arXiv:2610.27881v1 [cs.CY] 20 Oct 2026

EVALUATING LLM-GENERATED FEEDBACK ON HANDWRITTEN MATHEMATICS: FEEDBACK QUALITY AND SUBSEQUENT EXAM PERFORMANCE

Arne Vanhoyweghen[1,2,✉] **Vincent Holst**[1,2] **Melika Mobini**[1,2]

Lukas Van de Voorde[1,2] **Tibo Vanleke**[1,2] **Andres Algaba**[1,2] **Sam Verboven**[1]

Marie-Anne Guerry[1] **Filip Van Droogenbroeck**[1] **Vincent Ginis**[1,2,3]

[1]Data Analytics Laboratory, Vrije Universiteit Brussel, Brussels, Belgium.
[2]imec-SMIT, Vrije Universiteit Brussel, Brussels, Belgium.
[3]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA.

October 21, 2026

ABSTRACT

We evaluate the quality of feedback and practice exercises generated by GPT-5.1 from handwritten university mathematics answers and their grades, and examine whether receiving this feedback improves subsequent exam performance. We use six weekly bonus tests in an English ($n = 343$) and a Dutch ($n = 236$) cohort. Tests 1 and 3 used a crossover feedback design; otherwise, all attendees received feedback. To evaluate the feedback, six reviewers assessed 157 messages from three English-cohort tasks using six criteria scored from 0 to 3. The practice exercises received the highest proportion of strong ratings (95.5%), while diagnostic specificity (41.1%) and clarity and tone (33.4%) received fewer. Feedback for the lowest-scoring answers was rated as less actionable, suggesting that these students may need more targeted guidance from instructors. Comparisons of related exam questions showed mixed results, with no statistically significant differences after adjustment across four comparisons. Greater attendance and feedback exposure were associated with higher exam scores, but the contribution of feedback could not be separated from attendance and practice. Our findings show the potential of an integrated framework merging instructor guidance with automated feedback generation to provide personalised learning support at scale.

Keywords Personalised Feedback · Handwritten assessment · Large language models

1 Introduction

A grade tells a student how well they performed, but it does not necessarily tell them what went wrong or how to improve. This creates an important gap between a course's evaluative and educational intent. Feedback is the most common way to close the gap between evaluation and learning, by helping students identify their mistakes and understand what they need to work on [1, 2, 3, 4, 5]. However, providing such feedback for handwritten answers is time-consuming and makes short feedback loops non-viable when the number of students increases. Previous work has used data from students' interactions with online learning activities to provide personalised feedback in large courses [6]. LLMs can assist human instructors by offering a way to quickly generate individual comments and practice exercises after an answer has been graded [7, 8, 9].

✉ Corresponding author: arne.vanhoyweghen@vub.be

We address two questions. First, can GPT-5.1 provide accurate feedback and practice exercises suited to the level demonstrated in a student's handwritten answer? Second, does receiving this feedback improve performance on related exam questions? A model might correctly explain how to solve an exercise but still provide little guidance to a student who does not know where to begin. It might also, because of hallucinations, offer advice on reasoning the student never used [10, 11, 12]. Additionally, LLMs might provide useful feedback, but the feedback might fail to engage students to use it in the learning process [13, 14]. To investigate the usefulness of the feedback, we therefore need to consider both what students receive and how they perform afterwards.

Handwritten mathematics poses an additional challenge compared with typed or structured answers. Before generating feedback, the model must interpret the notation, spatial structure, corrections, and order of the student's reasoning. Errors in this interpretation can carry into the diagnosis and feedback [15, 16, 17]. Research in this area has mainly focused on recognition, error detection, and grading [18, 19]. We extend this work by using handwritten university mathematics answers to generate individual feedback during a course. We assess the quality of this feedback and use a crossover design to examine whether receiving it improves later performance on exam questions testing the same mathematical skills.

We build on the pipeline detailed in [20], in which handwritten assessments were scanned, graded with LLM support, and verified by human assessors. Once the grades were finalised, we used a separate generation step to provide comments on the student's answer and two practice exercises. The feedback was then sent by email, together with the grade and an image of the student's work. This allowed us to add individual feedback to the existing assessment process.

In this study, we describe how we implemented this approach and evaluate its use across six weekly bonus tests. We first assess a sample of the feedback using criteria that distinguish mathematical correctness from the quality of the diagnosis, guidance, and presentation. We then compare performance on related exam questions between students who did and did not receive feedback on the corresponding bonus test. Finally, we consider how attendance and the number of tests with feedback relate to overall exam performance.

2 Methods

2.1 Participants and setting

We collected data from the English- and Dutch-taught cohorts of the same undergraduate mathematics course at the Vrije Universiteit Brussel, with 343 and 236 students with bonus-test records, respectively. Both cohorts shared many similarities; they were taught by the same professor and followed the same course content and structure. However, they also differed in important ways, including teaching assistants and backgrounds: the Dutch-speaking students came from the same schooling system, while the English-speaking students came from a wide range of backgrounds.

To investigate performance on related content, we linked specific bonus-test questions to final-exam questions at the start of the semester. Each cohort took a separate exam, but the topics and question types were closely matched. We linked the bonus-test data to exam records for 343 students in the English cohort and all 236 students in the Dutch cohort. Of these, 331 and 236 students participated in the first session exam.

2.2 Bonus test design and A/B version groups

Students could participate in six weekly bonus tests during the exercise sessions. Each test lasted 10 minutes, covered material from a previous exercise session, and consisted of two handwritten exercises. Participation was voluntary. The six tests could add up to 1.0 point to the final exam grade, which was graded out of 20.

Each test contained a common question (Q1) and a version-specific question (Q2). Within a class group, students assigned to versions A and B received the same Q1. We used this question as an anchor to assess comparability across version groups. Across class groups, the questions covered the same topic at a similar level of difficulty, but could differ in their wording or problem instance.

The version of Q2 depended on the parity of the student's identifier ($n_A^{\text{ENG}} = 159$, $n_B^{\text{ENG}} = 185$; $n_A^{\text{NL}} = 121$, $n_B^{\text{NL}} = 115$). Versions A and B covered the same topic but required different solution strategies. Consequently, when comparing performance on a linked exam question, we only include students who received the corresponding version of Q2.

2.3 Grading and feedback pipeline

Grading. We build on the grading pipeline set out in [20]. After scanning and anonymising the answer sheets, we used GPT-5.1 (Reasoning High) to score each response five times against a rubric-style grading key. We took the

maximum score as the provisional grade, which a human grader then verified before release. We recorded all scores on a 0–10 scale.

Feedback randomisation. We varied feedback assignment on tests 1 and 3. On each of these tests, students were assigned to receive feedback on both questions or on neither question. The assignment followed a fixed crossover: students who received feedback on test 1 served as controls on test 3, and vice versa. We limited the randomised withholding of feedback to these two tests because extending it would have increased differences in the support students received on assessed course content. On the remaining tests, all attendees received feedback. Table 1 reports the numbers in each group.

Table 1: Number of attendees assigned feedback or control on each bonus test. Assignment crossed over between tests 1 and 3. All attendees received feedback on the remaining tests. The duplicate Dutch record has been excluded.

	ENG		NL	
Test	Treated	Control	Treated	Control
1	145	143	98	104
2	277	0	172	0
3	120	109	86	88
4	235	0	146	0
5	211	0	122	0
6	196	0	123	0

Feedback generation. After verifying the grades, we used GPT-5.1 to generate feedback for each student's answer. We generated each feedback message only once. The model used the final grade to customize its comments and practice exercises to the student's performance, without altering the grade itself. For each question, we asked it to identify strengths and mistakes in the answer and provide two practice exercises.

To generate this feedback, we provided GPT-5.1 with the question text, a cropped image of the student's answer, the official worked solution, and the final numerical grade. No names or student identifiers were included. We used the following prompt for both cohorts:

> You are supposed to give personalized feedback to a student's solution. Point out what they did well and what they could improve on. Be kind, constructive, and to the point. Use the solution provided to guide your feedback; you should not reference it explicitly. Treat the method used in the provided solution as the preferred method and keep your feedback and practice exercises in line with this method. You are not allowed to mention the final score. You can only use it to tailor your feedback. In this course, we only work with real numbers. So if the student does not mention complex numbers, neither should you. Moreover, it's important to provide actionable feedback in the form of two sample exercises which they can use to improve their skills and practice more. For students with a score of 10, provide two similar exercises that are slightly more challenging. For students who did not attempt to solve the question, reassure them that it is not a problem and provide two simpler exercises to help them get started. For students who made a different exercise than the one asked, remind them to make the correct exercise in the future and give them two practice exercises. For all practice exercises, include only the final solution. Do not provide hints, explanations, or intermediate steps for these practice exercises. Students will not be able to respond to the feedback, or interact with you in any way. So please don't prompt them to interact with you.
>
> *[Optional: Your feedback should be in Dutch.]*
>
> **Question:** {question text}
> **Solution:** {solution text}
> **Student's score:** {student score}
> **Student answer:** {anonymised answer image}

Post-processing and delivery. To ensure students could understand the feedback, we adapted the LLM-generated feedback, using `pylatexenc` and custom regular expressions to convert LaTeX expressions and formatting into unformatted text. We also screened the generated feedback for potentially unintended or dangerous content and found none. Students received their final grade, an image of their answer, and, where assigned, the feedback. The email stated that the feedback was AI-generated and should be read critically. Students in the control group received only their grade. We delivered the feedback approximately one and a half week after each test.

2.4 Evaluation of the generated feedback

To evaluate the generated feedback, we sampled messages for three tasks from the English cohort: test 1, version B, Q2; test 4, version A, Q2; and test 5, Q1 in both versions. These tasks provided 82, 108, and 210 eligible messages, respectively. Each message concerned one student's answer to one question. We stratified the sample by the verified grade to include feedback for students at different performance levels. The score bands were $[0, 1)$, $[1, 4)$, $[4, 7)$, $[7, 10)$, and $\{10\}$, with target counts of 5, 15, 15, 15, and 5 per task. When a band contained fewer messages than the target, we included all of them. This resulted in 51, 51, and 55 messages from the three bonus tests.

Reviewers assessed each message alongside the student's answer, the question, the worked solution, and the verified grade. We used six criteria to distinguish mathematical accuracy, diagnostic specificity, actionability, method alignment, clarity and tone, and the quality of the practice exercises (Table 2). Each criterion was scored from 0 to 3, with 0 indicating deficient, 1 weak, 2 adequate, and 3 strong feedback. The rubric provided a separate description of each category for each criterion. Reviewers also completed a checklist indicating whether the feedback mentioned a strength, proposed an improvement or extension, gave a concrete next step, and included two practice exercises.

Table 2: Criteria and scoring guidelines used to evaluate the feedback. Scores were assigned relative to the student's answer and demonstrated level.

Criterion	0: Deficient	1: Weak	2: Adequate	3: Strong
A1: Mathematical accuracy Correct evaluation of the student's work	Incorrect evaluation; invents errors or misses major ones	Partially correct evaluation with minor misjudgements	Correct overall, with minor gaps	Fully correct evaluation of the student's solution
A2: Diagnostic specificity Identification of strengths and weaknesses	No meaningful identification; generic or absent	Identifies the topic, but remains vague	Identifies specific elements that are correct or incorrect	Precisely identifies and explains why elements are correct or incorrect
A3: Actionability Usefulness of the guidance for further learning	Provides no guidance	Provides generic advice that the student cannot readily act on	Provides some concrete guidance, such as a correction or extension	Provides clear and targeted next steps for remediation or further learning
A4: Method alignment Consistency with the expected solution method	Conflicts with the expected method	Partially aligns with the expected method	Mostly aligns with the expected method	Fully aligns with the expected method
A5: Clarity and tone Quality of the communication	Confusing or inappropriate	Understandable, but poorly structured or verbose	Clear and reasonably concise	Clear, concise, and constructive
A6: Practice-exercise quality Relevance, correctness, and calibration of the exercises	Missing, incorrect, or irrelevant	Targets the wrong concept or level of difficulty	Relevant and correct, but weakly calibrated	Precisely targeted and well calibrated to the student's needs

2.5 Linking bonus tests to exam questions

At the start of the semester, we linked specific bonus-test items to final-exam questions covering the same mathematical skill. For example, version B of Q2 in Test 1 concerned irrational inequalities, including domain restrictions and the conditions under which an inequality can be squared. The corresponding exam question assessed these same skills using a different inequality. Similarly, version A of Q2 in Test 3 concerned limits at positive infinity involving fractions with square roots and polynomials. These links allowed us to examine later performance on the specific content addressed by the feedback.

The English- and Dutch-language exams were aligned and assessed the same types of mathematical questions, using different expressions. Table 3 summarises the topics and methods linking the bonus-test items to the exam questions. For Q2, the relevant link depended on the student's assigned version.

Table 3: Mathematical topics and methods linking bonus-test items to final-exam questions. Question references follow the English-language scoring file.

Bonus-test item	Exam item	Topic or method
Test 1, Q2, B	Q6	Solving irrational inequalities involving square roots, with attention to domain and sign restrictions.
Test 3, Q2, A	Q10A	Evaluating limits at positive infinity involving square roots and polynomial growth.
Test 4, Q2, A	Q7B	Differentiating functions with a variable base and exponent, using logarithmic differentiation or an equivalent exponential form.
Test 4, Q2, B	Q7C	Applying the chain rule to composite functions with an outer square root.
Test 5, Q1	Q11A.ii	Determining vertical and oblique asymptotes of rational functions with a cubic numerator and quadratic denominator.
Test 5, Q2, A	Q11B.ii	Determining intervals of convexity and concavity from the sign of a given second derivative.
Test 5, Q2, B	Q11B.i	Determining increasing and decreasing intervals from the sign of a given first derivative.
Test 6, Q2, A	Q9A	Evaluating indefinite integrals by substitution.
Test 6, Q2, B	Q9B	Evaluating indefinite integrals by integration by parts.

3 Results

We first report participation and the evaluation of the feedback, followed by comparisons on linked exam questions and the descriptive results for overall exam performance.

3.1 Sample characteristics

Table 4 summarises the sample composition and participation patterns in the two cohorts. Students attended approximately four of the six bonus tests on average, with similar attendance across cohorts. Median final-exam performance was somewhat higher in the English cohort, which is likely a result of the fact that both courses were graded by different people, albeit with the same grading key. Nevertheless, the overall participation structure was comparable across the two language groups.

Table 4: Sample characteristics by cohort.

	ENG	**NL**
Students with bonus-test records	343	236
Version A / Version B	159 / 185	121 / 115
Numerical total exam scores	331	236
Mean tests attended (SD)	4.17 (1.79)	3.98 (1.81)
Exam Total/20: median (IQR)	7.65 (3.91–11.33)	6.38 (2.71–10.96)
Crossover students (n)	199	147
$\rightarrow$ T1-treated / T3-treated	95 / 104	75 / 72

3.2 Evaluation of feedback quality

Figure 1 shows the distribution of ratings for each criterion and assignment. We find the highest ratings for the practice exercises (A6): 300 of the 314 ratings (95.5%) were strong, with one weak rating and no deficient ratings. For mathematical accuracy (A1), 195 ratings (62.1%) were strong, while 26 (8.3%) were deficient. Thus, although the practice exercises were generally rated highly, the feedback did not always correctly evaluate the student's answer.

The proportion of strong ratings was lower for diagnostic specificity (A2) and clarity and tone (A5), at 129/314 (41.1%) and 105/314 (33.4%), respectively. Most of the remaining ratings were adequate. However, clarity and tone also received 61 weak ratings and one deficient rating, together accounting for 19.7% of ratings. The comments give some indication of what reviewers found lacking. For example, feedback sometimes credited reasoning that was absent from the student's answer, missed an error, or suggested unnecessary improvements to a correct solution.

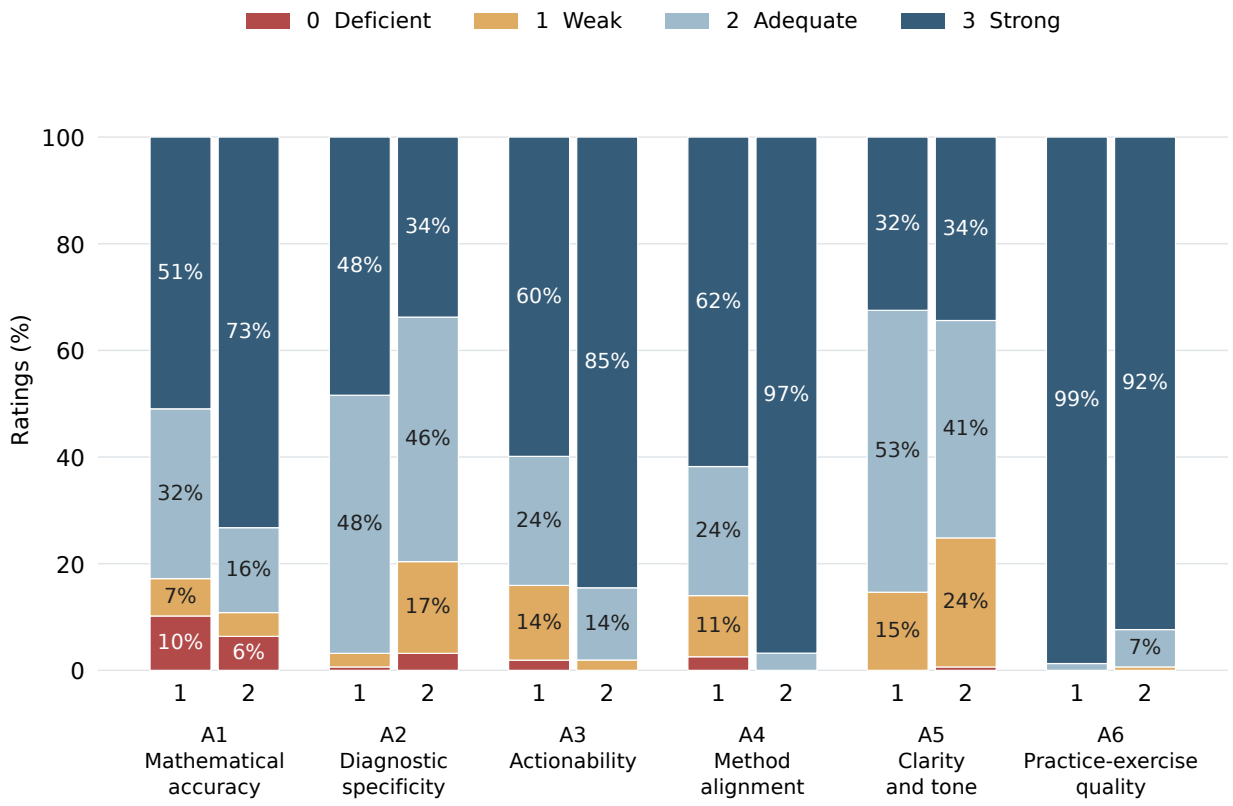


Figure 1: Distribution of scores (0–3) by criterion. Each of the 157 messages was rated independently by two of the six reviewers; bars 1 and 2 show the first and second rating of each message (157 ratings per bar). The y-axis gives the percentage of ratings per score.

Referees differed the most for actionability and method alignment. For actionability, the first rating was strong for 131/155 messages (84.5%), compared with 94/157 (59.9%) for the second rating. For method alignment, these counts were 150/155 (96.8%) and 97/157 (61.8%), respectively. Moreover, we also observed lower actionability ratings for students who scored 0 or 1 out of 10. Within this group, the first rater assigned the highest rating to 16/28 retained messages (57.1%), compared with 115/127 (90.6%) for grades above 1/10. For the second rater, the corresponding proportions were 14/30 (46.7%) and 80/127 (63.0%). This difference may reflect that students who scored 0 or 1 on a question generally wrote very little, resulting in feedback that was more encouraging than specific.

Alongside the feedback rubrics, we also included compliance checks. The model generally included the requested components. Reviewers recorded a strength in 933/942 reviews (99.0%), an improvement or extension in 921/942 (97.8%), concrete next steps in 902/942 (95.8%), and two practice exercises in 932/942 (98.9%).

3.3 Crossover analysis on linked exam questions

Table 5 reports the primary crossover comparisons for the two linked exam questions derived from tests 1 and 3. The estimated differences are mixed in direction across cohorts and questions, and at the level of individual linked exam questions the randomized evidence does not indicate a stable feedback effect.

This pattern is not entirely surprising in light of the broader feedback literature. Feedback effects are known to be sensitive to timing, specificity, and uptake [21, 3], and a single feedback exposure tied to one topic may simply be too light of an intervention to produce a reliable gain on one specific exam question asked months later.

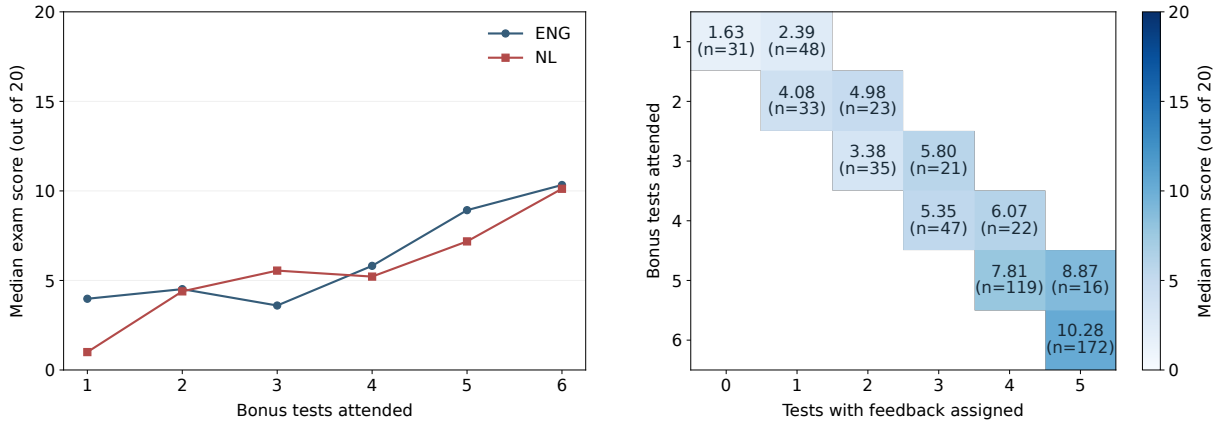
3.4 Descriptive patterns in attendance, feedback exposure, and exam performance

When considering overall exam performance, we observe a positive association with attendance in both cohorts (Figure 2a; Spearman $\rho_{\text{ENG}} = 0.49$, $\rho_{\text{NL}} = 0.54$, both $p < 0.001$). Students who attended more bonus tests tended

Table 5: Linked exam-question scores (out of 10) among students attending both crossover tests. Tests are two-sided Mann–Whitney comparisons, with Holm adjustment across the four rows.

		Feedback		Control			
Cohort	Link	n	Median	n	Median	p	p_{Holm}
ENG	T1→Q6	52	3.50	52	5.00	0.148	0.443
ENG	T3→Q10A	51	2.00	41	2.00	0.670	0.670
NL	T1→Q6	40	4.25	31	2.00	0.026	0.104
NL	T3→Q10A	40	1.75	33	5.00	0.230	0.461

to obtain higher exam scores. However, they also had more opportunities to practise and may have differed in their engagement with the course.



(a) (b)

Figure 2: **(a)** Median exam score out of 20, before bonus points, by the number of bonus tests attended, showing an attendance effect in both cohorts. **(b)** Median exam score, before bonus points, as a joint function of bonus-test attendance and cumulative exposure to AI-generated feedback. Within all attendance levels, greater feedback exposure is associated with higher median exam performance.

Furthermore, figure 2b compares the median exam performance within attendance levels as a joint function of attendance and cumulative feedback exposure. Within all attendance strata, students with more feedback exposure also achieved higher median exam scores. A simple descriptive summary of these within-attendance contrasts suggests a conditional mean advantage of about 0.5 grade points and a conditional median advantage of about 0.9 grade points for the more-exposed groups. Because feedback exposure outside the crossover tests was driven by which bonus tests students attended, these within-attendance contrasts cannot be cleanly attributed to feedback alone. Even so, the pattern is consistent with the possibility that repeated AI-generated feedback could support learning gains across the semester, rather than only performance on a single later exam question.

4 Discussion

In this study, we examined both the quality of AI-generated feedback on handwritten mathematics and its contribution to subsequent exam performance. The clearest positive result concerned the practice exercises: 95.5% of ratings placed them in the highest category for relevance, correctness, and calibration to the student's needs. The feedback itself received more mixed assessments, particularly for diagnostic specificity and clarity. These findings show that the model could provide suitable additional practice even when its explanation of the student's answer needed improvement. Informal comments collected from students attending the final lecture were consistent with these ratings. Several students reported that the feedback helped them identify what to work on or recognise what they had done well. Others described explanations as unclear or superficial, pointed out errors, or objected to the use of AI. Some also expressed disappointment when feedback was absent. Although these comments come from a self-selected group and cannot be taken as representative of either cohort, they point to the same weaknesses as the reviewer ratings.

The lower ratings for diagnostic specificity and clarity echo previous trepidations with respect to AI feedback [22]. Reviewers described feedback that attributed reasoning to students which they had not used, proposed unnecessary corrections, or gave overly extensive explanations. These observations are relevant to the prompt, which required the model to identify both strengths and areas for improvement. When a student has answered correctly, this instruction may encourage the model to look for a weakness where little needs to be changed. For students who scored 0 or 1, the problem can be different. As R1 explained, a correct account of how to solve the entire exercise may still leave the student unsure how to proceed. We believe this distinction between correctness and usefulness deserves particular attention when generating feedback for students who struggle with the material. The practice exercises in our setup included only final answers, as the prompt explicitly excluded hints and intermediate steps. Providing a worked example or a first step could help these students begin practising. These observations suggest that the guidance should depend on how much progress the student has made. Students who have barely started may need help with a first step, while those with a correct answer may benefit more from an extension than from suggestions for improvement. Whether adapting the prompt in this way improves the feedback and subsequent performance remains to be tested.

The positive relation between attendance, feedback exposure, and overall exam scores may suggest a direct benefit from receiving AI feedback. However, students who attended more tests also practised more, while differences in exposure at a given attendance level depended on which tests they attended. These observations therefore cannot resolve the inconclusive results of the randomised comparisons. We would need a design that varies repeated feedback exposure more extensively to assess whether its effects accumulate across the semester.

Several limitations should be considered when interpreting our findings. Each feedback rubric was graded by only two referees, who mostly aligned, but important differences existed, especially in the specificity and actionability of the feedback. Within the exam analysis, the comparison groups were relatively small and restricted to students who attended both crossover tests and had an observed outcome. Conditioning on attendance may affect the comparability created by random assignment, and feedback on one topic could also directly affect performance on another subject; for instance, feedback on limits could impact students' ability to recognise oblique asymptotes. Students were informed that the generated feedback was AI-generated, which in turn could have lowered their confidence or reliance on the provided feedback [23, 24]; some informal student comments indicated objections to the use of AI.

Our findings show that GPT-5.1 can generate practice exercises from handwritten mathematics answers that reviewers judge relevant, correct, and suited to the student's level. They also identify where the accompanying feedback needs improvement, particularly in diagnosing mistakes and explaining how to proceed. The crossover comparisons did not establish statistically significant improvements in later exam performance. The positive associations with repeated feedback exposure warrant further investigation in a design that further separates feedback from attendance and practice. Recording whether students read the feedback and attempt the exercises would also help explain when the assessed quality of the material translates into learning.

Declaration of Conflicting Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research was supported by funding from the Vrije Universiteit Brussel Research Council (VUB-OZR) and the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program. Andres Algaba acknowledges support from the Francqui Foundation (Belgium) through a Francqui Start-Up Grant and a fellowship from the Research Foundation Flanders (FWO) under Grant No.1286924N. Vincent Ginis acknowledges support from Research Foundation Flanders under Grant No.G032822N and G0K9322N.

Data Availability

The computer data has been made available online at: https://figshare.com/s/9208da7e9bc80b2cd01f.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT-5.6 to improve the manuscript's readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Ethics approval and consent to participate

This study retrospectively analyses pseudonymised educational assessment data generated through routine teaching practice from an existing formative assessment structure embedded in the course curriculum. Under institutional guidance on the reuse of educational/secondary data, no separate ethical review was required.

References

[1] John Hattie and Helen Timperley. The power of feedback. *Review of Educational Research*, 77(1):81–112, 2007.

[2] Paul Black and Dylan Wiliam. Assessment and classroom learning. *Assessment in Education: principles, policy & practice*, 5(1):7–74, 1998.

[3] Valerie J Shute. Focus on formative feedback. *Review of educational research*, 78(1):153–189, 2008.

[4] David J Nicol and Debra Macfarlane-Dick. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education*, 31(2):199–218, 2006.

[5] Benedikt Wisniewski, Klaus Zierer, and John Hattie. The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in psychology*, 10:3087, 2020.

[6] Abelardo Pardo, Jelena Jovanovic, Shane Dawson, Dragan Gašević, and Negin Mirriahi. Using learning analytics to scale the provision of personalised feedback. *British journal of educational technology*, 50(1):128–138, 2019.

[7] Jennifer Meyer, Marlene Steinbach, Ronja Schiller, Ute Mertens, Nils-Jonathan Schaller, Andrea Horbach, Johanna Fleckenstein, Olaf Köller, Rene F Kizilcec, and Thorben Jansen. Comparing teacher and ai-generated feedback in the writing classroom: Experimental results from secondary school classrooms. In *Proceedings of the Thirteenth ACM Conference on Learning@ Scale*, pages 264–276, 2026.

[8] Wei Dai, Yi-Shan Tsai, Jionghao Lin, Ahmad Aldino, Hua Jin, Tongguang Li, Dragan Gašević, and Guanliang Chen. Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence*, 7:100299, 2024.

[9] Sami Baral, Eamon Worden, Wen-Chiang Lim, Zhuang Luo, Christopher Santorelli, Ashish Gurung, and Neil Heffernan. Automated feedback in math education: A comparative analysis of llms for open-ended responses. *arXiv preprint arXiv:2411.08910*, 2024.

[10] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

[11] Ethan Mollick and Lilach Mollick. Assigning ai: Seven approaches for students, with prompts. *arXiv preprint arXiv:2306.10052*, 2023.

[12] Marlene Steinbach, Shreya Bhandari, Jennifer Meyer, and Zachary A Pardos. When llms hallucinate: Examining the effects of erroneous feedback in math tutoring systems. In *Proceedings of the Twelfth ACM Conference on Learning@ Scale*, pages 139–150, 2025.

[13] Danielle R Thomas, Conrad Borchers, Shambhavi Bhushan, Erin Gatz, Shivang Gupta, and Kenneth R Koedinger. Llm-generated feedback supports learning if learners choose to use it. In *European Conference on Technology Enhanced Learning*, pages 489–503. Springer, 2025.

[14] David Carless and David Boud. The development of student feedback literacy: Enabling uptake of feedback. *Assessment & evaluation in higher education*, 43(8):1315–1325, 2018.

[15] Oikantik Nath, Hanani Bathina, Mohammed Safi Ur Rahman Khan, and Mitesh M Khapra. Can vision-language models evaluate handwritten math? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14784–14814, 2025.

[16] Tianyi Liu, Julia Chatain, Laura Kobel-Keller, Gerd Kortemeyer, Thomas Willwacher, and Mrinmaya Sachan. Ai-assisted automated short answer grading of handwritten university-level mathematics exam. *Teaching Mathematics and Its Applications*, 45(1):84–105, 2026.

[17] Gerd Kortemeyer, Julian Nöhl, and Daria Onishchuk. Grading assistance for a handwritten thermodynamics exam using artificial intelligence: An exploratory study. *Physical Review Physics Education Research*, 20(2):020144, 2024.

[18] Dingjie Song, Tianlong Xu, Yi-Fan Zhang, Hang Li, Zhiling Yan, Xing Fan, Haoyang Li, Lichao Sun, and Qingsong Wen. Can mllms read students' minds? unpacking multimodal error analysis in handwritten math. In *International Conference on Artificial Intelligence in Education*, pages 110–124. Springer, 2026.

[19] Weiyu Sun, Liangliang Chen, Yongnuo Cai, Huiru Xie, Yi Zeng, and Ying Zhang. Edu-circuit-hw: Evaluating multimodal large language models on real-world university-level stem student handwritten solutions. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 15281–15314, 2026.

[20] Arne Vanhoyweghen, Vincent Holst, Melika Mobini, Lukas Van de Voorde, Tibo Vanleke, Bert Verbruggen, Brecht Verbeken, Andres Algaba, Sam Verboven, Marie-Anne Guerry, et al. Human-in-the-loop llm grading for handwritten mathematics assessments. *arXiv preprint arXiv:2603.13083*, 2026.

[21] Avraham N Kluger and Angelo DeNisi. The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin*, 119(2):254, 1996.

[22] Niklas Scholz, Manh Hung Nguyen, Adish Singla, and Tomohiro Nagashima. Partnering with ai: a pedagogical feedback system for llm integration into programming education. In *European Conference on Technology Enhanced Learning*, pages 243–248. Springer, 2025.

[23] Tanya Nazaretsky, Paola Mejia-Domenzain, Vinitra Swamy, Jibril Frej, and Tanja Käser. Who gives feedback matters: Student biases towards human and ai-generated formative feedback. *Journal of Computer Assisted Learning*, 42(1):e70153, 2026.

[24] Michael Henderson, Margaret Bearman, Jennifer Chung, Tim Fawns, Simon Buckingham Shum, Kelly E Matthews, and Jimena de Mello Heredia. Comparing generative ai and teacher feedback: student perceptions of usefulness and trustworthiness. *Assessment & Evaluation in Higher Education*, 51(5):863–878, 2026.