Scaling, Demonstrating and Personalizing: the Benefits of LLMs in Educational Contexts

Eamon Worden
Worcester Polytechnic Institute
100 Institute Road
Worcester, Massachusetts
elworden@wpi.edu

Neil Heffernan
Worcester Polytechnic Institute
100 Institute Road
Worcester, Massachusetts
nth@wpi.edu

ABSTRACT

The benefits of explanations and feedback on learning are well established. However, delivering effective feedback at scale remains a substantial challenge due to various logistical and cost-related constraints. This paper reports on my randomized controlled trials (RCTs) evaluating the effectiveness of AI-generated content for middle school mathematics. My first studies include RCTs comparing AI-generated explanations and feedback with business-as-usual (no support), as well as AI-generated content with teacher-written content. Both used a similar pipeline to generate feedback, going through an iterative prompt engineering process with experienced educators, followed by LLM-as-a-judge and a limited human review. One RCT found that students receiving AI-generated feedback were 16% more likely to correct their current answer and 7% more likely to succeed on the subsequent problem. I then moved to live content generation, with LLMs providing feedback to short-answer math problems. I used a cross-over design to find the impact of using GPT-4o to provide short, long, or affective feedback, and also analyzed a model fine-tuned on teacher-written feedback. I found affective feedback was the only feedback that was consistently helpful in both the short- and medium-term transfer learning for students. This informs my future research, aiming to generate and cache various styles of content using LLMs, then personalize it to students' learning characteristics. My future work builds on this by aiming to personalize the style, length, and tone of feedback to students at scale based on both student and misconceptions features at scale, which no prior work has managed to achieve.

Keywords

Large Language Models, AI-generated feedback, Randomized Controlled Trial, Mathematics Education, Learning at Scale, Automated Content Generation

1. INTRODUCTION

Eamon Worden, and Neil T. Heffernan. Scaling, Demonstrating and Personalizing: the Benefits of LLMs in Educational Contexts. In Anthony Botelho, Maria Mercedes T. Rodrigo, Adish Singla, Hiroaki Ogata, Hyojeong So, and Young Hoan Cho (eds.) Proceedings of the 19th International Conference on Educational Data Mining, Seoul, Republic of Korea, June, 2026, pp. 884–888. International Educational Data Mining Society (2026).

https://doi.org/10.5281/zenodo.21039954

Large Language Models (LLMs) are increasingly deployed in educational contexts for tasks such as chatbots [21], feedback generation [24], and question generation [13]. However, equally important to creating these systems is evaluating their effectiveness and fairness. LLM-generated content needs to be grounded in pedagogy and delivered effectively, otherwise it risks decreasing critical thinking by removing the opportunity for productive struggle [4]. However, if implemented effectively, it can personalize content such as on-demand support, feedback, and more.

Given the risks and potential of LLMs, my focus has been on using established pedagogical techniques and advanced prompt engineering to generate content at scale before empirically demonstrating the effectiveness. Working with the ASSISTments platform [8], I have conducted RCTs using AI-generated content with tens of thousands of students in grades 6-8, Illustrative Math. Using techniques such as RAG [12], few-shot prompt [2], LLM-as-a-judge [27], and chain-of-thought [22], I have generated on-demand supports to aid students, which had previously been an expensive and time-consuming effort requiring the recruitment of teachers, who themselves are limited in time. This has informed my research into how LLMs can be used to scale content in digital learning platforms (DLPs) at an affordable and unprecedented rate. I am now working to generate various styles of content, which, with the correct bandits algorithm, aims to provide personalized content on a new level.

My works so far has scaled on-demand explanations and feedback. Explanations are a form of instructional support that provides all the steps necessary to walk through a solution to a problem, often presented as step-by-step solutions. They are well-established as effective tools for student learning [20, 17]. Notably, explanations are more effective for students with less expertise than their higher-skilled peers as described by the "expertise reversal effect" [9]. One major theory for why explanations are helpful for all students, but particularly helpful for low-performing students, is cognitive load theory [18], which posits that explanations reduce the mental effort required to process information, thereby allowing students to focus their cognitive resources on acquiring and remembering mathematical knowledge.

The other content I've been working to generate, feedback, is among the most effective instructional interventions documented in the learning sciences [6, 7]. Meta-analyses consistently show that well-designed corrective feedback supports

error detection, improves problem-solving strategies, and promotes self-regulation, particularly for lower-achieving students [23, 10]. Despite this evidence, delivering personalized feedback at scale remains a persistent challenge in digital learning platforms (DLPs). Prior approaches using crowdsourcing have shown promise but suffer from quality variability [15, 5]. Recent advances in large language models (LLMs) present a new opportunity. LLMs can generate coherent, context-sensitive feedback at low marginal cost. However, whether AI-generated feedback improves learning outcomes over no feedback—and for which students—remains an open empirical question with limited large-scale evidence in authentic classroom settings.

This informs my three main research questions for my last studies. **Research Plan 1** investigates "Under what conditions does feedback style moderate learning outcomes in middle school mathematics?" This aims to analyze the impact of various feedback styles inspired by Hattie & Timperley [6] and which works best by looking at student characteristics and classifying wrong answers. With sufficient data, this will enable **Research Plan 2**, which asks "Can a bandit or similar algorithm effectively personalize feedback style for students in a live DLP, and what student features are necessary to do so?" This investigates personalizing at scale within ASSISTments, which requires understanding students and constantly measuring learners' characteristics and updating latent student traits [19]. Lastly, moving to a more dynamic setting, **Research Plan 3** asks "Does personalized LLM-generated feedback on open-ended responses improve learning outcomes compared to no feedback, and does it do so sustainably over a full academic year?" Open-ended problems are very important as they promote deeper and more critical thinking, but scaling personalized feedback for tens of thousands of students can be expensive.

2. PUBLISHED WORKS

During my PhD, I have published several works on my various studies, aiming to generate and evaluate the effectiveness of content. I have first-author submissions to both Learning at Scale [25] and AIED [24] and an under-review submission at Learning at Scale this year.

Studies 1 & 2: Scaling Static Explanations and Feedback [25, 26]. In these works, I have empirically evaluated the effectiveness of LLM-generated explanations and static feedback messages through RCTs in ASSISTments. For these RCTs, I collaborated with former teachers, curriculum designers at ASSISTments, as well as other PhD students within my lab, to engineer effective prompts for explanations and feedback, before caching them in the ASSISTments database and engineering the middleware needed to randomly select the LLM-generated content, which was displayed to students as shown in figures 1 and 2. These studies have been run *en vivo* in ASSISTments by randomly selecting to deliver nothing, LLM-generated content, or teacher-authored content, and comparing their effectiveness, with tens of thousands of students. To analyze our results, I used multiple mixed-effects linear and logistic regression models with various outcome metrics. Our results were consistent across the studies: LLM-generated content was better than nothing, and there was no reliable difference between LLM-generated content and teacher-authored content. Also consistent with the expertise reversal effect, I found these supports were more effective for lower-knowledge students. Given the incredibly cheap cost of scaling this content with LLMs (about 5 cents per content), I feel this confirms that LLMs can aid students by scaling content without harming student performance.

Study 3: Scaling Feedback Generation in a Live Setting [24]. While scaling static content is valuable as done in **studies 1 & 2**, providing dynamic feedback to short answers is equally valuable. For **study 3** I worked with ASSISTments to engineer a student interface that could deliver feedback when they submit short-answer responses to open-ended problems. Prior to our work, of the 50 million instances of students submitting a short-answer, non-computer-gradeable response in ASSISTments, only 1 million (2%) actually received feedback, and often 2-3 days later. As such, I am laying the groundwork for ASSISTments to provide short-answer feedback at scale.

In this study, along with collaborators at WPI and the ASSISTments foundation, I recruited 11 teachers with a total of 322 students, and employed a crossover design to evaluate the effectiveness of LLM-generated feedback of various styles, short feedback, long feedback, affective (self-level) feedback, or feedback from a model which was fine-tuned on teacher-written feedback in my prior work [1]. Our findings suggested that only affective feedback was useful to students in both the next-problem and medium transfer range, whereas other feedback styles had minimal help, or harmed student learning due to giving away excessive information and removing the opportunity for productive struggle. This was a surprising result, given that affective supports are generally thought to be ignored by students in DLPs. However, it also inspires my future work on providing personalized feedback to all students, so that students who benefit more from a specific feedback style receive that personalized feedback.

3. ONGOING WORK

I intend for this to be the main component of my dissertation. Having shown LLMs can generate effective content, I aim to generate personalized content at scale. This involves three studies, the first of which is an RCT aiming to explain when different styles of feedback are useful, and the second, which aims to personalize in a live setting using a bandit's algorithm [16], and the last, which repeats this for open-response questions.

Research Plan 1 & 2. This study, which has already started, aims to build upon my prior work generating wrong answer feedback for students with LLMs. I found LLM-generated feedback was effective; however, inspired by Hattie & Timperley's styles of feedback [6], I have generated 6 styles of feedback: Feedback; Feed-up; Feedback & Feed-up; Feedback & Self-regulation; Feedback, Feed-up & Self-regulation; and Self-level feedback. These feedback styles differ in the type of information provided, such as whether a small hint is given, whether self-regulatory behavior is encouraged, or whether the feedback is affective. The goal is to determine which type of feedback works best, and answer "when?" and "why?".

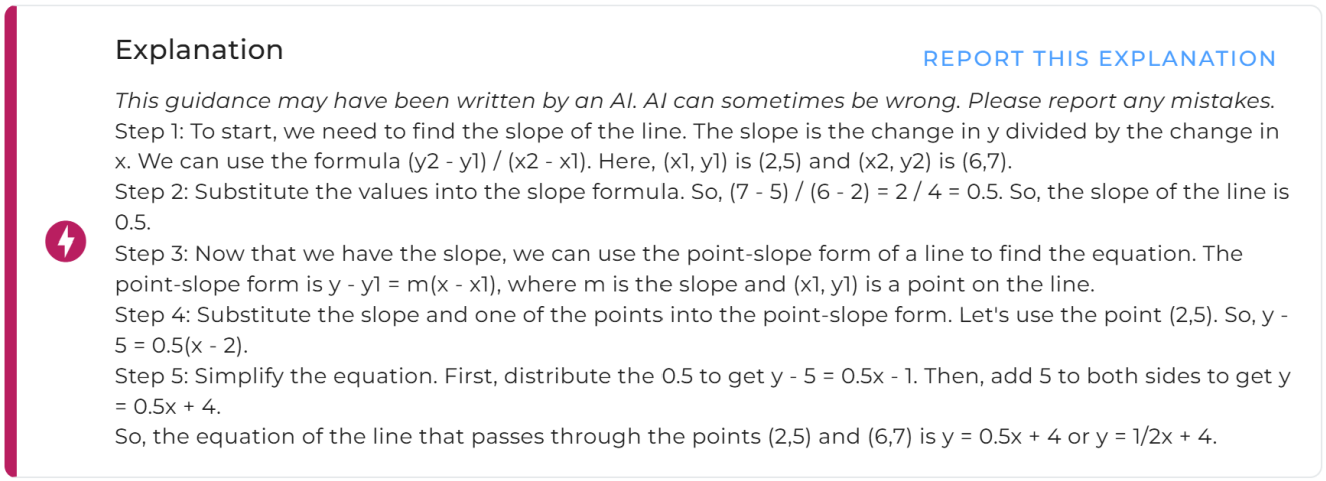

Figure 1: An example explanation with the report button and warning message.

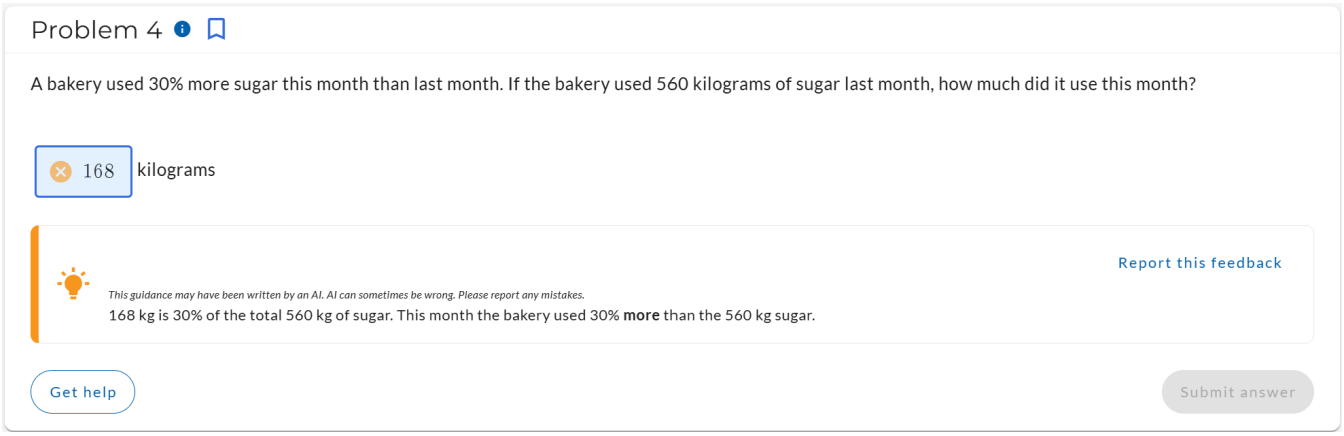

Figure 2: An example feedback message as it appears in ASSISTments.

The "when?" refers to different types of wrong answers and student characteristics. For instance, for the problem **"A bakery used 30% more sugar this month than last month. The bakery used 560kg of sugar last month. How much did it use this month?"**, I hypothesize that the best style of feedback for a student who said "168" (forgot to add it to the original amount) will be very different from the best type of feedback for a student who said "590" (does not understand percentages). For the student who answered "168", I hypothesize self-regulatory feedback, which emphasizes that reading the problem carefully will best help the student, whereas for the student who answered "590", a combination of feedback and feed-up will best guide them towards a better understanding of percentages. The same can be true for student characteristics. Reinforcement learning has shown that incorperating student features into a bandit model can improve learning outcomes, which I intend to do once I have collected sufficient data [11].

As mentioned, my first research project comparing the styles of feedback is currently running, however, I could use guidance on the analysis plan. I hypothesize that the key to understanding which styles of feedback are helpful, and when, is categorizing styles of misconceptions into different types. Categories such as "forgot last step", "conceptual misunderstanding", "procedural error" or more may reveal deeper insights into when each style of feedback may be more or less helpful. Determining which categories to include and how to classify them is a challenge I am facing. However, there are other approaches I have considered. Using Bayesian Knowledge Tracing or a similar algorithm to classify a student's knowledge of various topics, or even their individual slip rate, appears feasible based on prior research [14], which could inform both analysis and bandits models, which is my second project relating to this work. However, defining the best methodological approach and analysis plan has been a difficult task for me, and I would appreciate the insights of experienced researchers in the field to improve and solidify this methodology.

Research Plan 3. Building on my feedback research, my other work involves generating dynamic feedback using LLMs for open-response problems. These are problems that require students to type 1-3 sentences to answer a question,

as opposed to fill-in problems, which require a number or short expression. This makes caching feedback infeasible, and before LLMs, attempts to scale feedback proved expensive and difficult [3]. However, in a live, undefined setting, I have more options for how I might generate feedback. I may generate feedback, feed-up, self-regulation, and affective feedback as in **Study 1**. However, by being able to see, rather than just infer, a student's work, I can go beyond basic feedback. I may provide targeted interventions, such as providing sentence starters or clarifying specific portions of answers. I have worked with ASSISTments and we intend to enable this feature for 50% of teachers beginning in the fall of 2026 and lasting for the complete 2026-2027 school year. This will include around 500 teachers with tens of thousands of students in the study.

This work will include building on my prior work (**Study 3**)by opening new opportunities for longitudinal analysis. This will analyze the effect of repeated interactions over the course of the school year. I will analyze whether there are diminishing returns, increased disengagement, or increased unproductive behavior (such as gaming) as a result of adding LLM-feedback to ASSISTments. Alternatively, I will be able to investigate whether there is a cumulative advantage or other sustained learning improvements for students who receive feedback on their open-ended questions. Last, I hope to gather sufficient data to fine-tune a custom LLM to generate feedback for students in ASSISTments, as I have done in prior work [1].

4. CONCLUSION

Personalizing feedback for students at scale remains a consequential and open problem in educational data mining. While the learning sciences have long established that feedback is among the most effective instructional interventions, we have lacked both the tools to deliver it affordably at scale and the empirical understanding of when different feedback styles are most effective. My completed work addresses this using three large-scale RCTs across 250,000 observations and 50,000 students, which demonstrate that LLM-generated content is effective and cost-effective. My remaining work addresses a harder yet more impactful problem. Categorizing student misconceptions is a prerequisite for understanding the causal structure of feedback effectiveness. If we cannot distinguish a student who made a procedural slip from one with a conceptual misunderstanding, we cannot know which feedback style to deliver, nor why it worked when it did. This represents a contribution to the EDM and learning science community beyond ASSISTments: a framework for misconception-aware feedback personalization that can inform how any DLP reasons about learner state. Combined with a bandit-based personalization system and longitudinal data from tens of thousands of students, this work aims to produce a large-scale empirical study of personalized feedback in middle school mathematics.

5. REFERENCES

[1] S. Baral, E. Worden, W.-C. Lim, Z. Luo, C. Santorelli, A. Gurung, and N. Heffernan. Automated feedback in math education: A comparative analysis of llms for open-ended responses. *arXiv preprint arXiv:2411.08910*, 2024.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] J. A. Erickson, A. F. Botelho, S. McAteer, A. Varatharaj, and N. T. Heffernan. The automated grading of student open responses in mathematics. In *Proceedings of the tenth international conference on learning analytics & knowledge*, pages 615–624, 2020.

[4] G. P. Georgiou. Chatgpt produces more" lazy" thinkers: Evidence of cognitive engagement decline. *arXiv preprint arXiv:2507.00181*, 2025.

[5] A. Gurung, S. Baral, M. P. Lee, A. C. Sales, A. Haim, K. P. Vanacore, A. A. McReynolds, H. Kreisberg, C. Heffernan, and N. T. Heffernan. How common are common wrong answers? crowdsourcing remediation at scale. In *Proceedings of the Tenth ACM Conference on Learning@ Scale*, pages 70–80, 2023.

[6] J. Hattie and H. Timperley. The power of feedback. *Review of educational research*, 77(1):81–112, 2007.

[7] K. Haughney, S. Wakeman, and L. Hart. Quality of feedback in higher education: A review of literature. *Education Sciences*, 10(3):60, 2020.

[8] N. T. Heffernan and C. L. Heffernan. The assistments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International journal of artificial intelligence in education*, 24(4):470–497, 2014.

[9] S. Kalyuga. The expertise reversal effect. In *Managing cognitive load in adaptive multimedia learning*, pages 58–80. IGI Global, 2009.

[10] A. N. Kluger and A. 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.

[11] M. P. Lee and N. T. Heffernan. Improving student support personalization with historical data and theoretically informed feature choice. In *International Conference on Artificial Intelligence in Education*, pages 421–426. Springer, 2025.

[12] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[13] S. S. Mucciaccia, T. M. Paixão, F. W. Mutz, C. S. Badue, A. F. de Souza, and T. Oliveira-Santos. Automatic multiple-choice question generation and evaluation systems based on llm: A study case with university resolutions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2246–2260, 2025.

[14] Z. A. Pardos and N. T. Heffernan. Modeling individualization in a bayesian networks implementation of knowledge tracing. In *International conference on user modeling, adaptation, and personalization*, pages 255–266. Springer, 2010.

[15] T. Patikorn and N. T. Heffernan. Effectiveness of crowd-sourcing on-demand assistance from teachers in

online learning platforms. In *Proceedings of the Seventh ACM Conference on Learning@ Scale*, pages 115–124, 2020.

[16] E. Prihar, A. Haim, A. Sales, and N. Heffernan. Automatic interpretable personalized learning. In *Proceedings of the Ninth ACM Conference on Learning@ Scale*, pages 1–11, 2022.

[17] A. Rourke and J. Sweller. The worked-example effect using ill-defined problems: Learning to recognise designers' styles. *Learning and Instruction*, 19(2):185–199, 2009.

[18] J. Sweller. Cognitive load theory, 2011.

[19] L. Tetzlaff, F. Schmiedek, and G. Brod. Developing personalized education: A dynamic framework. *Educational psychology review*, 33(3):863–882, 2021.

[20] P. W. Van Gerven, F. G. Paas, J. J. Van Merriënboer, and H. G. Schmidt. Cognitive load theory and aging: Effects of worked examples on training efficiency. *Learning and instruction*, 12(1):87–105, 2002.

[21] R. E. Wang, A. T. Ribeiro, C. D. Robinson, S. Loeb, and D. Demszky. Tutor copilot: A human-ai approach for scaling real-time expertise. *arXiv preprint arXiv:2410.03017*, 2024.

[22] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[23] B. Wisniewski, K. Zierer, and J. Hattie. The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in psychology*, 10:487662, 2020.

[24] E. Worden, M. Lee, A. Siedahmed, A. Sales, J. Zhang, R. Shraga, and N. Heffernan. Short, long, or affective: Evaluating llm-generated feedback styles for student learning. In *International Conference on Artificial Intelligence in Education*, 2026.

[25] E. Worden, K. Vanacore, A. Haim, and N. Heffernan. Scaling effective ai-generated explanations for middle school mathematics in online learning platforms. In *Proceedings of the Twelfth ACM Conference on Learning@ Scale*, pages 40–49, 2025.

[26] E. e. a. Worden. Under review. In *Conference on Learning@Scale*, 2026.

[27] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.