Do Tutors Learn from Equity Training and Can Generative AI Assess It?

Danielle R Thomas
Carnegie Mellon University
Pittsburgh, PA, USA
dchine@andrew.cmu.edu

Conrad Borchers
Carnegie Mellon University
Pittsburgh, PA, USA
cborcher@cs.cmu.edu

Sanjit Kakarla
Carnegie Mellon University
Pittsburgh, PA, USA
sanjit.kakarla@gmail.com

Jionghao Lin
Carnegie Mellon University
Pittsburgh, PA, USA
jiongh.lin@gmail.com

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

Boyuan Guo
Carnegie Mellon University
Pittsburgh, PA, USA
boyuang@andrew.cmu.edu

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

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

Abstract

Equity is a core concern of learning analytics. However, applications that teach and assess equity skills, particularly at scale are lacking, often due to barriers in evaluating language. Advances in generative AI via large language models (LLMs) are being used in a wide range of applications, with this present work assessing its use in the equity domain. We evaluate tutor performance within an online lesson on enhancing tutors' skills when responding to students in potentially inequitable situations. We apply a mixed-method approach to analyze the performance of 81 undergraduate remote tutors. We find marginally significant learning gains with increases in tutors' self-reported confidence in their knowledge in responding to middle school students experiencing possible inequities from pretest to posttest. Both GPT-4o and GPT-4-turbo demonstrate proficiency in assessing tutors ability to *predict* and *explain* the best approach. Balancing performance, efficiency, and cost, we determine that few-shot learning using GPT-4o is the preferred model. This work makes available a dataset of lesson log data, tutor responses, rubrics for human annotation, and generative AI prompts. Future work involves leveling the difficulty among scenarios and enhancing LLM prompts for large-scale grading and assessment.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Applied computing** → **Computer-managed instruction**; • **Computing methodologies** → **Artificial intelligence**.

LAK 2025, Dublin, Ireland

ACM ISBN 979-8-4007-0701-8/25/03
https://doi.org/10.1145/3706468.3706531

Keywords

Tutor Training, Generative AI, Large Language Models, Assessment, Equity

ACM Reference Format:
Danielle R Thomas, Conrad Borchers, Sanjit Kakarla, Jionghao Lin, Shambhavi Bhushan, Boyuan Guo, Erin Gatz, and Kenneth R Koedinger. 2025. Do Tutors Learn from Equity Training and Can Generative AI Assess It?. In *LAK25: The 15th International Learning Analytics and Knowledge Conference (LAK 2025), March 03–07, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3706468.3706531

1 Introduction

Equity is a core concern of learning analytics (LA) [21], however, applications that teach and assess equity competencies are lacking. Intelligent tutoring systems, a core learning analytics application, have improved learning in STEM domains with well-defined rules for assessment. However, equity competencies are taught and assessed through language which are harder to grade, requiring advanced assessment methods. Advances in large language models (LLMs), which demonstrate text comprehension in various domains [6], could help bring the benefits that make tutoring systems effective (e.g., automated assessment and immediate feedback) to the equity domain. To this end, this work introduces the use of generative AI to evaluate human tutors' open-ended responses involving approaches to equity, a novel and under-researched LA application. In addition, we contribute a dataset of lesson log data, human annotation rubrics, and generative AI prompts to enhance transparency, reproducibility, and collaboration within the LAK community.

Equality within education refers to providing all students with the same resources and opportunities, regardless of circumstances and despite any inherent advantages or disadvantages. *Equity*, on the other hand, focuses on tailoring support to meet the individual needs of the student, recognizing that the background and challenges of each student are unique [11]. While equality treats every student alike, equity adjusts resources to ensure all students have

the chance to achieve similar success. Social justice education focuses on raising students' consciousness about inequity in everyday social and educational situations [10, 15]. A tutor telling a student, "if you work hard enough, you will be successful" may be great advice but only if the student has access, human support, and opportunities to engage with the learning materials. For example, a student without books at home may find it difficult to read when out of school. Tutors can help secondary students manage inequities in their learning by assisting them with recognizing possible inequities and supporting them in advocating for themselves–but how does a tutor go about providing students equity–based support? How can tutors learn and develop advocacy skills? This present work uses human coding and generative AI to evaluate tutor learning within a lesson to help students manage inequities.

The evaluation of open-ended responses within assessments by human graders is costly and time consuming [3]. Leveraging large language models to automatically assess tutors' open responses holds promise for transforming small-scale instructional activities into large-scale and personalized training programs [25, 34]. In this study, we discuss evaluating tutors' open-ended responses using LLMs. Previous research has explored the use of LLMs to assess tutor learning on skills related to social emotional learning, such as giving effective praise [16] and providing content support such as reacting to students making math errors [20]. Here, we leverage this method and adapt it to equity-focused tutor training. Ultimately, this present work uses generative AI, specifically the large language models GPT-4 and GPT-4o, to develop a method to assess open responses of tutors while participating in scenario-based training. We intend to develop a systematic, scalable approach to provide real-time assessment.

The present work is of great interest to the LAK community through the identification of evidence of learning skills and their assessment. Using lesson log data, we use open-ended responses and multiple-choice selections of the tutor to analyze and determine tutor learning gains. Adding data from self-reported tutor surveys, we determine the construct validity of equity-focused training and the perceptions of tutors about their learning. Expanding the LAK methodological toolbox, we leverage a novel use of generative AI allowing LLMs to assess open-ended responses of teachers in the equity domain. Currently, we know of very little past work using generative AI to assess tutor responses within scenario-based training. In this work, we address the following research questions:

RQ1: Is the scenario-based lesson effective in teaching tutors new skills for responding to students possibly experiencing inequities?

RQ2: How does tutors' self-reported confidence of their knowledge attending to students experiencing potential inequities change from pretest to posttest, and do tutors feel they can apply what they have learned?

RQ3: How effective are large language models GPT-4o and GPT-4-turbo in assessing tutors' actions in responding to students managing possible inequity?

RQ4: How do the large language models GPT-4o and GPT-4-turbo compare in performance, efficiency, and cost?

2 Related Work

2.1 The Role of Scenario-based Learning in Tutor Development

Scenario-based learning integrates educational activities within real-world contexts, promoting rapid development of skills through situational activities [2]. This instructional strategy has been successfully applied across multiple disciplines, such as medical education, fostering prospective thinking among high school students, and enhancing the growth of pre-service teachers (e.g., [1, 19, 28]). Furthermore, scenario-based models, including digital simulations, offer novice teachers and tutors valuable low-risk practice environments to gain situational experience [7, 8, 36]. Authentic scenarios, coming from real-life tutoring situations, are used by tutors to practice "learning by doing," which is an instructional approach that emphasizes active participation and hands-on practice to acquire knowledge and skills through direct experience. The intention is that this learning transfers to real-life tutoring environments. Learning by doing requires the application of skills that model the needs of the real world [23], facilitating the transfer of new knowledge to similar experiences that trigger recall [31].

Past studies have shown approximately 20% learning gain from pretest to posttest in scenario-based lessons covering tutor topics such as giving effective praise to students; reacting when a student makes an error; and determining what students know [34]. Both training and transfer scenarios (also known as pretest and posttest) follow a modified predict-observe-explain (POE) approach, theoretically connected to Gibbs' Reflective Cycle, a cyclical instructional model providing structure for learning by doing to individual learning experiences [12]. In line with Figure 1, tutors respond to a training scenario, or pretest, (that is, a student possibly experiencing an inequity) by asking them to *predict* how to best respond within 1) an open-ended question and 2) a multiple-choice question (MCQ). Then tutors *explain* their prediction via 3) an open-ended question and 4) a MCQ. Tutors then 5) observe the given research recommendation and receive feedback before they 6) *explain* the reasoning behind what they observed. Finally, the tutors complete the transfer scenario or posttest, following the same pattern of predicting the best responses and explaining their reasoning (7-10). The pretest and posttest each contain a maximum of four points (two MCQs and two open-ended questions), with tutor learning gains determined by subtracting tutor pretest score from the posttest score. This process provides a method for assessing the transfer of learning and, ultimately, the tutor's learning gain [7, 34].

2.2 Addressing Educational Inequities through Scenario-based Learning

Ensuring equity remains one of the greatest challenges in education today [11, 13]. For example, a common situation that educators and tutors face is a student who cannot complete an assignment or access instructional materials outside of school because they do not have the Internet at home. Among school-age children worldwide, an estimated two-thirds do not have access to the Internet at home [37]. Many students do not recognize such inequities when they experience them. A student without the internet at home tasked with completing an assignment that requires broadband access

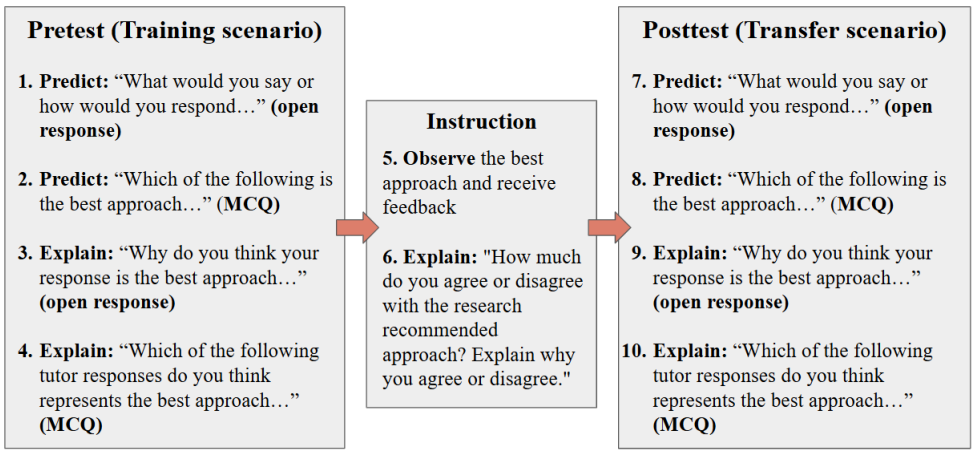

Figure 1: The modified predict-observe-explain cycle for the pretest and posttest scenarios.

will simply not do the assignment. They may not recognize the lack of access as inequitable and frankly unfair. Educators and tutors can play an important role in promoting equity. Rapid-cycle learning, in the form of short instructional activities for educators and tutors, is gaining traction to exercise applying strategies to advance equitable practices [11]. Rapid-cycle learning, exemplified by the situational judgment tests presented in this work, offers a method of providing instructional support to educators and tutors in attending to social justice and equity concerns [32]. Through brief, scenario-based learning activities, tutors participate in low-risk opportunities to pilot strategies to assist students with diverse experiences and needs, whether or not they are related to systemic inequities. Providing students with high-impact and personalized tutoring is becoming a potential solution to narrow the opportunity gap among underserved students [14]. However, there is a shortage of experienced tutors [24], along with limited training in supporting social justice and ensuring equitable learning environments. This work aims to evaluate the learning of the tutor from such activities on how to help middle school students manage potential inequitable situations.

2.3 The *Helping Students Manage Inequity* Lesson

One approach for promoting equity-responsive practices among secondary students is instructing students on collaborating with adults and advocating for themselves [13]. The lesson draws on previous research to identify the key competencies of effective tutoring, with scenario-based learning activities developed to align with key competencies [34]. We strive to determine tutor learning gains similar to [34] while focusing specifically on how tutors respond to students experiencing needs, potentially indicative of inequities. The lesson objectives include: recognizing when a student may be experiencing inequity related to their learning; and applying strategies to help students manage inequities by assisting students to advocate for themselves.

In one of two scenarios (used interchangeably as a pretest or posttest), student Jeremiah could not complete his homework because he did not have access to the Internet at home. In this specific situation, the tutor does not know if the student did not have the internet for temporary reasons, such as a random system outage, or if the lack of internet access is associated with a systemic disparity affecting the student's ability to succeed, such as socioeconomic status. However, the research-recommended approach of the tutor is to help Jeremiah recognize his need and empower him to advocate for himself by exercising his voice [13]. Figure 2 illustrates the scenario involving student Jeremiah showing the open-ended question asking tutors to *predict* the best approach (shown), which is followed by a selected-response question with options for tutors to choose how they would respond (not shown). The tutors are then asked to *explain* why they chose their choice within an open response and multiple-choice questions.

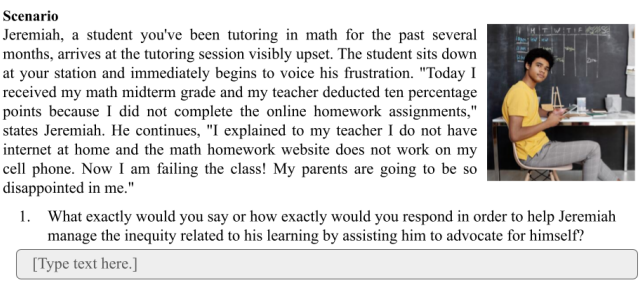

Figure 2: The scenario involving student Jeremiah with the open-ended question prompting a tutor to predict the best approach.

Similarly, an analogous scenario details the situation of Alexis shown in Figure 3, who expresses to her tutor that she earned a bad grade on a math assignment because she sits in the back of the classroom and cannot hear. Both the Jeremiah and Alexis scenarios and all questions are available in the Digital Appendix. Alexis seating issue, such as Jeremiah's lack of Internet access, might seem minor,

but could indicate deeper systemic inequities in education, such as inadequate support for diverse learning needs or accessibility challenges. Recognizing and addressing these issues is vital for tutors, as it involves understanding broader educational barriers and developing inclusive strategies to support every student's success. This approach represents a fundamental step towards achieving educational equity, ensuring that all students receive the necessary opportunities and support to thrive, regardless of their background or circumstances.

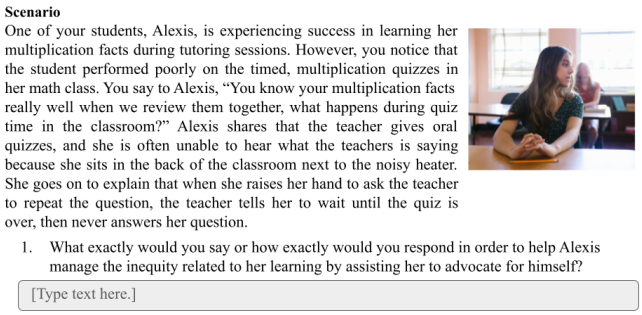

Figure 3: The scenario involving student Alexis with the open-ended question prompting a tutor to predict the best approach.

2.4 Benefits of Multiple-choice and Open-Ended Responses

Multiple-choice questions (MCQs) are a type of closed-response question often used in assessments due to their efficiency and objective grading [5]. However, they come with challenges such as encouraging reliance on test-taking strategies, issues with face validity, and difficulty in generating high-quality distractors or "incorrect" options [5]. In contrast, open-ended questions challenge students to develop their own responses, most often through textual language that reduces the influence of guessing, offer improved face validity, and can require higher-order thinking skills more readily [5]. Despite these advantages, grading open-ended responses is resource intensive (e.g., human graders often need to assess learner responses and provide feedback), making them less practical for large-scale use [3]. Recent advances in automatic short answer grading (ASAG) have shown potential in automating this process, but these models have historically faced challenges when used in different educational settings. Subtle differences between tasks (e.g., tutor training in promoting advocacy and ensuring equity in nuanced social situations) can significantly affect how well the model performs in diverse educational contexts [39].

2.5 Using Generative AI to Assess Open-ended Tutor Responses

Early ASAG approaches for assessing open-ended responses relied on traditional machine learning techniques, such as feature extraction and bag-of-words models, which provided easily interpretable results [18, 27, 39]. However, these early models struggled with domain shifts, where subtle differences in assessment tasks significantly impaired their performance [17]. More recently, studies have turned to deep learning models for automated assessment of open-ended responses [9, 33, 40]. For example, [9] explored ASAG using Sentence-BERT (SBERT) to measure textual similarity within learner response grading. Although SBERT showed promise for generalization, it encountered difficulties with unseen questions, revealing the need for models capable of deeper contextual understanding and adaptability, particularly in more complex domains such as equity training.

Recent advances in generative AI, particularly large language models (LLMs) such as GPT-4, have revolutionized ASAG by enabling models trained on extensive datasets to interpret and assess nuanced textual responses [6], with applications in tutor training [16, 20, 26]. LLMs offer flexibility, as they can be fine-tuned for specific educational tasks or used in their pre-trained form. Studies show that, with effective prompting, LLMs can capture the subtleties in learner responses, improving grading efficiency and scalability [17]. Nevertheless, challenges persist, as LLMs operate as "black-box" models, making their outputs difficult to interpret [6]. Furthermore, LLMs lack consistent knowledge of the pedagogy and content and are prone to hallucinations, generating confidently inaccurate or nonsensical information [39]. There remains an urgent need to explore how LLMs can be trained to assess equity-focused tutor responses, such as recognizing when tutors advocate for underserved students or support them in addressing inequities. These advancements could significantly improve tutor training to promote equitable education.

3 Method

3.1 Tutor Participants, Lesson Delivery, and Construct Validity

There were 81 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, the tutors reported an experience level of 3.2 (SD = 1.22). Furthermore, measures of tutor self-perceptions of their confidence in the topic were surveyed before pretest and after posttest using a 5-point Likert scale ranging from 1 (*not at all confident*) to 5 (*extremely confident*). After the posttest, tutors were surveyed regarding the application of the lesson content using a 5-point Likert scale (1- *strongly disagree*; 5- *strongly agree*): *I am confident I can apply what I learned.* The lesson was crafted by a university research team specializing in learning science, in collaboration with an equity-focused consulting firm, enhancing construct validity. The lesson was delivered through an online tutoring platform and is consistent with the research-shown competencies of effective tutoring within the area of *Advocacy* [30, 34]. We prioritized maintaining the privacy and confidentiality of tutors, adhering to all Institutional Review Board (IRB) requirements.

3.2 Human Open-ended Response Coding

Open-ended questions were coded, with "correct" tutor responses designated as "1" and "incorrect" as "0." Two experienced researchers coded participant responses to assess interrater reliability. The responses were deemed correct if the tutors demonstrated understanding of the lesson objectives and the research-recommended approach. Tables 1 and 2 present responses sourced from the learners with a rationale for the coding to predict and explain the responses, respectively. For *predicting* the best response, or correct response (score = 1), tutors should apply the strategy of assisting the middle or high school student in encouraging them to advocate for themselves by talking directly with their teacher or person in charge. In contrast, incorrect responses (score = 0) involved the tutor suggesting that the student find a way to solve the problem on their own or made suggestions that do not directly involve the student advocating for themselves in attending to their needs. Incorrect responses also include responses where the tutor merely indicates to a student to speak with a teacher without encouraging active advocacy. For *explaining* their chosen approach, tutor responses must demonstrate that the tutor recognizes that the student needs support in advocating for themselves and encouraging the student to act.

Table 1: Learner-sourced responses for *predicting* the best approach with rationale. Utterances aligned with correct, or "desired," rationale are highlighted in green.

Tutor Response	Coding Rationale
Let's work together to come up with a solution and find a way to address this issue with your teacher, so you have a fair opportunity to succeed.	Correct (1): This response assists the student with recognizing possible inequity related to their learning and helps the student in advocating for themselves.
I would provide him with other possible ways to access the internet that helps him do the homework equally. For example, he could do the homework at school.	Incorrect (0): This response recommends to the student alternative approaches to solve the problem but does not directly promote student advocacy.
I am very sorry to hear this. I understand that you are upset and I am upset about it too. However, unfair things happen in life, so we should find ways to resolve them in our own situations.	Incorrect (0): Although this response validates the student's feelings, it does not help the student recognize possible inequity nor help the student advocate for themselves.

Table 2: Learner-sourced responses for *explaining* their chosen approach with rationale. Utterances aligned with correct, or "desired," rationale are highlighted in green.

Tutor Response	Coding Rationale
It will allow Alexis to go over the problem herself and allows her to feel empowered.	Correct (1): This response indicates the tutor recognizes the importance of the student advocating for themselves.
It encourages him to make a plan to try and fix the problem, rather than just encouraging him to "work harder" when hard work simply will not resolve the issue.	Correct (1): This response indicates the tutor recognizes the importance of the student advocating for themselves.
Actively trying to come up with a solution is more important than simply expressing sorrow.	Incorrect (0): This response does not indicate the tutor recognizes the importance of student advocacy but merely focuses on problem solving with student.

3.3 Inter-rater Reliability Among Human Graders

Human coders assessed the responses of all 81 tutors for the *predict* and *explain* open-ended responses. The results indicated a relatively high agreement in inter-rater reliability between the coders using the binary coding system (i.e., 0, 1). For responses requiring tutors to *predict* the best course of action, there was 89% agreement and a Cohen's κ of 0.75. For responses asking tutors to *explain* their rationale, there was an 87% agreement rate and a Cohen's κ of 0.73. Both reflect substantial agreement, supporting the reliability of the coding process.

3.4 Determining Tutor Learning Gains

We employed a mixed-effects ANOVA to examine the impact of lesson scenarios on tutor performance. The *scenarios* (i.e., Jeremiah or Alexis) served as the between-subjects factor, while time, specifically *pretest* and *posttest*, served as the within-subjects factor. Treating *scenario* as a fixed effect helps to determine if there is an imbalance in difficulty between the two scenarios, while considering the test *time* as a random effect accounts for variation within the subjects. The ANOVA was run on data of students who completed all training and test items, resulting in a reduced sample of 81 students. A split-half reliability analysis adjusting for a time factor between pre and post-test was used to determine the reliability of the employed test battery. Specifically, we correlated person parameters of Rasch models across all possible splits of items. This resulted in an average reliability of 0.489 across all assessment items, which is acceptable for a test including only eight assessment items, although lower than typical test batteries which feature more items [22].

3.5 Prompt Engineering to Evaluate Tutor Responses

Drawing from recent advancements in prompt engineering, particularly within the domain of tutor training [25, 35], we developed prompts to leverage GPT-4 and GPT-4o for evaluating the correctness of open-ended tutor responses. Model temperature was set to 0 to ensure the model operated deterministically, selecting the most likely next word (or token) based on the input, reducing variability. This approach leads to more predictable and generally more cautious responses. The output was limited to 300 tokens to avoid excessive verbosity. Using zero-shot and few-shot prompting approaches, we provided the models with examples to guide their assessment of the predict and explain responses. The creation of these prompts followed an iterative process, with several rounds of adjustments informed by feedback from initial model outputs. Table 3 and 4 demonstrate the specific few-shot learning prompts designed for these tasks.

We employed several prompt engineering strategies to enhance model performance, including chain-of-thought prompting [38] to encourage step-by-step reasoning by prompting the model to provide the rationale. We used few-shot prompting (Brown et al., 2020), by supplying the model with relevant examples of correct and incorrect tutor responses to guide its output. Last, we used contextual priming to help the model understand the task within a specific context (i.e., *"You are a tutor evaluator…"*). Together, these techniques enabled the models to assess tutor responses with greater accuracy and consistency, aligning outputs more closely with human evaluations and enhancing overall model performance.

Table 3: Complete prompt for GPT-4o (few shot) used for the task of assessing tutors in *predicting* the best approach.

SCORING_PROMPT_START = """

Please assess a tutor's response in a tutor training scenario involving a middle school student struggling to understand a math problem.
-if the tutor's response helps the student identify that they have a need and provides the student support on how to remedy, by dealing with the issue and promoting student advocacy, or the tutor assists the student in helping them talk with their teacher without providing an alternative solution or work around, score with a 1. Examples of responses scoring a 1 are: "Jeremiah can advocate for improved technology access"; and "Jeremiah, you are doing a good job trying to communicate with your teacher. As she doesn't care, I can provide you several ways to let her not treating you unfairly."
-if the tutor's response does not demonstrate that the tutor understands that the student needs support in advocating for themselves, score with a 0. Sample responses scoring a 0 include: "That would be the right approach because I am helping the student solve the problem rationally"; "This can let him feel supported and at the same time, gives him a solution of this problem"; "Because it teaches Alexis a possible way to avoid similar problems in the future; and "Discussing the plan with the student will provide them how to stand up for the inequity."
Response Start ---

"""

FORMAT_PROMPT =

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

Table 4: Complete prompt for GPT-4o (few shot) used for the task of assessing tutors in *explaining* the best approach.

SCORING_PROMPT_START = """

Please assess a tutor's response in a tutor training scenario involving a middle school student struggling to understand a math problem.
-if the tutor's response demonstrates that they recognize that the student needs support in advocating for themselves and encourages the student to act, score with a 1. Sample responses scoring a 1 include: "It practices critical hope. It encourages her to advocate for herself by speaking to the teacher about it, and I offer my own help to assist her"; "It prepares Alexis to solve the problem herself by practicing in a low stress environment. It would make her more confident to talk to her teacher and do better in class."
-if the tutor's response does not demonstrate that the tutor understands that the student needs support in advocating for themselves, score with a 0. Sample responses scoring a 0 include: "That would be the right approach because I am helping the student solve the problem rationally"; "This can let him feel supported and at the same time, gives him a solution of this problem"; "Because it teaches Alexis a possible way to avoid similar problems in the future; and "Discussing the plan with the student will provide them how to stand up for the inequity."
Response Start ---

"""

FORMAT_PROMPT =

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

4 Results

4.1 RQ1: Is the scenario-based lesson effective in teaching tutors new skills for responding to students possibly experiencing inequities?

The analysis revealed a marginally significant main effect of *time* on tutor performance, $F(1, 79) = 3.20$, $p = .078$, indicating an overall improvement in tutors' performance from pretest to posttest. This suggests a general learning effect or improved skills. This main effect was qualified by a significant interaction between *scenario* and *time*, $F(1, 79) = 4.31$, $p = .041$. This significant interaction implies that the degree of improvement from pretest to posttest varied between the two scenarios. There was no statistically significant main effect of the specific *scenario* on tutor performance, $F(1, 79) = 0.54$, $p = .465$, indicating that the difficulty level between the Jeremiah and Alexis scenarios did not differ significantly in terms of overall tutor performance. *Post-hoc contrasts* were conducted to examine differences between pretest and posttest scores across the two scenario orders based on marginal means. Shining light on the significant interaction, learning gains in the *Alexis:Jeremiah* scenario order were not statistically significant, $M = -0.01$, $SE = 0.05$, $t(79) = -0.02$, $p = .846$. However, learning gains were significant in the *Jeremiah:Alexis* scenario order, $M = 0.12$, $SE = 0.04$, $t(79) = 3.07$, $p = .001$. The differences in assessment scores between the two scenario order conditions are visualized in Figure 4.

4.2 RQ2: How does tutors' self-reported confidence of their knowledge attending to students experiencing potential inequities change from pretest to posttest, and do tutors feel they can apply what they learned?

We sought to determine if there was a change in tutors' self-reported confidence in their knowledge of the topic from pretest to posttest, as well as their perceived ability to apply what they learned. We compared tutors' Likert scale survey results collected prior to pretest and after posttest. Prior to beginning the lesson, tutors reported an average confidence level of 3.44 ($SD = 1.09$) on a scale ranging from 1 (*not at all confident*) to 5 (*extremely confident*). Following participation in the lesson, this confidence level increased

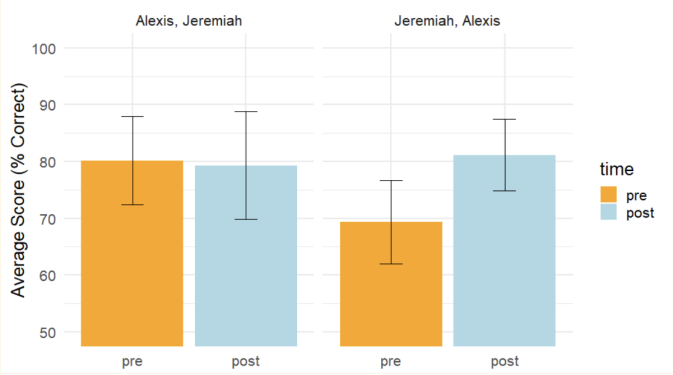

Figure 4: Mean pretest and posttest scores between scenario order conditions and measurement points.

to 4.51 (*SD* = 0.56). Among 81 tutors, 35 completed the post-lesson survey of confidence level. A paired sample *t*-test of those 35 tutors revealed a statistically significant difference between pretest and posttest confidence level (t(34) = 6.82, p < .001), indicating a reliable improvement in tutors' self-reported confidence in their knowledge attending to students experiencing potential inequities. Additionally, tutors were asked to rate their confidence in applying what they learned after completing the short lesson. The average self-reported score for this measure was 4.71 (*SD* = 0.46), indicating a high level of perceived ability to apply the acquired knowledge from the scenario to their own real-life tutoring.

4.3 RQ3: How effective are large language models in assessing tutor's actions in responding to students managing possible inequity?

GPT-4 and GPT-4o showcased proficiency in evaluating tutor's actions in responding to students managing inequity, with better performance on the open responses tasking tutors to *predict* the best approach compared to the open-response questions tasking tutors to *explain* the rationale behind their provided approach. Table 5 displays the absolute performance of GPT-4 and GPT-4o for *predict* and *explain* question types employing both zero- and few-shot prompting methods. Accuracy measures the proportion of correct predictions out of all predictions, providing an overall sense of a model's performance. AUC (Area Under the ROC Curve), often used with imbalance datasets, assesses how well the model distinguishes between classes, while the F1 score balances precision and recall, offering a measure of the model's effectiveness in handling both false positives and false negatives Given the same LLM, few-shot prompting outperformed zero-shot prompting. However, F1 scores for all models performed well ranging from 0.79 to 0.92.

Figure 5 illustrates the average open responses scores (2 pts total) at pretest and posttest by scenario for each of the LLM models compared to human graders. Aligning with performance measures from Table 5, few-shot learning models more closely aligned with human graders compared to zero-shot learning models.

Table 5: Absolute model performance for open response questions prompting the models to assess tutor's ability to predict and explain the best tutor response.

Model	Predict			Explain		
	Acc.	AUC	F1	Acc.	AUC	F1
`GPT-4o (zero-shot)`	0.87	0.81	0.91	0.71	0.68	0.79
`GPT-4o (few-shot)`	0.89	0.88	0.92	0.88	0.87	0.89
`GPT-4-turbo (zero-shot)`	0.85	0.90	0.79	0.80	0.78	0.83
`GPT-4-turbo (few-shot)`	0.89	0.87	0.92	0.89	0.89	0.90

4.4 RQ4: How do the large language models GPT-4o and GPT-4-turbo compare in performance, efficiency, and cost?

Both GPT-4o and GPT-4-turbo performed well. For educational purposes at scale, it becomes necessary to determine the practical balance of performance, efficiency, and cost. For it doesn't matter how well a model performs if it is not feasible due to practicality and cost. Table 6 provides a comparison of model performance, efficiency, and estimated cost for the assessment of 1,000 tutor completions. Input and output token length includes pretest and posttest for the assessment of all open responses.[1]

5 Discussion

This study investigated the efficacy of various instructional conditions on tutor learning, the impact of assessment methods on learning outcomes, and the role of generative AI in evaluating tutor performance. Several important insights emerged, which offer a comprehensive understanding of this present work.

5.1 Tutors demonstrate new learning on applying equity-focused skills, however, improved balance of scenario difficulty is needed.

Tutors displayed evidence of new learning when responding to students possibly experiencing inequities from pretest to posttest. However, the practice related to learning gains in both item order conditions included the same activities, hence differences in learning gains by scenario order condition can be attributed to differences in students or the assessments of both scenario order conditions. One hypothesis for why learning gains were significantly higher in one scenario-order condition is that one of the scenario batteries is systematically different from the other. Tutors who had the Jeremy scenario followed by the Alexis scenario demonstrated a 12% increase from the pretest to the posttest, respectively. This is not as high as previous work [34], however, equity training is a more nuanced and ill-defined domain. The Alexis test battery had a higher average student score at pretest than the Jeremiah test batter and was therefore easier (79. 9% compared to 72. 7%), although not significantly so based on a two-sided *t*-test, $t(12.87) = 0.88, p = .393$. Still, we tested the robustness of the main

[1]Input and output cost per token for GPT-4o ($5/1M and $15/1M) and GPT-4-turbo ($10/1M and $30/1M) as of Sept. 1, 2024 (https://openai.com/)

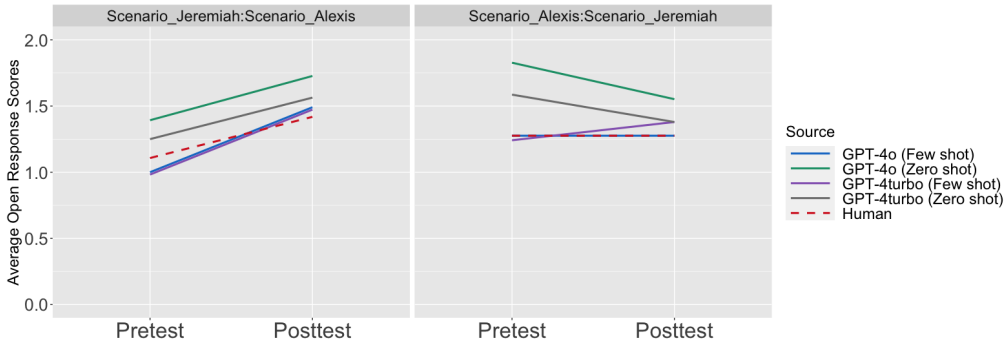

Figure 5: Average open response scores (2 pts total) at pretest and posttest by scenario for each LLM model and human graders.

Table 6: Comparison of estimated cost for 1,000 lesson completions (includes pretest and posttest for all open-ended questions)

Model	input tokens	input cost*	avg. output tokens	output cost*	total cost
GPT-4o (zero-shot)	696	$3.48	188	$2.82	$6.30
GPT-4o (few-shot)	1176	$5.88	198	$2.97	$8.85
GPT-4-turbo (zero-shot)	696	$6.96	271	$8.15	$15.11
GPT-4-turbo (few-shot)	1176	$11.76	282	$8.46	$20.36

*OpenAI API pricing as of Sept. 1, 2024 (https://openai.com/)

effect of time (i.e., learning gains) using an adjusted test score by applying z-score transformations within test-battery groups at pretest. That analysis consolidated a marginally significant learning gain between both scenario order conditions $F(1, 79) = 2.82, p = .097$. The interaction between scenario order condition and time remained robust after adjustment, $F(1, 79) = 3.96, p = .050$. Similarly, a Rasch model adjusting for individual item-difficulty and a fitted effect of time to test learning gains resulted in a marginally significant time effect representing learning gains, $\beta = 0.42, p = .066$.

5.2 Tutors reported improved confidence from pretest to posttest on applying equity-focused skills and are generally confident in applying what they have learned.

The results indicate a significant improvement in tutors' self-reported confidence from pretest to posttest, highlighting the effectiveness of the lesson in increasing their confidence of attending to students experiencing potential inequities. The statistically significant increase in confidence suggests that the lesson had a meaningful impact on tutors' perceptions of their own competence. Furthermore, the high average confidence in applying the acquired knowledge (4.71) suggests that tutors feel well prepared to transfer what they learned into practice. However, does the increase in self-perceived confidence in applying learned skills translate into actual learning gains from pretest to posttest? To explore the relationship between tutor self-reported confidence in applying the skills taught in the lesson and their individual learning gains, we found no statistically significant relationship between the two. One possibility is that our statistical tests were underpowered given the post-survey dropout, and the true effect of the lessons on learning small.

5.3 Overall, Generative AI performed well, but not perfect, on assessing tutor performance.

The results show that GPT-4 and GPT-4o are proficient at evaluating tutor responses, particularly in predicting the best approach when addressing student inequities. Few-shot prompting consistently outperformed zero-shot prompting across all metrics, demonstrating the value of providing examples to guide model responses [4]. Strong F1 scores in all models (ranging from 0.79 to 0.92) suggest generally good performance.

However, some notable challenges persist, particularly when evaluating responses that require subjective interpretation. Despite iteratively modifying the prompts, the models occasionally scored tutor responses differently from human graders. For instance, the following explanation was scored as incorrect (0) by human graders but coded as correct (1) by all models: *"It'll help Jeremiah learn to take agency over his life."* This response raises the question: Does the tutor's statement show that they recognize Jeremiah's need for support and encouragement in advocating for himself? According to GPT-4o's rationale, *"The tutor's response indicates that they recognize the importance of Jeremiah learning to take agency over his life, which implies encouraging him to advocate for himself."* This divergence in scoring highlights a fundamental difficulty in grading subjective responses. While the model's rationale is logical, it differs from the human graders' interpretation, potentially because human graders may consider broader contextual factors or expect more explicit language about advocacy and support. Alternatively, it's possible that the human grader's reasoning could be faulty, or influenced by biases or errors in judgment, which can happen in subjective evaluation scenarios. Determining the "source of truth" becomes especially challenging in such cases, as subjective responses often involve nuanced human judgment that LLMs, even

with advanced prompting techniques, may struggle to fully capture [9]. Moreover, human error can introduce inconsistencies, leading to disagreements between evaluators or between human and AI assessments. Therefore, developing more objective and transparent methods for determining the "source of truth" is advantageous.

5.4 Balancing performance, cost, and speed, GPT-4o with few-shot learning is the optimal choice.

Currently, GPT-4o with few-shot learning emerges as the preferred model when considering performance, cost, and speed, especially compared to GPT-4-turbo. While GPT-4-turbo offers only marginal performance improvements, it is nearly five times slower (20 tokens/sec vs. 109 tokens/sec for GPT-4o) and significantly more expensive [29]. Slow processing time would make it difficult to implement its use in providing automated assessment and immediate feedback to learners. Figure 6 displays the performance, cost, and processing time for each of the four models among 1,000 lesson completions. Using processing speeds and multiplying by the number of tokens (for simplicity combining input tokens and the average output tokens) for 1,000 lesson completions, we calculated the time it takes the models to assess tutor performance. GPT-4o using few-shot learning (represented by the green bubble in Figure 6) exhibits the most favorable balance of performance, cost, and processing speed. In comparison, GPT-4-turbo models are substantially more expensive and slower. Using GPT-4o (few-shot), assessing the performance of 1,000 tutors completing the lesson would cost $8.85 and take 3.5 hours. In contrast, What would it cost for humans to perform this same task? From our experience, human graders require around 15 seconds per response (1 minute per lesson for 4 open responses), resulting in 16.7 hours to assess 1,000 lessons. At a rate of $30 per hour for a skilled human grader, the cost would be $500. This comparison underscores the growing interest within learning analytics and the LAK community in leveraging generative AI for scalable, cost-effective grading and assessment.

6 Limitations

This study has several limitations that merit consideration. First, the sample size of 81 tutor participants, while providing some insights, remains relatively small, potentially influencing the observed differences in learning gains across item counterbalancing conditions. Although we ruled out item difficulty as a primary factor, the small sample size raises the possibility that randomization differences between students could contribute to the gain differences. Future process-to-gain analyses could provide further clarity on this issue. Additionally, the reliability of the assessment used was lower compared to larger test batteries, potentially limiting the ability to detect learning gains with high precision. Despite this, we did observe marginally significant learning outcomes based on the available data. A notable challenge was the disparity in difficulty between the scenarios used in the pretest and posttest; while we applied an easiness correction to address this, creating scenarios that are analogous in learning objectives yet different enough to capture genuine learning gains remains a complex task. Regarding self-reported tutor data, only 35 out of 81 tutors completed the post-lesson survey, which presents a limitation to the generalizability of the findings for RQ2. The relatively low response rate may introduce bias, as the results could reflect the experiences of a subset of tutors who were more or less engaged or had a more positive or negative experience with the lesson. Another limitation stems from the nuanced nature of the scenarios, making human coding difficult, particularly in the alignment between human and AI evaluations for explain responses. Despite the challenges of human coding, the inter-rater reliability was high. There were possible ceiling effects, particularly among the selected-response questions. Using high-frequency, learner-sourced responses that were deemed "incorrect," could be used as multiple-choice options [34]. Using high frequency and incorrect open responses as multiple-choice options may greatly increase lesson difficulty by capturing common misconceptions.

Only 35 out of 81 tutors completed the post-lesson survey, which presents a limitation to the generalizability of the findings for RQ2. The relatively low response rate may introduce bias, as the results could reflect the experiences of a subset of tutors who were more or less engaged or had a more positive or negative experience with the lesson. Another limitation stems from the nuanced nature of the scenarios, making human coding difficult, particularly in the alignment between human and AI evaluations for *explain* responses. Despite the challenges of human coding, the inter-rater reliability was high. There were possible ceiling effects, particularly among the selected-response questions. Using high-frequency, learner-sourced responses that were deemed "incorrect," could be used as selected-response options [34]. Using common and incorrect open-ended responses as selected response options may greatly increase lesson difficulty by capturing common misconceptions.

7 Future Work and Conclusion

Future work should address several key areas. First, increasing the complexity of scenarios, particularly the Alexis scenario, could provide deeper insights into tutor performance. Furthermore, research could investigate how increased self-reported confidence translates into real-world tutoring effectiveness and long-term student outcomes. Further studies should also explore the performance and variability of all combinations of few-shot prompts, rather than only reporting the best iteration, to better understand the impact of prompt design. Another important direction is determining the predictive validity of lessons and tutor learning transfer by analyzing real-life tutor-student interactions and transcription data. Finally, exploring more objective metrics, such as multiple-choice questions, as a potential "source of truth" could enhance the robustness of assessment methods.

In conclusion, this study demonstrates the potential of generative AI, particularly GPT-4o with few-shot prompting, as a valuable tool to assess the performance of the tutor in equity-focused online lessons. Using a mixed methods approach, we evaluated the performance of 81 undergraduate remote tutors, combining quantitative analysis of learning gains and self-reported confidence with the assessment of tutor responses. Tutors showed marginally significant learning gains and reported increased confidence from pretest to posttest in applying skills to address inequitable situations, underscoring the effectiveness of the training. While generative AI models

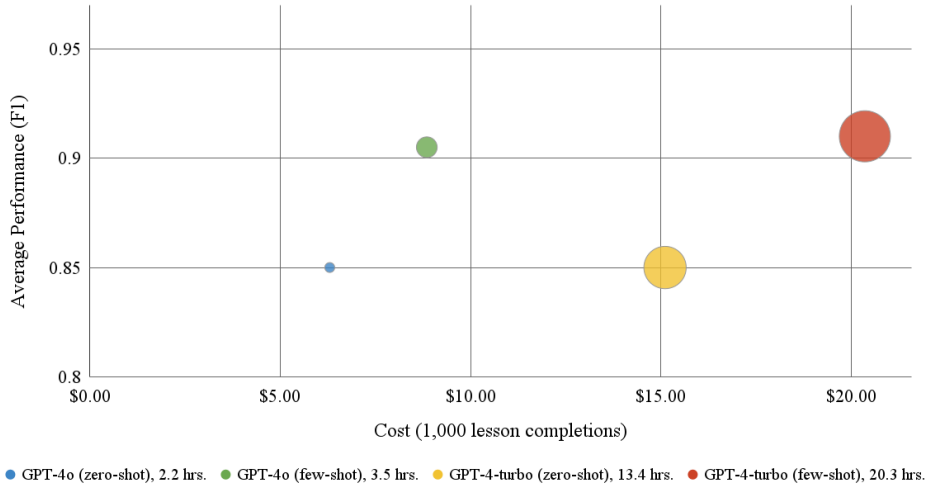

Figure 6: Average performance and cost of GPT models for 1,000 lesson completions. Bubble size is proportional to processing time.

performed well in evaluating tutor responses, challenges remain, particularly in improving alignment between scenario difficulty and model assessment accuracy. By balancing cost, performance, and speed, GPT-4o emerged as the most effective model for large-scale assessment. Future work should focus on refining scenario complexity and optimizing LLM prompts to enhance tutor training and AI-driven assessments. The release of the data set from this study offers a valuable resource for further research on equity-focused learning and assessment.

Acknowledgments

This work was made possible with the support of the Learning Engineering Virtual Institute. The opinions, findings and conclusions expressed in this material are those of the authors.

References

[1] Reem H Alattar. 2019. The effectiveness of using scenario-based learning strategy in developing EFL eleventh graders' speaking and prospective thinking skills. *The Islamic University of Gaza, Palestine* (2019).

[2] Lisa Bardach, Robert M Klassen, Tracy L Durksen, Jade V Rushby, Keiko CP Bostwick, and Lynn Sheridan. 2021. The power of feedback and reflection: Testing an online scenario-based learning intervention for student teachers. *Computers & Education* 169 (2021), 104194.

[3] Paul Black and Dylan Wiliam. 2009. Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of personnel evaluation in education)* 21 (2009), 5–31.

[4] T Brown, B Mann, N Ryder, M Subbiah, JD Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, et al. 2020. Language models are few-shot learners Advances in neural information processing systems 33. (2020).

[5] Andrew C Butler. 2018. Multiple-choice testing in education: Are the best practices for assessment also good for learning? *Journal of Applied Research in Memory and Cognition* 7, 3 (2018), 323–331.

[6] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology* 15, 3 (2024), 1–45.

[7] 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.

[8] Danielle R. Chine, Pallavi Chhabra, Adetunji Adeniran, Joseph Kopko, Cindy Tipper, Shivang Gupta, and Kenneth R. Koedinger. 2022. Scenario-based training and on-the-job support for equitable mentoring. In *Proceedings of The Learning Ideas Conference 2022*. Springer, Cham, Switzerland, 581–592.

[9] Aubrey Condor, Max Litster, and Zachary Pardos. 2021. Automatic Short Answer Grading with SBERT on Out-of-Sample Questions. *International Educational Data Mining Society* (2021).

[10] Jeffrey Duncan-Andrade. 2009. Note to educators: Hope required when growing roses in concrete. *Harvard educational review* 79, 2 (2009), 181–194.

[11] The Annie E. Casey Foundation. 2015. Race Equity and Inclusion Action Guide: Embracing Equity: 7 Steps to Advance and Embed Race Equity and Inclusion Within Your Organization. https://assets.aecf.org/m/resourcedoc/AECF_EmbracingEquity7Steps-2014.pdf Accessed: 2024-09-05.

[12] Graham Gibbs. 1988. Learning by doing: A guide to teaching and learning methods. *Further Education Unit* (1988).

[13] Taucia González, Kate M McCabe, and Carolina Lobo De Castro. 2017. An Equity Toolkit for Inclusive Schools: Centering Youth Voice in School Change. *Equity Assistance Center Region III, Midwest and Plains Equity Assistance Center* (2017).

[14] Jonathan Guryan, Jens Ludwig, Monica P Bhatt, Philip J Cook, Jonathan MV Davis, Kenneth Dodge, George Farkas, Roland G Fryer Jr, Susan Mayer, Harold Pollack, et al. 2023. Not too late: Improving academic outcomes among adolescents. *American Economic Review* 113, 3 (2023), 738–765.

[15] Zaretta Hammond. 2021. Liberatory education: Integrating the science of learning and culturally responsive practice. *American Educator* 45, 2 (2021), 4.

[16] Zifei FeiFei Han, Jionghao Lin, Ashish Gurung, Danielle Thomas, Eason Chen, Conrad Borchers, Shivang Gupta, and Ken Koedinger. 2024. Improving Assessment of Tutoring Practices using Retrieval-Augmented Generation. In *AI for Education: Bridging Innovation and Responsibility at the 38th AAAI Annual Conference on AI.*

[17] Owen Henkel, Libby Hills, Adam Boxer, Bill Roberts, and Zach Levonian. 2024. Can Large Language Models Make the Grade? An Empirical Study Evaluating LLMs Ability To Mark Short Answer Questions in K-12 Education. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*. 300–304.

[18] Wen-Juan Hou and Jia-Hao Tsao. 2011. AUTOMATIC ASSESSMENT OF STUDENTS'FREE-TEXT ANSWERS WITH DIFFERENT LEVELS. *International Journal on Artificial Intelligence Tools* 20, 02 (2011), 327–347.

[19] Cigdem Hursen and Funda Gezer Fasli. 2017. Investigating the efficiency of scenario based learning and reflective learning approaches in teacher education. *European Journal of Contemporary Education* 6, 2 (2017), 264–279.

[20] Sanjit Kakarla, Danielle Thomas, Jionghao Lin, Shivang Gupta, and Kenneth R Koedinger. 2024. Using large language models to assess tutors' performance in reacting to students making math errors. *arXiv preprint arXiv:2401.03238* (2024).

[21] Mohammad Khalil, Paul Prinsloo, and Sharon Slade. 2023. Fairness, trust, transparency, equity, and responsibility in learning analytics. *Journal of Learning Analytics* 10, 1 (2023), 1–7.

[22] Seonghoon Kim and Leonard S Feldt. 2010. The estimation of the IRT reliability coefficient and its lower and upper bounds, with comparisons to CTT reliability statistics. *Asia Pacific Education Review* 11 (2010), 179–188.
[23] Kenneth R Koedinger, Jihee Kim, Julianna Zhuxin Jia, Elizabeth A McLaughlin, and Norman L Bier. 2015. Learning is not a spectator sport: Doing is better than watching for learning from a MOOC. In *Proceedings of the second (2015) ACM conference on learning@ scale*. 111–120.
[24] Matthew A Kraft and Grace T Falken. 2021. A blueprint for scaling tutoring and mentoring across public schools. *Aera Open* 7 (2021), 23328584211042858.
[25] Jionghao Lin, Eason Chen, Zifei Han, Ashish Gurung, Danielle R. Thomas, Wei Tan, Ngoc Dang Nguyen, and Kenneth R. Koedinger. 2024. How Can I Improve? Using GPT to Highlight the Desired and Undesired Parts of Open-ended Responses. In *Proceedings of the 17th International Conference on Educational Data Mining*, Benjamin PaaÃŸen and Carrie Demmans Epp (Eds.). International Educational Data Mining Society, Atlanta, Georgia, USA, 236–250.
[26] 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. *International Journal of Artificial Intelligence in Education* (2024), 1–27.
[27] Nitin Madnani, Jill Burstein, John Sabatini, and Tenaha O'Reilly. 2013. Automated Scoring of Summary-Writing Tasks Designed to Measure Reading Comprehension. *Grantee Submission* (2013).
[28] Susan F McLean. 2016. Case-based learning and its application in medical and health-care fields: a review of worldwide literature. *Journal of medical education and curricular development* 3 (2016), JMECD–S20377.
[29] OpenAI. 2024. OpenAI API Pricing. https://openai.com/api/pricing/ Accessed: 2024-09-01.
[30] PLUS- Personalized Learning Squared. 2024. PLUS- Personalized Learning Squared. https://tutors.plus/ Accessed: 2024-09-05.
[31] Roger C Schank, Tamara R Berman, and Kimberli A Macpherson. 2013. Learning by doing. In *Instructional-design theories and models*. Routledge, 161–181.
[32] T. Shenbanjo, A. Buonaspina, A. Bhagwat, and S. Baumgartner. 2024. Championing Change: A Practitioner Guide for Leading Inclusive and Equity-Infused Rapid-Cycle Learning. Mathematica. https://www.mathematica.org/ Accessed: 2024-09-05.
[33] Hongye Tan, Chong Wang, Qinglong Duan, Yu Lu, Hu Zhang, and Ru Li. 2023. Automatic short answer grading by encoding student responses via a graph convolutional network. *Interactive Learning Environments* 31, 3 (2023), 1636–1650.
[34] 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.
[35] Danielle R Thomas, Jionghao Lin, Shambhavi Bhushan, Ralph Abboud, Erin Gatz, Shivang Gupta, and Kenneth R Koedinger. 2024. Learning and AI Evaluation of Tutors Responding to Students Engaging in Negative Self-Talk. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*. 481–485.
[36] Meredith Thompson, Kesiena Owho-Ovuakporie, Kevin Robinson, Yoon Jeon Kim, Rachel Slama, and Justin Reich. 2019. Teacher Moments: A digital simulation for preservice teachers to approximate parent–teacher conversations. *Journal of Digital Learning in Teacher Education* 35, 3 (2019), 144–164.
[37] Unicef et al. 2020. *How many children and young people have internet access at home?: estimating digital connectivity during the COVID-19 pandemic*. Technical Report. Unicef.
[38] 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.
[39] Mengxue Zhang, Sami Baral, Neil Heffernan, and Andrew Lan. 2022. Automatic short math answer grading via in-context meta-learning. *arXiv preprint arXiv:2205.15219* (2022).
[40] Yuan Zhang, Rajat Shah, and Min Chi. 2016. Deep Learning+ Student Modeling+ Clustering: A Recipe for Effective Automatic Short Answer Grading. *International Educational Data Mining Society* (2016).

A Digital Appendix

All analysis code, study materials, and log data references can be found in the study's supplementary GitHub repository: https://github.com/CMU-PLUS/LAK2025-Inequity