HaLLMos: Constitutional AI for Feedback on Mathematical Proofs

Sarah Sword
Education Development Center, Inc.

Vince Vatter
University of Florida

Jay Pantone
Marquette University

Ryota Matsuura
St. Olaf College

Learning to construct and critique proofs is central to advanced mathematics but challenging for students. This paper reports on HaLLMos*, an AI model that helps students learn to write mathematical proofs. HaLLMos uses a constitutional approach along with expert-curated feedback exemplars to generate responses to students' proof efforts. Guided by Schoenfeld's Goals–Orientations–Resources framework and perspectives on proof as discourse, we investigate the tool's alignment with its constitutional principles and student perspectives on HaLLMos' value. Preliminary data from interviews and system logs suggest that HaLLMos supports students' independence while also preparing them to participate in mathematical conversations about proof. Preliminary findings suggest HaLLMos can expand resources for students and faculty while raising new questions about how AI-generated feedback can support students' mathematical interactions and their sense of belonging in mathematics.*

Keywords: Mathematical Proof, Feedback, Constitutional AI

Learning to construct proofs is one of the most significant challenges students face in the transition to postsecondary mathematics (Selden, 2012; Moore, 1994; Clark & Lovric, 2009). Proofs are not only logical structures, but also discourse practices, shaped by the norms of justification, explanation, and persuasion defined by the mathematical community (Stylianides, 2007, 2018; Weber, 2014). Many undergraduates struggle with both the technical and discursive demands of proof (e.g., Selden & Selden, 2008; Dreyfus, 1999) and how to participate in mathematical discourse (Sword et al., in press; Herbst & Chazan, 2010).

Mathematics and mathematics education faculty who seek to support students in learning to write proofs face the ongoing challenge of providing personalized (Pedemonte, 2018), immediate (Bokhove & Drijvers, 2012), iterative (Barker & Pinard, 2014), and well-framed (Alvidrez, Louie, & Tchoshanov, 2022) feedback on students' mathematical proofs. There remains a gap between what research suggests students need and what professors can provide, given the constraints of time and resources (Stylianides & Stylianides, 2018), a gap that widens where faculty are under-resourced, and for students who may be coming to college with less experience with proof methods (Fyfe & Rittle-Johnson, 2017).

To close this gap, HaLLMos, an AI tool, takes a *Constitutional AI* approach (Askell et al., 2021; Bai et al., 2022). To build HaLLMos, we convened a group of experts to create a *constitution* guiding the development and training of a Large Language Model specifically trained to give students feedback in line with constitutional principles. This tool is not intended to replace faculty as teachers and mentors but can offload some of the work of feedback on proofs so that faculty and students can focus on human aspects of mathematics (Jones & Herbst, 2012; Selden & Selden, 2008). HaLLMos.com is a prototype website for students to receive feedback on their proofs and a setting for research on how HaLLMos serves students and faculty.

Drawing on Schoenfeld's Goals–Orientations–Resources framework and perspectives on proof as discourse, we investigate the following questions: How does feedback from HaLLMos align with its constitutional principles? How do students use HaLLMos, and how do they perceive its value in building their capacity to read, critique, and construct proofs? How does use of the tool affect students' preparation for mathematical discourse?

How HaLLMos Works

On HaLLMos.com, students enter a proof for a pre-loaded or custom-written exercise and receive feedback from the model. Unlike general-purpose chatbots, HaLLMos does not provide complete proofs or engage in open-ended conversation; it responds only to student proof attempts, encouraging revision rather than replacement.

To generate feedback, the model acts in two stages. In the first "critic" stage, the model analyzes each step of the proof for correctness, relevance, and support. This critique is not shared with the student but is fed into a second "coach" stage, in which the model, explicitly guided by constitutional principles and fine-tuned on over 350 expert-curated feedback exemplars, generates the feedback students receive. Sample constitutional principles include: the feedback should begin by identifying one or more correct elements of the proof, if any; feedback should always be connected to the content of the work, not about the person who wrote it; and feedback should not be overly prescriptive in how issues should be fixed, preferring instead to give the student enough information to fix problems on their own.

HaLLMos does not have "conversations" the way ChatGPT does. Students get iterative feedback by revising their proofs and resubmitting. Notably, unlike existing LLMs, HaLLMos does *not* provide correct proofs after giving feedback. Figure 1 illustrates this process, showing five iterations of a student proof with changes between iterations highlighted.

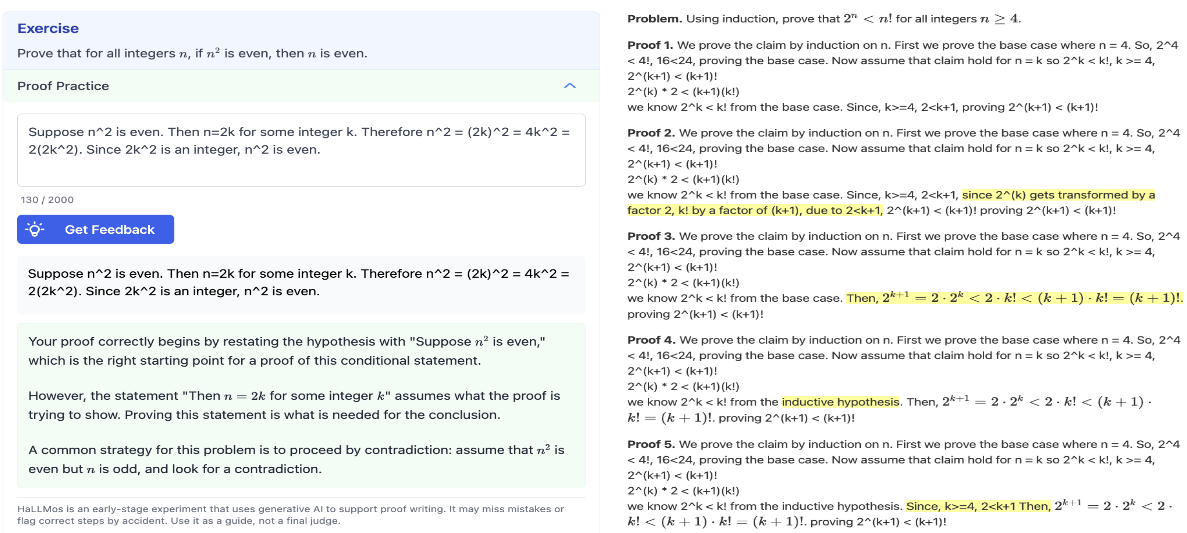

Figure 1. A screenshot from HaLLMos.com (left) and an example of a student iterating on a proof (right).

What Makes This Paper Preliminary

We are writing this paper at the end of a year of work largely dedicated to developing HaLLMos. At the time of writing, we have collected over 14,000 user-input proofs, surveyed students at four pilot sites, interviewed three self-selected students and one faculty member. In addition, we have had extensive informal conversations with pilot faculty, all of which have informed our interview protocol design. Multiple pilots of the tool are scheduled for spring 2026. We will conduct faculty interviews, as well as additional student surveys and interviews. While the current size of the study limits the conclusions we can draw, we anticipate that the emerging themes will interest the mathematics education community as the field considers how to incorporate AI tools responsibly and productively into college mathematics.

As we continue this work, we pose the following questions for discussion with the RUME community: (1) What additional frameworks might illuminate the relationship between AI-mediated and human-mediated mathematical discourse? (2) How might we better assess alignment between AI feedback and constitutional principles at scale? (3) What safeguards should guide implementation of AI feedback tools in proof-based courses?

Conceptual Frameworks

This study is framed by Schoenfeld's theory of decision-making (2010, 2011) and the discourse perspective on proof (e.g., Stylianides, 2007, 2018; Weber, 2014; Herbst & Chazan, 2010). Together, these frameworks help us conceptualize HaLLMos as a pedagogical actor whose effectiveness lies not only in correctness but also in its capacity to realign decision-making resources and support students' proof discourse.

Schoenfeld (2010, 2011) describes teachers' instructional choices arising from the interaction of goals, orientations, and resources. We extend this framework to HaLLMos itself. HaLLMos' *goals* are to provide immediate, useful feedback and reduce faculty feedback load. Its *orientations* are encoded in the constitution, which emphasizes humanizing feedback that centers students' ideas. Its *resources* include extensive training exemplars. This framework allows us to evaluate HaLLMos' alignment with its constitution through its outputs to students.

Stylianides (2007, 2018) argues that learning proof requires appropriation of the sociomathematical norms governing acceptable justification. Herbst and Chazan (2010) demonstrate that classroom discourse norms determine who can participate and how, with students often excluded less by logical incapacity than by unfamiliarity with discourse expectations. Feedback thus functions as more than correction: it apprentices students into proof discourse by modeling the language, structure, and norms of mathematical argumentation.

Methods

HaLLMos was piloted in multiple Introduction to Proofs courses at R1 universities as an optional feedback tool for homework assignments. Data sources included website logs from student submissions, interviews with three self-selected students (majoring in engineering, statistics, and mathematics), and course materials. Interviews lasted approximately 30–45 minutes, were audio-recorded, and transcribed for analysis. Website logs captured student proof submissions, model feedback, and revision patterns. Interview protocols focused on students' experiences with HaLLMos, perceptions of feedback quality, and changes in mathematical interactions.

We employed template analysis (Brooks et al., 2015) informed by the GOR and proof-as-discourse frameworks. Initial codes were drawn deductively from both frameworks, with additional emergent codes identified through iterative analysis. Analysis focused on instances where students described their interactions with HaLLMos and how those interactions shaped subsequent mathematical activity.

Preliminary Findings

Our analysis of three focal student cases reveals themes illuminating how HaLLMos aligns with its constitutional principles, how students perceive its value, and how it reshapes their preparation for mathematical discourse.

Theme 1: Alignment with Constitutional Principles

Students consistently reported that HaLLMos feedback was useful in identifying errors and suggesting improvements: “My biggest complaint was that inputting my proofs took a while... But once I got all the way in there, the feedback was, for the most part—I never got any, like, *bad* feedback.” Another student compared HaLLMos to other AI tools: “The tone voice was always good... HaLLMos definitely gave me better answers compared to ChatGPT.”

These responses reflect alignment with key constitutional principles. Students did not report feeling personally criticized, consistent with the principle that feedback should address the work rather than the person. Descriptions of HaLLMos helping students “further develop the proof” rather than simply providing answers suggest feedback avoids being overly prescriptive.

However, students also identified areas where alignment could be strengthened: “It would help diagnose those problems. But when it had just one minor thing that really didn’t need to be changed, it would still tell me to change it... if it could differentiate between the major fundamental problems and the minor... that would be helpful.” Another noted, “Sometimes I would submit my proof... and maybe submit the same thing again, and I would get different feedback.” These observations point to opportunities for refining the model and highlight where human instructor judgment remains essential.

Theme 2: Building Confidence for Independent Work

Students described HaLLMos as an accessible “extra set of eyes” available outside office hours. One noted, “It definitely helped me with my homework, and the homework is where I learned.” Another began to think “more like a logician,” explaining, “As a mathematician, we have to basically explain and justify why our solutions work. Using HaLLMos helped me reinforce that understanding.”

One student pointed to the use of HaLLMos for getting started: “the biggest difference is when starting a proof by myself can be a little challenging... even if you start with a bare bones introduction to your proof in HaLLMos, it can provide feedback on how you can further develop the proof from there.” However, students consistently cautioned against over-reliance, emphasizing the importance of struggling through proofs. As one put it, “being stuck matters.”

Theme 3: Preparation for Human Interactions and Mathematical Discourse

One student pointed to the value of using HaLLMos before talking to a professor: “I didn’t want to just say, ‘I don’t know, I’m doing *anything!* Help me!’ So I wanted to have some understanding, and HaLLMos helped me with the first steps.” Another described the value of having a “peek behind the curtain” of mathematics that HaLLMos afforded him. Together, these findings point to HaLLMos as both a decision-making resource (in the Schoenfeld framework) and an apprenticeship tool for discourse, with potential to humanize students’ experience of learning to write proofs.

Faculty Perspectives

Faculty emphasized the pedagogical value of immediacy. As one instructor noted, "Anything with feedback that comes much later is not going to motivate students. But if the feedback is immediate, that helps students a lot." Another elaborated on HaLLMos' role in the iterative process of proof-writing: "Writing proofs—there's many ways you can go. A few right ways and many more wrong. Getting real-time feedback like 'don't go that way, try this'... that's absolutely beneficial." These observations align with the literature on feedback timing (Bokhove &

Drijvers, 2012) while suggesting that faculty see HaLLMos as addressing a gap they experience but cannot fill with existing resources.

Although our faculty data collection is ongoing, early interviews and informal communications from pilot instructors echo the student themes. One email exchange, while not part of our formal research data, has shaped our thinking on what kinds of things we can learn in a study like this one. The faculty member shared that a student "came to him after class and said she is so thankful for him sharing HaLLMos with her, because it doesn't judge her ideas." This unsolicited comment points to an affective dimension we had not fully anticipated: for some students, the non-evaluative nature of AI feedback may create space to take risks that feel too costly in human interactions.

We want to emphasize two important points about what the faculty member shared: he valued this student's experience of *feeling* not judged, not because he himself was judging her, but because she saw that aspect of HaLLMos as valuable. Second, as a team, we deeply value the human relationships that spring up around mathematics. We are not sharing that quotation because we believe HaLLMos could or should replace those human interactions. Our hope for HaLLMos is that it offer students and faculty more common ground for mathematical interactions, not less, lowering the barriers that keep some students from showing up to mathematical conversations at all. When routine feedback is offloaded to a non-judgmental machine, faculty and students can spend their time together on the aspects of mathematics that most require human presence: intuition, meaning, and community.

Implications

HaLLMos can relieve some of the burden on faculty by providing timely, principled feedback on proofs. This enables faculty to reallocate time towards activities often underemphasized in transition-to-proof courses, such as conjecturing, developing mathematical intuition, and cultivating mathematical habits of mind (Lin et al., 2012; Cuoco, Goldenberg, & Mark, 1996; Thurston, 1994). Furthermore, by modeling proof language and structure, HaLLMos lowers the threshold for students to participate in office hours and class discussions. Student reports that HaLLMos helped them know "what to ask" highlight the tool's role in apprenticing learners into proof discourse.

AI tools are proliferating quickly across education. There is a real need for research that informs the field about how to build and implement those tools responsibly, and for oversight about tools' effects on students. Such research should be situated in current literature rather than treated as "AI education research" independent of existing work.

Beyond studying the use of AI tools in undergraduate mathematics, this research offers a lens on proof learning. The constitutional process required us to work with our advisors to articulate criteria for high-quality feedback that often remain tacit in instructional practice. Website logs reveal the iterative, nonlinear nature of student proof development in ways that submitted homework cannot. And student responses, especially around students' perspectives on non-judgment and freedom to make mistakes, surface affective dimensions of proof learning that may be difficult to access through direct questioning about human instructors. AI tools, when designed and studied carefully, can function not only as interventions but also as instruments that make visible aspects of mathematical learning that are otherwise hard to see.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 2336775. Any opinions, findings, and conclusions or recommendations expressed in this

material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Alvidrez, M., Louie, N., & Tchoshanov, M. (2022). From mistakes, we learn? Mathematics teachers' epistemological and positional framing of mistakes. *Journal of Mathematics Teacher Education.* https://doi.org/10.1007/s10857-022-09553-4

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T., Clark, J., … Kaplan, J. (2021). *A general language assistant as a laboratory for alignment*. arXiv. https://arxiv.org/abs/2112.00861

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., & Kaplan, J. (2022). *Constitutional AI: Harmlessness from AI feedback* (arXiv:2212.08073). arXiv. https://arxiv.org/abs/2212.08073

Barker, M., & Pinard, M. (2014). *Closing the feedback loop? Iterative feedback between tutor and student in coursework assessments*. *Assessment & Evaluation in Higher Education, 39*(8), 899–915. https://doi.org/10.1080/02602938.2013.875985

Bokhove, C., & Drijvers, P. (2012). Effects of feedback in digital environments for algebra. *International Journal of Technology in Mathematics Education, 19*(2), 61–75.

Brooks, J., McCluskey, S., Turley, E., & King, N. (2015). The utility of template analysis in qualitative psychology research. *Qualitative Research in Psychology, 12*(2), 202–222.

Clark, M., & Lovric, M. (2009). Understanding secondary–tertiary transition in mathematics. *International Journal of Mathematical Education in Science and Technology, 40*(7), 927–944.

Cuoco, A., Goldenberg, E. P., & Mark, J. (1996). Habits of mind: An organizing principle for mathematics curricula. *Journal of Mathematical Behavior, 15*(4), 375–402.

Dreyfus, T. (1999). Why Johnny can't prove. *Educational Studies in Mathematics, 38*(1-3), 85–109.

Fyfe, E. R., & Rittle-Johnson, B. (2017). Feedback both helps and hinders learning: The causal role of prior knowledge. *Journal of Educational Psychology, 109*(7), 994–1009.

Herbst, P., & Chazan, D. (2010). Exploring the practical rationality of mathematics teaching through conversations about videotaped episodes. *For the Learning of Mathematics, 30*(3), 2–8.

Jones, K., & Herbst, P. (2012). Proof, proving, and teacher-student interaction. In G. Hanna & M. de Villiers (Eds.), *Proof and Proving in Mathematics Education* (pp. 261–277). Springer.

Lin, F.-L., Yang, K. L., & Lee, Y. S. (2012). The roles of conjecturing and proving in mathematics learning. *International Journal of Science and Mathematics Education, 10*(4), 865–889.

Moore, R. C. (1994). Making the transition to formal proof. *Educational Studies in Mathematics, 27*(3), 249–266.

Pedemonte, B. (2018). Proof as a tool for learning mathematics. In T. Dreyfus et al. (Eds.), *Developing research in mathematics education* (pp. 249–261). Routledge.

Schoenfeld, A. H. (2010). *How we think: A theory of goal-oriented decision making and its educational applications.* Routledge.

Schoenfeld, A. H. (2011). Toward professional development for teachers grounded in a theory of decision making. *ZDM – The International Journal on Mathematics Education, 43*(4), 457–469.

Selden, A. (2012). Transitions and proof in mathematics education. *International Journal of Mathematical Education in Science and Technology, 43*(7), 873–884.

Selden, A., & Selden, J. (2008). Overcoming students' difficulties in learning to understand and construct proofs. In M. Carlson & C. Rasmussen (Eds.), *Making the Connection: Research and Teaching in Undergraduate Mathematics* (pp. 95–110). MAA.

Stylianides, A. J. (2007). Proof and proving in school mathematics. *Journal for Research in Mathematics Education, 38*(3), 289–321.

Stylianides, A. J. (2018). Proof and proving in mathematics education research. In T. Dreyfus et al. (Eds.), *Developing research in mathematics education* (pp. 237–248). Routledge.

Stylianides, G. J., & Stylianides, A. J. (2018). The role of instructional engineering in reducing the uncertainties of ambitious teaching. *Cognition and Instruction, 36*(4), 296–332.

Sword, S., Westine, C., & Young, M. (in press). Recruiting and retaining successful doctoral students from underrepresented groups: five unhelpful narratives and one helpful one.

Thurston, W. (1994). On proof and progress in mathematics. *Bulletin of the American Mathematical Society, 30*(2), 161–177.

Weber, K. (2014). Mathematics majors' perceptions of conviction, validity, and proof. *Mathematical Thinking and Learning, 16*(4), 341–373.