Does Multiple Choice Have a Future in the Age of Generative AI? A Posttest-only RCT

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
boyuan@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

The role of multiple-choice questions (MCQs) as effective learning tools has been debated in past research. While MCQs are widely used due to their ease in grading, open response questions are increasingly used for instruction, given advances in large language models (LLMs) for automated grading. This study evaluates MCQs effectiveness relative to open-response questions, both individually and in combination, on learning. These activities are embedded within six tutor lessons on advocacy. Using a posttest-only randomized control design, we compare the performance of 234 tutors (790 lesson completions) across three conditions: MCQ only, open response only, and a combination of both. We find no significant learning differences across conditions at posttest, but tutors in the MCQ condition took significantly less time to complete instruction. These findings suggest that MCQs are as effective, and more efficient, than open response tasks for learning when practice time is limited. To further enhance efficiency, we autograded open responses using GPT-4o and GPT-4-turbo. GPT models demonstrate proficiency for purposes of low-stakes assessment, though further research is needed for broader use. This study contributes a dataset of lesson log data, human annotation rubrics, and LLM prompts to promote transparency and reproducibility.

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/10.1145/3706468.3706530

Keywords

Tutoring, Generative AI, Human-AI tutoring, AI-assisted tutoring, Assessment

ACM Reference Format:
Danielle R Thomas, Conrad Borchers, Sanjit Kakarla, Jionghao Lin, Shambhavi Bhushan, Boyuan Guo, Erin Gatz, and Kenneth R Koedinger. 2025. Does Multiple Choice Have a Future in the Age of Generative AI? A Posttest-only RCT. 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/10.1145/3706468.3706530

1 Introduction

The effectiveness of multiple-choice questions (MCQs) in learning is the subject of much debate [3, 16, 18]. Although MCQs are often criticized for lack of depth, they remain a common feature in K-12 and higher education, due to their ease of grading [3]. However, their potential as instructional tools, rather than just assessment tools, meaning that they provide feedback from which students can learn, has received less attention. In contrast, open-response questions are frequently used in assignments such as homework, under the assumption that they promote deeper learning [3, 16]. However, open responses can be more time consuming for learners and resource intensive to grade [3], although recent advancements in the field have made the automated grading of these responses more feasible. This study evaluates the effectiveness of MCQs in relation to open-response questions, both individually and in combination, as learning-by-doing activities. These learning-by-doing activities are embedded in six tutoring lessons that involve advocacy training. To investigate scalability of autograding open-ended responses, we use generative AI to evaluate tutors' open responses. The contributions of this work are twofold: *theoretically*, it offers insights into the learning benefits of MCQs compared to open responses in learning-by-doing instruction; and *practically*, it provides implications for optimizing tutor training by determining the most efficient method of instructing and assessing tutors, as measured in the completion

time of instruction and accuracy of automated open-response grading. Furthermore, this study contributes a dataset of lesson log data, human annotation rubrics, and AI-generated generative commands to improve transparency, reproducibility, and collaboration within the learning analytics community.

The design of instructional materials plays a critical role in promoting effective learning. MCQs are often favored for their efficiency–they can be administered and graded quickly and require less time for learners to respond, making them appealing in large-scale settings [3, 16, 18]. However, MCQs are sometimes criticized for promoting surface-level learning, as they may encourage guessing and recognition rather than deeper understanding [3]. In contrast, open-response questions require tutors to construct their answers, thereby engaging in higher-order thinking and reflection. Although open-response questions can be powerful pedagogically, they are also more time-consuming to complete and evaluate [3]. A key question is whether MCQs can be designed to be as pedagogically effective as open-response questions within the context of teaching tutors advocacy skills, particularly in scenarios where scalability is a concern [16]. Advocacy, an emerging area of instruction aimed at improving equity and inclusion in tutoring, is particularly suited for a comparison of MCQ to open-ended responses, because the skills it requires–such as critical thinking and ethical reasoning–are potentially more effective for comprehension when practiced through open-ended formats [3] rather than MCQs where generating distractors can pose a challenge [13]. Compared to STEM learning, where many closed-form grading systems exist (e.g., tutoring systems), there is a need in learning analytics to study what forms of instruction are effective in novel, less structured domains like advocacy training.

Generative AI, particularly large language models (LLMs), have the capability to evaluate tutors' textual, open responses in real-time. LLMs such as GPT-4 [31], Claude [5], and LLaMa [37] have demonstrated remarkable performance in a variety of linguistic tasks. These modern LLMs are built on a large-scale transformer architecture and trained on extensive datasets [2, 14]. As a result, LLMs have attracted substantial interest from researchers across various fields, including education, because of their potential to perform reasoning tasks at scale and with reduced costs. Generative AI systems can evaluate human tutor responses across a wide range of scenarios, providing feedback and assessment at a scale that would be impossible for human evaluators alone. Importantly, LLMs may have the potential to make situational judgments, assessing not only the correctness of a response but also underlying reasoning [2]. This capability is crucial in scenario-based training, where tutors must navigate complex real-world situations. However, despite their potential, LLMs also have limitations, such as the tendency to generate nonsensical or factually incorrect outputs [42], and bias and fairness issues [14]. Using generative AI for tutor evaluation, this study explores the potential of AI to support large-scale, effective tutor training, ultimately improving tutor quality and accessibility.

This work addresses the need for effective and scalable tutor training by evaluating tutors' posttest performance across six scenario-based lessons focused on advocacy skills. Using a posttest-only randomized experimental design, we analyze tutor performance and time spent across three learning conditions: MCQ only, open-response questions only, and a combination of both. We evaluated the scalability of using generative AI to evaluate tutor responses, comparing the performance of GPT-4-turbo and GPT-4o with human graders. This study addresses the following research questions:

RQ1: What differences exist in tutor learning, as evidenced by posttest performance, across the learning-by-doing activities, i.e., MCQ only, open-response questions only, or both?

RQ2: In what contexts do MCQs, open-response questions, or a combination of both yield the highest accuracy and efficiency, thereby optimizing the impact of the lesson?

RQ3: How effective are LLMs, namely GPT-4o and GPT-4-turbo, in assessing tutors' open responses at posttest?

2 Related Work

2.1 Tutor Advocacy Skills and Scenario-based Training

Tutoring is widely recognized as one of the most effective interventions for improving student learning outcomes [17, 30, 32]. Research consistently shows that personalized support from skilled human tutors can significantly boost student academic performance, particularly among struggling students [34]. However, ensuring access to adequately trained tutors is challenging [6, 24], with many tutoring organizations relying on paraprofessionals. Many paraprofessional tutors have college education, but lack formal training and in providing instruction and building quality relationships with students [6, 30]. In addition, very limited instructional materials are available for tutors on attending to students' social-emotional needs. The process of training human tutors presents substantial scalability challenges, such as the need for human evaluators to assess tutor performance. Traditional methods of tutor training and evaluation are both time-consuming and resource-intensive, limiting the ability to scale tutoring programs to meet the needs of all students. Tutoring is more effective when delivered by teachers or well-trained professional tutors [30]. Currently, limited instructional materials are available to prepare and provide situational experiences to inexperienced tutors.

The lessons draw from previous research that identified impactful competencies of effective tutoring within the area of Advocacy [6, 9, 35]. Past studies have internally validated the construct validity by demonstrating 20% learning gain from pretest to posttest on similarly-structured lessons covering topics related to: giving effective praise to students; reacting when a student makes an error; and determining what students know [35]. Using the same scenario-based structure as in previous work [8, 9, 35, 36], our goal is to optimize tutor learning focusing on tutor lessons that instruct tutors in advocacy skills. There are very limited instructional and training materials available to tutors in the area of advocacy. Advocacy in teaching and tutoring encompasses a range of skills that promote student success by addressing their academic, social-emotional, and equity-related needs [6, 12]. Key areas include: promoting equity and inclusion; fostering cultural awareness; and challenging unconscious bias and assumptions [6, 35].

2.2 Learning Engineering and Instructional Design

The design of instructional materials is essential for effective tutor training, with multiple-choice questions (MCQs) often preferred for their efficiency in large-scale settings, although they may promote surface-level learning [3, 16, 18]. In contrast, open-response questions foster deeper reflection and higher-order thinking, but are more time-consuming, raising the question of whether MCQs can be made pedagogically effective to teach advocacy skills in scalable scenarios [3, 16]. This present work applies a learning engineering approach to investigate the learning efficiency of the following type of questions: open response, which encourages deeper cognitive engagement; MCQs which can provide structured assessment and objective grading; and a combination leveraging the strengths of both [3, 16]. Central to this present work is the "learn-by-doing" methodology, which emphasizes active participation in the learning process. This approach aligns with "doer" philosophy, advocating for hands-on, practical experiences to enhance understanding and retention [23]. An example in practice is the integration of computer-based Cognitive Tutors, whereby students are required not only to complete tasks but also to articulate (analogous to the ability of tutors to explain in this current work) their reasoning, reinforcing their comprehension and retention [1]. This dual emphasis on *doing* and *explaining* has been shown to significantly improve learning outcomes, fostering deeper comprehension and critical thinking skills [1, 35].

Brief scenario-based lessons were strategically designed using the learning-by-doing approach that provides actionable feedback and requires tutors to apply what they learned within the learning-by-doing conditions and the instruction phase by applying their learning in analogous tutoring situations at posttest. Fig. 1 illustrates the instructional design of the lesson. First, the tutors are presented with a scenario (Scenario 1), whereby they are prompted to *predict* the best approach, followed by being asked to *explain* their rationale or reasoning behind their chosen approach. There are three possible learning-by-doing conditions: multiple choice only, open response only, or both. Multiple choice questions begin with *"Which of the following…,"* followed by four options for the tutor to choose. Open response questions start with *"What would you say or how would you respond…"* for predicting the best approach and *"Why do you think your response is the best approach…"* for explaining the rationale behind their chosen approach. The tutors then engage in the instruction phase where the tutors observe the research-recommended approach and explain their reasoning in support or not of the best approach. Finally, the tutors complete a posttest, which is the same for all tutors and uses both MCQs and open responses. This instructional design is considered a modified predict-observe-explain (POE) approach and is theoretically related to Gibbs' Reflective Cycle, a cyclical instructional model that provides structure for learning by doing to individual learning experiences [15].

2.3 Using Generative AI to Evaluate Responses

Traditional automated assessment methods, such as natural language processing and traditional machine learning, often struggle to capture the complexity of open-ended responses due to their limited ability to understand context, subtle language variations, and complex semantics [10, 29]. Additionally, these models require extensive labeled training data to perform accurately, which can be time-consuming and costly to gather, especially in specialized domains like education. Automated assessment of open-ended responses is a significant task in learning analytics, as it can evaluate the quality of learners' understanding of various topics, such as student responses to college readiness [10]. Previous studies have used language models such as BERT [29] and Sentence-BERT [10] to develop automated evaluation systems for open-ended responses. Although these models have shown promising results, they have limitations in their ability to understand deeper contextual nuances and adapt to complex domains. These machine-learning models often struggle to fully capture the nuanced linguistic and semantic meanings in tutor responses.

In response to the limitations of traditional machine learning methods, recent advancements in generative AI, particularly LLMs, have demonstrated significant potential to assess open-ended responses. For example, [25] applied GPT-3.5 and GPT-4 to automatically score student-written responses in science assessments. They found that few-shot learning approaches significantly improved scoring accuracy, especially when paired with contextual item stems and scoring rubrics. Their study demonstrated that GPT-4 outperformed GPT-3.5, highlighting the potential of generative AI to provide more accurate and explainable automatic scores in educational contexts. [28] used few-shot prompting strategies with GPT-4 to assess the correctness of the responses of novice tutors in various tutoring strategies, such as giving praise, responding to student errors, and understanding student knowledge levels. [41] used GPT-4-turbo to evaluate the understanding of novice tutors of essential tutoring practices, including encouraging active student learning and fostering a respectful community. These studies illustrate the growing capacity of generative AI to offer more nuanced and contextually aware assessments of open-ended educational responses.

2.4 Prompt Engineering

Prompt engineering is a crucial technique in using LLMs to produce more accurate and contextually relevant output. The goal is to provide better context and structure in prompts to guide the LLMs toward the desired responses. Prompt engineering is particularly important when nuanced understanding and reliability are essential. One key technique in prompt engineering is few-shot prompting [2], where a set of exemplars is provided in the input prompt to demonstrate the ideal model behavior. By showing the model examples of the desired output format, few-shot prompting helps the LLM generalize to similar but unseen tasks. This approach has been used effectively in various automated assessment tasks, such as evaluating tutoring practice [19, 22] and student explanation on computer science questions [4]. Another approach is chain-of-thought (CoT) prompting [40], which instructs the model to "think step by step" by outlining intermediate reasoning steps. This technique helps the model handle more complex tasks by breaking the problem down into manageable parts, leading to more coherent and accurate responses. Such techniques are also used effectively to assess responses in educational settings [21, 28].

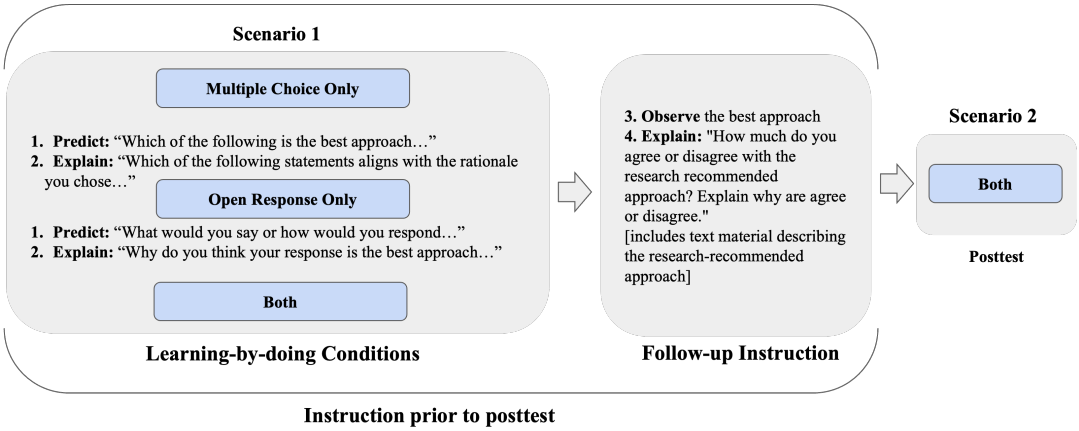

Figure 1: Instructional design sequence of the lessons illustrating the three learning-by-doing conditions, then the follow-up instruction phase, and concluding with posttest.

In addition to structuring prompts, techniques such as self-consistency [39] enhance the reliability of the model's outputs. Rather than modifying the prompt itself, self-consistency involves sampling multiple outputs and using a majority vote to determine the final response. This approach reduces the likelihood of hallucinations and improves the reliability of the outputs generated. For example, when evaluating open-ended responses, employing self-consistency can ensure that the model consistently interprets nuanced answers in a similar manner across multiple runs. Furthermore, prompting for rationale can help in obtaining more interpretable outputs. By explicitly asking the model to explain its reasoning or decision-making process, researchers can gain insight into how the model arrived at its conclusions, thus improving the transparency and interpretability of the assessment. This is particularly valuable in educational contexts, where understanding the reasoning behind a student's answer is as important as the answer itself. Overall, prompt engineering strategies such as few-shot prompting [2], chain-of-thought prompting [40], and self-consistency [39] play a vital role in enhancing the performance and reliability of LLMs in educational assessments. By carefully designing prompts and leveraging these techniques, we can guide LLMs to provide more accurate, nuanced, and consistent evaluations of open-ended educational responses.

3 Method

Six scenario-based lessons were created and designed to align with the tutoring competencies within the area of Advocacy [6]. The lesson content taken from the tutoring platform and formatted as documents can be found in Digital Appendix. The lesson titles and learning objectives are listed below.

- *Addressing Microaggressions*: define the term microaggression; identify microaggressions that occur in tutoring settings; and apply equity-focused strategies to help students address microaggressions.
- *Avoiding Unconscious Assumptions*: identify unconscious assumptions; and apply strategies to prevent making unconscious assumptions while tutoring.
- *Building Cultural Competence*: identify when students have different cultural backgrounds and experiences than your own; practice strategies to build cultural competence, supporting and engaging students across cultures.
- *Exploring Implicit Bias*: identify implicit, or conscious bias; and apply strategies to counter the effects of your own implicit biases.
- *Narrowing Opportunity Gaps*: define the term opportunity gap; identify examples of opportunity gaps in tutoring settings; and explain strategies to narrow opportunity gaps in tutoring settings.
- *Helping Students Manage Inequity*: recognize when a student is experiencing inequity related to their learning; and apply strategies to help students manage inequities by assisting students to advocate for themselves.

3.1 Tutor Participants & Lesson Delivery

There were 234 tutors, mainly college students, who completed any number of the six lessons for a total of 790 lesson completions. The tutors were undergraduate college students 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. Before starting a lesson, participants provided their levels of self-reported tutoring experience 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 3.56 (SD = 1.09) in all lessons. Tutors also self-reported their perceived knowledge of the lesson topic using a similar Likert scale with 1 (little to no knowledge of the topic) and 5 (expert level knowledge). In all lessons, the tutors reported an average knowledge of lesson topics of 3.59 (SD = 1.00). Table 1 displays the number of participants by condition with the average self-reported knowledge of the topic for each lesson. The advocacy lessons were developed in collaboration with tutoring supervisors, who are responsible for training tutors and working with students, and a university research team specializing in learning science, thereby enhancing construct validity. The lessons were delivered

via an online tutoring platform and align with the research-shown competencies of effective tutoring [6, 35]. For broad dissemination and use across varying tutoring organizations, all lessons have a Flesch-Kincaid readability index measure that ranges from grades 6-9 (defined as spanning from *easy to read* to *average reading level*). This ensures that all tutors, regardless of their individual reading level, can understand the content of the lesson. Each participant was randomly assigned one of three conditions in the learning-by-doing phase of the lesson: *multiple choice only*, *open response only*, or *both*. All participants received one of two randomly chosen scenarios used in the posttest for counterbalancing of scenario difficulty.

3.2 Human Open-Response Annotation and Inter-rater Reliability

Two experienced researchers scored participant responses to assess inter-rater reliability. Open-response questions tasking tutors to predict and explain the best approach were binary-coded. Correct responses (score = 1) align with the research-recommended approach of the lesson. Conversely, incorrect responses (score = 0) do not align with research-driven tutoring best practices. Human annotation rubrics for scoring tutor responses to *predict* and *explain* the best approach, along with learner-sourced examples of coded responses are located within the Digital Appendix. Two experienced researchers scored tutor responses to assess inter-rater reliability. Table 2 presents the inter-rater reliability between two experienced researchers for each lesson, assessing both predict and explain responses.

Tutor performance on the open-response questions ranged from 45% to 89% on the predict questions and 51% to 68% for the explain question types. To gain perspective on lesson and question difficulty, Table 3 displays the percentage of correct responses for each lesson by question type. Overall, tutor performance was lower on questions prompting tutors to explain the rationale behind their predictions of the best approach.

3.3 Prompting Large Language Models

Drawing on prior research in prompt engineering and large language models [2, 27], we developed a method for utilizing LLMs to evaluate the correctness of open-ended responses at posttest. Specifically, we implemented a few-shot learning approach, which has been shown to enhance performance in natural language understanding tasks [2]. In this method, the model was provided with a set of learner-generated responses along with human-scored examples to help it generate accurate assessments. The prompts were designed to assess two distinct types of responses: (1) "predict" responses, where the participants predicted the best course of action, and (2) "explain" responses, where they justified their decisions. Tables 4 and 5 present the specific prompts used to evaluate the predict and explain responses in the *Addressing Microaggressions lesson*, respectively. Prompts for all lessons and question types (i.e., predict and explain) are located within the Digital Appendix.

The development of these prompts was iterative, with multiple rounds of refinement informed by feedback on initial outputs. We employed various prompt engineering strategies including: content framing by providing context-specific details (e.g., *"You are a tutor evaluator…"*); randomly selecting human-coded examples of correct, partially correct, and incorrect tutor responses; and prompting the model to give rationale through chain-of-thought prompting [40]. To ensure deterministic outputs, the model's temperature was set to 0, guiding it to select the most probable word (or token) based on the input, thereby minimizing randomness and producing consistent, often more conservative responses. The output length was limited to 300 tokens to avoid unnecessary verbosity. These strategies allowed us to create an efficient and accurate method for assessing textual responses, with absolute performance metrics presented in the *Results*. The process also raised critical considerations for deploying LLMs in assessment contexts, which are further discussed in the *Discussion*.

3.4 Research Design & Analysis Plan

We employed a posttest-only randomized experimental design to evaluate tutor performance across three distinct conditions within the learning-by-doing phase: multiple-choice questions only, open-response questions only, or a combination of both. Participants were randomly assigned to one of these conditions to ensure that any differences observed in the posttest outcomes could be attributed to the specific question format rather than pre-existing differences among participants. A posttest-only design was chosen to avoid potential biases that might arise from pretesting, such as testing effects or sensitization [38]. Assessing tutor performance solely on the posttest scenario, we aimed to obtain a clear measure of the impact of the different learning-by-doing conditions on the tutor's subsequent performance [38]. Attending to **RQ1**, we employed an ANOVA to examine the impact of lesson, conditions, and scenario order on tutor performance, with all factors being between subjects. Regarding **RQ2**, we replicated the ANOVA used for RQ1 on the time it took students to complete the learning-by-doing and follow-up instruction in each condition. For answering **RQ3**, we used prompt engineering of large language models, GPT-4 and GPT-4o, to evaluate tutors' open responses. Then employed the same ANOVA model from RQ1 to determine if the results are synonymous with the analysis of human-graded tutor responses. We then report the absolute performance of both LLM models.

3.4.1 Lesson Log Data. Student responses to individual practice questions, survey questions, and other forms of instruction (e.g., responses to multiple-choice options, Likert scales, and open-ended responses) were recorded in PSLC DataShop, an open repository for educational log data commonly used for tutoring systems in learning analytics research [23]. Specifically, data was recorded in transaction format, which means that we analyzed individual interactions of students with timestamps. We prioritized maintaining the privacy and confidentiality of tutors, adhering to all Institutional Review Board (IRB) requirements. The lesson log data can be accessed within the Digital Appendix.

To measure performance on the posttest (RQ1), we aggregated the accuracy of student responses on the posttest, where students completed two multiple-choice and two open-response questions, with the latter graded by two experienced researchers and LLM models (RQ3). To measure the time students took to complete the instruction (RQ2), we calculated the difference between the instruction start time, as recorded in log data, and the last student

Table 1: Number of tutors by lesson for each condition with the average self-reported knowledge of the lesson topic (1-5).

Lesson	MCQ only (n)	Topic Knowledge	Open response only (n)	Topic Knowledge	Both (n)	Topic Knowledge
Addressing Microaggressions	39	3.55	42	3.42	41	3.97
Avoiding Unconscious Assumptions	34	3.68	37	4.07	39	3.49
Building Cultural Competence	41	3.86	28	3.84	51	3.42
Exploring Implicit Bias	38	3.79	41	3.54	39	3.66
Helping Students Manage Inequities	72	3.32	71	3.49	82	3.37
Narrowing Opportunity Gaps	37	3.73	30	3.68	28	3.64

Table 2: Inter-rater reliability between human evaluators for each lesson.

Lesson	*Predict responses*		*Explain responses*	
	Agreement (%)	Cohen's Kappa (κ)	Agreement (%)	Cohen's Kappa (κ)
Addressing Microaggressions	85.1%	0.70	94.7%	0.89
Avoiding Unconscious Assumptions	94.4%	0.88	94.4%	0.87
Building Cultural Competence	93.8%	0.73	87.8%	0.72
Exploring Implicit Bias	90.3%	0.55	84.0%	0.68
Helping Students Manage Inequities	96.7%	0.86	96.5%	0.93
Narrowing Opportunity Gaps	93.7%	0.86	96.2%	0.92

Table 3: Percentage of correct open responses for each lesson broken out by predict and explain responses.

Lesson	Correct (%)	
	Predict	*Explain*
Addressing Microaggressions	45.5%	51.8%
Avoiding Unconscious Assumptions	63.3%	68.5%
Building Cultural Competence	87.6%	67.4%
Exploring Implicit Bias	89.3%	56.4%
Helping Students Manage Inequities	86.4%	55.6%
Narrowing Opportunity Gaps	63.3%	58.2%

response to questions associated with each lesson's instruction. In cases where students completed lessons over multiple sessions with substantial breaks between them, we excluded the breaks from the total lesson completion time. Due to the right-skewed distribution of completion times, which are always greater than zero, we applied a logarithmic transformation for ANOVA and statistical tests that assume normally distributed outcomes. This assumption was confirmed in the logarithmic transformation data through visual inspection of standard diagnostic plots (e.g., residual Q-Q plots). However, for ease of interpretation, we re-transformed the averages and confidence intervals from the log scale back to the standard time scale (i.e., minutes) for presentation in plots.

4 Results

4.1 Learner Performance Across Conditions

The overall average results of the posttest are shown in Fig. 2. As is visually apparent, there is no consistent pattern illustrating that one instructional condition produces better learning outcomes than others across the conditions. Indeed, in an analysis of variance, we did not find a statistically significant main effect of condition, $F(2, 717) = 0.27$, $p = .765$. There was a significant interaction between condition and lesson $F\ (10, 717) = 2.20$, $p = .012$, which means that the posttest scores differed significantly by condition depending on what lesson the tutors completed. Furthermore, a significant main effect of the lesson suggested substantial accuracy differences by lesson, which means that some lessons were harder than others, on average, $F(5, 717) = 10.18$, $p < .001$.

We did *post hoc* contrasts using marginal means to estimate which condition-level differences within lessons were reliable [26]. Significant differences were found in only two lessons. In the *Addressing Microaggressions* lesson, the *Both* condition produced higher posttest scores than the *Open-response Only* condition (estimate = 0.136, SE = 0.054, $p = .031$) with the *MCQ Only* condition ambiguously in between (i.e., not statistically different from the other two conditions, p-values > 0.12). The results for the *Exploring Implicit Bias* lesson were essentially the opposite, consistent with the overall interaction. In the *Exploring Implicit Bias* lesson, the *Both* condition produced lower posttest scores than the *Open-response Only* condition (estimate = -0.175, SE = 0.054, $p = .004$). In this lesson, the *Both* condition also produced lower posttest scores than the *MCQ Only* condition (estimate = -0.134, SE = 0.056, $p = .046$). All other comparisons were not significant ($p > .299$).

4.2 RQ2: Optimizing Lesson Impact

There was no significant interaction between condition and lesson when comparing *instruction time prior to posttest* (See Fig. 1), $F(10, 716) = 13.46$, $p = .199$, indicating that the time taken did not significantly differ by condition depending on the lesson students completed. However, there were significant main effects of both condition and lesson. A significant main effect of condition suggested that the time taken differed substantially by condition, $F(2, 716) = 12.56$, $p < .001$, while a significant main effect of lesson indicated that some lessons took longer to complete than others, on average, $F(5, 716) = 8.61$, $p < .001$. Across lessons, total instruction time prior to posttest ranged between $M = 2.65$ minutes (*Avoiding Unconscious Assumptions*) to $M = 5.60$ minutes (*Helping Students*

Table 4: The LLM prompt for scoring *predict* responses from *Addressing Microaggressions.*

SCORING_PROMPT_START = """

You are a tutor evaluator. Please score the following tutor response to a tutor training scenario involving a middle school student facing a microaggression as follows:
-if the tutor's response acknowledges a student facing a microaggression by providing support to the student, by giving the student space to talk, or facilitating an apology, score with a 1. Sample responses scoring a 1 are "Nori, I heard what your teacher said about the exit ticket. It's okay to find problems challenging. Let's work on it together during our session. Your efforts are important, and I'm here to support you." and "Nori, I noticed what your teacher said to you could have been upsetting and I am here if you would like to talk about it."; and "Ashley, that was not a kind or respectful thing to say. Please apologize to your classmate."; "I would acknowledge Ashley's mean comment and tell her what she did wasn't right"; "Nori, I notice that you look sad on your face. Is it because what your teacher just said? Do you want to talk about it with me?"
-if the tutor's response does not acknowledge the microaggression or does not provide any support to the student, score with a 0. Yes or no questions, if they don't acknowledge the student's experience are scored with a 0. Sample responses scoring a 0 are "Nothing is too difficult if you decide you want to do it Nori. You are capable"; "Do you want to talk about what the teacher said to you"; "Nori, I heard what your teacher said to you about the exit ticket. Do you want to discuss it with me"; and "I would address the problem."
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 5: The LLM prompt for scoring *explain* responses from *Addressing Microaggressions.*

SCORING_PROMPT_START = """

You are a tutor evaluator. Please assess a tutor's response within a tutor training scenario involving a tutor instructing a middle school student who has faced a microaggression: The tutor is explaining the rationale behind their response. Assess and score the tutor's response, as follows:
-if the tutor's response demonstrates that they understand how to recognize and acknowledge a microaggression by providing the student support or issuing an apology, score with a 1. Sample responses scoring a 1 include: "Acknowledging the student's feelings and naming the microaggression, the teacher's comment, will provide an opportunity to address the microaggression"; "This approach will acknowledge the microaggression because it directly addresses Nori's feelings and opens up a supportive dialogue. By acknowledging that the teacher's comment may have been hurtful, it validates Nori's experience and gives her the opportunity to express her emotions."
-if the tutor's response does not demonstrate that the tutor recognizes how they should acknowledge microaggressions, score with a 0. Sample responses scoring a 0 include: "Telling the student she is capable of solving the problem will boost her confidence and addressing the problem will help to boost the students emotional status"; "It encourages the student to work on the problem"; and "This will provide her with a safe space to communicate."
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}."

Manage Inequity) (Fig. 3). Further, the *MCQ Only* condition took students the shortest (M = 3.83 minutes) while the *Both* condition took them the longest time (M = 5.87 minutes), although not substantially longer than the *Open-response Only* condition (M = 5.38 minutes). Hence, the *Both* condition took students less time than the sum of the open and MCQ conditions. Based on marginal mean comparisons, these differences were statistically significant, such that the *MCQ Only* condition took learners significantly shorter than the *Both* condition, estimate = −0.37, SE = 0.07, $p < .001$, and the *Open-response Only* condition, estimate = -0.30, SE = 0.07, $p < .001$. However, there was no significant difference in completion time between the *Open-response Only* and *Both* conditions, estimate = -0.07, SE = 0.07, $p = .642$.

Although the overall interaction between condition and lesson was not significant, we report marginal mean contrasts of conditions within lessons similar to RQ1. The *MCQ Only* condition took significantly less time than the *Both* condition in the *Building Cultural Competence* and *Narrowing Opportunity Gaps* (estimate = -0.55, SE = 0.19, $p = .011$). In the *Building Cultural Competence* lesson, *MCQ Only* was also significantly faster than the *Open-response Only* (estimate = -0.50, SE = 0.19, $p = .025$) and *Both* conditions (estimate = -0.76, SE = 0.16, $p < .001$). Notably, in no lesson was the *Open-response Only* condition significantly different from the *Both* condition (p-values > .269). Overall, these findings suggest that the *Both* condition took learners significantly longer than the *MCQ Only* condition in 2/6 lessons, but not longer than the *Open-response Only* condition.

4.3 Large Language Model Performance

The absolute performance of the LLMs was determined in assessing tutors' open response at posttest. Tables 6 and 7 display the comparison of absolute performance for GPT-4o and GPT-4-turbo, respectively, across lessons for *predict* and *explain* open responses. The LLM prompts used for each lesson can be accessed in the Digital Appendix. Both GPT-4o and GPT-4-turbo showcased proficiency in evaluating tutors' responses. Accuracy measures the proportion of correct predictions out of all predictions, providing an overall sense of a model's performance. *AUC* (Area Under the Curve), often used with imbalanced datasets, assesses how well the model distinguishes between classes, while the F_1 score balances precision and recall, offering a measure of the model's effectiveness in

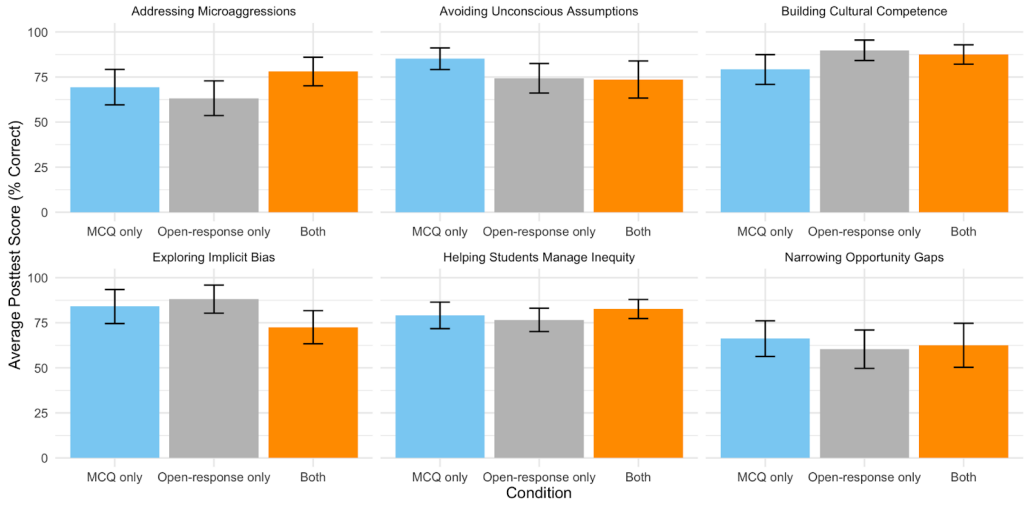

Figure 2: Average posttest scores compared across learning-by-doing conditions: *MCQ Only, Open-response Only,* or *Both*. No significant differences were found in posttest scores between conditions. Error bars represent 95% confidence intervals.

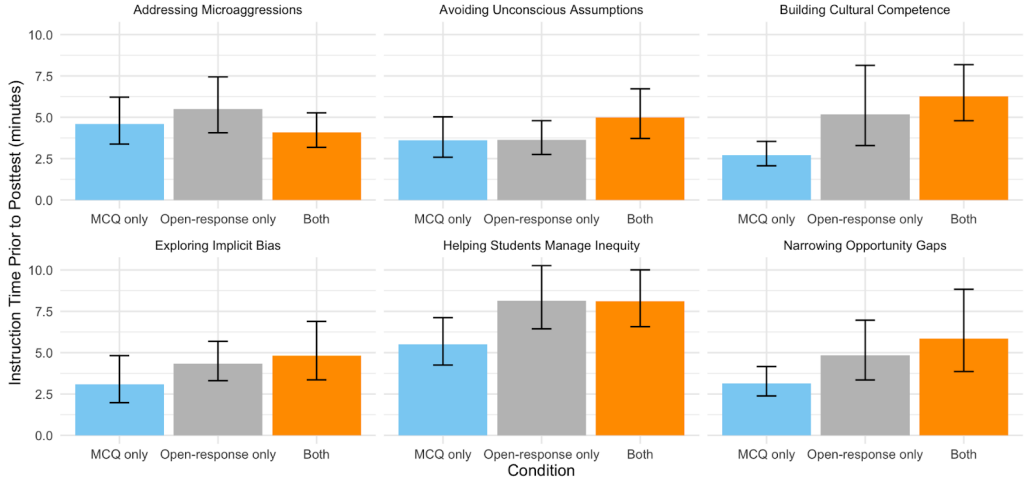

Figure 3: Average instruction time prior to posttest compared across learning-by-doing conditions: *MCQ Only, Open-response Only, or Both*. Although *MCQ Only* took less time on average, no overall significant differences in instruction time prior to posttest were found between the conditions. Error bars represent 95% confidence intervals.

handling both false positives and false negatives. In general, both models demonstrated proficiency in performance, with accuracy ranging from 71% to 91% for *predict* responses and 71% to 87% for *explain* responses—with some exceptions in GPT-4-turbo. GPT-4-turbo demonstrated proficiency across all lessons with poorer performance relative to the other lessons for assessing *predict* responses in *Helping Students Manage Inequity* (AUC = 0.17, F_1 = 0.45). The AUC was low for scoring other lessons: *predict* response in *Exploring Implicit Bias* (AUC = 0.43); *explain* responses in *Building Cultural Competence* (AUC = 0.45), suggesting that the model performs poorly in distinguishing between correct and incorrect responses.

5 Discussion

This study investigated differences in tutor learning across conditions that align with varying learning-by-doing activities (i.e., MCQ only, open response only, or both) and assessed the scalability of using generative AI for evaluating tutor responses. Several important insights emerged, which offer a comprehensive understanding of this present work.

5.1 No overall condition differences in learning outcomes.

In summary, there was no main effect of learning-by-doing condition on posttest scores, or learning outcomes. However, there

Table 6: Comparison of absolute performance for `GPT-4o` across lessons for *predict* and *explain* open responses.

Lesson (Evaluation by `GPT-4o`)	*Predict responses*					*Explain responses*				
	Accuracy (%)	Precision	Recall	*AUC*	F_1 score	Accuracy (%)	Precision	Recall	*AUC*	F_1 score
Addressing Microaggressions	79%	0.80	0.71	0.78	0.75	76%	0.92	0.59	0.77	0.72
Avoiding Unconscious Assumptions	75%	0.81	0.75	0.75	0.78	76%	0.65	0.74	0.44	0.69
Building Cultural Competence	91%	0.86	0.89	0.45	0.88	71%	0.83	0.74	0.72	0.78
Exploring Implicit Bias	71%	0.98	0.69	0.79	0.81	74%	0.85	0.66	0.76	0.75
Helping Students Manage Inequity	81%	0.85	0.84	0.46	0.84	87%	0.84	0.94	0.86	0.89
Narrowing Opportunity Gaps	82%	0.93	0.78	0.84	0.85	85%	0.54	0.78	0.42	0.64

Table 7: Comparison of absolute performance for `GPT-4-turbo` across lessons for *predict* and *explain* open responses.

Lesson (Evaluation by `GPT-4-turbo`)	*Predict responses*					*Explain responses*				
	Accuracy (%)	Precision	Recall	*AUC*	F_1 score	Accuracy (%)	Precision	Recall	*AUC*	F_1 score
Addressing Microaggressions	77%	0.85	0.60	0.76	0.70	75%	0.92	0.56	0.75	0.70
Avoiding Unconscious Assumptions	70%	0.57	0.67	0.47	0.61	78%	0.98	0.72	0.84	0.83
Building Cultural Competence	79%	0.86	0.78	0.43	0.82	59%	0.64	0.56	0.45	0.60
Exploring Implicit Bias	85%	0.87	0.83	0.43	0.85	79%	0.85	0.76	0.79	0.80
Helping Students Manage Inequity	57%	0.68	0.34	0.17	0.45	88%	0.59	0.93	0.56	0.72
Narrowing Opportunity Gaps	75%	1.00	0.60	0.80	0.75	73%	0.88	0.63	0.76	0.73

was a significant interaction between the condition and the lesson indicating some potential heterogeneity, meaning that the learning outcomes differed significantly by condition depending on the lesson the tutors completed. Two of the three significant pairwise comparisons (comparing lesson and condition) were quite close to the p-value threshold of 0.05 and with 18 such comparisons (3 comparisons for each of 6 lessons). So, perhaps these are chance occurrences. Nevertheless, 3 out of 18 significant differences are substantially more than the expected false positive rate of 1 in 20 comparisons and therefore worth further exploration. The pairwise comparison indicating that *Open-response Only* is significantly better than *Both* for *Exploring Implicit Bias* is hard to interpret. It suggests that adding MCQ questions may harm learning in this lesson, and thus the *MCQ Only* condition should be worse. However, the difference in learning outcomes between *MCQ Only* and *Open-response Only* was not significant ($p = .719$). Inspection of the *Exploring Implicit Bias* did not reveal any substantial differences from other lessons in the nature of the MCQ or open-response questions. Overall, we have not found a consistent and generalizable explanation for the observed condition by lesson interaction. Perhaps the best explanation for this is random variability. As a content-general conclusion, our evidence suggests that replacing open-response learning tasks with MCQ learning tasks produces generally equivalent learning outcomes. Certainly, we found no substantial evidence against MCQs as learning tasks. Indeed, we found no overall statistically reliable evidence for a decrease in learning outcomes due to the use of MCQs as learning tasks. In addition, there were no consistent trends across the six lessons in this direction. Indeed, the *MCQ only* condition had the highest average posttest in two of the six lessons and the lowest only once. In contrast, the *Open-response Only* condition had the highest posttest for two lessons, but the lowest for three.

5.2 The *MCQ Only* condition requires less time.

There was no significant interaction between *instruction time spent prior to posttest*, or time spent, and condition on learning outcomes. However, there was a significant interaction between the condition and the lesson, indicating that some lessons took longer to complete than others. In general, and not surprisingly, the *MCQ Only* condition took the tutors the shortest time. However, the *Both* condition took the longest time of the tutors, though not substantially longer than *Open-response Only*, with the *Both* condition taking the students less time than the sum of the open and MCQ conditions. Why did it not take learners longer to complete both forms of practice (*Open-response* and *MCQ*) compared to *Open-response Only*? One possible explanation can arise from differences in the speed with which students in each condition complete the Follow-Up Instruction (Fig. 1). Conducting a *post hoc* ANOVA on follow-up instruction time revealed significant differences in follow-up instruction by condition, although that instruction included the same material in each condition ($F(2, 576) = 8.52$, $p < .001$). The mean comparisons indicated that learners in the *Both* condition were faster to complete the follow-up instruction ($M = 0.155$ minutes), followed by the *MCQ Only* condition ($M = 0.239$ minutes) and the *Open-response Only* condition ($M = 0.347$ minutes). These differences were statistically significant: the *Both* condition was significantly faster in processing the follow-up instruction than the *MCQ Only* condition, estimate = -0.47, $SE = 0.18$, $p = .020$, as well as the *Open-response Only* condition, estimate = -0.78, $SE = 0.19$, $p < .001$.

5.3 LLMs demonstrate proficiency, but more research is needed for wide-scale assessment.

Similar to [25], we found the GPT models to be comparable to each other and overall demonstrated proficiency. However, GPT-4-turbo

exhibited variability across lessons, with poorer performance in the *Helping Students Manage Inequity* (AUC = 0.17) and *Exploring Implicit Bias* (AUC = 0.43) lessons. The low AUC scores in these cases suggest that the GPT-4-turbo struggled to classify responses in more complex or nuanced topics. This highlights the need for further refinement of LLMs to enhance their ability to assess open responses in content areas that require situational reasoning. In addition, more research is needed on the nuances within each lesson. The interrater reliability between the human graders was high for some lessons (κ = 0.93), so it is necessary to revisit the human grade and annotation rubrics for the lessons and determine the sources of disagreement between the GPT models and humans (and between the raters) to better understand the results. This work adds to the recent literature on the potential use of LLMs for low stakes assessment tasks in different domain areas (Henkel et al. [20]), adding to the area of tutor training in advocacy skills.

5.4 Limitations

While this study used a posttest-only randomized experimental design, which offers several advantages, there are inherent limitations to this approach. One limitation is the lack of baseline data due to the absence of an analogous pretest scenario, which prevents measuring individual learning gains from pretest to posttest. This limitation makes it challenging to quantify the exact effect size of the lessons and to track individual progress. However, the primary purpose of this study was not to determine individual learning gain but to identify whether any of the learning-by-doing conditions–multiple choice questions, open-response questions, or both–led to better learning outcomes. By focusing on posttest results and employing random assignment to the condition, we were able to directly compare the relative learning effectiveness of each condition. In addition, the study aimed to find the most efficient lesson design by analyzing the time it took learners to complete each condition. Although the absence of pretest data can sometimes reduce statistical power, our sample size was sufficient to detect meaningful differences across conditions in time taken. Furthermore, our previous study [35], which implemented the *Both* condition with a counter-balanced pre-post design, demonstrated that our assessments are sufficient to detect significant pre-post learning gains. Finally, while the lack of baseline data can complicate the replication of the study in different contexts, the posttest only design was specifically chosen to minimize the risk of pretest sensitization, which can enhance the validity of the experimental results [11, 38].

6 Future Work and Conclusion

We found evidence for learning efficiency benefits of multiple choice questions (MCQs) with feedback relative to the use of open-response questions with feedback or the combination of both. Practically, this result deserves particular attention given the widespread use of open-response questions in homework assignments as a practice task across content areas in college and K-12 education. This benefit of MCQs as learning tasks was demonstrated in a large sample of learners (n=235) and a variety of instructional lessons (6) and was revealed in generally equivalent learning outcomes achieved in significantly less time, corresponding to a 29% practice time reduction compared to the open-response only and 35% reduction compared a condition with both forms of instruction. Although these results favor the use of MCQs for instruction, they do not have any direct bearing on whether MCQs or open-response questions are better for assessing student learning. We used both in our posttest, and we intend to keep open-response questions to continue to do so to test that instructional practice with MCQ transfers to performance on open-ended questions.

Our finding that MCQ practice transfers to open-ended performance is an important general result. The use of MCQs is often the target of criticism in assessment and instructional design. This criticism of MCQs as a shallow form of practice is perhaps more common when used for content that is less well defined, such as is the case for the learning goals of the tutoring lessons used in these studies (i.e., advocacy). Our evidence does not support such criticism and, further, provides evidence for the instructional benefits of MCQs relative to open-response questions in requiring less time for learners to reach the same outcome. Practically, this finding provides support for greater use, or at least greater exploration, of MCQs as practice tasks during instruction. Theoretically, this result has implications for the refinement of general frameworks for instructional design. In particular, the ICAP framework [7] makes a general prediction in the subject matter domain that constructive learning tasks, such as open-response questions, should produce better learning outcomes than active (but not constructive) learning tasks, such as feedback-based MCQs. Our learning outcome evidence is inconsistent with this prediction, especially as a generalization across learning content.

We observed evidence of effect heterogeneity, with some lessons that included MCQ learning tasks yielding better learning outcomes, while others showed improved outcomes when MCQ tasks were excluded. This content-treatment interaction has been found and well explained in previous research (e.g., [33]); however, in this case, the explanation is not clear, and we suggest future work to probe whether there is a replicable finding here and, if so, what theory might explain it. Finally, given how prevalent open-response questions are as learning tasks in homework assignments in school and college education, we recommend further investigation of the potential for more efficient and equally effective learning from the use of multiple-choice questions as learning tasks.

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] Vincent A Aleven and Kenneth R Koedinger. 2002. An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive science* 26, 2 (2002), 147–179.

[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] 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.

[4] Dan Carpenter, Wookhee Min, Seung Lee, Gamze Ozogul, Xiaoying Zheng, and James Lester. 2024. Assessing Student Explanations with Large Language Models Using Fine-Tuning and Few-Shot Learning. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Ekaterina Kochmar, Marie Bexte, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Anaïs Tack, Victoria Yaneva, and Zheng Yuan (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 403–413.

[5] 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* 21 (2024).

[6] 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), San Diego, CA, USA, 1812–1817.

[7] Michelene TH Chi and Ruth Wylie. 2014. The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist* 49, 4 (2014), 219–243.

[8] 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*. Association for Computing Machinery, New York, NY, USA, 469–471.

[9] 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.

[10] 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).

[11] Thomas D Cook and Donald T Campbell. 2007. *Experimental and quasi-experimental designs for generalized causal inference*. Figures.

[12] Alison Cook-Sather. 2020. Student voice across contexts: Fostering student agency in today's schools. *Theory into practice* 59, 2 (2020), 182–191.

[13] Andreea Dutulescu, Stefan Ruseti, Denis Iorga, Mihai Dascalu, and Danielle S McNamara. 2024. Beyond the Obvious Multi-choice Options: Introducing a Toolkit for Distractor Generation Enhanced with NLI Filtering. In *International Conference on Artificial Intelligence in Education*. Springer, 242–250.

[14] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics* (2024), 1–79.

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

[16] Ashish Gurung, Kirk Vanacore, Andrew A Mcreynolds, Korinn S Ostrow, Eamon Worden, Adam C Sales, and Neil T Heffernan. 2024. Multiple Choice vs. Fill-In Problems: The Trade-off Between Scalability and Learning. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. 507–517.

[17] Jonathan Guryan, Jens Ludwig, Monica P Bhatt, Philip J Cook, Jonathan Davis, Kenneth Dodge, George Farkas, Roland G Fryer Jr, Susan Mayer, Harold Pollack, et al. 2021. Not Too Late: Improving Academic Outcomes among Adolescents (Working Paper 28531). *National Bureau of Economic Research* (2021).

[18] Thomas M Haladyna. 2004. *Developing and validating multiple-choice test items*. Routledge.

[19] 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*.

[20] 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.

[21] Dollaya Hirunyasiri, Danielle R Thomas, Jionghao Lin, Kenneth R Koedinger, and Vincent Aleven. 2023. Comparative Analysis of GPT-4 and Human Graders in Evaluating Human Tutors Giving Praise to Students.. In *Human-AI Math Tutoring@ AIED*. 37–48.

[22] 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).

[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*. Association for Computing Machinery (ACM), Vancouver, BC, Canada, 111–120.

[24] MA Kraft and G Falken. 2021. A blueprint for scaling tutoring across public schools (EdWorkingPaper No. 21–335). Annenberg Institute at Brown University.

[25] Gyeong-Geon Lee, Ehsan Latif, Xuansheng Wu, Ninghao Liu, and Xiaoming Zhai. 2024. Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence* 6 (2024), 100213.

[26] Russell V. Lenth. 2024. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. https://CRAN.R-project.org/package=emmeans R package version 1.10.2.

[27] 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*. Atlanta, Georgia, USA, 236–250.

[28] 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.

[29] Jionghao Lin, Danielle R Thomas, Feifei Han, Shivang Gupta, Wei Tan, Ngoc Dang Nguyen, and Kenneth R Koedinger. 2023. Using Large Language Models to Provide Explanatory Feedback to Human Tutors. In *Human-AI Math Tutoring@ AIED*. 12–32.

[30] Andre Nickow, Philip Oreopoulos, and Vincent Quan. 2020. The impressive effects of tutoring on prek-12 learning: A systematic review and meta-analysis of the experimental evidence. (2020).

[31] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/2303.08774

[32] John F Pane, Beth Ann Griffin, Daniel F McCaffrey, and Rita Karam. 2014. Effectiveness of cognitive tutor algebra I at scale. *Educational Evaluation and Policy Analysis* 36, 2 (2014), 127–144.

[33] Napol Rachatasumrit, Paulo F Carvalho, Sophie Li, and Kenneth R Koedinger. 2023. Content matters: A computational investigation into the effectiveness of retrieval practice and worked examples. In *International Conference on Artificial Intelligence in Education*. Springer, Tokyo, Japan, 54–65.

[34] Carly D Robinson and Susanna Loeb. 2021. High-impact tutoring: State of the research and priorities for future learning. *National Student Support Accelerator* 21, 284 (2021), 1–53.

[35] 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*. Association for Computing Machinery (ACM), Arlington, TX, USA, 250–261.

[36] 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.

[37] 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).

[38] William MK Trochim and James P Donnelly. 2001. *Research methods knowledge base*. Vol. 2. Atomic dog publishing Cincinnati, OH.

[39] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*.

[40] 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.

[41] Joy Yun, Yann Hicke, Mariah Olson, and Dorottya Demszky. 2024. Enhancing Tutoring Effectiveness Through Automated Feedback: Preliminary Findings from a Pilot Randomized Controlled Trial on SAT Tutoring. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*. 422–426.

[42] 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).

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-Advocacy