

Bot or not? Exploring if maths lecturers can detect AI solutions

Declan Manning^{}, Shane O Rourke^{}, Violeta Morari^{}

Department of Mathematics, Munster Technological University, Ireland

How to cite: Manning, D.; O Rourke, S.; Morari, V. (2026). Bot or not? Exploring if maths lecturers can detect AI solutions. In: 12th International Conference on Higher Education Advances (HEAd'26). Valencia, 15-18 June 2026. <https://doi.org/10.4995/HEAd26.2026.21590>

Abstract

This study investigates whether mathematics lecturers can distinguish between student-written and AI-generated mathematical solutions. Twenty-four university lecturers classified eight solutions to first-year mathematics problems. Four solutions were based on genuine student work, two were copied verbatim from ChatGPT-4o outputs, and two were “Maths Only” solutions containing only the mathematical statements extracted from a ChatGPT-4o response.

Across 192 classifications, overall accuracy was 57.8%. AI solutions written with full explanatory text were easily identified (93.8%), but “Maths Only” AI solutions were rarely detected (18.8%). This indicates that mathematical structure alone does not allow for reliable distinction between AI and student work, even for experienced lecturers.

Participants had the opportunity to change their classifications following a short group discussion. However, this led to a net decrease in classification accuracy (10 incorrect switches compared to 8 correct switches).

Keywords: Generative AI; AI Detection; Mathematics

1. Introduction

The mathematical problem-solving ability of generative AI has advanced significantly in recent years. OpenAI regularly evaluates its models using the American Invitational Mathematics Exam (AIME), an advanced set of problems used in selecting the US Mathematics Olympiad team. ChatGPT-4o (released May 2024) achieved a score of 13.4% on the 2024 AIME, while ChatGPT-o1 (released December 2024) scored 83.3% on the same exam. ChatGPT-5.2 (released December 2025) is reported to have achieved a perfect score of 100% on the 2025 AIME. Specialised problem-solving models such as Google’s Gemini Deep Think have achieved remarkable results. At the 2025 International Mathematical Olympiad, Gemini Deep Think achieved a gold-medal score when graded under official competition criteria. Experimental evidence from Ye et al. (2025) further indicates that AI models not only can

correctly solve problems, but that they are capable of producing novel and creative mathematical solutions.

The increasing prevalence of generative AI presents significant challenges to academic integrity in third-level education. Take-home assignments and essays are vulnerable to being partially or completely generated by AI. Lecturers effectively face a modern Turing Test (Turing, 1950): when reviewing a submission, lecturers must judge whether the work was produced by the student or by AI. With the aid of sophisticated AI detection tools, some studies report high success rates in identifying AI-generated content. Cingillioglu (2023) formed a dataset containing both AI and human essays. They found that GPTZero, a widely used AI detection tool, obtained a 96% classification accuracy. In contrast, Weber-Wulff et al. (2023) tested twelve popular AI detection tools and reported that none achieved an average accuracy above 80%. There is concern that detection becomes more difficult as AI models improve. Elkhatat et al. (2023) tested five AI detection tools and found that they achieved an accuracy of 94.7% for ChatGPT-3.5 outputs, compared to a 50.7% accuracy for ChatGPT-4 outputs.

Without the use of detection tools, humans struggle to distinguish between AI and human generated content. Matthews and Volpe (2023) presented participants with two paragraphs side by side, informing them that one was written by a human and the other generated by AI. The sources were correctly identified in 50% of cases, equivalent to random guessing. Similarly, Chein et al. (2024) reported a 57% accuracy when 194 participants attempted to identify the origin of human/AI texts.

While AI detection has been widely studied in text-based contexts, relatively little work has focused on mathematical content. Jia et al. (2024) asked participants to classify mathematical problem-solving tips as being AI or human generated. The classification accuracy for the AI tips was under 40%, highlighting the difficulty of the task. Notably, only 3 of the 121 participants reported mathematics teaching experience.

2. Method

2.1. Study design

The primary objective of the study was to obtain an unbiased estimate of how accurately mathematics lecturers can distinguish between student-written and AI-generated mathematical content. The study was conducted on the 14th of May 2025, during a 45-minute session at a staff day for the Munster Technological University maths department.

We identified three primary goals for the study design; participants should provide their judgement independent from the influence of others, the number of clues arising from superficial features of the solutions should be minimized, and there should be a diverse range

of question types and solution styles. With this in mind, we chose the format of a printed handout that participants could complete individually. The 45-minute session was structured as follows: a 10-minute presentation to introduce the task, 25 minutes for participants to independently complete the handout in silence, and a 10-minute group discussion period to conclude.

The handout contained several maths solutions, some based on genuine student work and others derived from answers generated by an AI model. The solutions were not labelled with their source, and participants were tasked with classifying each solution as “Student Work” or “Generative AI response”.

2.2. Problem choice

When selecting the problems, we focused on exam questions that had been attempted by students earlier in the semester. This allowed us to draw on authentic student work for the student solutions. To ensure the mathematical content was accessible to all maths staff, we selected exam problems from a first-year introductory maths module which covered core topics such as algebra and trigonometry. Due to the limited time available to run the study, we selected 2 problems, presenting 4 different solutions for each. The first problem was chosen to be purely algebraic in nature whereas the second involved the use of broader problem-solving skills:

Problem 1 (Log Equation): Solve for x in the equation $9 \log_6(x + 8) = 18$

Problem 2 (Area of Rectangle): In a rectangle, the length is 16 cm longer than the width. If the area of the rectangle is 800 cm², find the length and width of the rectangle. The answers should be rounded to 3 decimal places.

To help protect student identities, the numerical values in the questions were modified from those on the original exam. Aside from the specific calculations involved, this surface-level adjustment did not alter the sequence of problem-solving steps required.

2.3. Solution creation

To ensure variety in the solutions, we selected four different sources:

- One student solution that followed the conventional solution style taught in class.
- One student solution that adopted a more unconventional approach.
- One “Full Workings” AI solution, copied verbatim from an AI response.
- One “Maths Only” AI solution, containing only the mathematical statements extracted from an AI response.

All presented solutions were numerically correct, as we wanted participants to focus on the reasoning and logic rather than on correctness. To maintain a consistent solution style, one team member handwrote all solutions by transcribing the respective source material. When

transcribing the student solutions, care was taken to preserve the original layout and sequence of problem-solving steps exactly. Apart from substituting the updated numerical values, the transcription was an exact reproduction of the student's exam work.

The AI solutions were generated on the 30th of April 2025 using the freely available version of ChatGPT at the time (ChatGPT-4o). The prompt consisted solely of the problem statement, with LaTeX notation used as necessary for superscripts and subscripts. To obtain two solutions per problem, we entered the same prompt twice in two separate, logged-out chat sessions. The first response was copied verbatim to create the "Full Workings" AI solution. For the second response, the mathematical statements only were transcribed to create the "Maths Only" AI solution. Due to ChatGPT's stochastic nature, the mathematical statements in the two outputs did not match, which provided useful variation in the solutions. The four solutions to the Log Equation problem are displayed in Figure 1.

GenAI (Full Workings) To solve for x in the equation: $9 \log_6(x+8) = 18$ Step 1: Isolate the logarithmic expression First, divide both sides of the equation by 9: $\log_6(x+8) = \frac{18}{9}$ This simplifies to: $\log_6(x+8) = 2$ Step 2: Convert the logarithmic equation to exponential form. Recall that $\log_b(a) = c$ is equivalent to $a = b^c$ Applying this to our equation: $x+8 = 6^2$ Step 3: Simplify the exponential expression Since $6^2 = 36$, we have: $x+8 = 36$ Step 4: Solve for x Subtract 8 from both sides: $x = 36 - 8$ $x = 28$	GenAI (Maths Only) $9 \log_6(x+8) = 18$ $\log_6(x+8) = \frac{18}{9} = 2$ $6^2 = x+8$ $36 = x+8$ $x = 36-8$ $x = 28$ Student (Conventional) $\frac{9 \log_6(x+8)}{9} = \frac{18}{9}$ $\log_6(x+8) = 2$ $6^{\log_6(x+8)} = 6^2$ $x+8 = 36$ $x = 28$ Student (Unconventional) $9 \log_6(x+8) = 18$ $\log_6((x+8)^9) = 18$ $(x+8)^9 = 6^{18}$ $x+8 = \sqrt[9]{6^{18}}$ $x+8 = 36$ $x = 28$
---	--

Figure 1. The 4 "log equation" solutions that were presented in the handout

3. Results

3.1. Classification accuracy

Twenty-four participants each classified eight maths solutions as AI-generated or student-written, resulting in 192 classifications. Participants received a score from 0 to 8 based on the number of correct classifications (see Table 1). The mean score was 4.625, with a standard deviation of 1.28. Table 2 summarises classification accuracy by question.

Table 1. Classification score (by participant, n = 24)

Correct Classifications	0	1	2	3	4	5	6	7	8
Number of Participants	0	0	1	2	10	6	2	3	0
Relative Frequency	0%	0%	4.2%	8.3%	41.7%	25%	8.3%	12.5%	0%

Table 2. Classification accuracy (by solution, presented in handout order).

	<i>Log Equation</i>				<i>Area of Rectangle</i>			
Source	AI	AI	Student	Student	Student	AI	Student	AI
Type	<i>Full Workings</i>	<i>Maths Only</i>	<i>Unconv.</i>	<i>Conv.</i>	<i>Conv.</i>	<i>Maths Only</i>	<i>Unconv.</i>	<i>Full Workings</i>
Correct Classifications	22	2	9	18	19	7	11	23
Accuracy	91.7%	8.3%	37.5%	75%	79.2%	29.2%	45.8%	95.8%

Overall, participants made 111 correct classifications out of 192 (57.8%). The AI Full Workings solution was classified with high accuracy for both the Log Equation and Area of Rectangle questions. When these two solutions are excluded, accuracy falls sharply to 66 of 144 (45.8%), which is worse than random guessing. The average classification accuracy was higher for the Area of Rectangle solutions (62.5%) than for the Log Equation solutions (53.1%).

3.2. Confidence ratings

After each classification, participants selected one of five confidence ratings. Table 3 shows the frequency of each rating and the number of associated correct classifications.

Table 3. Confidence ratings and associated accuracy

	<i>Certain</i>	<i>Very Confident</i>	<i>Moderately Confident</i>	<i>Slightly Confident</i>	<i>Completely Unsure</i>	Blank
Total Frequency	7	32	43	52	47	11
Correct Classifications	7	24	25	25	24	6
Accuracy	100%	75%	58.1%	48.1%	51.1%	54.5%

The data indicates a positive relationship between confidence and classification accuracy, with higher confidence generally associated with greater accuracy. All seven *Certain* classifications were correct. In contrast, across the *Slightly Confident* and *Completely Unsure* categories, only 49 of 99 classifications were correct (49.5%).

3.3. Academic integrity actions

Alongside their confidence rating, participants selected one of three actions they would take in relation to academic integrity if the solution appeared in a take-home assignment where generative AI use was explicitly prohibited. Table 4 summarises the results, split by whether the solution was classified as AI or student.

Table 4. Academic integrity actions

	<i>Raise infringement concern with head of department</i>	<i>Ask a colleague for their opinion</i>	<i>Take no action</i>	Blank
Classified as AI	18	45	18	12
Classified as Student	0	10	77	12

In the 18 cases where a solution was classified as AI-generated and the action *Raise an infringement concern with the head of department* was selected, the solution was actually student work in 4 cases (false positive rate of 22.2%). *Take no action* was selected in the vast majority of cases where the solution was classified as student-generated. When a solution was classified as AI-generated, the most common action was *Ask a colleague for their opinion*.

3.4. Explanations

Participants were asked to explain the reasoning behind their classification. Table 5 presents two explanations for each solution, including justifications for correct and incorrect decisions.

Table 5. Selected explanations. Associated classification (AI/Student) provided in brackets.

1. Log Equation GenAI (Full Work.)	(AI) Comments are very detailed, proper grammar etc. Too perfect
	(AI) A typical student would not give level of explanation in their work.
2. Log Equation GenAI (Maths Only)	(Student) This solution is typical of what I see in a closed book assessment.
	(Student) Look like a logical student answer.
3. Log Equation Student (Unconv.)	(AI) Very unusual method so probably GenAI
	(Student) It is an unusual approach to the solution which makes me think student solution. I feel AI would take the most direct route to solve the question.
4. Log Equation Student (Conv.)	(AI) Students don't usually show the first line.
	(Student) I have no reason to suspect A.I.
5. Area of Rectangle Student (Conv.)	(AI) I think the student wouldn't work with $\sqrt{3456}$
	(Student) No fluff in solution. All relevant steps included.
6. Area of Rectangle GenAI (Maths Only)	(AI) A good bit of "algebra" work was shown and this is uncharacteristic of students.
	(Student) It looks like responses I might get from giving a similar question to students.
7. Area of Rectangle Student (Unconv.)	(AI) No workings given
	(Student) Gen AI would have produced better work. Student can/will be penalised anyway for failure to show workings.
8. Area of Rectangle GenAI (Full Work.)	(AI) Explanation too good. Unusual for a student to give such detail.
	(AI) A student wouldn't annotate their answer to that extent (in my experience)

3.5. After discussion changes

After independently reviewing the eight solutions, the lecturers took part in group discussions of four to six people. The final page of the handout asked participants to indicate if they wished to change any of their answers following this discussion (results in Table 6).

Table 6. After discussion changes

Source	Log Equation				Area of Rectangle			
	AI	AI	Student	Student	Student	AI	Student	AI
Type	Full Workings	Maths Only	Unconv.	Conv.	Conv.	Maths Only	Unconv.	Full Workings
Correct Switches	1	1	0	2	2	1	1	0
Incorrect Switches	0	0	3	2	0	2	3	0

Eighteen answers were changed: eight from incorrect to correct and ten from correct to incorrect, resulting in a net decrease in overall accuracy. A large number of responses were left blank (76 of 192), likely because participants ran out of time.

4. Discussion and conclusions

This study demonstrates that easily implemented edits to AI-generated output can substantially affect its detectability. The *Full Workings* AI solutions were identified with high accuracy

(93.8%), with many participants noting that their level of detail was uncharacteristic of student work. In contrast, the *Maths Only* AI solutions were rarely classified correctly (18.8%). Rather than copying a full AI response verbatim, a student attempting to cheat on a take-home assignment could submit work resembling the *Maths Only* solutions, resulting in a low chance of detection. These findings suggest that take-home maths assignments where AI use is prohibited may no longer be appropriate.

A major challenge for third-level lecturers is the growing prevalence of generative AI output in student submissions. Seeking advice from a colleague is viewed as a helpful strategy, reflected by the frequent selection of the *Ask a colleague for their opinion* action for AI classified solutions. However, in this study, post-discussion changes led to a net decrease in classification accuracy, suggesting that consulting colleagues may be less effective than perceived.

An interesting finding of this study is that lecturers had differing opinions on how to classify an unusual solution. The two *Unconventional Student* solutions were classified as student work in 37.5% and 45.8% of cases respectively - closer to a 50:50 split than any of the other six solutions. This disagreement is highlighted by the contrasting comments made on the same solution: one participant remarked “Very unusual method so probably GenAI” while another stated “It is an unusual approach to the solution which makes me think student solution.”

One limitation of this study is the simplicity of the two problems, which were both drawn from an introductory first-year mathematics module. More complex problems with richer solution steps would likely provide lecturers with more clues for classification. Accuracy was higher for the problem-solving *Area of Rectangle* question (62.5%) than for the algebraic *Log Equation* question (53.1%), suggesting that increased problem complexity may improve classification accuracy.

In conclusion, this study provides empirical evidence that experienced mathematics lecturers could not reliably distinguish between student-written and AI-generated solutions when surface-level linguistic clues are removed, raising serious implications for assessment and academic integrity in mathematics.

References

- Chein, J.M., Martinez, S.A. & Barone, A.R. Human intelligence can safeguard against artificial intelligence: individual differences in the discernment of human from AI texts. *Sci Rep* 14, 25989 (2024). <https://doi.org/10.1038/s41598-024-76218-y>
- Cingillioglu, I. (2023). Detecting AI-generated essays: the ChatGPT challenge. *The International Journal of Information and Learning Technology*, 40(3), 259-268. <https://doi.org/10.1108/IJILT-03-2023-0043>

- Elkhatat, A., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, 19, 1-16. <https://doi.org/10.1007/s40979-023-00140-5>
- Jia, J., Wang, T., Zhang, Y., & Wang, G. (2024). The comparison of general tips for mathematical problem solving generated by generative AI with those generated by human teachers. *Asia Pacific Journal of Education*, 44, 8 - 28. <https://doi.org/10.1080/02188791.2023.2286920>
- Matthews, J., & Volpe, C. R. (2023). Academics' perceptions of ChatGPT-generated written outputs: A practical application of Turing's Imitation Game. *Australasian Journal of Educational Technology*, 39(5), 82–100. <https://doi.org/10.14742/ajet.8896>
- Turing, A. M. (1950) Computing machinery and intelligence. *Mind*, 59(236), 433–460. <https://doi.org/10.1093/mind/LIX.236.433>
- Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023) Testing of detection tools for AI-generated text. *International Journal for Educational Integrity* 19,26. <https://doi.org/10.1007/s40979-023-00146-z>
- Ye, J., Gu, J., Zhao, X., Yin, W., & Wang, G. (2025). Assessing the Creativity of LLMs in Proposing Novel Solutions to Mathematical Problems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24), 25687-25696. <https://doi.org/10.1609/aaai.v39i24.34760>