated output with more marks available and crucially, more multi-part questions. Claude and Copilot by contrast consistently allocated similar numbers of marks. It is worthy of note that Copilot frequently assigned a final mark which we considered superfluous, and if this mark was ignored, the same order from least to most seen in the word count would also be visible in the number of marks allocated.

There may be some interplay between the number of marks and whether the output was mathematically correct, as analysis of the individual questions produced by ChatGPT revealed that in all cases, only one part of the multi-part mark schemes were incorrect, suggesting the apparent inaccuracy could plausibly be a function of the increased opportunities for the model to make mistakes. Prompts designed to define the number of marks allocated may improve the apparent inaccuracy of ChatGPT and this would be interesting to explore further in future research.

Conclusion

Collins et al (2023) suggest important implications for teachers working with AI: it is better to use models which communicate uncertainty, respond well to user corrections, and are interpretable and concise; equally, humans should be aware of language models' propensity to give mathematically incorrect solutions, and use them carefully on that basis. They also suggest that a static model of evaluation – that is, evaluating LLMs with one set of inputs and outputs, rather than using them dialogically – "fails to account for the essential interactive element in LLM

deployment, and therefore limits how we understand language model capabilities" (p.1), which is an important limitation of this study. In line with these ideas, we suggest that future research is focused on:

- Analysing responses received when purposes/aims are specified to the LLM that focus on more than just the content
- Detailing the possibilities conferred when dialogic models are used to elicit tasks from LLMs by human agents with mathematics education expertise
- Exploring ways to better understand the types of error made by LLMs, and supporting development in creating feedback loops for LLMs to error check and/or give uncertainty information.

In summary, we see this small study as a catalyst to consider in more detail ways in which dialogic AHIC (artificial and human intelligence collaboration) can be used to co-design mathematics education questions which are correct, interesting, and most of all fulfil pre-agreed purposes beyond just those regarding content – i.e. making explicit assumptions and drivers regarding the philosophy of mathematics underpinning the design.

References

Abdelghani, R., Wang, Y.-H., Yuan, X., Wang, T., Lucas, P., Sauzéon, H., & Oudeyer, P.-Y. (2023). GPT-3-driven pedagogical agents for training children's curious question-asking skills. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-023-00340-7

Baidoo-Anu, D., & Ansah, L. O. (n.d.). *Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning*.

Collins, K. M., Jiang, A. Q., Frieder, S., Wong, L., Zilka, M., Bhatt, U., Lukasiewicz, T., Wu, Y., Tenenbaum, J. B., Hart, W., Gowers, T., Li, W., Weller, A., & Jamnik, M. (2023). *Evaluating Language Models for Mathematics through Interactions*. https://doi.org/10.48550/ARXIV.2306.01694

Dijkstra, R., Genç, Z., Kayal, S., & Kamps, J. (2022). *Reading Comprehension Quiz Generation using Generative Pre-trained Transformers*. https://dare.uva.nl/search?identifier=a1109043-92d4-4c63-be33-6e238780d3b7

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. https://doi.org/10.1016/j.lindif.2023.102274

Kieran, C., Doorman, M., & Ohtani, M. (2015). Frameworks and principles for task design. In A. Watson & M. Ohtani (Eds.), *Task Design in Mathematics Education: An ICMI study 22* (pp. 19–41). Springer.

Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2023). Auditing large language models: A three-layered approach. AI Ethics.

Prusak, N., Hershkowitz, R., & Schwarz, B. B. (2013). Conceptual learning in a principled design problem solving environment. *Research in Mathematics Education*, *15*(3), 266–285. https://doi.org/10.1080/14794802.2013.836379