EasyChair Preprint

№ 13351

Learning and AI Evaluation of Tutors Responding to Students Engaging in Negative Self-Talk

Danielle R Thomas, Jionghao Lin, Shambhavi Bhushan,
Ralph Abboud, Erin Gatz, Shivang Gupta and Kenneth Koedinger

May 18, 2024

Learning and AI Evaluation of Tutors Responding to Students Engaging in Negative Self-Talk

Danielle R. Thomas
Carnegie Mellon University
Pittsburgh, USA
drthomas@cmu.edu

Jionghao Lin
Carnegie Mellon University
Pittsburgh, USA
jionghao@cmu.edu

Shambhavi Bhushan
Carnegie Mellon University
Pittsburgh, USA
shambab@andrew.cmu.edu

Ralph Abboud
Learning Engineering Virtual Institute
Oxford, UK
ralph@ralphabb.ai

Erin Gatz
Carnegie Mellon University
Pittsburgh, USA
egatz@andrew.cmu.edu

Shivang Gupta
Carnegie Mellon University
Pittsburgh, USA
shivang@cmu.edu

Kenneth R. Koedinger
Carnegie Mellon University
Pittsburgh, USA
koedinger@cmu.edu

ABSTRACT

Addressing negative self-talk by students, such as responding to a student when saying, "I am dumb" or "I can't do this" can be difficult for even the most experienced tutor. Despite potential tutor learning from scenario-based lessons on this topic, human-graded assessment remains time-consuming. Leveraging generative AI for evaluating textual responses in online training presents a scalable solution. Research suggests a tutor validates student's feelings when they speak negatively of themselves, e.g., by a tutor responding, "I understand how you feel" or "I recognize this is difficult." This ongoing work assesses the performance of 60 undergraduate tutors within an online lesson on enhancing tutors' abilities to respond to students engaging in negative self-talk. We find statistically significant tutor learning gains from pretest to posttest. Additionally, we describe a method of using generative AI for assessing tutors' responses to *predict* the best approach and subsequently *explain* the rationale behind it. Using the large language model GPT-4, we find high absolute performance when evaluating tutor responses involving *predicting* ($F1 = 0.85$) and *explaining* ($F1 = 0.83$) the best approach. Minor improvements are needed to the lesson itself. A future goal of this work is to fully develop automated systems of assessing tutor learning attending to barriers to students' motivation and doing so at scale.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing** → **Artificial intelligence**.

KEYWORDS

Tutor Training, Generative AI, Large Language Models, Assessment

1 INTRODUCTION

Addressing barriers to learning, such as negative self-talk among students, is a critical aspect of effective tutoring [4, 10, 17]. This work leverages tutor training using research-driven strategies, previously demonstrated to be effective [5, 18], to equip tutors with the necessary skills to effectively respond to students expressing self-doubt or feelings of incompetence. While scenario-based lessons offer a promising avenue for learning [15, 18], the evaluation process traditionally relies heavily on human graders, posing limitations to scalability. Additionally, learners receiving real-time corrective feedback on their open, or constructed, responses has been known to improve learners' performance [13], which is not feasible using human grading. To address these challenges, the integration of generative AI for evaluating textual responses in online training emerges as a viable solution, promising to streamline the evaluation process and enhance scalability.

Drawing from existing research emphasizing the importance of validating students' feelings when expressing negativity towards themselves [7, 23], this study evaluates the efficacy of an online lesson designed to enhance tutors' abilities. Analyzing the lesson performance of 60 undergraduate remote tutors, we strive to determine tutors' learning gains. Additionally, we describe a methodological approach of employing generative AI for assessing tutors' open textual responses. Leveraging LLMs, we hypothesize that responses evaluated by GPT-4 will be comparable to that of human graders. Ultimately, we strive to create a real-time and scalable approach to evaluating tutor actions in training. Within this present work, we aim to answer the following:

RQ1: Is the scenario-based lesson effective in teaching tutors new skills in addressing students when they are engaging in negative self-talk?
RQ2: Can large language models be used to assess tutors' open responses, thus creating an automated system of evaluating tutors' overall performance within the lesson?

2 BACKGROUND

Students engage in self-talk, also known as inner speech, as a tool to regulate their thinking and behavior [8]. Self-talk assists students

with learning efficiently and can improve social-emotional health. However, negative self-talk characterized by the individual's engagement with pessimistic self-evaluations through negative statements can be counterproductive to learning [8, 9, 14]. Repeatedly expressing pessimistic emotions and negative self-talk may serve as a risk factor for emotional health concerns [25].

When addressing negative self-talk, it is recommended to use a research-supported approach to consistently uplift students and enhance their self-confidence [23]. An appropriate tutor response to students engaging in negative self-talk involves initially validating and acknowledging the student's feelings while promptly reframing the situation, highlighting their strengths and capabilities. Furthermore, tutors should counter negative dialogue with positive affirmations and exemplary behavior, emphasizing past successes [16]. By employing these strategies collectively to combat negative self-talk, students can cultivate a growth mindset [7].

2.1 The *Responding to Negative Self-Talk* Lesson

The efficacy of scenario-based lessons as a training tool for enhancing tutoring skills, such as responding to student errors and delivering effective praise, has previously demonstrated 20 percent increases in tutor learning from pre-instruction to post-instruction [18]. These lessons also facilitate the collection of valuable data directly from tutors, which can be used to inform the development of generative AI models aligned with specific competencies. The *Responding to Negative Self-Talk* lesson presents two scenarios to instruct a human tutor on handling negative self-talk. Digital Appendix A (https://tinyurl.com/b82ze9mc) provides the lesson description, details presented to tutors commencing the lesson, and the research-recommended approach provided to tutors.[1] In one scenario, a tutor hears the student Eduardo say to himself, "I am so stupid," while solving a math problem. In this situation, the research-backed approach recommends the tutor to recognize the student's emotions instead of dismissing them while reminding the student of their capabilities through affirmations [16]. Fig. 1 shows the scenario involving Eduardo with an open response question, asking tutors to input their response to the situation along with rationale in future questions. Another scenario involves student Geetika remarking to herself and another student, "I am dumb and I will never be able to do this" with similar open-response questions. Modeling [5] and [18], each scenario contains two open-responses and two multiple-choice questions (MCQ). First, tutors are tasked to *predict* the best approach, answering an open response and MCQ. Then they are asked to *explain* the rationale behind their approach in an open response and MCQ. Tutors observe the research-recommended best approach upon completion of the first scenario and answering of questions. Then tutors are presented a second analogous scenario containing open response and MCQ, which is used as a posttest. The order of the scenarios is random, ensuring counterbalancing when determining learning gains. Both scenarios and subsequent questions can be accessed in the Digital Appendices B and C, respectively.

[1]Access the Digital Appendix: https://tinyurl.com/b82ze9mc

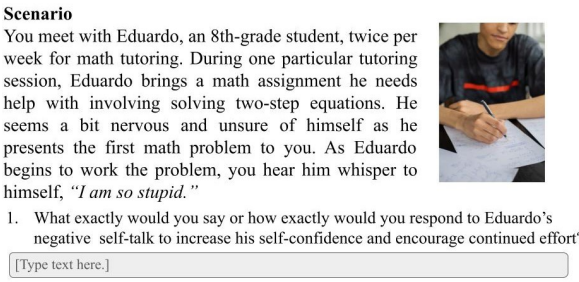

Figure 1: The scenario involving student Eduardo with the corresponding open response question asking a tutor to *predict* the best approach.

2.2 Using Generative AI to Evaluate Responses

Generative AI, in particular large language models, possess the ability to assess tutor's textual, open responses in real-time. Large language models (LLMs), such as GPT-4 [1], Claude [3], and LLaMa [19], have recently achieved breakthrough performance on a wide variety of linguistic tasks. Modern LLMs are based on a large-scale transformer backbone [20], which is trained on vast amounts of open data with the goal of maximizing its likelihood and ultimately generating new content. LLMs have garnered substantial attention from researchers in several areas, including education, as a means to perform reasoning at scale and at a lower cost. However, despite their promise, LLMs suffer from key limitations. First, LLMs are black-box models whose internal function remains largely opaque, leading to safety concerns. Second, LLMs tend to "hallucinate" [24] and create plausible sounding but inaccurate content that misleads end users. To alleviate the latter limitation, researchers have developed techniques to address hallucination by manipulating the input text to the LLM, a technique commonly referred to as prompt engineering [12]. In prompt engineering, the goal is to provide better context and/or structure to the LLM so as to guide it to the correct outputs. For instance, few-shot prompting [2] provides a set of exemplars to the LLM in the input prompt to demonstrate the ideal model behavior. Moreover, chain-of-thought (CoT) prompting [22] encourages the model to "think step by step" and better emulate intermediate reasoning steps. In this work, we use and describe several prompt engineering techniques to leverage GPT-4 in evaluating tutors' open responses.

3 METHOD

3.1 Tutor Participants & Lesson Delivery

There were 60 college-student participants who completed the lesson, employed as paid tutors for a remote tutoring organization supporting middle-school students. While the demographics of the tutors were undisclosed, they exhibited cultural and racial diversity. Tutors' self-reported tutor experience levels were assessed using a 5-point Likert scale with 1 indicating little to no experience (novice) and 5 indicating an expert tutor. On average, tutors reported an experience level of 3.3 (SD = 1.25). The lesson was developed in collaboration with tutoring supervisors, who reported considerable negative self-talk among their students, and a university research

Table 1: Sample learner-sourced responses for *predicting* the best approach with coding and rationale (green).

Tutor response	Coding and rationale
Eduardo, it is normal to feel these feelings of self-doubt. You have the ability to solve this problem with the skills I have seen you demonstrate.	Correct (1): This response validates the student's feelings by recognizing the student's doubt and provides positive affirmation by stating the student has the ability to solve the problem.
Eduardo, please don't be so harsh on yourself. You are so smart and hardworking. I love how you are so consistent. Let's look at the problem together.	Partially correct (0.5): This response provides positive affirmation by complimenting the student's work ethic. However, it does not explicitly validate the student's feelings.
Hey, you're not dumb. Why don't you see what you and your partner can come up with if you work together?	Incorrect (0): This response does not validate the student's feelings nor does it provide positive affirmation.

team specializing in learning science, thereby enhancing construct validity. The lesson was delivered via an online platform and aligns with research-shown competencies of effective tutoring [4, 18].

3.1.1 Human Open-Response Coding and Inter-rater Reliability. Two experienced researchers coded participant responses to assess inter-rater reliability. Open-response questions tasking tutors to predict the best approach were open-coded. Correct responses (score=1) need to: 1) acknowledge the student's feelings and validate them, such as a tutor saying, "I understand you may be frustrated" or "I realize this is hard for you"; and 2) remind the student of their strengths by modeling positive self-talk, such as saying, "I have seen you solve problems similar to this before." Partially correct (score=0.5) responses apply only one of these two strategies. Incorrect responses (score=0) apply neither strategy. Tutor responses tasking tutors to *explain* the rationale behind their chosen approach were binary coded. Table 1 illustrates sample learner-sourced responses for *predicting* the best approach with coding and rationale. Highlighted utterances align with correct rationale. Appendices D and E display the annotation rubric and learner-sourced responses with rationale for predictions and explanations, respectively.

There was relatively high agreement in inter-rater reliability between the two human coders. For responses requiring tutors to *predict* the best course of action, there was 87% agreement and weighted Cohen's Kappa of 0.80. For responses asking tutors to *explain* their rationale, there was a 96% agreement rate and a Cohen's Kappa of 0.91. Both reflect substantial agreement, supporting the reliability of the coding process.

3.1.2 Determining Tutor Learning Gains. We employed a mixed-effects ANOVA to examine the impact of lesson scenarios on tutor performance. The *scenarios* (i.e., Eduardo or Geetika) served as the between-subjects factor, while *time*, specifically *pretest* and *posttest*, served as the within-subjects factor. Treating the *scenario* as a fixed effect aids in determining if there exists an imbalance in difficulty between the two scenarios while considering test *time* as a random effect accounts for within-subjects variation.

3.1.3 Prompting Generative AI to Evaluate Open Responses. We draw from past research in the field [6, 11] and prompt engineering techniques to effectively prompt GPT-4. Two prompts were developed using zero-shot learning and prompt chaining. Table 2 illustrates the prompt used for the task of assessing tutors in *predicting* the best approach. Appendices F and G illustrate the complete prompts and code for GPT-4 to evaluate tutor's responses *predicting* the best approach and *explaining* their rationale, respectively.

Table 2: Prompt used for the task of assessing tutors in *predicting* the best approach.

```
FILTER_PROMPT = """

Please assess a tutor's response in a tutor training scenario involving
a middle school student learning math. The student is engaging in
negative self-talk by saying negative comments about themselves, such
as "I am dumb" or "I will never be able to do this." Assess and score the
tutor's response, as follows:

-if the tutor's response acknowledges the student's feelings by validating, or acknowledging, them AND provides positive affirmation or
encouragement, score with a 1.
-if the tutor's response acknowledges the student's feelings by validating, or acknowledging, them OR provides positive affirmation or
encouragement, score with a 0.5.
-if the tutor's response does not validate the student's feelings by validating them NOR provide positive affirmation, score with a 0.

Response Start ---
"""
SCORING_FORMAT_PROMPT = """

--- Responses End. Given the earlier response, please return a JSON
string following the format,
 {\"Rationale\": \"your reasoning here\", \"Score\":0/0.5/1}.

"""
```

4 RESULTS & DISCUSSION

RQ1: Is the scenario-based lesson effective in teaching tutors new skills on addressing students when they are engaging in negative self-talk? The analysis revealed a statistically significant main effect of time on tutor performance, $F(1, 58) = 4.897$, $p < .05$, indicating an overall improvement in tutors' performance from pretest to posttest regardless of the scenario order. This suggests a general learning effect or improvement of skills. There was no statistically significant main effect of the specific *scenario* on tutor performance. Furthermore, the interaction between *scenario* and *time* was found to be not statistically significant. These non-significant findings suggest that the scenario difficulty was similar and the extent of tutoring improvement from pretest to posttest did not vary significantly according to the scenario order. *Post-hoc* analysis revealed that tutors who had the Eduardo scenario in the pretest followed by the Geetika scenario in the posttest, *Eduardo:Geetika*, although lower scoring at pretest ($M = 2.66$; $SD = 1.13$), demonstrated larger learning gains at posttest ($M = 3.08$; $SD = 1.04$). Tutors who had the Geetika scenario at pretest ($M = 3.02$; $SD = 0.91$) followed by the Eduardo scenario at posttest ($M = 3.09$; $SD = 1.1$), *Geetika:Eduardo*, did not improve as much. This indicates that the Geetika scenario was relatively easier. The *Eduardo:Geetika* condition demonstrated greater tutor learning. Fig. 2 illustrates the pretest to posttest scores based on the order of the scenarios.

MCQs were determined to be too easy. The maximum score for selected and open responses was two points. The mean MCQ score at pretest, M = 1.67 (SD = 0.62), was relatively high leaving little room for tutors to demonstrate learning gains, with posttest scores, M = 1.76 (SD = 0.53), suggestive of possible ceiling effects. Tutor performance on open responses demonstrated larger gains evident by the substantial increase (approx. 25%) from pretest scores, M = 1.05, SD = 0.70, to posttest scores, M = 1.32, SD = 0.72.

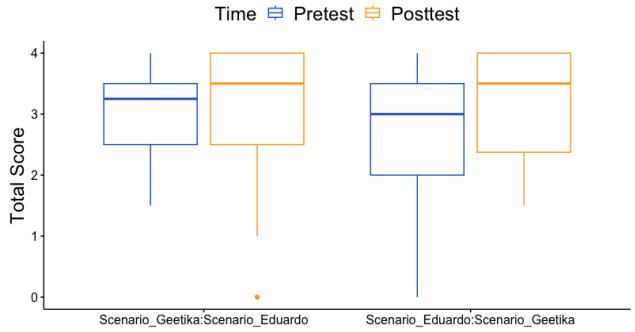

Figure 2: Pretest and posttest scores by scenario order with open responses graded by human experts.

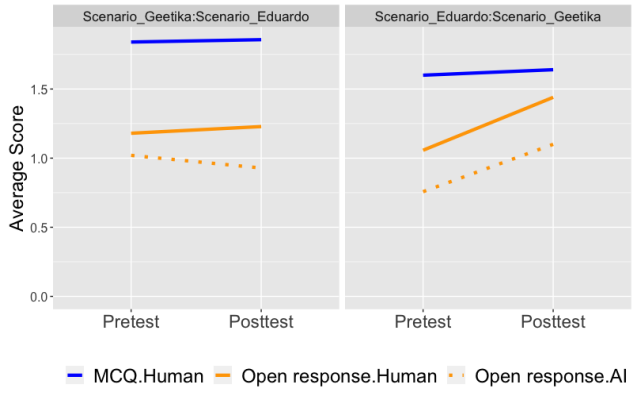

Figure 3: Average pretest and posttest scores by scenario order for MCQ and open responses, illustrating a ceiling effect among MCQs. Responses scored by GPT-4 were overall lower than human graders.

RQ2: Can large language models be used to assess tutor performance on open responses, thus creating an automated system of evaluating tutors' performance? The GPT-4 model performed considerably well on the task of coding tutor responses for *predicting* the best approach (weighted Cohen's Kappa = 0.69; $F1$ = 0.85) and *explaining* (Cohen's Kappa = 0.65; $F1$ = 0.83) the rationale behind their chosen approach. Appendices G and H display the confusion matrices and performance measures for *predict* and *explain* open responses, respectively. These results demonstrate the promising efficacy of generative AI in evaluating the quality and correctness of tutor responses within scenario-based training. Performing similar ANOVA analyses using AI-coded data, which was used with human-coded data for RQ1, we found no statistically significantly learning effect from pretest to posttest. However, there was a statistically significant interaction between scenario and time, $F(1, 58) = 7.44$, $p<.05$, indicating that the scenario order had an effect on learning gains. Performing pairwise t-test comparisons for each scenario order: *Eduardo:Geetika* demonstrated strong and positive statistically significant learning gains, $t(4.15, p<.05)$; and *Geetika:Eduardo* demonstrated no statistically significant differences indicative of learning gains.

Table 3: Comparison of pretest and posttest scores between humans and GPT-4. Standard deviations in parentheses.

	Mean Pretest	Mean Posttest	Mean Gain
Human	2.73 (1.09)	3.08 (1.06)	0.35 (0.03)
GPT-4	2.51 (1.00)	2.77 (0.98)	0.26 (0.02)

We posit that the lack of statistically significant learning gains was not due to poor performance of the LLM model, as the absolute performance was considerably high, i.e., F1= 0.85, but due to the effectiveness of the lesson itself. Among human-graded responses, tutors demonstrated small, yet significantly positive, gains. Generative AI-coded responses were consistent with human-coded responses, but overall slightly lower in score. Fig. 3 illustrates this finding, particularly among the *Eduardo:Geetika* scenario, indicated by the parallel lines for human- and AI-coded open responses, with the latter 25% lower. Table 3 provides a comparison of pretest to posttest scores between human graders and the GPT-4 model.

Both the LLM model and humans showed sensitivity to response length. For instance, there were several very short (<80 characters) open responses whereby human graders deemed the response correct (score = 1) and the LLM model incorrect (score = 0), e.g., "it acknowledges his negative self-talk but then encourages him to keep going." In other words, when the human grader says its correct, the LLM model was more likely to reject it, or mark incorrect, if it was shorter than if it was longer in length. Inversely, longer responses (>150 characters) posed challenges as well, with humans more likely to mark longer responses as correct when they were identified as incorrect by the AI. There were a few cases of the LLM model performing poorly when encountering a "double negative" statement, (e.g., "…*don't* worry so much which *doesn't* actually help…"). We hypothesize that few-shot learning approaches, chain-of-thought prompting, and other techniques will assist with these challenges and improve model performance.

5 LIMITATIONS, FUTURE WORK, & CONCLUSION

The strategy of employing small, "micro-moment" scenarios within training might seem granular and perhaps idealistic when aiming for comprehensive tutor training. Considering the broader objective of enduring solutions to teaching tutors skills, we argue that the incremental benefits of repeated situational immersion among tutors may exhibit increasing returns over time. There were possible

ceiling effects among the MCQs. Using high-frequency, learner-sourced responses that were deemed "incorrect," could be used as selected-response options [18, 21]. Using incorrect, open responses occurring in high frequency as multiple-choice options may greatly increase lesson difficulty by capturing common misconceptions [21]. In addition, lesson design modifications, such as improving the Geetika scenario to achieve similar effectiveness as the Eduardo scenario is warranted. Lastly, we described a process of iteratively applying prompt engineering techniques while concurrently assessing model effectiveness. This present work showcases the early steps of this process. The prompts will continue to be improved by exploring few-shot learning and chain-of-thought prompting techniques [2, 22]. Future work involves increasing scale by leveraging generative AI for assessing tutor lesson performance by order of magnitude from 60 tutors to say 600 tutors. In summary, this work highlights the potential of using scenario-based online training to enhance tutors' skills and leveraging generative AI for large-scale tutor evaluation.

A DIGITAL APPENDIX

https://tinyurl.com/b82ze9mc

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[3] Loredana Caruccio, Stefano Cirillo, Giuseppe Polese, Giandomenico Solimando, Shanmugam Sundaramurthy, and Genoveffa Tortora. 2024. Claude 2.0 Large Language Model: tackling a real-world classification problem with a new Iterative Prompt Engineering approach. *Intelligent Systems with Applications* (2024), 200336.

[4] Pallavi Chhabra, Danielle Chine, Adetunji Adeniran, Shivang Gupta, and Kenneth Koedinger. 2022. An evaluation of perceptions regarding mentor competencies for technology-based personalized learning. In *Society for Information Technology & Teacher Education International Conference*. Association for the Advancement of Computing in Education (AACE), 1812–1817.

[5] Danielle R Chine, Pallavi Chhabra, Adetunji Adeniran, Shivang Gupta, and Kenneth R Koedinger. 2022. Development of scenario-based mentor lessons: an iterative design process for training at scale. In *Proceedings of the Ninth ACM Conference on Learning@ Scale*. 469–471.

[6] Wei Dai, Yi-Shan Tsai, Jionghao Lin, Ahmad Aldino, Hua Jin, Tongguang Li, Dragan Gaševic, and Guanliang Chen. 2024. Assessing the Proficiency of Large Language Models in Automatic Feedback Generation: An Evaluation Study. (2024).

[7] Carol S Dweck. 2006. *Mindset: The new psychology of success*. Random house.

[8] Róisín M Flanagan and Jennifer E Symonds. 2022. Children's self-talk in naturalistic classroom settings in middle childhood: A systematic literature review. *Educational Research Review* 35 (2022), 100432.

[9] Philip C Kendall, Bonnie L Howard, and Rebecca C Hays. 1989. Self-referent speech and psychopathology: The balance of positive and negative thinking. *Cognitive therapy and research* 13 (1989), 583–598.

[10] Mark R Lepper and Maria Woolverton. 2002. The wisdom of practice: Lessons learned from the study of highly effective tutors. In *Improving academic achievement*. Elsevier, 135–158.

[11] Jionghao Lin, Zifei Han, Danielle R Thomas, Ashish Gurung, Shivang Gupta, Vincent Aleven, and Kenneth R Koedinger. 2024. How Can I Get It Right? Using GPT to Rephrase Incorrect Trainee Responses. *arXiv preprint arXiv:2405.00970* (2024).

[12] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023), 1–35.

[13] Santosh A Mathan and Kenneth R Koedinger. 2018. Fostering the intelligent novice: Learning from errors with metacognitive tutoring. In *Computers as Metacognitive Tools for Enhancing Learning*. Routledge, 257–265.

[14] Beverly D Payne and Brenda H Manning. 1991. Self-talk of student teachers and resulting relationships. *The Journal of Educational Research* 85, 1 (1991), 47–51.

[15] Justin Reich. 2022. Teaching drills: Advancing practice-based teacher education through short, low-stakes, high-frequency practice. *Journal of Technology and Teacher Education* 30, 2 (2022), 217–228.

[16] Cheska Robinson. 2018. Listserv Roundup: Guest Speakers and Mentors for Career Exploration in the Science Classroom. *Science Scope* 41, 8 (2018), 18–21.

[17] Peter Schaldenbrand, Nikki G Lobczowski, J Elizabeth Richey, Shivang Gupta, Elizabeth A McLaughlin, Adetunji Adeniran, and Kenneth R Koedinger. 2021. Computer-supported human mentoring for personalized and equitable math learning. In *International Conference on Artificial Intelligence in Education*. Springer, 308–313.

[18] Danielle Thomas, Xinyu Yang, Shivang Gupta, Adetunji Adeniran, Elizabeth Mclaughlin, and Kenneth Koedinger. 2023. When the tutor becomes the student: Design and evaluation of efficient scenario-based lessons for tutors. In *LAK23: 13th International Learning Analytics and Knowledge Conference*. 250–261.

[19] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[21] Xu Wang, Srinivasa Teja Talluri, Carolyn Rose, and Kenneth Koedinger. 2019. UpGrade: Sourcing student open-ended solutions to create scalable learning opportunities. In *Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale*. 1–10.

[22] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[23] Huseyin Yaratan and Rusen Yucesoylu. 2010. Self-esteem, self-concept, self-talk and significant others' statements in fifth grade students: Differences according to gender and school type. *Procedia-Social and Behavioral Sciences* 2, 2 (2010), 3506–3518.

[24] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the AI ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219* (2023).

[25] Nikos Zourbanos, Athanasios Papaioannou, Evaggelia Argyropoulou, and Antonis Hatzigeorgiadis. 2014. Achievement goals and self-talk in physical education: The moderating role of perceived competence. *Motivation and Emotion* 38 (2014), 235–251.

Received 16 April 2024; accepted 9 May 2024