Proceedings of Machine Learning Research vol 257:77–84, 2024 AI for Education Workshop at AAAI 2024

# Using Large Language Models to Assess Tutors' Performance in Reacting to Students Making Math Errors

**Sanjit Kakarla** SANJIT.KAKARLA@GMAIL.COM
**Danielle R. Thomas** DRTHOMAS@CMU.EDU
**Jionghao Lin** JIONGHAO@CMU.EDU
**Shivang Gupta** SHIVANG@CMU.EDU
**Kenneth R. Koedinger** KOEDINGER@CMU.EDU
*Human-Computer Interaction Institute*
*Carnegie Mellon University*
*5000 Forbes Ave.*
*Pittsburgh, PA 15213, USA*

## Abstract

Research suggests that tutors should adopt a strategic approach when addressing math errors made by low-efficacy students. Rather than drawing direct attention to the error, tutors should guide the students to identify and correct their mistakes on their own. While tutor lessons have introduced this pedagogical skill, human evaluation of tutors applying this strategy is arduous and time-consuming. Large language models (LLMs) show promise in providing real-time assessment to tutors during their actual tutoring sessions, yet little is known regarding their accuracy in this context. In this study, we investigate the capacity of generative AI to evaluate real-life tutors' performance in responding to students making math errors. By analyzing 50 real-life tutoring dialogues, we find both GPT-3.5-Turbo and GPT-4 demonstrate proficiency in assessing the criteria related to reacting to students making errors. However, both models exhibit limitations in recognizing instances where the student made an error. Notably, GPT-4 tends to overidentify instances of students making errors, often attributing student uncertainty or inferring potential errors where human evaluators did not. Future work will focus on enhancing generalizability by assessing a larger dataset of dialogues and evaluating transfer learning. Specifically, we will analyze the performance of tutors in real-life scenarios when responding to students' math errors before and after lesson completion on this crucial tutoring skill.

**Keywords:** tutoring; tutor evaluation; real-time feedback; math learning; LLMs; GPT-4

## 1. Introduction

Personalized tutoring remains a consistently effective academic intervention, significantly benefiting student learning (Kraft and Falken, 2021; Reynolds and McDonell, 2021). However, the scarcity of human tutors and lack of essential skills among those available pose a challenge (Thomas et al., 2023). Recent advancements in pre-trained large language models (LLMs) offer promise in real-time assessment of tutor performance (Chen et al., 2022). This present work investigates generative AI's capacity to evaluate tutors' effectiveness in addressing students' making math errors. Traditionally, intelligent tutoring systems employing knowledge tracing methods have a reputation for swiftly and directly intervening when students make mistakes (Merrill et al., 1992, 1995; Mathan and Koedinger, 2018). In

contrast, expert human tutors have demonstrated success through subtler, indirect guidance, enabling students to find their own errors and repair them (Lepper and Woolverton, 2002). Exploring the assessment of this specific tutoring strategy using LLMs becomes crucial, given its importance in fostering independent error correction among students.

Recent advances in LLMs offer a host of possibilities aimed to enhance learning, including furnishing explanatory feedback to learners (Dai et al., 2023), with prompt engineering of models leaning towards more of an art than a science (Wei et al., 2022). Among these, Generative Pre-trained Transformer (GPT) models, particularly the latest iteration, GPT-4, exhibits notable improvements over GPT-3.5. GPT-4 excels in tackling more complex tasks, learns faster, and demonstrates reduced potential for biased or offensive responses (OpenAI, 2023). However, it suffers from slower response and generation times, presenting a notable challenge in handling large transcriptions. Despite its advancements, GPT-4 comes at a considerably higher cost of $0.03/1K tokens compared to GPT-3.5, which costs $0.0015/1K tokens for input, marking a 20-fold increase in expense (OpenAI, 2023). Considering the speed and cost-effectiveness of GPT-3.5, our focus lies in evaluating its suitability for assessing tutor performance in practical, real-world tutoring settings. To this end, we aim to investigate the following research questions: **RQ1**: Can large language models accurately assess components of effective human tutors' responses to students making errors? **RQ2**: What is the comparative accuracy and performance between GPT-3.5-Turbo and GPT-4 in assessing tutoring dialogues for how tutors respond to students making errors?

## 2. Related work

Human tutors are particularly effective when trained on building relationships and fostering rapport with students (Marshall et al., 2021). The online tutor lesson *Reacting to Errors* (Appendix A) is an inspiration for this work (Thomas et al., 2023). In this brief, scenario-based lesson, tutors practice responding to a student who has made a math mistake. *Reacting to Errors* and the associated criteria for responding to the student are described.

### 2.1. The *Reacting to Errors* Lesson and Associated Criteria

In *Reacting to Errors*, a tutor is required to respond to a student's mistake in solving a math problem according to specified elements of appropriate responses as highlighted in research (Lepper and Woolverton, 2002; Loewenberg Ball and Forzani, 2009). These elements encompass praising the attempt or effort; subtly drawing attention to the mistake, and guiding the student toward self-correction (Lepper and Woolverton, 2002). Any response that explicitly points out the student's error, instructs the student on what exactly to do, or simply provides the correct answer is not desirable (Lepper et al., 1993). To assess a tutor's response, we establish five criteria in line with the research-recommended approach. The tutor's response should be: 1) process- or effort-focused, acknowledging the student's effort; 2) motivating, avoiding negative language and encouraging the student to recognize the mistake on their own; 3) indirect in addressing the error, using leading questions without negative connotations, such as using words like "mistake" or "error" that may discourage the student; 4) immediate, remaining relevant to the problem at hand; and 5) sincere and accurate, such as ensuring truthful praise to the situation (e.g., praising a student for working hard only when they actually put forth effort) and being mathematically correct.

Research suggests that students with high and low self-efficacy measures benefit differently from tutorial dialogue (Wiggins et al., 2017). Low-efficacy students may benefit more from an indirect and more subtle approach when reacting to errors than high-efficacy students, who may not be bothered by tutors calling direct attention to their mistakes. The proposed tutoring strategy is recommended for low-efficacy students, or students who historically struggle in math indicated by a history of low performance and proficiency.

### 2.2. Using Large Language Models (LLMs) to Assess Tutor Moves

Large Language Models (LLMs) are models trained on and exposed to a variety of information - almost all that exists on the Internet - using artificial neural networks as part of deep learning to process information and create outputs in written language familiar to human text. GPT (Generative Pre-trained Transformer) models, 3.5-Turbo and 4, are examples of LLMs; GPT-3.5-Turbo only accepts text inputs while GPT-4 accepts text alongside image inputs (Espejel et al., 2023). An area that captivates researchers is whether these LLMs have the capability to assess particular criteria or performance on highly nuanced and humanistic interactions, such as coaching teachers (Wang and Demszky, 2023) and providing feedback to educators and students (Dai et al., 2023; Kupor et al., 2023). Past work analyzed the potential of LLMs to guide math tutors in remediating student errors, concluding the best-performing model falls short compared to skilled math teachers (Wang et al., 2023). We expand on past work by: analyzing the ability of generative AI to recognize if an error has been made by the student and, if so, the tutor's response to the student; and comparing different GPT models to determine cost effectiveness at scale. The purpose of our investigation is to understand if GPT models can assess tutor responses to students. When using LLMs, prompt engineering, or the field of carefully and tactically crafting prompts for generative AI, has a huge impact on the nature of the responses generated (Wei et al., 2022). Prompt engineering is a paradigm within itself, an area researchers do not fully comprehend, as features of effective prompts and prompting techniques in particular situations are unknown (Reynolds and McDonell, 2021). Prompting techniques such as few-shot prompting, where examples are provided, and zero-shot prompting, where no examples are given, are often employed and tested to gauge the model's responses to eventually develop a prompt that stimulates the model to generate the responses the user desires (Espejel et al., 2023).

## 3. Method

Initially, we focus on prompting GPT-4 for the creation of synthetic dialogues to calibrate human assessment and determine inter-rater reliability. Although we understand there is no exact substitute for original dialogues from actual tutor-student interactions, we use these synthetic dialogues to determine human graders' reliability in assessing tutor's feedback to errors. The synthetic dialogues serve as a proxy for ensuring consistency and reducing bias. We employ GPT-4 to generate 50 tutor-student dialogues prompting the LLM to provide a range in tutor performance. There were 156 words ($SD$ = 45.9) and 8.6 utterances ($SD$ = 2.7), on average, per dialogue. Appendix B displays the prompt used to generate the synthetic tutoring dialogues.

### 3.1. Human Grader's Identification of Criteria

Two human annotators with experience in tutoring assessed tutor performance. As a prerequisite to annotating, graders completed the *Reacting to Errors* lesson and referenced the annotation guide (Appendix C), containing explanations of criteria (Section 2.1): *process-focused*, acknowledging student effort; 2) *motivating*, encouraging the student to find their own mistake; 3) *indirect*, not calling direction attention to the student's error; 4) *immediate*; and 5) *accurate*, mathematically correct and sincere. If the student did not make a math error, graders coded the dialogue as *no error*. Inter-rater reliability between the two annotators is shown in Table 1 (Wan et al., 2015). Appendix D illustrates a synthetic tutoring dialogue that scored five points by both graders.

Table 1: Agreement Among Human Graders

| Criteria | Agreement Score | Cohen's Kappa | Interpretation |
|---|---|---|---|
| *process-focused* | 68.8% | 0.59 | moderate |
| *motivating* | 79.2% | 0.57 | moderate |
| *indirect* | 73.9% | 0.50 | moderate |
| *immediate* | 100.0% | 1.00 | perfect |
| *accurate* | 91.7% | 0.54 | moderate |
| *no error* | 100.0% | 1.00 | perfect |

### 3.2. Corpus Description, Data Pre-Processing, & Prompting GPT

The corpus consists of an unknown number of online tutors (a tutor could be represented in more than one dialogue) who were college students at a Pennsylvanian university. The students were middle school students, ranging from 6th-8th grade, from two schools. The student-level demographics represented in the corpus are unknown, however, the school-level demographics consisted of 52% Latinx from a California public school and 100% Black and male from a Pennsylvania charter school. Math proficiency is low at both schools, with one school at zero percent proficiency suggesting the majority of students have low self-efficacy in learning math. Tutoring was performed remotely using Pencil as a remote communication platform with audio recordings transcribed within the platform. Individual dialogue recordings ranged in size from 100 bytes to 37KB. Transcriptions between 2KB and 8KB were used to provide sufficient utterances to assess dialogue while not overloading and slowing down the processing of the GPT models. In addition, we strive to keep costs low, particularly when scaling to more transcriptions. For all transcriptions, the tutor was the first utterance in the dialogue, which was used as a guide in diarization. We prompt GPT-3.5-Turbo and GPT-4 using the prompt created (shown in Appendix E), with the temperature at 0. Running the prompt on 50 real-life tutoring dialogues, we report the absolute performance of each model compared to a human grader. Appendix E displays the prompt used to assess real-life dialogues.

## 4. Results and Discussion

**Large language models demonstrate proficiency in identifying criteria on how to best respond to students making an error, however, both models exhibit limitations in recognizing if an error was made.** In responding to RQ1 on the ability of LLMs to effectively identify criteria for tutors reacting to students making errors, both GPT-3.5-Turbo and GPT-4 performed fairly well. Table 2 displays the absolute performance of both models in assessing the criteria. GPT-3.5-Turbo and GPT-4 both did particularly well on the criteria of *immediate* and *accurate*, with F1 scores equal to or greater than 0.80 for both these criteria. We posit that these models were able to achieve high levels of accuracy in identifying these criteria as *immediate* was quite straightforward to recognize. The tutor typically discussed situations relevant to the problem, and the *accurate* criteria could be graded by simply ensuring well-established mathematical principles were followed. For the *process-focused* and *indirect* criteria, both models struggled a bit more, as shown by lower F1 scores relative to the rest of the criteria. We believe the models interpreted *indirect* feedback to a student making an error as primarily focused on not giving away the correct response, allowing the terms "mistake" and "error" (e.g., a tutor saying, *"You're close, but there may be a small error"*), which are phrases our human graders found as discouraging in a tutoring session. The annotation guide for human graders (Appendix C) states, "The tutor avoids the use of words such as *mistake* or *error*." However, the few-shot LLM prompt does not explicitly state tutors should avoid such terms and only provides examples of met criteria, such as "*You have the right idea*" and "*Explain to me what you did here.*"

Both models encountered challenges in discerning when a student had made an error. Specifically, GPT-3.5-Turbo accurately identified dialogues where the student had not made an error 54% of the time, while GPT-4 exhibited slightly better performance at 63%. GPT-4 had 23 usable responses for assessing absolute performance while GPT-3.5-Turbo yielded 17 responses. To evaluate the models' absolute performance in assessing tutors against the five criteria, calculations were carried out exclusively on the dialogues where both the GPT model and human grader concurred that an error had indeed been made by the student.

Table 2: Performance Comparison of GPT-3.5-turbo and GPT-4

| | `GPT-3.5-turbo` | | | `GPT-4` | | |
|---|---|---|---|---|---|---|
| **Criteria** | Precision | Recall | F1 Score | Precision | Recall | F1 Score |
| `process-focused` | 0.46 | 0.67 | 0.55 | 0.57 | 0.73 | 0.64 |
| `motivating` | 0.64 | 0.82 | 0.72 | 0.71 | 0.75 | 0.73 |
| `indirect` | 0.58 | 0.64 | 0.61 | 0.59 | 0.67 | 0.63 |
| `immediate` | 0.74 | 0.88 | 0.80 | 1.0 | 0.87 | 0.93 |
| `accurate` | 0.74 | 0.88 | 0.80 | 1.0 | 0.91 | 0.95 |

**GPT-4 performed better in assessing all criteria compared to GPT-3.5-Turbo.** In response to RQ2, the GPT-4 model saw slightly greater performance levels for all five criteria. For criteria such as `immediate` and `accurate`, there was an F1 score difference of around 0.10, indicating that GPT-4 was able to more accurately assess the more straightforward criteria. For the rest of the criteria, both models have very similar

F1 scores. This raises an important question: is the substantial 20-fold increase in the cost of GPT-4 justified by its enhanced performance compared to GPT-3.5-Turbo (OpenAI, 2023)? We believe the performance of **GPT-3.5-Turbo is sufficient for our purposes** and upon enhancing the prompt for identifying when an error was made by the student and `process-focused` and `indirect` criteria, we can improve performance. However, despite future improvements, we are unsure if reliance on LLMs alone to evaluate and guide tutors is sufficient, with past work concluding similar results when using LLMs to provide step-by-step remediation of math errors (Wang et al., 2023).

**GPT-4 shows evidence of inferring and recognizing uncertainty.** For given dialogues where the human indicated no student error was made and GPT-4 said there was an error, we asked GPT-4, *"Point out the area of text where the student made an error."* GPT-4's notable responses included *"The error made by the student is not explicitly mentioned... However, it can be inferred from the dialogues... where the student expresses some uncertainty" and "The student is confused about how to rewrite a fraction... 'Divided by what?' showing uncertainty."* These responses show that GPT-4 has the capability of reasoning that the student is confused and has consequently likely made an error. The model also associated uncertainty around a problem as being an error, indicating that it is more likely to indicate situations as errors when compared to a human grader who is more prone to simply analyzing the situation without making any inferences while also understanding when exactly a student is feeling doubtful by understanding the context of the tutoring situation.

## 5. Limitations, Future Work, & Conclusion

This investigation contains several limitations. First, the dataset comprises a rather small number of dialogues. Increasing the number of dialogues will enhance the generalizability of our findings. Second, the real-life transcriptions we employ include fragmented conversations with many students inputting responses in the chat or providing non-verbal communication through video. We had the initial goal of keeping our costs low and running a single prompt to assess dialogues. Prompting the LLM to check if an error was not made before assessing the criteria may be beneficial. Third, the few-shot prompt had on average three correct examples of tutor responses per criteria. Including incorrect examples of tutor responses by criteria may clear up the ambiguity on what constitutes an error made by the student (e.g., a student displaying uncertainty). Future work consists of firstly analyzing more transcriptions. Second, we will prioritize the synchronization of chat and audio within real-life dialogues and employ more effective data-cleaning techniques. Third, we plan on employing LLMs to assess other specific tutoring skills (Chhabra et al., 2022) with past work demonstrating tutor learning (Thomas et al., 2023). Ultimately, assessing evidence of transfer learning from the completion of scenario-based lessons is our goal (Macaulay and Cree, 1999). For example, among tutors who completed the lesson *Reacting to Errors*, do we find evidence of tutors modifying their behavior after completing the lesson? Presently, we have more than 20 online tutor lessons aligned with specific tutoring skills such as *Responding to Negative Self-Talk* and *Determining What Students Know*. With new assessments of tutors, we intend to modify our prompts to fit a diverse array of tutoring settings. Despite the changed prompting methods in these instances, the criteria

warranting desired tutoring responses would remain unchanged. We would aim to test if chain-of-thought prompting and strategies such as Self-Reflection for LLMs would improve performance in assessment. Eventually, we would want to explore the capabilities of other non-GPT models such as Google Bard and llama-70b in this task to understand which LLM has optimal performance.

In this study, we investigated the capability of GPT-3.5-Turbo and GPT-4 to identify criteria for how tutors should respond to students making math errors. Our findings indicate that both LLMs perform adequately in assessing certain criteria, but they encounter challenges in accurately determining whether the student has made an error. GPT-4 exhibits a slight advantage over GPT-3.5-Turbo, demonstrating the ability to make inferences and judge uncertainty, however, GPT-3.5-Turbo appears to suffice for scalable dialogue assessment while maintaining cost-efficiency. Future research will involve using a larger dataset, exploring transfer learning, and applying this method across different tutoring skills.

**Acknowledgments**
This work was made possible with support from the Richard King Mellon Foundation (Grant No. 12126) and the Learning Engineering Virtual Institute. Any opinions, findings, and conclusions expressed in this material are those of the authors.

## Appendix A. Digital Appendix

https://tinyurl.com/bdfht7jp

## References

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*, 2022.

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

Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. Can large language models provide feedback to students? a case study on chatgpt. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 323–325. IEEE, 2023.

Jessica López Espejel, El Hassane Ettifouri, Mahaman Sanoussi Yahaya Alassan, El Mehdi Chouham, and Walid Dahhane. Gpt-3.5, gpt-4, or bard? evaluating llms reasoning ability in zero-shot setting and performance boosting through prompts. *Natural Language Processing Journal*, 5:100032, 2023.

Matthew A Kraft and Grace T Falken. A blueprint for scaling tutoring and mentoring across public schools. *AERA Open*, 7:23328584211042858, 2021.

Ashlee Kupor, Candice Morgan, and Dorottya Demszky. Measuring five accountable talk moves to improve instruction at scale. *arXiv preprint arXiv:2311.10749*, 2023.

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

MR Lepper, M Woolverton, D Mumme, and J Gurtner. Motivational techniques of expert human tutors: Lessons for the design of computer-based tutors. computers as cognitive tools. sp lajoie, & sj, derry, hillsdate, 1993.

Deborah Loewenberg Ball and Francesca M Forzani. The work of teaching and the challenge for teacher education. *Journal of teacher education*, 60(5):497–511, 2009.

Cathlin Macaulay and Viviene E Cree. Transfer of learning: Concept and process. *Social work education*, 18(2):183–194, 1999.

Lydia Marshall, Jonah Bury, Robert Wishart, Rebekka Hammelsbeck, and Emily Roberts. The national online tuition pilot. *Education Endowment Foundation. https://educationendowmentfoundation. org. uk/projects-and-evaluation/projects/online-tuition-pilot*, 2021.

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

Douglas C Merrill, Brian J Reiser, Michael Ranney, and J Gregory Trafton. Effective tutoring techniques: A comparison of human tutors and intelligent tutoring systems. *The Journal of the Learning Sciences*, 2 (3):277–305, 1992.

Douglas C Merrill, Brian J Reiser, Shannon K Merrill, and Shari Landes. Tutoring: Guided learning by doing. *Cognition and instruction*, 13(3):315–372, 1995.

OpenAI. OpenAI, 2023. URL https://www.openai.com/.

Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021.

Danielle Thomas, Xinyu Yang, Shivang Gupta, Adetunji Adeniran, Elizabeth Mclaughlin, and Kenneth Koedinger. 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*, pages 250–261, 2023.

Tang Wan, Hu Jun, Hui Zhang, Wu Pan, and He Hua. Kappa coefficient: a popular measure of rater agreement. *Shanghai archives of psychiatry*, 27(1):62, 2015.

Rose E Wang and Dorottya Demszky. Is chatgpt a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. *arXiv preprint arXiv:2306.03090*, 2023.

Rose E Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. Step-by-step remediation of students' mathematical mistakes. *arXiv preprint arXiv:2310.10648*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Joseph B Wiggins, Joseph F Grafsgaard, Kristy Elizabeth Boyer, Eric N Wiebe, and James C Lester. Do you think you can? the influence of student self-efficacy on the effectiveness of tutorial dialogue for computer science. *International Journal of Artificial Intelligence in Education*, 27:130–153, 2017.