ASM | Journal of Microbiology & Biology Education

Check for updates

| Tips and Tools

# Developing a generative AI chatbot to support students' higher-order thinking in STEM education

**Lindsey Engle Richland,[1] Shuyuan Yu,[1] Banu Karyagdi,[1] Hongyang Zhao,[2] Shayan Doroudi,[1] Joseph T. Wong,[1] Ella Rose,[1] Maryam Cheraghi,[1] Katherine Trieste Rhodes[1]**

**AUTHOR AFFILIATIONS** See affiliation list on p. 7.

**ABSTRACT** Higher-order thinking is a key component of science, technology, engineering, mathematics (STEM) reasoning and supports students' ability to reflect on their ideas, make connections, and generalize across contexts. We report a new approach to the use of generative AI in STEM education, in which we are developing a chatbot that aims to engage students in higher-order thinking, rather than giving instructional content. We describe a proof of concept test of whether interaction with our generative artificial intelligence chatbot could elicit students' higher-order thinking talk (HOTT) in conversation as they reflected on their solutions to proportional reasoning problems. Thirty-three fifth- and sixth-grade students received a 30-min lesson on symbolic and number line strategies for comparing ratios and completed an assessment following the lesson. Students then engaged in dialog with the UCI Math ThinkBot (afterward *Math ThinkBot*) to explain their solutions and compare strategies. These transcripts were examined qualitatively and quantitatively for HOTT, revealing that HOTT appeared in both chatbot prompts and children's responses—which is important because eliciting student HOTT can be a challenge in classroom discourse. In addition, children's HOTT was positively associated with the HOTT produced by the chatbot. These findings suggest that human-artificial intelligence dialog can lead to instructional HOTT discourse and may offer a promising approach for supporting higher-order thinking and STEM reasoning.

**KEYWORDS** higher-order thinking, generative artificial intelligence, chatbot, reflective dialogue, mathematics discussion, STEM

Downloaded from https://journals.asm.org/journal/jmbe on 14 October 2026 by 128.195.95.37.

Higher-order thinking (HOT), defined as the ability to connect representations, make inferences, and reason about relationships (1), is central to science, technology, engineering, mathematics (STEM) reasoning. HOT supports engagement with mathematical or scientific phenomena much as an expert might, such that students focus on inferential reasoning, including reasoning about how concepts function, developing causal explanations, and drawing links between concepts or problems (2–5). Higher-order thinking talk (HOTT) (6), or language expressing relations, is a powerful way to engage in this reasoning. By requiring students to intentionally communicate about links and articulate conceptual or theoretical relationships, these interactions can become rich opportunities for students to both hone their HOT skills and their understanding of STEM disciplines (7). Despite its value, teachers often describe engaging students in HOTT as a challenge (8). Limited engagement by many students in a whole-class environment, a lack of understanding by a student of what constitutes a strong HOTT response to a HOTT question, or teachers' challenges in posing high-quality HOTT questions may all present barriers to encouraging HOTT in STEM learning environments. These challenges are particularly evident in mathematical learning environments where students often

**Editor** Thiru Vanniasinkam, Charles Sturt University, Wagga Wagga, New South Wales, Australia

Address correspondence to Lindsey Engle Richland, l.richland@uci.edu.

The authors declare no conflict of interest.

This research was presented at the Conference on Advancing Research on Quantitative Reasoning in STEM Education (2026).

**Received** 15 May 2026
**Accepted** 15 July 2026
**Published** 14 October 2026

Copyright © 2026 Richland et al. This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International license.

struggle to produce meaningful mathematical explanations, and leading mathematics discussions is challenging for teachers (8).

We describe a novel approach to providing students opportunities to engage in HOTT during STEM instruction through a generative artificial intelligence (GenAI) chatbot designed specifically to prompt students' HOTT (9–11). The project focuses on mathematics, since eliciting HOTT in mathematical talk is particularly challenging from an instructional standpoint (8) and yet mathematical reasoning is a foundation for many STEM fields. Mathematical reasoning and providing mathematical explanations are also specific components of the common core practice standards required to be taught in the U.S. mathematics curriculum at all grade levels (2, 12). We build on a large language model approach to develop a chatbot that can be predictive and responsive to students' produced language and mathematical problem solving.

Downloaded from https://journals.asm.org/journal/jmbe on 14 October 2026 by 128.195.95.37.

## PROCEDURE

### Development of the Math ThinkBot

#### *Reflective dialogue*

We developed a chatbot, Math ThinkBot, to support fifth- and sixth-grade students in engaging in higher-order mathematical thinking and reasoning about ratios and proportions, a cognitively challenging mathematical context because it involves reasoning about relationships. Rather than providing direct instruction or answers, it was meant to prompt strategy description, comparison, scaffold reasoning, encourage explanations, and help students reflect on their strategies. These designs were intended to maximize reflection, support conceptual clarity, and elicit higher-order thinking talk from both the child and the chatbot.

#### *Training materials*

The Math ThinkBot was developed using OpenAI's GPT-4o through Playlab, a website supporting educators and researchers in designing GenAI-powered chatbots. To ensure that the Math ThinkBot engages students in higher-order thinking rather than completing the work for them, it was designed to prompt students to explain, compare, and reflect on how they solved the problems, not to provide step-by-step procedural instructions, and not to confirm or correct students directly. Key design features include avoiding answer generation, limiting the number of questions and words per conversational turn, and using encouraging, student-friendly language. To further guide the chatbot's instructional behavior, we provided training inputs to guide it to use HOTT, including explicit instructions and researcher-developed sample student-chatbot dialogs designed to encourage HOTT based on previous research (6, 13). Training materials also focused on pedagogical content knowledge including reference documents on ratio and proportion concepts, multiple problem-solving strategies, common student misconceptions, prohibited strategies, assessment problems, and solution strategies. See File S2 for model parameters, configurations, instructions, and all training materials.

### Study design

#### *Participants*

Thirty-three diverse fifth and sixth graders (age mean = 11.4 yr, SD = 0.6, 55% boys, 33% White, 45% Asian, 9% African American, 9% Latine, 3% Native Hawaiian or Pacific Islander, 6% other) recruited from a research database participated in the study.

#### *Study procedure*

The data reported here are part of a larger study in which students completed a mathematics lesson on proportional reasoning followed by an assessment. Then, students interacted with the Math ThinkBot one-on-one to reflect on and compare the

strategies they used to complete the assessment (see Fig. S1 and S2 for sample problems and corresponding students' work). The study was approved by the Institutional Review Board of the University of California, Irvine. Parents' consents and children's assents were obtained.

### *Coding*

Total words and HOTT cue words were quantitatively counted in child and chatbot turns. Based on previous research (6, 13), HOTT cue words were defined as follows: *Why, How, Think, More, Less, Compare, Easier, Similar, Similarity, Difference, Different, Differently, Bigger, Smaller, If, Then, Just like, Same, Both,* and *Like*. We used this cue word list as an estimate of HOTT frequency, following earlier studies in which these emerged as the most frequent HOTT words (6, 13). See Fig. 1 and 2 for sample dialogs with keywords highlighted.

## CONCLUSION

### Results

Qualitative and frequency-based analyses illustrate how the Math ThinkBot supported higher-order thinking talk. As shown in Fig. 1 and 2, the Math ThinkBot was able to prompt children to explain and justify their procedures and to compare alternative strategies, thereby providing rich opportunities for children to connect between procedures, problems, and strategies. In general, the Math ThinkBot was able to consistently encourage children to move beyond simple, procedural accounts of their solution strategies and instead increasingly demonstrate conceptual understanding and strategic problem-solving. For example, Fig. 1 is an excerpt from a transcript in which the child justifies why a strategy works and describes different solution strategies. The child in this example has just described their solution, which was correct. Instead of the chatbot moving on to the next problem, the child was prompted to explain the reasoning underlying their strategy (i.e., why scaling both quantities by the same

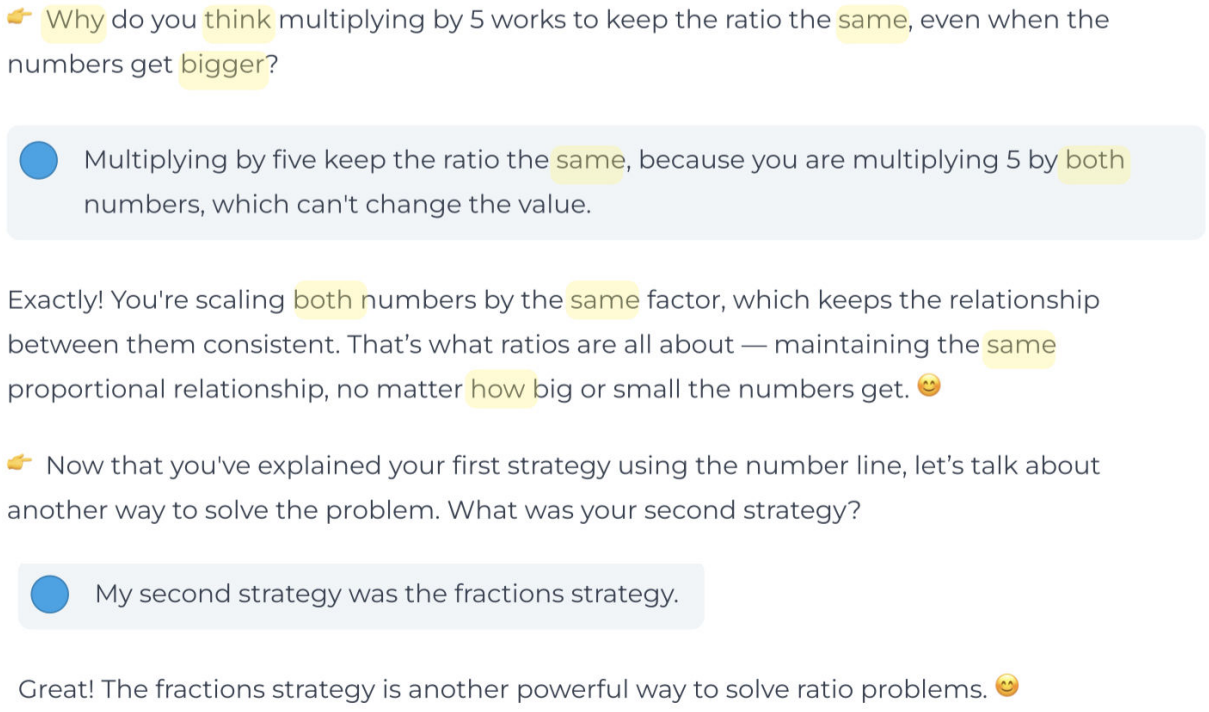


**FIG 1** Transcript excerpt: the Math ThinkBot prompts the child to explain and justify their problem-solving strategy. Texts on a white background represent the Math ThinkBot's conversational turns. Texts beginning with a blue dot represent the child's conversational turns. Higher-order thinking talk cue words are highlighted in yellow.

Downloaded from https://journals.asm.org/journal/jmbe on 14 October 2026 by 128.195.95.37.

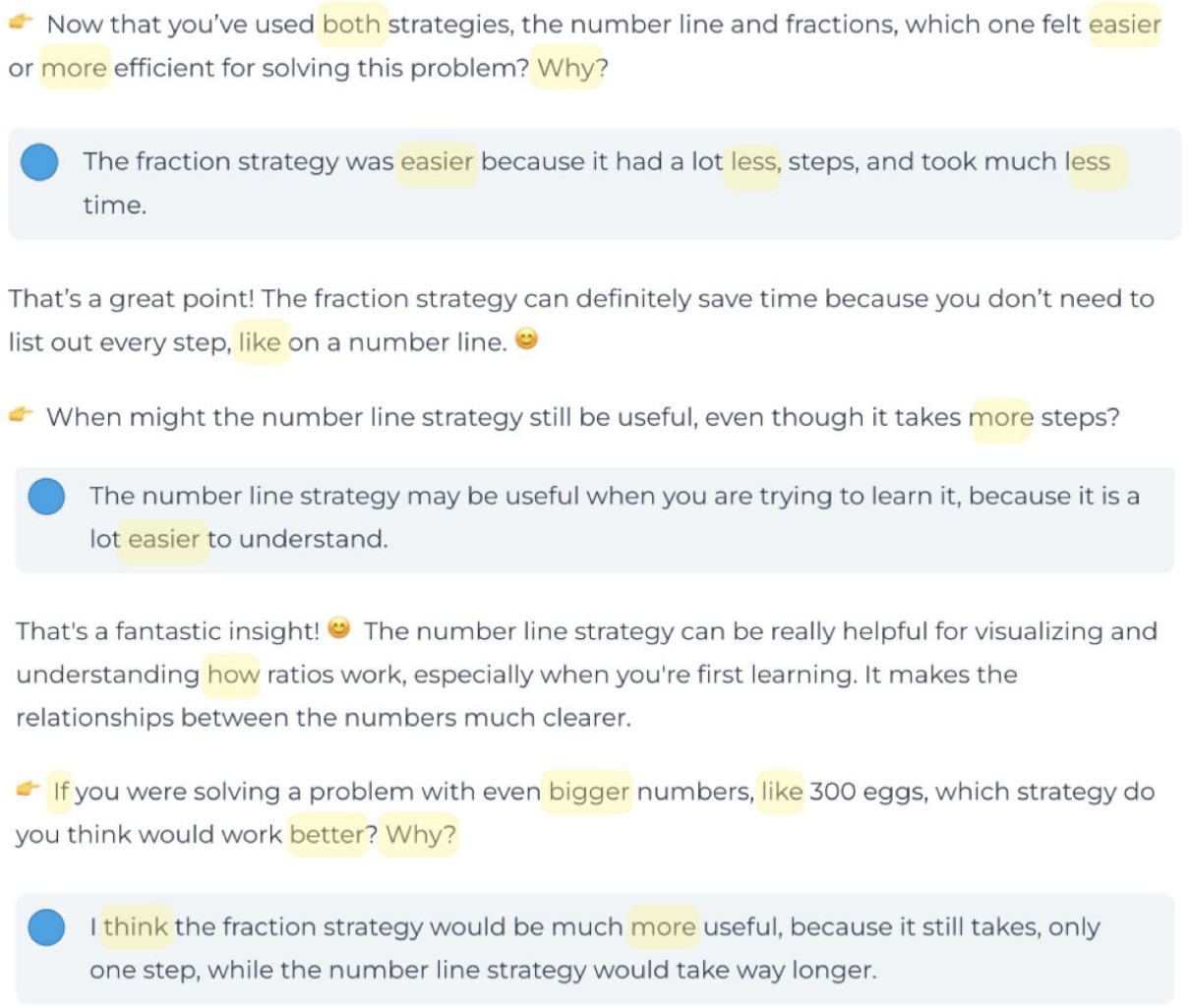


**FIG 2** Transcript excerpt: the Math ThinkBot prompts the child to compare alternative strategies. Texts on a white background represent the Math ThinkBot's conversational turns. Texts beginning with a blue dot represent the child's conversational turns. Higher-order thinking talk cue words are highlighted in yellow.

Downloaded from https://journals.asm.org/journal/jmbe on 14 October 2026 by 128.195.95.37.

number preserves the ratio), which was then reinforced by the chatbot through a restatement of the key conceptual relationship.

Additionally, an examination of the texts shows that children were encouraged to articulate their strategy preferences for efficiency and utility, thereby engaging in complex comparisons of optimal problem-solving strategies (Fig. 2). Such attention to efficiency is important in mathematics learning, as there are often multiple solution strategies to a single problem, and efficiency can provide a leverage point for discussing the conceptual underpinnings of mathematical relationships. Training students to think about mathematics as a system of relationships that can be compared for different problem-solving goals (efficiency, clarity, generalization, etc.) could have benefits not only for understanding this specific content area of ratio and proportion but also for developing expert-like mathematical thinking approaches.

Quantitative analyses further revealed that both the Math ThinkBot and children reliably produced HOTT during the interaction. All children produced some HOTT, with an average of 13.8 HOTT cue words (SD = 9.4) in their interaction with the Math ThinkBot. Similarly, the Math ThinkBot always produced some HOTT, generating an average of 60.3 HOTT cue words (SD = 30.96). This was itself an important finding, since relational reasoning has been posited to be a uniquely human skill (14), but the chatbot demonstrated at least a strong proficiency in effective usage of the relational HOTT language.

Downloaded from https://journals.asm.org/journal/jmbe on 14 October 2026 by 128.195.95.37.

The HOTT cue words or phrases were a small proportion of the total word count generated by each entity (the Math ThinkBot vs Children, 5.9% vs 4.8%), but this was anticipated since they are cues and language frames, and so there might be a single HOTT cue word embedded in a full sentence that all articulated a HOTT concept. For example, consider the Math ThinkBot's HOTT statement in Fig. 1: *"That's a great point! The fraction strategy can definitely save time because you don't need to list out every step,* ***like*** *on a number line."* That single HOTT cue word, "like," will only be counted as one word, because it articulates that this is a comparative explanation, showing how even small numbers of HOTT words are impactful and shape large discourse production.

Importantly, the quantity of HOTT cue words produced by the Math ThinkBot and children was significantly correlated [$r$ (31) = 0.52, 95% CI (0.21, 0.73), $p$ = 0.002, Fig. 3], indicating the Math ThinkBot and children were successful at engaging in HOTT interactionally. Sometimes, this took the form of the chatbot asking a HOTT question in which the child's answer revealed HOTT through its response. Other times, it responded to children's HOTT, either by answering the child's HOTT question or by reiterating a HOTT statement introduced by the child. Note that these findings are correlational due to the nature of conversational data. Alternative explanations include the possibility that longer conversations may naturally contain a higher frequency of HOTT cue words, or that children simply repeated HOTT cue words that appeared in the chatbot's prompts rather than independently generating HOTT. However, these alternative explanations are consistent with the intended function of the Math ThinkBot—to engage students with HOTT by repeatedly raising HOTT questions and modeling the use of HOTT.

Comparison between children with low vs high frequencies of HOTT cue words indicated that the measure reflects the quality of higher-order thinking. For example, the conversation between the Math ThinkBot and a child with code name "Tarantula," who used few HOTT cue words, was primarily filled with descriptions of problem-solving procedures and step-by-step calculations (see Fig. 4A for a transcript excerpt and 3.1 in File S3 for the full transcript). The child's responses were dominated by asking clarification questions, stating numbers that appeared in the problem, and recalling addition or multiplication facts, rather than engaging in reasoning or justification. The child also disengaged when prompted to compare different problem-solving strategies. In contrast, a different child with code name "Elephant," who used a high frequency

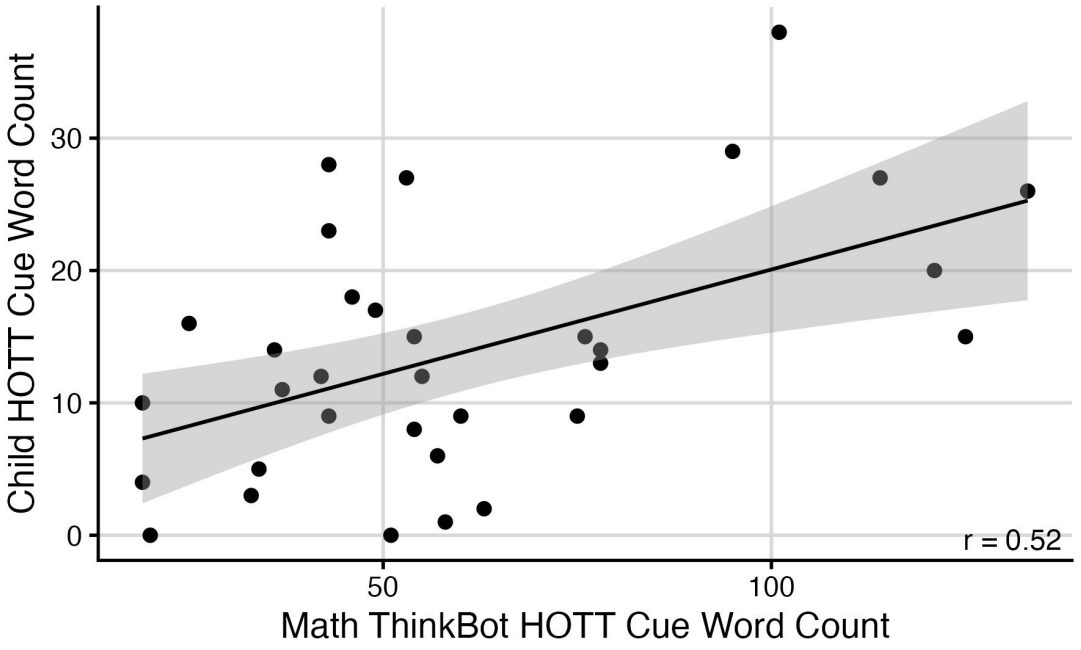


**FIG 3** Scatterplot illustrating the association between the number of higher-order thinking talk (HOTT) cue words produced by the Math ThinkBot and by children. Note that the axes are not symmetrical because the Math ThinkBot generated a significantly larger number of words than students.

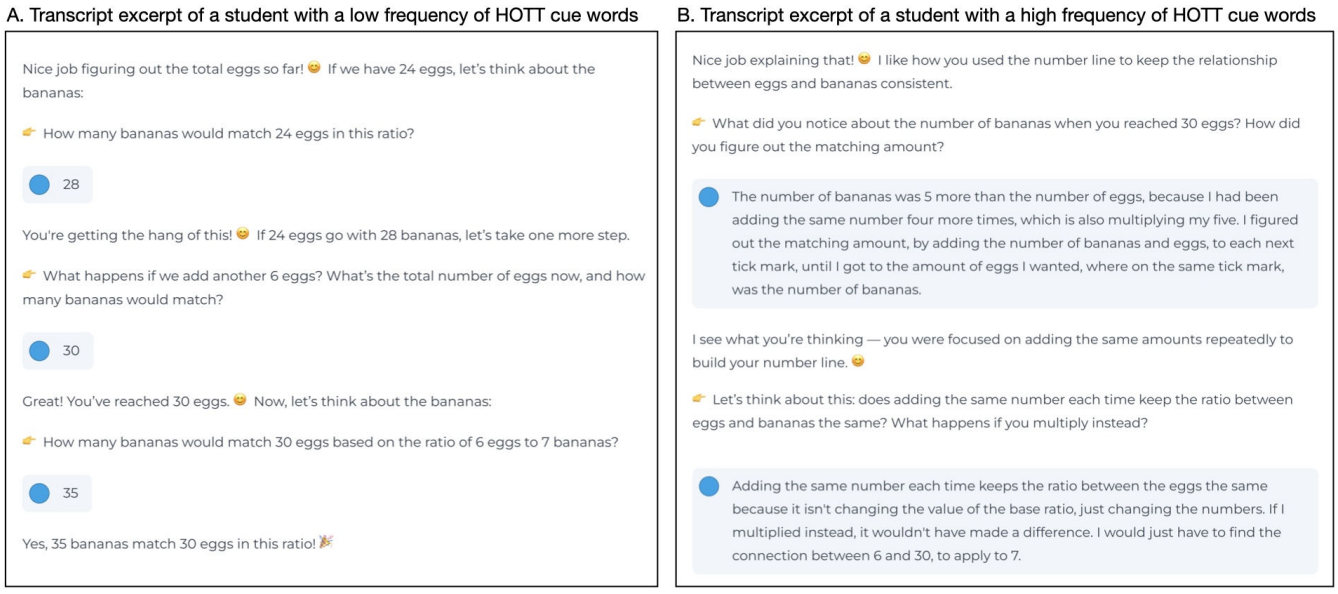


**FIG 4** Transcript excerpts of students with low (A) vs high (B) frequencies of higher-order thinking talk (HOTT) cue words.

of HOTT cue words, articulated not only the steps taken to reach a correct answer but also more sophisticated reasoning throughout the conversation, including strategy comparison and abstraction (see Fig. 4B for a transcript excerpt and 3.2 in File S3 for the full transcript). Importantly, after the student provided a solid explanation of his/her strategy, the Math ThinkBot was able to follow up with deeper HOTT questions, prompting the student to compare two solution strategies.

## Limitations and future directions

The current study focuses on mathematical thinking, though there are clear opportunities to expand this to other STEM domains, and future studies should examine these applications. Additionally, our participants were recruited from a research database which was developed through community events, school district lists, and sign-ups from prior research. The database was designed to represent the diversity of the geographical region in whch the research took place, though future studies would benefit from larger, more inclusive samples. Moreover, although our study showed the potential of the Math ThinkBot for supporting students' STEM reasoning, we were not able to examine learning gains in mathematical understanding because the Math ThinkBot interaction took place after the assessment as a reflective activity. Future studies are planned using randomized control trials to test causal relationships between using the Math ThinkBot and students' higher-order thinking talk and learning outcomes. Finally, the frequency of HOTT cue words serves as a rough estimate of the presence of higher-order thinking. Future research should involve coding the types and quality of higher-order thinking talk based on the semantic meaning of the transcripts.

## Summary

Taken together, these findings show the potential of human-GenAI dialog as a novel approach for fostering and assessing higher-order thinking in STEM learning. They further highlight the distinct roles that interactive AI partners may play in shaping children's mathematical reflection. Future research involves conducting HOTT coding that captures the meaning of entire entries rather than simply frequencies of HOTT cue words to categorize the discourse with more cognitive precision. We are also exploring how different prompts relate to the quality of children's higher-order thinking to generate improved training materials to inform both chatbots and teachers. Finally, we

Downloaded from https://journals.asm.org/journal/jmbe on 14 October 2026 by 128.195.95.37.

are also co-refining the chatbot with teachers to create a tool with wide STEM usage capability, first for use at more points within the mathematics curriculum, and then potentially in other STEM domains. Together, this work will allow us to investigate the Math ThinkBot as a generative AI application that can be integrated into classroom instruction to ultimately scaffold key reasoning skills within STEM learning at scale. Our aim is to hone generative AI tools to support children's development of STEM reasoning skills rather than to provide a resource that would outsource the higher-order thinking itself to the AI.

## ACKNOWLEDGMENTS

This research was supported by the Institute of Education Sciences, U.S. Department of Education, under grant #R305A240287 to the University of California, Irvine. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

## AUTHOR AFFILIATIONS

[1]School of Education, University of California, Irvine, Irvine, California, USA
[2]College of Education, University of Illinois Urbana-Champaign, Champaign, Illinois, USA

## AUTHOR ORCIDs

Lindsey Engle Richland http://orcid.org/0000-0003-1514-6013
Shuyuan Yu http://orcid.org/0000-0002-5993-1681
Banu Karyagdi http://orcid.org/0000-0003-2805-2868
Hongyang Zhao http://orcid.org/0000-0003-4711-0314
Shayan Doroudi http://orcid.org/0000-0002-0602-1406
Joseph T. Wong http://orcid.org/0000-0003-1890-6284

## AUTHOR CONTRIBUTIONS

Lindsey Engle Richland, Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review and editing | Shuyuan Yu, Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review and editing, Data curation | Banu Karyagdi, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Writing – review and editing, Writing – original draft, Software | Hongyang Zhao, Investigation, Methodology, Project administration, Software, Supervision, Writing – review and editing | Shayan Doroudi, Conceptualization, Funding acquisition, Investigation, Methodology, Software, Supervision, Writing – review and editing | Joseph T. Wong, Investigation, Methodology, Software | Ella Rose, Investigation, Methodology, Project administration | Maryam Cheraghi, Investigation, Project administration | Katherine Trieste Rhodes, Conceptualization, Formal analysis, Funding acquisition, Supervision, Writing – review and editing

## ADDITIONAL FILES

The following material is available online.

### Supplemental Material

**File S1 (jmbe00101-26-s0001.docx).** Assessment problems (i.e., problems 2 and 3) used in the ThinkBot reflection with sample student work.
**File S2 (jmbe00101-26-s0002.docx).** ThinkBot design prompts and parameters.
**File S3 (jmbe00101-26-s0003.docx).** Sample transcripts of children with low vs high frequencies of higher-order thinking talk (HOTT) cue words.

Downloaded from https://journals.asm.org/journal/jmbe on 14 October 2026 by 128.195.95.37.

## REFERENCES


1. Richland LE, Simms N. 2015. Analogy, higher order thinking, and education. Wiley Interdiscip Rev Cogn Sci 6:177–192. https://doi.org/10.1002/wcs.1336
2. Richland LE, Stigler JW, Holyoak KJ. 2012. Teaching the conceptual structure of mathematics. Educ Psychol 47:189–203. https://doi.org/10.1080/00461520.2012.667065
3. Lombrozo T. 2012. Explanation and abductive inference, p 260–276. In Holyoak KJ, Morrison RG (ed), The Oxford handbook of thinking and reasoning. Oxford University Press, Oxford, UK.
4. Rittle-Johnson B, Star JR. 2007. Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. J Educ Psychol 99:561–574. https://doi.org/10.1037/0022-0663.99.3.561
5. Rittle-Johnson B, Star JR, Durkin K. 2017. The power of comparison in mathematics instruction: experimental evidence from classrooms, p 273–296. In Geary DC, Berch VB, Koepke KM (ed), Mathematical cognition and learning. Elsevier.
6. Frausel RR, Silvey C, Freeman C, Dowling N, Richland LE, Levine SC, Raudenbush S, Goldin-Meadow S. 2020. The origins of higher-order thinking lie in children's spontaneous talk across the pre-school years. Cognition 200:104274. https://doi.org/10.1016/j.cognition.2020.104274
7. Gentner D, Holyoak KJ, Kokinov B, eds. 2001. The analogical mind: perspectives from cognitive science. MIT Press, Cambridge, MA.
8. Kazemi E, Hintz A. 2014. *Intentional Talk: How to Structure and Lead Productive Mathematical Discussions*. 1st ed. Routledge.
9. Borge M, Smith BK, Aldemir T. 2024. Using generative AI as a simulation to support higher-order thinking. Int J Comput Support Collab Learn19:479–532. https://doi.org/10.1007/s11412-024-09437-0
10. Lee AVY, Tan SC, Teo CL. 2023. Designs and practices using generative AI for sustainable student discourse and knowledge creation. Smart Learn Environ 10:59. https://doi.org/10.1186/s40561-023-00279-1
11. Li S, Zeng C, Liu H, Jia J, Liang M, Cha Y, Lim CP, Wu X. 2025. A meta-analysis of AI-enabled personalized STEM education in schools. Int J STEM Educ12:58. https://doi.org/10.1186/s40594-025-00566-y
12. Common Core State Standards Initiative. 2010. Common core state standards for mathematics. Washington, DC. National Governors Association Center for Best Practices and Council of Chief State School Officers.
13. Frausel RR, Richland LE, Levine SC, Goldin-Meadow S. 2021. Personal narrative as a "breeding ground" for higher-order thinking talk in early parent–child interactions. Dev Psychol 57:519–534. https://doi.org/10.1037/dev0001166
14. Penn DC, Holyoak KJ, Povinelli DJ. 2008. Darwin's mistake: explaining the discontinuity between human and nonhuman minds. Behav Brain Sci 31:109–130. https://doi.org/10.1017/S0140525X08003543


Downloaded from https://journals.asm.org/journal/jmbe on 14 October 2026 by 128.195.95.37.