CASE STUDY

Generating maths solutions with ChatGPT

Declan Manning, Department of Mathematics, Munster Technological University, Cork, Ireland.
Email: declan.manning@mtu.ie

Abstract

Creating step-by-step maths solutions takes significant time and effort. Starting with a ChatGPT-generated draft and proceeding to carefully review and improve it can lead to significant time savings. In this case study, solution documents were created for two past exam papers in a second-year undergraduate maths module. Using a ChatGPT-generated draft as a starting point led to a total creation time of 2 hours and 36 minutes, compared to 4 hours and 31 minutes without the assistance of ChatGPT. This article explains the procedure for obtaining the ChatGPT draft, provides the background for the study, and presents the findings. It highlights key strengths of using ChatGPT for this purpose, including its speed, accuracy and quality of explanation. Limitations are also discussed, such as the risk of calculation errors, incorrect workings or over complicated answers.

Keywords: Generative AI, ChatGPT, Maths solutions, Time efficiency

1. Introduction

The growing popularity of generative AI models like ChatGPT has forced educators to reconsider various aspects of maths teaching. These models offer potential educational benefits such as personalised and adaptive instructional support (Ahmad, Murugesan and Kshetri, 2023). ChatGPT also has the potential to help students overcome learning barriers and strengthen their ability to transfer knowledge into new contexts (Mollick and Mollick, 2022).

However, caution is necessary as generative AI becomes increasingly prominent in classrooms. Bastani, et al. (2024) investigated if students given access to a ChatGPT-4 based tutor during study sessions performed better than those who weren't. Access to the tutor increased study performance by 48%, but decreased performance by 17% in a subsequent exam when tutor access was unavailable. The authors suggest that students may become overly reliant on the tutor during practice, preventing them from effectively learning key problem-solving skills.

The problem-solving abilities of generative AI models are advancing rapidly, with significant improvements observed between ChatGPT-3.5 (released in November 2022) and ChatGPT-4 (released in March 2023). Newton and Xiromeriti (2023) conducted a scoping review on ChatGPT's performance in multiple-choice questions across various subject areas. Out of 18,862 tested questions, ChatGPT-3.5 answered 49.5% correctly, significantly lower than ChatGPT-4's 75.5% accuracy rate. To assess advancements in mathematical reasoning ability, Frieder, et al. (2024) tested ChatGPT models on a novel dataset featuring exercises from graduate-level textbooks on probability theory, topology and functional analysis, as well as holes-in-proofs exercises and symbolic integration tasks. Despite performing below the level of an average graduate student, ChatGPT-4 significantly outperformed older versions of ChatGPT. Newer models are showing continuous improvements in advanced reasoning, with ChatGPT-o1 (released in December 2024) scoring 83% on the American Invitational Mathematics Exam, a qualifying exam used in the selection process for the US Maths Olympiad team (OpenAI, 2024).

ChatGPT's responses are mostly correct, with newer models making fewer and fewer mistakes. However, it has yet to achieve perfection - and likely never will. Alkaissi and McFarlane (2023) found that ChatGPT fabricated references, remarking, "*While ChatGPT can write credible scientific essays, the data it generates is a mix of true and completely fabricated ones.*" Giray (2024) calls on academics to carefully verify AI-generated content and develop a deep understanding of the limitations and risks of AI tools.

Other large language models have also made significant progress in problem-solving over the past few years. Claude 3 Opus (released in March 2024) outperformed its main competitors across a range of mathematical domains, including grade school math, undergraduate knowledge, and graduate-level reasoning (Anthropic, 2024). AlphaGeometry2, a specialised model developed by Google DeepMind, recently outperformed gold medal standards in Math Olympiad geometry (Chervonyi, 2025).

This case study, conducted in summer 2024, explores the efforts of a lecturer in an engineering maths module to develop solution documents for two past exam papers. For the first exam paper, ChatGPT-4o (released in May 2024) generated draft solutions, which the lecturer then verified, corrected, and refined. For the second exam paper, solutions were created without the assistance of generative AI. ChatGPT was selected based on the lecturer's personal preference, though other large language models like Claude or Gemini would have been equally suitable for this task. This article provides background details on the maths module in question, details the input and output methods used to generate the draft, presents the case study findings, and analyses ChatGPT's strengths and limitations for this application.

2. Background

Exam solution documents were created for a second-year undergraduate engineering maths module at Munster Technological University. The exam paper was a two-hour closed book written assessment with four questions covering vectors, matrices, differentiation and integration. The module introduced fundamental concepts in each of these areas, including:

- Addition, subtraction and scalar multiplication of vectors;
- The dot product and cross product for vectors;
- Addition, subtraction and multiplication of matrices;
- Finding the determinant of a matrix;
- The inverse matrix method for solving matrix equations;
- Parametric differentiation, implicit differentiation and partial derivatives;
- Integration by substitution, integration by parts and integration with partial fractions.

3. Generating draft solutions with ChatGPT

This section outlines the input method, output method and prompt design used to obtain draft solutions from ChatGPT.

3.1 Input Method

In subjects such as English and business, text-based prompts are well suited since questions can easily be typed using a standard keyboard. However, entering maths questions in plain text is challenging due to mathematical notation such as fractions and integral signs. One option is to type prompts using LaTeX syntax, which ChatGPT can accurately interpret. However, this approach is time-consuming and prone to errors. If a handwritten or digital copy of the question is available, a more efficient approach is to upload an image or screenshot. ChatGPT-4o supports image uploads

and is effective at interpreting the mathematical content within them. If a PDF file containing multiple questions is available, it can be uploaded directly. This method was used to create the draft solutions in this case study.

3.2 Output Method

ChatGPT's default behaviour is to provide answers in a chat-based format, with mathematical expressions embedded directly in the conversation where necessary. As the chat history grows, it can become difficult to navigate. While useful for quickly reviewing responses, this format is not ideal for sharing or editing the output. Taking screenshots is useful for capturing a key section of the chat, but this method lacks the flexibility needed for editing responses. In this case study, ChatGPT was prompted to output its responses in the form of LaTeX code. The response was copied into a LaTeX editor, reviewed for correctness, and modified as needed. The final version was exported as a PDF to facilitate efficient sharing and distribution.

3.3 Prompt Design

In addition to uploading the PDF of the exam paper, a written prompt was given to ChatGPT to guide its response. It took three iterations of prompt refinement to obtain a satisfactory result; however, it is worth noting that the prompt could still be improved further. The first prompt attempted was: "Provide written solutions to the given exam paper. Present your answer in a latex document". This resulted in solutions that lacked detail and were too brief for the lecturer's liking. In a fresh chat, the exam paper PDF was reuploaded with the revised prompt: "Provide written solutions to the given exam paper, *breaking things down into small steps*. Present your answer in a latex document". The output initially looked promising, but ChatGPT stopped unexpectedly after generating solutions for the first half of the paper. The third iteration used the prompt: "Provide written solutions to the given exam paper, breaking things down into small steps. *Solve every single question part in detail.* Present your answer in a latex document". This time, ChatGPT produced full, detailed solutions to the entire paper.

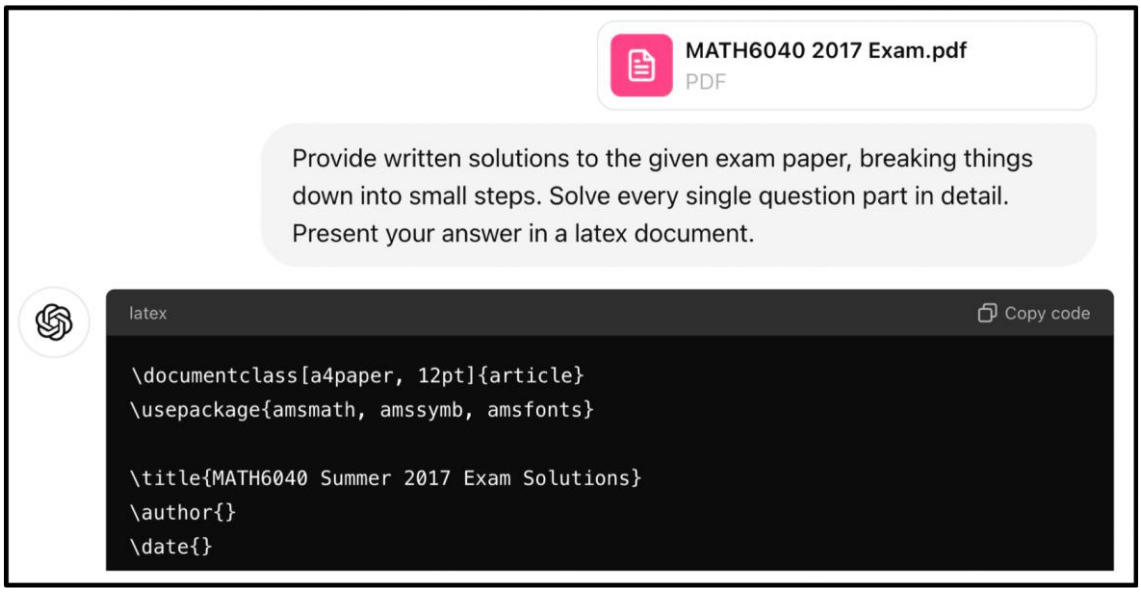

Figure 1. Uploading a PDF file of a maths exam and asking ChatGPT to return LaTeX code containing solutions.

4. Findings

Using the input and output methods outlined in section 3, ChatGPT was given the 2017 summer exam paper for the module. It generated a complete LaTeX solution document in 3 minutes and 36 seconds. For two question parts that required the drawing of a graph, the LaTeX document included a textual description rather than an image. A follow up prompt was provided explicitly requesting image files, which ChatGPT successfully generated. After careful review of the solutions, ChatGPT was found to have made errors in 6 of the 28 question parts: two miscalculations, one misread question, two solutions with incomplete final steps, and one instance where the correct answer was obtained but the workings were incorrect. When graded by the module's lecturer, ChatGPT's solutions would have achieved a score of 86%.

The code was copied into a LaTeX editor and refined by the module's lecturer with two main objectives: ensuring mathematical correctness and aligning the solution style with examples presented in class. To resolve the misread question error, an image file of the question was reuploaded, which ChatGPT correctly interpreted on the second attempt. If ChatGPT's original solution method deviated significantly from classroom examples, the question was reuploaded with a prompt which specified the preferred approach. The process of reviewing, reprompting as needed, editing, and formatting took a total of 2 hours and 32 minutes.

For comparison, the lecturer also created step-by-step solutions to the 2018 summer exam paper without assistance from ChatGPT. To minimise the time spent on the task, the lecturer decided to electronically handwrite the solutions on an iPad, the same method that had been used for creating the modules lecture notes. Since many exam questions closely resembled examples from the lecture notes, existing content could be copied and edited to produce the exam solutions. Creating the handwritten solution document to the 2018 exam paper took a total of 4 hours and 31 minutes. Due to the time savings from copy and pasting existing handwritten content, manually producing the solutions in LaTeX would likely have taken significantly longer.

It should be noted that the experiment design has several limiting factors: the comparison involved two different source exam papers, the output formats were not consistent, and the exams' questions focused only on specific engineering maths topics. This experiment serves as a proof of concept that utilising generative AI can lead to significant time savings, but further investigation is required to determine how broadly this result applies.

5. Strengths of ChatGPT Generated Solutions

This section explores the advantages of using ChatGPT for generating maths solutions. The reflections here, as well as in the next section, come from various interactions with ChatGPT extending beyond the case study in section 4.

5.1 Speed

Models such as ChatGPT-4o generate output remarkably quickly. In this case study, ChatGPT-4o produced full solutions to an exam paper in 3 minutes and 36 seconds, requiring just 3% of the two-hour time limit available to students when sitting the exam.

5.2 Accuracy

Each mistake in the generated solutions requires time and effort to correct, so fewer errors result in a lower overall time commitment. In this case study, ChatGPT-4o demonstrated strong accuracy, producing mathematically correct solutions for 78.6% of the questions on the exam paper. As more advanced models are developed, this accuracy is likely to improve.

5.3 Quality of Explanation

For years, tools like WolframAlpha have been able to solve mathematical problems accurately. However, the step-by-step solutions provided are typically written in dense mathematical notation that can be challenging for novice learners to understand. A major strength of ChatGPT is its ability to explain solutions in simple language while carefully sequencing solution steps in a clear and structured manner. In Figure 2, ChatGPT's is tasked with solving a linear equation. The solution is well structured and effectively explains each step of the process in plain English.

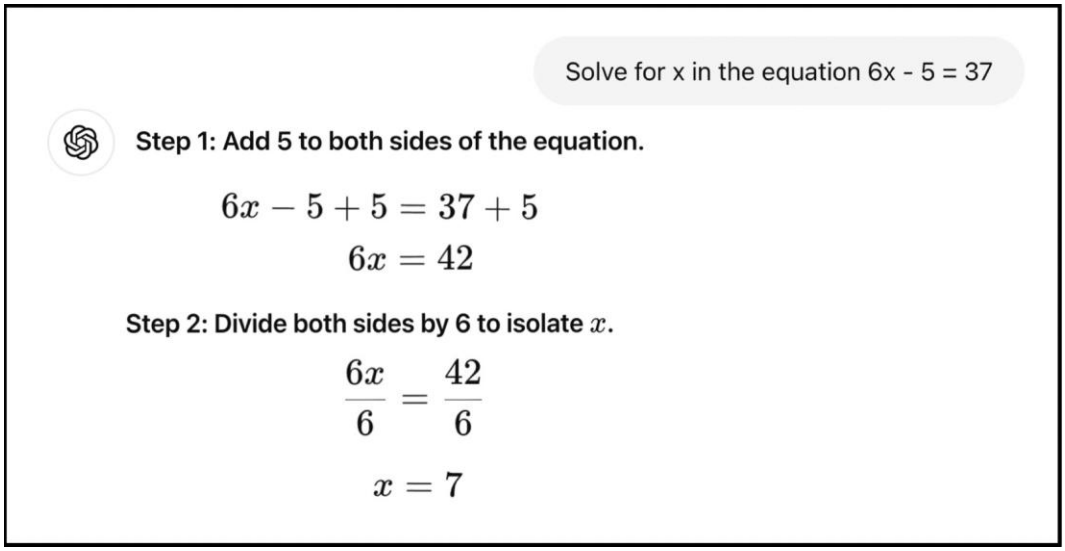

Figure 1. Example of ChatGPT-4o correctly solving a linear equation.

In Figure 3, ChatGPT is asked to differentiate an expression requiring the product rule. Instead of immediately performing the calculation, it first explains the method it is going to use. This puts emphasis on the key mathematical insight needed to solve the problem. ChatGPT proceeds to split the product rule application into three clear steps, then performs and explains each one.

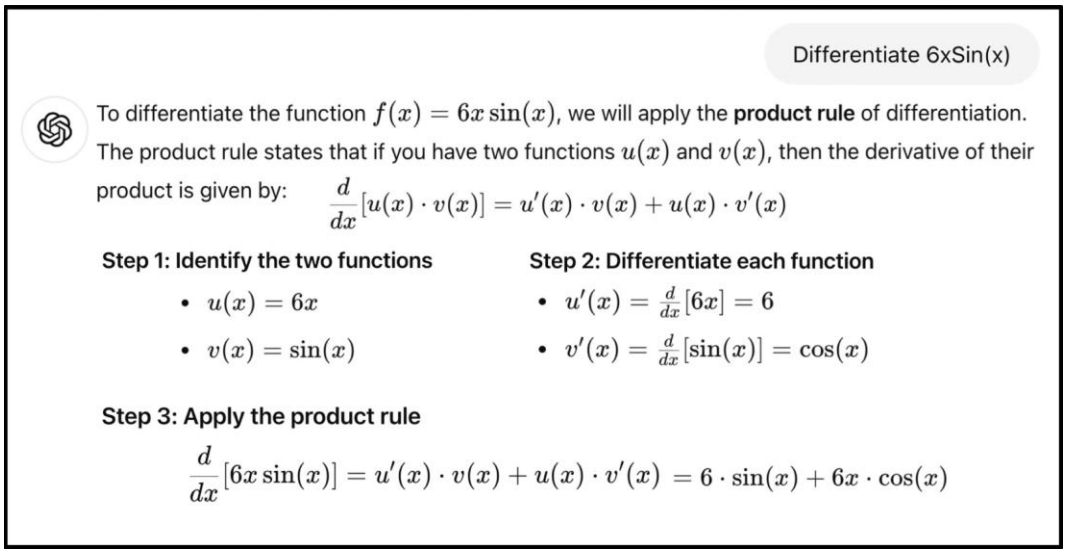

Figure 3. Example of ChatGPT-4o correcly applying the product rule.

6. Limitations of ChatGPT generated solutions

A review of ChatGPT's mathematical errors reveals several recurring patterns. This section highlights three common pitfalls, each explained and illustrated with an example.

6.1 Calculation Errors

A major flaw of older ChatGPT models is their inability to reliably perform numerical computations. In Figure 4, ChatGPT-3.5 was prompted to calculate 0.72 raised to the power of 9 three different times. It yielded three different results, none of which were correct.

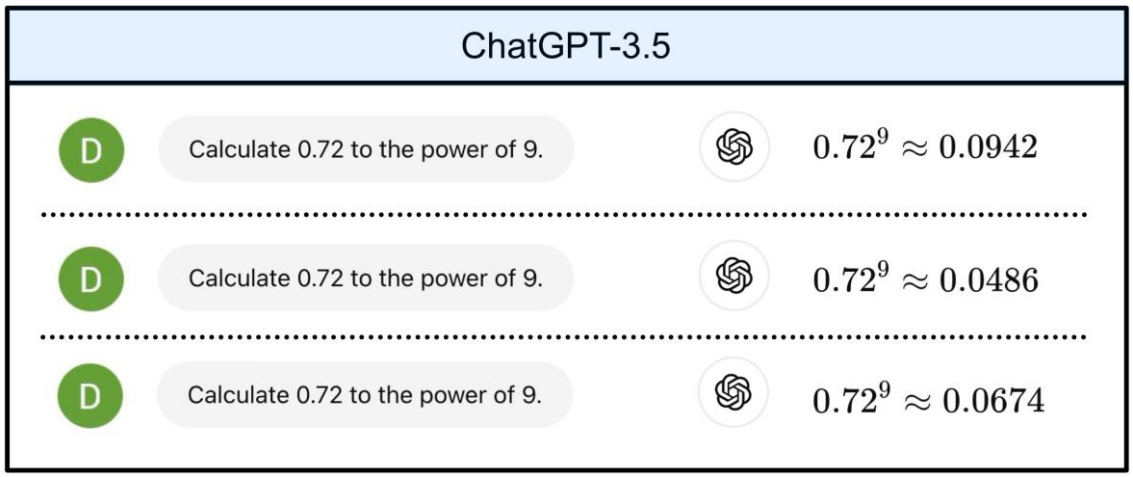

Figure 4. ChatGPT-3.5 struggles with basic computations.

The computational limitations of earlier models, such as ChatGPT-3.5, are well documented. Raftery (2023) found that manually correcting ChatGPT-3.5's calculation errors using a hand calculator improved its average performance on a series of online quizzes from 35% to 72%.

Newer models, such as ChatGPT-4o, incorporate a code interpreter that utilises Python to perform numerical computations. Given Python's reliability in handling such calculations, computation errors are effectively eliminated in models equipped with this functionality. Figure 5 showcases ChatGPT-4o's code interpreter window, which is used to accurately compute 0.72 raised to the power of 9.

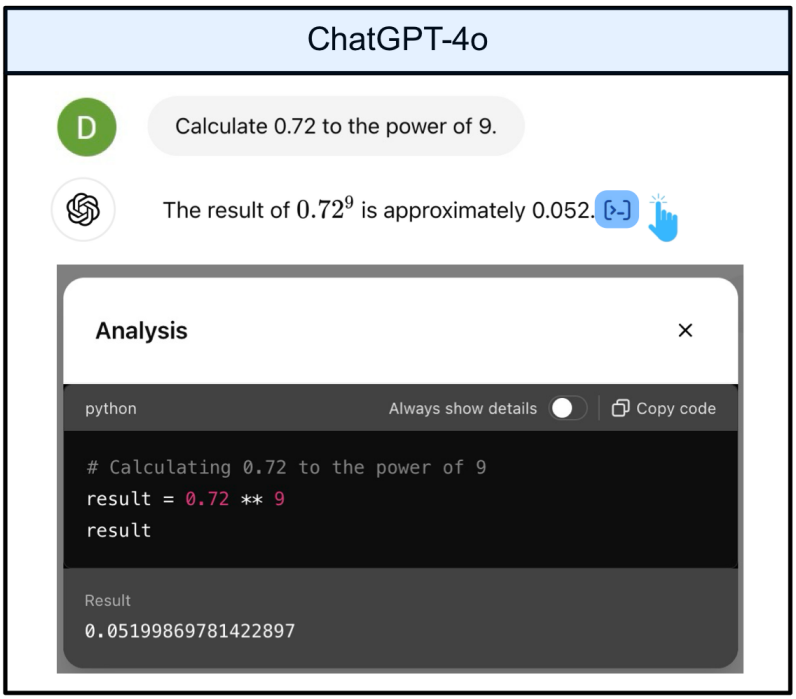

Figure 5. ChatGPT-4o utilises Python to accurately perform numerical calculations.

When using ChatGPT for numerical computations, it is important to check whether the model version includes a code interpreter, as this significantly affects calculation accuracy.

6.2 Incorrect Workings

In some cases, the solution provided by ChatGPT starts correctly, produces the correct answer, but contains mistakes in the intermediate steps. In Figure 6, ChatGPT is prompted to calculate the determinant of a 3x3 matrix. It correctly states a formula for calculating the determinant and arrives at the correct answer of 0. However, the intermediate workings contain multiple errors.

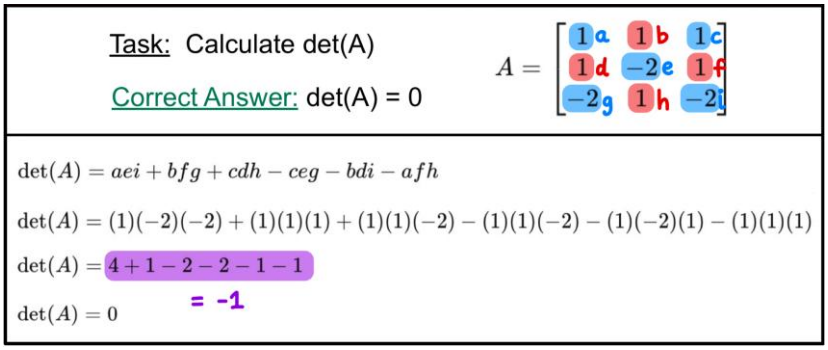

Figure 6. ChatGPT starts correctly and arrives at the correct final answer, but has incorrect workings inbetween.

ChatGPT's text-based responses are known to be highly persuasive, even when conveying information that is factually incorrect. This concern extends to ChatGPT's mathematical reasoning. Even if a solution appears correct at first glance, it is important to thoroughly verify every step to ensure correctness.

6.3 Overcomplicated Answers

The more powerful ChatGPT models become, the greater the risk of them overcomplicating their answers. Figure 7 compares ChatGPT-3.5 and ChatGPT-4o's responses to the following prompt: “A person has the option of taking one of three routes to work, A, B or C. The probability of taking route A is 35%, and B is 25%. The probability of being late for work if she goes by route A is 10% and similarly by route B is 5% and route C is 2%. Draw a tree diagram to illustrate the outcomes and their probabilities.”

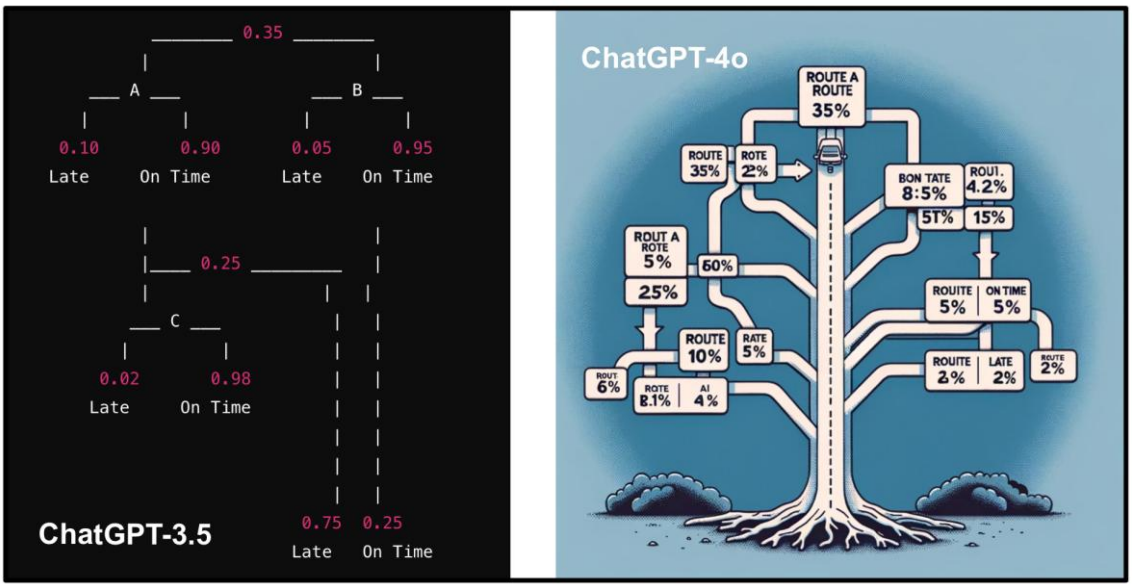

Figure 7. An example where ChatGPT-4o overcomplicates its response, leading to worse performance than ChatGPT-3.5.

While not entirely correct, ChatGPT-3.5 makes a reasonable attempt at the problem. Notably, it infers information beyond what is directly stated in the question, such as correctly determining that if the probability of being late on route A is 10% (0.1), then the probability of being on time must be 90% (0.9). ChatGPT-4o generates a visually appealing but mathematically useless depiction of a tree with a car driving down its trunk. In most cases, this issue can be resolved by reprompting for a simpler answer or specifying the image generation method, such as using the Matplotlib library in Python.

7. Conclusions

Using a ChatGPT-4o-generated draft as a starting point for an exam solutions document led to a total creation time of 2 hours 36 minutes, compared to 4 hours 31 minutes without using the assistance of ChatGPT - a 42% reduction in the time needed.

The format of each solution document was selected according to the lecturer’s preference, with the aim of minimising the total time required. This resulted in two different formats - a typed LaTeX document with ChatGPT’s assistance versus electronically handwritten solutions without. A fairer comparison would involve creating LaTeX documents in both cases, though it is believed this would result in an even greater reduction in the time required.

An issue not addressed in this case study is the quality of the solution documents. While every effort was made to ensure correctness, their effectiveness in supporting learning was not assessed. Future

work could involve surveying students who used these resources for revision to determine their perceived usefulness.

ChatGPT-4o is highly effective at generating maths solutions, offering quick responses and well-structured explanations. However, it has limitations, including the possibility of calculation errors, incorrect workings, and overly complex answers. Given the risk of unnoticed errors hindering learning, it is advisable for a subject expert to verify the accuracy of ChatGPT-generated maths solutions before they are shared with novice learners.

8. References

Ahmad, N., Murugesan, S. and Kshetri, N., 2023. Generative Artificial Intelligence and the Education Sector. *Computer*, 56(6), pp. 72-76, http://doi.org/10.1109/MC.2023.3263576.

Alkaissi, H. and McFarlane, S.I., 2023. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus* 15(2), http://doi.org/10.7759/cureus.35179.

Anthropic, 2024. *Introducing the next generation of Claude* [Online]. Available at: https://www.anthropic.com/news/claude-3-family [Accessed 10 April 2025].

Chervonyi, Y., Trinh, T.H., Olšák, M., Yang, X., Nguyen, H., Menegali, M., Jung, J., Verma, V., and Le, Q.V., and Luong, T., 2025. *Gold-medalist Performance in Solving Olympiad Geometry with AlphaGeometry2*. Available at: https://doi.org/10.48550/arXiv.2502.03544 [Accessed 10 April 2025].

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö. and Mariman, R., 2024. Generative AI Can Harm Learning. *The Wharton School Research Paper*. http://doi.org/10.2139/ssrn.4895486.

Frieder, S., Pinchetti, L., Griffiths, R. R., Salvatori, T., Lukasiewicz, T., Petersen, P. and Berner, J., 2024. Mathematical capabilities of chatgpt. *Advances in neural information processing systems*, 36.

Giray, L., 2023. Authors Should be Held Responsible for Artificial Intelligence Hallucinations and Mistakes in their Papers. *Journal of the Practice of Cardiovascular Sciences*, 9(2), pp.161-163, http://doi.org/10.4103/jpcs.jpcs_45_23.

Mollick, E. R. and Mollick, L., 2022. *New Modes of Learning Enabled by AI Chatbots: Three Methods and Assignments.* Available at: http://doi.org/10.2139/ssrn.4300783.

Newton, P.M. and Xiromeriti, M., 2023. *ChatGPT Performance on MCQ Exams in Higher Education. A Pragmatic Scoping Review*. Available at: https://doi.org/10.35542/osf.io/sytu3 [Accessed 7 February 2025].

OpenAI, 2024. *Learning to reason with LLMs* [Online]. Available: https://openai.com/index/learning-to-reason-with-llms/ [Accessed 8 February 2025].

Raftery, D., 2023. Will ChatGPT pass the online quizzes? Adapting an assessment strategy in the age of generative AI. *Irish Journal of Technology Enhanced Learning*, 7(1), https://doi.org/10.22554/ijtel.v7i1.114.