ARTICLE

How Can I Get It Right? Using GPT to Rephrase Incorrect Trainee Responses

Jionghao Lin[1] · **Zifei Han**[1] · **Danielle R. Thomas**[1] · **Ashish Gurung**[1] · **Shivang Gupta**[1] · **Vincent Aleven**[1] · **Kenneth R. Koedinger**[1]

Abstract

One-on-one tutoring is widely acknowledged as an effective instructional method, conditioned on qualified tutors. However, the high demand for qualified tutors remains a challenge, often necessitating the training of novice tutors (i.e., trainees) to ensure effective tutoring. Research suggests that providing timely explanatory feedback can facilitate the training process for trainees. However, it presents challenges due to the time-consuming nature of assessing trainee performance by human experts. Inspired by the recent advancements of large language models (LLMs), our study employed the GPT-4 model to build an explanatory feedback system. This system identifies trainees' responses in binary form (i.e., correct/incorrect) and automatically provides template-based feedback with responses appropriately rephrased by the GPT-4 model. We conducted our study using the responses of 383 trainees from three training lessons (*Giving Effective Praise*, *Reacting to Errors*, and *Determining What Students Know*). Our findings indicate that: 1) using a few-shot approach, the GPT-4 model effectively identifies correct/incorrect trainees' responses from three training lessons with an average F1 score of 0.84 and AUC score of 0.85; and 2) using the few-shot approach, the GPT-4 model adeptly rephrases incorrect trainees' responses into desired responses, achieving performance comparable to that of human experts.

Keywords Large language models · Feedback · Tutoring training · ChatGPT · GPT-4

Introduction

One-on-one tutoring has been recognized as a highly effective strategy for enhancing student learning, with substantial evidence supporting its impact (Kraft & Falken, 2021; Nickow et al., 2020). However, there are significant challenges associated with the scalability of one-on-one tutoring, primarily due to the scarcity of skilled tutors, including certified teachers and paraprofessionals. This shortage has left an estimated

---

16 million students in the United States in need of individualized support, as high-lighted by Kraft and Falken (2021). In response to this shortage, there has been a strategic shift towards effectively training novice tutors, including community volun-teers, retired individuals, and college students, to fulfill tutoring role (Nickow et al., 2020).

The growing demand for skilled tutors has resulted in the development of various professional development programs tailored to the unique needs of nonprofessional and novice tutors (Nickow et al., 2020). Driven by this need, researchers have explored the use of online scenario-based training to simulate real-life tutoring scenarios for novice tutors (Thomas et al., 2023) and pre-service teachers (Thompson et al., 2019). Figure 1 illustrates a scenario on *Giving Effective Praise*. It demonstrates how tutors can fail to appropriately acknowledge the student's efforts by providing outcome-based praise as opposed to effort-based praise. For instance, saying *"Kevin, good job getting the problem correct!"* fails to acknowledge the student's efforts and persistence. As indicated in previous research (Lin et al., 2023; Hirunyasiri et al., 2023), the availability of real-time explanatory feedback within the scenario-based training lessons can help tutors provide effective praise. Particularly, real-time feedback on learners' errors, similar to the feedback received while engaging in the deliberate practice of responding to situational judgment tests, is described as a favorable learning condition and can lead to better learning outcomes (Koedinger et al., 2023, p. 5).

While the benefits of real-time explanatory feedback in enhancing tutor learning outcomes are well-documented, crafting such feedback presents substantial challenges due to its labor-intensive nature. Traditionally, providing this level of specialized train-ing, replete with personalized explanatory feedback, warrants a substantial investment

Scenario

You're tutoring a student named Kevin. He is struggling to understand a math problem. When he doesn't get the answer correct the first time he wants to quit. After trying several different approaches, Kevin gets the problem correct. As Kevin's tutor, you want him to continue working through solving more problems on his math assignment.

Question 1 *

1. What exactly would you say to Kevin to provide effective praise that will increase his motivation to complete his math assignment and increase engagement?

Kevin, good job getting the problem correct!

956 chars remaining

Powered by OpenAI ⓘ   🔍 Review ③

Fig. 1 An example of a trainee (i.e., novice tutor) incorrectly responding to an open-ended question on how to best reply to a student by giving effective praise. In this particular example, the trainee is praising the student for getting the problem correct, which is achievement or outcomes-based praise and not based on effort

of effort and time. The process of providing personalized feedback to novice tutors requires considerable time and effort from skilled tutors to ensure feedback effectiveness and relevance. Moreover, beyond the substantial investment of time and effort, the feasibility of scaling such training protocols to meet the high demand across educational settings significantly compounds the challenge. However, recent breakthroughs in large language models (LLMs) offer a promising avenue for streamlining this process. Models such as the Generative Pre-trained Transformer (GPT) could potentially automate the generation of personalized, real-time feedback for tutors (Hirunyasiri et al., 2023; Dai et al., 2023). This automation not only has the potential to alleviate the resource burden but also to enhance the specificity and precision of the feedback by accurately identifying the personalized needs of the tutors (Hirunyasiri et al., 2023).

Currently, the quality of automated explanatory feedback is lacking, with many systems failing to provide learners with accurate feedback on their constructed responses (Lin et al., 2023; Hirunyasiri et al., 2023). We argue that the quality of feedback for tutor training can be further improved. Inspired by the feedback research (Henderson et al., 2019; Hattie & Timperley, 2007; Butler et al., 2013), where learners interpret performance-related information to enhance their understanding, we postulate that presenting desired tutoring responses within feedback to novice tutors can enhance the effectiveness of the training. However, rephrasing incorrect tutor responses into the correct or desired form often necessitates a substantial investment of time and effort from experienced tutors-hence introducing scalability constraints associated with tutor training. Thus, we aim to explore approaches to improve our ability and accuracy in providing tutors with explanatory feedback while also mitigating the time and effort requirements of human graders by automating the process of generating explanatory feedback and correction to their responses. The automation requires the development of classification systems that can effectively analyze tutor responses or, in other words, classification systems that determine the correctness of tutor responses to scenario-specific requirements of the learners. However, there is useful learner information within appropriate classified incorrect responses. These incorrect learner-sourced responses can be used to provide tutors corrective, explanatory feedback by taking an incorrect response and rephrasing or modifying it to make it a desired, or correct, response. Research supports when learners are given specific feedback related to their responses, such as taking incorrect tutor responses and personalizing them by making them correct, they gain a better understanding of their learning (Attali & Powers, 2010; Torres, 2022).

We aim to explore how GPT models can serve as supplementary tools to deliver synchronous feedback to tutors on their responses of how to best respond to specific training scenarios (e.g., praising a student for effort) leveraging useful tutor incorrect responses. We propose two **R**esearch **Q**uestions:

RQ1: Can a large language model accurately identify trainees' incorrect responses where trainees failed to effectively guide students in specific training scenarios?

RQ2: Can GPT-4 be harnessed to enhance the effectiveness of trainees' responses in specific training scenarios?

We initially developed a binary classifier to determine tutor's correct and incorrect responses from three training lessons: *Giving Effective Praise*, *Reacting to Errors*,

AI-generated feedback suggests that **your response could be more focused on praising the student for their effort**. Here is an AI-rephrased version of your response that more effectively focuses on effort:

Kevin, great effort on persisting through the challenge to get the problem correct!

| Feedback Helpful | Not Helpful | Feedback Incorrect |

Fig. 2 Explanatory feedback for novice tutor responses

and *Determining What Students Know*. We employed zero-shot and few-shot learning approaches to classify the trainees' responses. Our result demonstrated that the five-shot learning approach achieved acceptable performance in identifying the incorrect responses. Building upon the results of **RQ1**, we selected the incorrect responses identified by our optimal few-shot learning classifier, which is further used for the **RQ2**. We explored the idea of rephrasing incorrect trainees' responses to determine if we can prompt GPT-4 to effectively make them correct. An example of an incorrect response from the lesson *Giving Effective Praise* is shown in Fig. 1), e.g., *"Kevin, good job getting the problem correct!"*. Through extensive experiments, we obtained an effective prompt to secure the rephrased responses presented in an accurate form with minimal changes of the words from the original incorrect responses. Building upon the result from RQ1 and RQ2, we build a feedback system to provide explanatory feedback to the incorrect trainee's response shown in Fig. 2.

Related Work

Significance of Feedback on Learning

Feedback plays a crucial role in improving the students' learning outcomes and performance (Hattie & Timperley, 2007; Henderson et al., 2019; Lin et al., 2023). In the field of feedback research, theoretical models have been developed to explain the impact of feedback on learning and to identify the core principles that underpin effective feedback design. Hattie and Timperley (2007) defined feedback as the information about the correctness of a learner's actions or decisions, along with explanations about why those actions or decisions are right or wrong, underlines the significance of feedback. As emphasized in their work (Hattie & Timperley, 2007), the influence of feedback on learning varies based on the type and timing of its delivery.

Effective feedback should assist learners in understanding the rationale behind the feedback, which is crucial for deeper learning (Henderson et al., 2019). Moreover, including the correct answer within the feedback substantially enhances its efficacy by offering learners the information needed to correct their errors (Butler et al., 2013). This is especially relevant when learners answer open-ended questions, as simply knowing that their response is incorrect may not suffice to improve their understanding

(Butler et al., 2013). By presenting the correct answer (or correct responses to open-ended question) in the feedback, learners can compare their responses with the correct responses, identify areas for improvement, and gain guidance on how to approach similar questions in the future (Attali & Powers, 2010; Torres, 2022). To help learners identify their misconception in the open-ended question, we posit that it is necessary to include the correct responses in the feedback. However, providing timely explanatory feedback faces challenges since crafting effective explanatory feedback is often time-consuming and labor-intensive nature (Lin et al., 2023; Dai et al., 2023; Hirunyasiri et al., 2023). To address this issue, it is necessary to develop automated feedback generation system.

Feedback Generation

The development of automated feedback has received significant attention from educational researchers (Dai et al., 2023; Lin et al., 2023; Hirunyasiri et al., 2023; Pardo et al., 2018; Demszky et al., 2021). For example, Ontask (Pardo et al., 2018) is a rule-based feedback provision system designed to assist instructors in delivering personalized feedback based on specific conditions of learners (e.g., the duration spent on the learning system). Additionally, Demszky et al. (2021) developed a feedback system that automatically delivers explanatory feedback to instructors via email within two to four days after their tutoring sessions. Their study results (Demszky et al., 2021) indicate that timely explanatory feedback enhanced learners' satisfaction. Lin et al. (2023) used sequence labeling techniques to provide automated explanatory feedback, which demonstrated the potential of the large language models on identifying the effective components of feedback. Despite demonstrating the effectiveness of automated feedback systems, the provision of feedback with correct responses to open-ended question is still under-explored, which are needed to advance feedback systems

Using Large Language Models for Feedback Generation

Inspired by recent research on using large language models for feedback generation (Levonian et al., 2023; Lin et al., 2023; Hirunyasiri et al., 2023; McNichols et al., 2023; MacNeil et al., 2022), we posit that GPT-based large language models hold potential for advancing the development of automated feedback. For example, Dai et al. (2023) investigated the capability of GPT-3.5 model (ChatGPT) to generate feedback for students' writing assignment and they Dai et al. (2023) found that GPT-3.5 could produce feedback that was more readable than that of human instructors. Then, Hirunyasiri et al. (2023) leveraged the GPT-4 model to provide timely feedback for human tutors' training. Their results (Hirunyasiri et al., 2023) indicated that GPT-4 outperformed human educational experts in identifying a specific tutoring practice, giving effective praise. While these studies have demonstrated the feasibility of GPT-based models in feedback generation, none have ventured into generating explanatory feedback with correct responses to open-ended questions. Given that

GPT-4 has shown remarkable performance on various educational tasks (e.g., generating high-quality answer responses for middle school math questions (Levonian et al., 2023) and providing feedback for multiple-choice questions at the middle-school math level (McNichols et al., 2023)), our study also leveraged the GPT-4 model to further explore its capabilities in automatically generating explanatory feedback.

Method

Data

We developed an online learning platform[1] to facilitate training for the novice tutors in the form of brief scenario-based lessons. Within the scope of this study, we refer to the novice tutors participating in the training activities as *trainees*. Aligning with previously demonstrated competencies of effective tutoring (Chhabra et al., 2022), each lesson presents scenario-based questions to facilitate an authentic and contextually relevant tutor learning opportunity. These scenarios challenged the tutors to apply their knowledge and skills by simulating real-world tutoring situations (see Fig. 1). We examined the trainees' performance and understanding across three lessons: *Giving Effective Praise*, *Reacting to Errors*, and *Determining What Students Know*. These lessons are based on the skillsets that were identified to be crucial for tutors in previous work (Chhabra et al., 2022; Thomas et al., 2023).

Each lesson consisted of two scenarios. Across all trainees, we collected 410 responses: 140 responses from the 70 trainees who took the *Giving Effective Praise* lesson, 118 responses from *Reacting to Errors* (59 trainees), and 152 responses from *Determining What Students Know* (76 trainees). Before analysis, we removed 10, 4, and 13 responses respectively from each lesson because they were either empty or contained incoherent or meaningless content (e.g., "*ad;fajkl*", "*test test test*" or "*I have no idea*"), resulting in a total of 383 analyzed responses. We also collected demographic information about the trainees, including their experience as tutors, as presented in Table 1. For each lesson, tutors provided self-reported demographic details, including information regarding their race, gender, age, and tutoring experience.

Annotation for Trainee's Responses

In the lesson *Giving Effective Praise*, trainees practice their skills in engaging students by offering effort-based praise. The praise provided by trainees should effectively acknowledge students' efforts and aim to enhance their motivation and desire to keep learning. A tutoring scenario was depicted where a student was struggling to persevere on an assignment (See the scenario in Table 2). The tutor trainee's responses were expected to show the components of effective praise as suggested by research recommendation (Thomas et al., 2023). Effective praise should be: 1) timely, positive, and sincere, 2) highlighting what student did well during the tutoring, 3) genuine and avoiding generic comments like *"great job"*, and 4) focus on the learning process

---

[1] https://www.tutors.plus/solution/training

Table 1 Demographic information of participants

Demographic Categories	Giving Effective Praise (n = 70)	Reacting to Errors (n = 59)	Determining What Students Know (n = 76)
Gender			
Female	27	23	32
Male	34	30	34
Non-binary	1	1	1
Unknown	8	5	9
Age			
18-24	8	8	11
25-34	11	6	9
35-50	12	9	9
51-64	21	22	25
65+	12	12	16
Unknown	6	2	6
Ethnicity			
Asian	12	10	18
White	34	30	34
Others	8	8	9
Unknown	16	11	15
Tutoring Experience*			
Level 1	4	3	4
Level 2	14	11	14
Level 3	22	18	24
Level 4	20	21	24
Level 5	5	4	5
Unknown	5	6	5

Tutoring Experience*: The tutors were asked to rate their prior tutoring experience on a five-point Likert scale, where Level 1 indicated a tutor with limited experience, and Level 5 signified an expert tutor

rather than on the student or the outcome. In short, correct praise responses should be supportive, positive, encouraging, and acknowledging the student's effort during the learning process. In Table 2, we demonstrate some praise responses with an explanation of the rationale for labeling responses as either **Correct** or **Incorrect**.

In the lesson *Reacting to Errors*, trainees practice their skills in responding to student errors. Trainees employ various pedagogical strategies aimed towards addressing gaps in the learners' knowledge through constructive feedback. Instead of overt criticism, the emphasis is on fostering a positive approach to errors. This approach seeks to shift students' perception towards errors by underscoring their importance in the learning process. A tutoring scenario was depicted where a student made a mistake in solving a problem (See the scenario in Table 3). The tutor trainee's responses to students' errors should help students develop their critical thinking skills and encourage students to

Table 2 Examples of correct and incorrect trainee responses for the lesson *Giving Effective Praise* with annotation rationale

Scenario

You're tutoring a student named Kevin. He is struggling to understand a math problem. When he doesn't get the answer correct the first time, he wants to quit. After trying several different approaches, Kevin gets the problem correct. As Kevin's tutor, you want him to continue working through solving more problems on his math assignment.

Trainee Response	**Interpretation**
	Correct Response
You are making steady progress and it is good to see the results of your efforts	The response expresses the sense of positive and sincere. The praise focuses on the student's perseverance and acknowledges the students for working hard and the process of learning.
	Incorrect Response
You did a great job, well done!	This response is sincere and positive but the praise does not focus on student learning efforts or learning actions (e.g., demonstrated the problem-solving procedural).

correct their mistakes. According to Thomas et al. (2023), to effectively respond to students' errors, one should: 1) indirectly inform students about their mistake in the problem-solving process, 2) guide the student towards self-correction, and 3) show praise for the student's effort or attempt. Responses that directly highlight the student's error or inform the student what to do are not desired in the tutoring practice (Thomas et al., 2023). In Table 3, we demonstrated some responses of reacting to errors with the explanation of the rationale for labeling responses as either **Correct** or **Incorrect**.

Table 3 Examples of both correct and incorrect trainee responses for the lesson *Reacting to Errors* with annotation rationale

Scenario

Imagine you are a mentor to a student, Aaron, who has a long history of struggling with math. Aaron is not particularly motivated to learn math. He just finished a math problem adding a 3-digit and 2-digit number and has made a common mistake (shown below).

Trainee Response:	**Interpretation:**
	Correct Response
Lucy, very well, but I have to point that we have another way of doing the math problem, we can repeat the math together, what do you think?	This response avoids using direct words about the student's mistake and implicitly clears up the misconception. Instead, the response encourages the student to make another attempt instead of explicit answers
	Incorrect Response
This is very close! I see one issue, can you walk me through the how you worked through the problem?	This response asks the student to walk through the steps but it still uses the word "issue", which may be frustrating.

In the lesson *Determining What Students Know*, this lesson is designed to enhance the tutor trainees' skills in discerning the current knowledge level of the students by distinguishing what the students have comprehended and what still needs to be learned. A tutoring scenario was depicted where a student was given a math problem they did not know how to solve (see the scenario in Table 4). The tutor trainee's responses were used to gauge the student's prior knowledge at the start of the session and provide instruction based on what students already know as a launching point for the rest of the session. According to Thomas et al. (2023), effective response of determining what students know should be: 1) prompting students to demonstrate what they have already done or explain what they know, 2) presenting in an open-ended form and avoiding asking student's understanding of specific knowledge concept, 3) guiding the tutoring conversation to locate the student's misunderstanding, 4) providing instructional support to help students find the correct answer. To summarize, correct response of determining what students know should assess a student's prior knowledge, guide the conversation to catch student's misconceptions or errors and support productive struggle. In Table 4, we demonstrated some responses of determining what students know with the explanation of the rationale for labeling responses as either **Correct** or **Incorrect**.

Identifying Desired Trainee Responses

One of the motivations for this study is the creation of a classifier capable of discerning desired attributes in a tutor's responses to scenario-based prompts. The goal is to determine whether the tutors can adapt to the specific scenarios and integrate scenario-specific instructional practices when supporting the learners. For instance, should a trainee fail to acknowledge the learner's effort when working on an activity requiring

Table 4 Examples of both correct and incorrect trainee responses for the lesson *Determining What Students Know* with annotation rationale

Scenario
You are working with a student named Cindy on her math homework. She is having trouble solving a geometry problem dealing with triangles. She shows you the following diagram displaying a triangle and states that she has to determine the value of angle x. Cindy says, "I don't know what to do."

Trainee Response:	**Interpretation:**
	Correct Response
What have you tried so far?	This response asks an open-ended question to understand what students have tried so far and gauge the student's knowledge
	Incorrect Response
Do you know what PEMDAS means?	Tutors' responses can contain questions to students but they must be open-ended and non-specific to assess student's knowledge of an individual knowledge component.

effective praise, the classifier would categorize the tutor's feedback as **Incorrect** (less desirable). Identifying these scenarios presents an opportunity to personalize training activities for trainees, enhancing their ability to learn from and rectify specific instructional methodologies.

In addressing **RQ1**, we first employed two expert raters, both specialists in educational instruction and feedback, to annotate trainees' responses as either **Correct** (desirable) or **Incorrect** (less-desirable). Using Cohen's κ, we determined inter-rater reliability, obtaining scores of $0.85, 0.81$, and 0.64 for *Giving Effective Praise*, *Reacting to Errors*, and *Determining What Students Know*, respectively. These scores of inter-rater reliability are considered sufficient (Neuendorf, 2017). Disagreements between the raters prompted input from a third expert to ensure consistency in annotations. Then, recognizing the typical need for a large amount of data when training classifiers from scratch for natural language processing tasks, we turned to recent advances in machine learning. As documented in Wang et al. (2020); Pourpanah et al. (2022), zero-shot and few-shot learning methods can effectively discern patterns in datasets, even when they are limited or absent. These methods leverage the inherent capability of pre-trained models, which is crucial for ensuring classification performance and generalizability. The principle mirrors human cognition, as explored in Wang et al. (2020); Pourpanah et al. (2022), where individuals apply their generalized knowledge to identify unfamiliar objects or concepts. Further details of these methods are described below:

- **Zero-shot Learning:** In zero-shot learning, the classifier is trained to perform tasks for which it has seen no labeled examples at all. This is achieved by transferring knowledge from related tasks and using semantic relationships between classes. The model's prior knowledge, often in the form of embeddings or representations that capture semantic meanings, is crucial for it to make predictions in unseen classes (Pourpanah et al., 2022).
- **Few-shot Learning:** In few-shot learning, the classifier is trained to perform tasks using a limited amount of labeled data. The underlying principle is to leverage the knowledge acquired by the model from previous and related tasks to facilitate effective generalization to a new task, even when provided with minimal data. The prior knowledge enables the classifier to adapt to new tasks with only a few examples (Wang et al., 2020). Additionally, given that our classifier is designed to categorize trainees' responses into two categories (i.e., correct or incorrect), the few-shot learning with two classification categories is commonly termed *"two-way few-shot learning"*. For instance, a two-way 2-shot contains two correct responses and two incorrect responses. Upon a thorough review of existing literature (Cao et al., 2019), we found that most studies implemented few-shot learning with the number of shots less than or equal to five. In line with this consensus, our study also sets five shots as the maximum threshold for the number of shots.

As described, both zero-shot and few-shot learning methods rely on a robust pre-trained model. The pre-trained models, having been exposed to extensive training corpora, inherently possess base knowledge that allows them to discern generalized

patterns even from minimal datasets. Inspired by the effectiveness of GPT-4 models on the existing educational tasks (Levonian et al., 2023; McNichols et al., 2023; Hirunyasiri et al., 2023), we adopted the state-of-the-art GPT-4 model (OpenAI, 2023) as the foundational model for conducting binary classification of trainees' responses. A GPT prompt is a sentence or phrase provided to the GPT model to produce a response (Dai et al., 2023; Li et al., 2023). Our prompt strategies are detailed in Table 5.

The prompt strategies are in the form of *Chat-Completion*, which refers to the generated response produced by the GPT-4 model during a conversation. When a user provides a prompt, GPT-4 processes the prompt and generates a relevant response, known as the *"Completion"*. The *Chat-Completion* is set up to generate the label for each trainee's response. For Zero-shot implementation, as presented in Table 5, the *Chat-Completion* has three different chat roles: **System**, **User**, and **Assistant**. The role of **System** represents the assigned default character for the machine. In our case, GPT-4 facilitates the role of a *"binary classifier"*. The role of **User** represents human input. The role of **Assistant** denotes a machine-generated response, which is to frame the prompting process as a conversation. Compared to the Zero-shot learning approach, the few-shot learning approach provides a limited number of correct and incorrect examples for the GPT-4 model to understand the classification patterns (Table 5). Subsequently, our proposed prompt requires specific inputs from the **User**. The input of {Lesson Principle} is based on the principles of a correct response from the lesson materials created by Thomas et al. (2023). The input of {Textual response} is the trainee's response. As there are three distinct lessons, the input of {Lesson Name} in the instruction prompt is substituted with the appropriate lesson name.

Table 5 Prompt strategies for a binary classifier

Zero Shot		Few-shot	
Role	Content	Role	Content
System	*"You are a binary classifier."*	**System**	*"You are a binary classifier."*
User	{Lesson Principle} + *"According to the lesson principle, please determine if the following response contains"* + {Lesson Name} + *"please respond YES; if not, please respond NO."*	**User**	{Lesson Principle} + *"According to the lesson principle, please determine if the following response contains"* + {Lesson Name} + *"please respond YES; if not, please respond NO."*
Assistant	*"Sure, please enter the response from tutor"*	**Assistant**	*"Please provide some examples of correct and incorrect response"*
User	{Textual response}	**User**	{Correct example} + {Incorrect example}
		Assistant	*"Sure, please enter the response from tutor"*
		User	{Textual response}

We used *Chat-Completion* to process the trainees' responses in batch

Enhancing the Trainee Responses by GPT Models

To explore **RQ2**, we used the GPT-4 model to rephrase incorrect responses into correct forms effectively. We designed the prompt strategies presented in Table 6. For the Zero-shot learning, we assigned a role with GPT-4 to rephrase the trainee's response (i.e., "*You are rephrasing tutor's response*"). For the role of **User**, similar to RQ1, we used {Lesson Principle} to enable GPT-4 to understand the correct form of tutor responses. To effectively rephrase the trainees' responses, we believe that providing context about the scenario in which the responses were given might lead GPT-4 to generate more accurate rephrased outputs. Thus, in the prompt, we also added the input of {Lesson Scenario}, which was the actual text of the scenario-based question, as demonstrated in Tables 2, 3, & 4. In the context of the few-shot learning approach, we supplied two examples of rephrased incorrect responses in their correct forms provided in the training lessons to help the GPT-4 model infer the rephrasing rules (see Table 5). The GPT-4 *Chat-Completion* is presented in Table 6.

Table 6 Prompt strategies for binary classifier

Zero Shot		Few-shot	
Role	Content	Role	Content
System	*"You are rephrasing tutor's response."*	**System**	*"You are rephrasing tutor's response."*
User	{Lesson Principle}+*"The provided response attempts to answer to the following scenario."*+{Lesson Scenario}+*"Please rephrase the tutor's response according to the principle mentioned above to create a better example of"*+{Lesson Name}+*"Retain words and ideas from the tutor's response. Limit changes to the original tutor's response to a minimum. Maintain the same length as the original tutor's response. Please rephrase as less words as possible from the original tutor's response. Highest priority is to make sure to follow the principle of the correct response when rephrasing."*	**User**	{Lesson Principle}+*"The provided response attempts to answer to the following scenario."*+{Lesson Scenario}+*"Please rephrase the tutor's response according to the principle mentioned above to create a better example of"*+{Lesson Name} + *"Retain words and ideas from the tutor's response. Limit changes to the original tutor's response to a minimum. Maintain the same length as the original tutor's response. Please rephrase as less words as possible from the original tutor's response. Highest priority is to make sure to follow the principle of the correct response when rephrasing."*
Assistant	*"Sure, please enter the response"*	**Assistant**	*"Please provide some examples of how you will rephrase the given incorrect response to make it correct"*
User	{Textual response}	**User**	{Rephrased examples}
		Assistant	*"Sure, please enter the response"*
		User	{Textual response}

We used *Chat-Completion* to process the trainees' responses in batch

Evaluation Approach

Evaluation for RQ1 We employ both the F1 score and the Area under the ROC curve (AUC) for evaluating the performance of our classification model. Furthermore, given our specific focus on identifying incorrect feedback, we incorporate two additional metrics: the Negative Predictive Value (NPV) and the True Negative Rate (TNR). These measures are crucial for determining the model's efficacy in minimizing false negatives and minimizing such errors is critical, as a false identification can result in incorrect feedback. Incorrect feedback can further undermine the training's effectiveness, potentially eroding trust and changing how trainees engage with the training activities. We provide the formulas for NPV and TNR in (1) and (2), respectively. Both NPV and TNR are metrics that range from 0 to 1, with higher values signifying a model's enhanced capability to correctly identify true negative instances.

$$Negative\ Predictive\ Value\ (NPV) = \frac{True\ Negative}{True\ Negative + False\ Negative} \tag{1}$$

$$True\ Negative\ Rate\ (TNR) = \frac{True\ Negative}{True\ Negative + False\ Positive} \tag{2}$$

Evaluation for RQ2 After rephrasing the trainee's responses, we evaluate the accuracy and quality of the rephrased responses. In order to achieve this, we first utilized the most effective binary classifier developed in RQ1 to classify the rephrase responses. Then, we compared the number of correct responses in rephrased responses and correct responses in original responses. Specifically, we wanted to investigate the extent to which the GPT-4 model has the capability to improve the accuracy of the trainee's responses. When the number of correct labels in rephrase responses is more than the correct responses in the original responses, it indicates that the GPT-4 model has the ability to accurately rephrase the trainee's responses and the classifier developed in RQ1 generally satisfied with the rephrased result. Additionally, we aim to compare the quality rephrased responses by GPT-4 with the ones by human expert. To do so, we first hired three experienced human tutors who completed the training for the three lessons. These three experts were asked to rephrase the incorrect responses based on the research recommendation provided in the lessons. Afterward, we invited a fourth human educational expert to assess the quality of rephrased responses in two dimensions: *Accuracy* and *Responsiveness*. The dimension of *Accuracy* was used to measure the correctness of the rephrased responses. Regarding the dimension of *Responsiveness*, it evaluates how the rephrased response selectively changes some words to improve the trainee's original response, while largely preserving the original words and ideas from the trainee's response. In our study, we designed the question for evaluating *Accuracy* by asking "*The rephrased response is a better example of* {Lesson Name} *than the trainee's response*" and the question for evaluating *Responsiveness* by asking "*The rephrased response changes some words to improve the trainee's response, but otherwise keeps words and ideas from the trainee's response*".

The educational expert answered the questions by using the five-point Likert scale (i.e., *Strongly Disagree* to *Strongly Agree*).

Results

Results for RQ1: Binary Classifier for Correct Responses

For RQ1, we explored the zero-shot and few-shot approaches to train a binary classifier using the GPT-4 model, as detailed in "Identifying Desired Trainee Responses" section. The classifier's performance is presented in Table 7. For the lesson *Giving Effective Praise*, the zero-shot approach resulted in an F1 score of 0.761 and an AUC of 0.743. When leveraging a two-way few-shot learning approach, we observed an improvement in the performance. The F1 scores remained consistently high, ranging from 0.856 to 0.872, with the 3-shot model achieving the peak performance. In parallel, the AUC scores were also robust, varying from 0.851 to 0.865, with the 5-shot model outperforming the others. Despite these improvements, the NPV and TNR metrics showed greater variability. The NPV spanned from 0.8 to 0.88, with the 3-shot model again taking the lead, whereas the TNR fluctuated between 0.744 to 0.851, with the 5-shot configuration achieving the strongest performance.

For the lesson on *Reacting to Errors*, the performance of the zero-shot learning approach resulted in an F1 score of 0.767 and an AUC of 0.768. It is worth noting that the zero-shot learning approach had an impressive NPV score of 0.911, the highest NPV score for feedback from *Reacting to Errors* activity, indicating the model's robustness in identifying true negative outcomes. When utilizing two-way few-shot

Table 7 Classification performance of the responses from three lessons

Lessons	Metrics	Zero Shot	Two-way Few-shot				
			1-shot	2-shot	3-shot	4-shot	5-shot
Giving Effective Praise	**F1**	0.761	0.870	0.845	**0.872**	0.856	0.860
	AUC	0.743	0.858	0.836	0.863	0.851	**0.865**
	NPV	0.666	0.841	0.853	**0.881**	0.841	0.800
	TNR	0.680	0.787	0.744	0.787	0.787	**0.851**
Reacting to Errors	**F1**	0.767	0.779	0.821	0.840	0.823	**0.867**
	AUC	0.768	0.778	0.819	0.838	0.822	**0.866**
	NPV	**0.911**	0.892	0.866	0.857	0.823	0.880
	TNR	0.585	0.622	0.736	0.792	0.792	**0.830**
Determining What Students Know	**F1**	0.660	0.712	0.718	0.747	0.798	**0.805**
	AUC	0.668	0.712	0.719	0.748	0.799	**0.806**
	NPV	0.630	0.714	0.733	0.733	0.818	**0.821**
	TNR	**0.828**	0.714	0.786	0.785	0.771	0.786

Note: **AUC** represents Area under the ROC Curve; **NPV** represent Negative Predicted Value; **TNR** represents True Negative Rate

learning approaches, the 5-shot learning approach presented the highest F1, AUC, and TNR scores at 0.867, 0.866, and 0.83, respectively.

Lastly, for the lesson on *Determining What Students Know*, the zero-shot learning approach resulted in an F1 score of 0.66 and AUC of 0.668, the lowest across the three lessons. Interestingly, the zero-shot model had a higher TNR score of 0.828, indicating that the model was adept at identifying true negative cases for this lesson. The performance across the F1, AUC, and NPV metrics presented a general uptick with the adoption of the two-way few-shot learning method, with the 5-shot variant demonstrating the highest enhancements, reflected by F1, AUC, and NPV scores of 0.805, 0.806, and 0.821, respectively.

Results for RQ2: Using GPT-4 to Rephrase Incorrect Responses

For RQ2, we examine the application of GPT-4 in transforming trainees' incorrect responses into a preferred format that exemplifies effective feedback, thereby demonstrating the correct manner to meet learner needs through feedback revision. To accomplish this, we utilized the most effective binary classifier identified from RQ1, the 5-shot classifier, to pinpoint incorrect responses within the three lessons. The identified responses were then compared with the responses identified by the expert human raters as described in "Identifying Desired Trainee Responses" section. The intersection of the responses identified as incorrect by both the classifier and the human rates resulted in 36 responses for *Giving Effective Praise*, 42 responses for *Reacting to Errors*, and 53 responses for *Determining What Students Know*. The overlap between the five-shot classifier and human raters was 85%, 83%, and 78.6% for *Giving Effective Praise*, *Reacting to Errors*, and *Determining What Students Know*, respectively, as indicated by the TNR scores for the 5-shot approach shown in Table 7.

As each training activity across the three lessons contained two paired examples to illustrate effective feedback in each scenario, we utilized the two paired examples per lesson to take a two-shot learning approach in exploring the effectiveness of GPT-4 in rephrasing student feedback. In this section, we report on the accuracy and responsiveness of the rephrased trainee responses by comparing the responses generated using zero-shot and two-shot GPT-4 models with responses rephrased by humans across the three lessons. The responses were assessed using a five-point Likert scale, i.e., *Strongly Disagree* (represented by -2), *Disagree* (represented by -1), *Neutral* (represented by 0), *Agree* (represented by 1), and *Strongly Agree* (represented by 2), as described in "Evaluation Approach" section. Given the ordinal nature of Likert scale data, we utilize the Mann-Whitney U test, a non-parametric statistical method, to ascertain if the accuracy and responsiveness of the rephrased responses are statistically different.

First, we examined the accuracy and responsiveness of the rephrased trainee responses for the lesson Giving Effective Praise, as presented in Fig. 3. We observed a higher median accuracy score of 1 for responses rephrased by GPT-4 (both Zero-shot and Few-shot) whereas the human rephrased responses received a median score of -1. As shown in Table 8, the accuracy scores of the rephrased responses generated using both GPT models (zero-shot and few-shot) were significantly higher than the responses rephrased by the humans ($p < 0.001$) indicating that the GPT-4 models
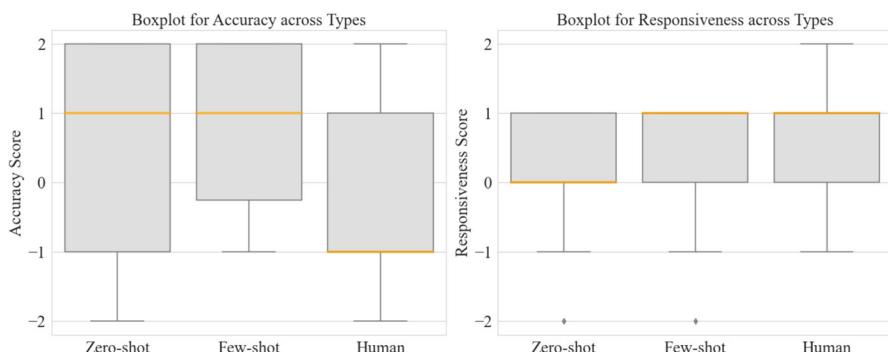
Fig. 3 Distribution of accuracy and responsiveness scores from the lesson *Giving Effective Praise*

Table 8 Statistics for rephrased responses from the lesson *Giving Effective Praise*

Metrics	Zero-shot	Few-shot	Human
Mean Accuracy	0.61	0.92	-0.36
Mean Responsiveness	0.22	0.44	0.44
words/response (mean)	17.28	21.72	12.28
words/response (SD)	6.28	18.94	5.88

were more effective at rephrasing the responses to the desired format in comparison to humans. While we did not observe a significant difference in the accuracy of the two GPT-based models, we observed a higher variance in the score of the zero-shot approach in comparison to the accuracy scores for the two-shot approach. When analyzing the responsiveness of the rephrased responses, we did not observe a significant difference between the responsiveness score of the GPT-4 rephrased responses and human rephrased responses; however, the human rephrased responses had a higher variance in comparison to the responsiveness scores of GPT-4 rephrased responses. The result demonstrated that the few-shot learning approach performed significantly better than the human in terms of the accuracy of the rephrased responses, while there was no significant difference in the responsiveness of the rephrased responses between the rephrased responses from the humans and the GPT-4 models. It indicated the effectiveness of few-shot learning on rephrasing the incorrect trainees' responses on the lesson of *Giving Effective Praise*.

Similarly, we evaluated the rephrased responses provided by both GPT-4 models and human for the *Reacting to Errors* lesson, presented in Fig. 4. The GPT-4-generated responses achieved a median accuracy score of 1, outperforming the human-revised responses, which held a median score of 0. Upon examining the rating further, as presented in Table 9, the accuracy of responses rephrased using the few-shot approach was significantly higher than those rephrased by humans ($p < 0.01$). Even the zero-shot rephrased responses were more accurate than human alterations ($p < 0.05$). As for the responsiveness, most of the scores from the GPT-revised and human-revised responses were clustered between 0 and 1, with no significant difference in responsiveness between them. Additionally, the table also indicated that the average

Fig. 4 Distribution of accuracy and responsiveness scores from the lesson *Reacting to Errors*

Table 9 Statistics for rephrased responses from the lesson *Reacting to Errors*

Metrics	Zero-shot	Few-shot	Human
Mean Accuracy	0.62	0.86	0.07
Mean Responsiveness	0.17	0.17	0.21
words/response (mean)	15.79	15.40	15.26
words/response (SD)	6.93	7.24	5.35

word count per response remained consistent between the GPT and human revisions, demonstrating that the GPT models, especially the few-shot approach, are adept at effectively rephrasing incorrect responses to *Reacting to Errors* without extensive modification to the original wording and sentence structure provided by the trainees.

Finally, our evaluation of the rephrased responses from the lesson *Determining What Students Know*, as illustrated in Fig. 5 and Table 10, revealed no significant difference in the dimensions of accuracy and responsiveness across the three approaches. Notably, unlike the accuracy in the other two chapters, the responsiveness scores from the few-shot method were marginally higher than those rephrased by humans ($p = 0.08$), indicating comparable performance between the automated few-shot and zero-shot approaches and human expertise. At the same time, no statistical significance was observed across conditions for responsiveness. Interestingly, it was in the *Determining What Students Know* lesson that the classification model had its weakest performance among the three lessons.

Discussion

Providing explanatory feedback is a fundamental requirement for delivering personalized feedback to learners. Our study explored the use of large language models (GPT-4 model) to automate the facilitation of explanatory feedback to novice tutors, where the main findings can be summarized in two folds: *Firstly*. GPT-4 models, especially for the few-shot approach, have the potential to accurately identify the correct and incorrect trainees' responses, which can be used to provide corrective feedback when
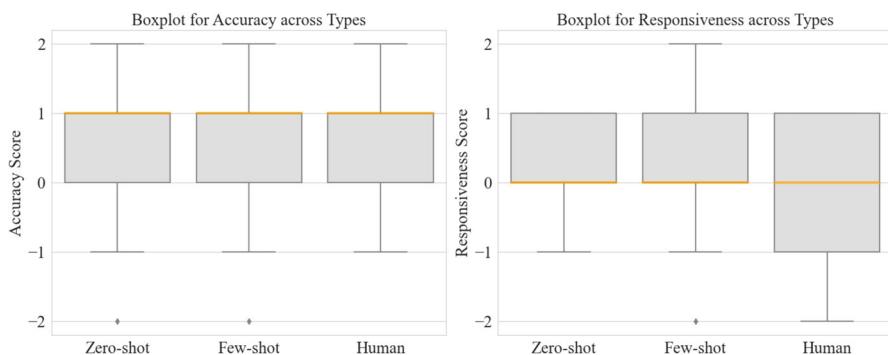
Fig. 5 Distribution of accuracy and responsiveness scores from the lesson *Determining What Students Know*

Table 10 Statistics for rephrased responses from the lesson *Determining What Students Know*

Metrics	Zero-shot	Few-shot	Human
Mean Accuracy	0.68	0.70	0.85
Mean Responsiveness	0.28	0.30	0.06
words/response (mean)	22.72	20.83	20.09
words/response (SD)	18.51	16.12	8.10

training novice tutors on the scenario-based tasks. Our results indicate that despite a limited number of samples, the GPT-4 model can accurately identify the incorrect trainees' responses across three different tutor training lessons (i.e., *Giving Effective Praise*, *Reacting to Errors*, and *Determining What Students Know*). By comparing the classification performance with zero-shot learning, the few-shot learning approach, especially with increasing shots, generally tends to improve the model's classification performance. This improvement suggests that more examples might increase GPT's capability to recognize the many different ways to express a target concept like effort-based praise (e.g., *"Good effort on solving the problem"*), and distinguish it from a related concept, like outcome-based praise (e.g., *"Good job"*). The implications of this finding is profound, especially when considered alongside existing research on neural network learning in humans. Previous research (Carvalho & Goldstone, 2022) has illustrated that both the quantity and diversity of examples play a significant role in the learning process, with optimal outcomes achieved through exposure to a range of examples that are internally diverse yet distinct from other categories. Applying this principle to the context of LLM training suggests a strategy where examples within a category (e.g., praising effort) are maximally diverse, whereas examples across categories are closely aligned (e.g., comparing praise for effort with praise for outcomes). Pursuing this line of inquiry in future research could yield valuable insights into the mechanisms underpinning effective learning in both human and artificial neural networks. By systematically exploring the interplay between example diversity and learning efficacy, we can refine our understanding of how best to structure training data for LLMs like GPT-4, ultimately enhancing their utility in educational applications.

Secondly, the capability of GPT-4, particularly when employing the few-shot learning approach, extends to effectively rephrasing trainees' incorrect responses into a desired format. Notably, GPT-4's performance in rephrasing incorrect responses to correct ones is on par with, and sometimes surpasses, that of experienced human tutors. This proficiency likely stems from GPT-4's advanced understanding of context and language nuances (OpenAI, 2023), enabling it to reconstruct trainees' incorrect responses to align more closely with the desired responses. The practical implications of the GPT4's capabilities are significant. The classified and rephrased responses generated by GPT-4 can be integrated into template-based feedback systems. Such integration facilitates the provision of real-time and explanatory feedback to novice tutors (or trainees) during their training sessions.

Implications

The incorporation of the binary classifier and its generalizability in terms of performance holds significant implications for providing explanatory feedback. The classified results (i.e., correct or incorrect responses) on trainee's responses can be further integrated into the provision of corrective feedback as shown in Fig. 2. Specifically, by identifying the incorrect responses, our feedback system can use the template-based feedback to provide suggestions for trainees to consider, i.e., "*AI-generated feedback suggests that your response could focus more on praising the student for their efforts in the learning process.* Providing corrective feedback is essential in the learning process for tutor training since it can assist the tutors in identifying their errors and improving the quality of their feedback (Butler et al., 2013).

Furthermore, this study demonstrated the potential of prompting GPT-4 models in rephrasing incorrect trainees' responses into the desired form. We measured the quality of rephrased responses from GPT-4 models and human experts in terms of their accuracy and responsiveness as described in "Evaluation Approach" section. Based on our observations, the rephrased responses consistently rated higher in accuracy while the responsiveness of the GPT-4 generated responses were comparable to the human-generated responses. For instance, a tutor response classified as incorrect, *"*Great Job*! But there is a tiny mistake, can you tell what was it ?"* from the lesson *Reacting to Errors* was rephrased in the following manner:

> **Zero-shot:** *"Good effort, Aaron! There's a small detail that needs tweaking. Can you spot what it might be?"*
> **Few-shot:** *"Good effort! Let's take another look at this, can you walk me through your process?"*
> **Human:** *"Great Job! But there is a tiny mistake, can you tell what was it by going over each step?"*

The incorrect response used the word "*mistake*" in the response, which failed to reframe the mistakes as a learning opportunity effectively and is considered to be an ineffective tutoring strategy as indicated by Chhabra et al. (2022); Thomas et al. (2023). In contrast, both zero-shot and few-shot responses were more effective in accounting for the student's errors when rephrasing the feedback. The zero-shot rephrased response avoids using direct words about the student's mistake and

implicitly mentions the misconception "*There's a small detail that needs tweaking.*" whereas the few-shot approach encourages the student to revisit the problem *Let's take another look at this* which is also in line with the effective practice of reacting to errors (Chhabra et al., 2022; Thomas et al., 2023). It should be noted that both zero-shot and few-shot rephrased responses also changed the praise on student's outcome (i.e., "*Great Job*") to student learning effort (i.e., "*Good effort*"). However, the human expert failed to rephrase the incorrect response into the desired form even though the expert changed the words from the original incorrect response. Additionally, when the accuracy of rephrased responses was rated at 1 (i.e., *Agree* on a five-point Likert scale) or 2 (i.e., *Strongly Agree*) but the responsiveness was rated from -2 (i.e., *Strongly Disagree*) to 0 (i.e., *Neutral*), the rephrased responses still were presented effective responses. For example, an example of incorrect response *"I think you are doing great."* from *Giving Effective Praise* was rephrased as below:

> **Zero-shot:** *"You're showing excellent effort in tackling your math work, Carla. Your perseverance is commendable."*
> **Few-shot:** *"I see you're making good progress with your efforts. Keep it up!"*
> **Human:** *"You are doing a great job!"*
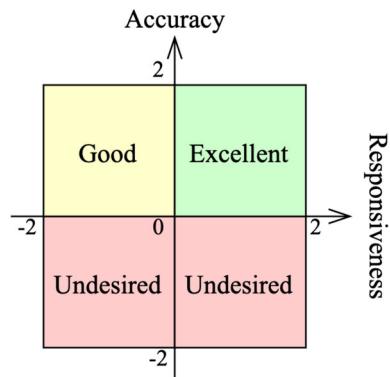
The incorrect response *"I think you are doing great."* failed to praise student on their learning efforts or learning actions but on their learning outcome, which is considered ineffective praise as indicated in Chhabra et al. (2022); Thomas et al. (2023). Both zero-shot and few-shot rephrased responses were rated accuracy of 2 (i.e., *Strongly Agree*) but responsiveness of 0 (i.e., *Neutral*). Both shot and few-shot rephrased responses demonstrated praise on the student learning efforts as underlined in the examples, but both responses changed many words, which was not similar to the original incorrect responses. In comparison, the responsiveness of human rephrased responses was rated at 1 since there was only several words were changed from the original incorrect response. However, the human expert failed to revise the praise correctly, and the rephrased response was rated at -1 (i.e., *Disagree*). The rephrased praise still focused on the student learning outcome (i.e., *great job!*) rather than their learning efforts, which is not considered an effective response for praising student as indicated by Thomas et al. (2023). As summarized by the evaluation results of both GPT-4 and human rephrased responses, we proposed a framework for determining the quality of the rephrased responses, shown in Fig. 6.

This framework (Fig. 6) aims to guide future work to understand the extent to which the rephrased responses are considered high quality. When the accuracy of the rephrased response is rated at 1 or 2, the rephrased responses are considered to be acceptable. Based on our observation, the optimal rephrased responses should be high in both accuracy and responsiveness (i.e., **Excellent** area in Fig. 6), which could guide the trainees to understand the desired form of the responses and also help them know where they did not perform well while providing their scenario specific feedback. Since the dimension of responsiveness aims to minimize the changes of words in the responses, we expect the trainees to be able to locate the parts of the sentence that are incorrect and rephrase them accordingly. Similarly, a high accuracy and lower responsiveness (i.e., **Good** area in Fig. 6) could guide the trainee to recognize the desired quality of the feedback. However, as shown in the above example, the low

Fig. 6 Framework for determining the quality of the rephrased responses

responsiveness of the rephrased responses is an indicator of the modifications required in the original incorrect responses, which may not be as helpful to the trainees if the rephrasing resulted in major structural and semantic changes that are harder to learn and retain. Finally, we defined responses in two areas as undesirable responses, as illustrated in Fig. 6. The undesirable responses, marked by a low accuracy score (≤ 0), undermine the effectiveness of the feedback (Thomas et al., 2023). While the rephrased responses might demonstrate high responsiveness, the low accuracy of the response is still detrimental to its effectiveness and, as such, is not desirable. The rephrased feedback (*"You are doing a great job!"*), as presented above, is an example of a rephrased response with a low accuracy but high responsiveness score.

Limitations and Future Work

Evaluating Impact of Proposed Feedback System on Tutoring Practice While our current findings demonstrated the potential of GPT models in providing explanatory feedback and appropriately rephrased responses, there is a need for a more comprehensive evaluation of the feedback's effectiveness in tutor training. In future work, we plan to investigate the influence of explanatory feedback on tutor practice. Specifically, we will examine the direct effects of our feedback on tutors' skill acquisition, retention, and application in real-world tutoring scenarios. By conducting longitudinal studies with both control and experimental groups, we aim to gain a clearer understanding of the long-term advantages and possible challenges of our approach. Such insights will not only shed light on the efficacy of our feedback system but also inform potential refinements to enhance the training process for novice tutors.

Using Advanced Prompt Strategies for Explanatory Feedback In our current study, we utilized zero-shot and few-shot prompt strategies to identify correct or incorrect trainees' responses (RQ1) and to rephrase these incorrect responses appropriately (RQ2). While our proposed prompting strategies demonstrated promising results, there is potential for further improvement. To push the boundaries of our research, we are considering the adoption of more advanced prompt strategies. Two such strategies that have caught our attention are the *Tree of Thoughts* (Yao et al., 2023) and *Graph*

of Thoughts (Besta et al., 2023). These prompting strategies are expected to offer a more nuanced and structured way of understanding the task context and generating relevant information, potentially leading to more accurate and insightful results. A comprehensive exploration of these advanced prompting strategies is beyond the scope of our current study. Thus, in future work, we aim to delve deeper into these prompt strategies to investigate their efficacy and potentials on the improvement of the quality of explanatory feedback.

Exploring Open-Sourced Models for Generating Explanatory Feedback Our investigation leverages GPT-4, a proprietary large language model by (OpenAI 2023), which, while powerful, is not open for flexible model adjustments. Driven by this limitation, we turn our attention to open-source large language models as potential alternatives for generating explanatory feedback. Open-source models such as LLaMA (Touvron et al. 2023) and Falcon (Penedo et al. 2023) present viable options. Multiple studies have substantiated their effectiveness in educational settings. For instance, one study showcases TRIPOST, an innovative algorithm that amplifies the performance of smaller language models in complex tasks like mathematics and reasoning by facilitating a cooperative dynamic with larger models, thus promoting self-evaluation and iterative improvement (Yu et al. 2023). Inspired by these advancements, we are committed to finding a harmonious balance between the efficiency and efficacy of large language models. Our objective is to enhance the practicality and expand the reach of our feedback system, thereby making it a more viable tool for educational purposes on a larger scale.

Generalizability Across Other Tutor Training Lessons While our study demonstrated promising results on providing explanatory feedback primarily from three lessons, to further explore the efficacy of our feedback system, broader evaluations of the feedback system on other lessons are also important such as *Using Motivational Strategies* and *Ensuring Conceptual Understanding*. All the lessons on our platform introduces tutors to unique teaching scenarios and challenges. Ensuring that our feedback system is equally adept at handling the intricacies of each lesson is crucial for its overall success. Thus, it is important to evaluate the efficacy of our developed feedback system across all lessons, ensuring that the feedback provided is accurate, relevant, and conducive to the emerging tutor training process, continuously guiding tutors towards pedagogical excellence.

Enhancing Explanatory Feedback through Sequence Labeling The primary objective of this study is to provide automatic explanatory feedback. We have demosntrated the demo of our developed explanaotry feedback system shown in Fig. 2. To further unlock the potential of automatic explanatory feedback, we propose a significant enhancement: the integration of sequence labeling method, as originally introduced in the work by Lin et al. (2023). In their research, they employed a color-coded highlighting approach to distinguish between the effective and ineffective component of trainee's responses, aiming to facilitate a clearer comprehension of correctness or incorrectness. By incorporating this sequence labeling approach in the provision of explanatory feedback, we expect that the feedback can demonstrate more corrective

information fostering a deeper understanding among trainees regarding the construction of effective responses.

Enhancing Trainee Response Evaluation Beyond Binary Classification Our study leveraged GPT-4's capabilities to categorize trainee responses into binary classes: correct or incorrect. However, this dichotomous approach may be overly simplistic and potentially limiting for real-world applications where a more nuanced understanding is required. Acknowledging this, we recognize the necessity of developing a more granular evaluation scale. A tiered ranking system, perhaps on a five- or ten-point scale, could provide a more detailed and effective assessment of trainee responses, aligning more closely with the complexities of real-world scenarios. This insight highlights a limitation in our current methodology and underscores the potential for future research to explore more sophisticated classification frameworks that can capture the varied spectrum of trainee performance more accurately.

Strategies for Safeguarding Privacy Information in Real-world Tutoring Our study observed that responses from trainee tutors across three different lessons often included the use of student names, as in "*Kevin, good job getting the problem correct!*" This pattern suggests a tendency among some tutors to personalize their feedback by mentioning students by name during actual tutoring sessions. To further evaluate the practices of novice tutors within real-world tutoring contexts, it is necessary to colect and archive transcripts of tutoring dialogues in our database. To protect data privacy, we intend to anonymize any sensitive information, such as names, locations, and ages, contained within these transcripts.

Enhancing Automated Explanatory Feedback Quality Through Human-in-the-loop Design In our future work, we aim to explore the enhancement of automated explanatory feedback quality through the incorporation of a human-in-the-loop design. This approach will involve integrating human interaction directly into the feedback loop, enabling a ranking system where responses generated by Large Language Models (LLMs) are reviewed and prioritized based on human judgment. Such a mechanism is

Table 11 Sample question for crowd sourcing the ratings of rephrased responses from the trainee tutors

Scenario: *What exactly would you say to Cindy to begin helping her solve the math problem?*

Response: *Are you familiar with the definition and notation for angle and side congruence?*

Rephrased Responses	Accuracy	Responsiveness
1. How would you define when angles or sides in a triangle are congruent? What does that mean to you?	4	3
2. Can you explain your understanding of angle and side congruence, and their notations?	4	–
3. What do you understand about the concept of congruence in relation to sides and angles?	–	–
4. Are you familiar with the definition and notation for angle and side congruence?	–	–

expected to provide stronger signals to the AI, guiding it towards producing outputs that are more aligned with human expectations.

Crowd Sourcing the Evaluation of Rephrased Responses from Trainees Inviting educational experts to evaluate the quality of rephrased responses is often time-consuming and impractical, especially when dealing with a large volume of tutor responses. To address this issue, we suggest a crowd-sourcing approach for rating the rephrased responses. we plan to include the question (shown in Table 11) into the lesson and invite tutor trainees to answer the question. Table 11 presents the Scenario question and a response from a previous trainee which was identified an incorrect response. We will employ the large language models to rephrase the incorrect trainee's response and also keep the original incorrect response in the question. The new trainees are invited to rate the quality of responses based on the accuracy, responsiveness in a five-point scale. We also incorporate the original response for trainee to rate their scores. Since our developed binary classifier was not perfect, misclassified incorrect might exist, we also want the trainee's to provide their rating on the original responses. By doing so, we can obtain their ratings of rephrased responses and we expect our trainees can obtain better understanding about the presence of the effective form of responses in different training lessons.

Explanatory Feedback for the Synchronous Tutoring Session Our study demonstrated the capability of GPT-4 models to provide explanatory feedback and adeptly rephrasing tutor responses into a desired format. As shown in "Results for RQ2: Using GPT-4 to Rephrase Incorrect Responses" section, our proposed few-shot learning approach could achieve performance comparable to human experts in rephrasing responses appropriately, which could help reduce the use of inappropriate instructional responses during the student learning process. Given our current findings, we expect the integration of our developed explanatory feedback system into synchronous text-based online tutoring could facilitate the tutoring process. Previous studies (Lin et al., 2022a, b, 2023) have emphasized the importance of showing effective responses to students. Given the growing demand for qualified tutors, our feedback system, when integrated with synchronous tutoring platforms, can equip novice tutors to deliver timely and appropriate instructional feedback. To assess the influence of our exploratory feedback system on tutoring, We recommend conducting randomized controlled experiments to examine the efficacy of our feedback system further. In the experiment setup, tutors in experimental group will use our explanatory feedback system to provide instructional response, whereas the tutors in the control group will follow business-as-usual tutoring. The investigation aims for a comprehensive understanding of the system's strengths and areas needing improvement.

Conclusion

We aimed to provide automatic explanatory feedback to enhance tutor training. Our study explored the potential of GPT-4 model in delivering real-time explanatory feedback for open-ended questions selected from three tutor training lessons. We first prompted the GPT-4 model to act as a binary classifier to identify incorrect tutor

responses. With well-designed prompting strategies, the GPT-4 model, using a few-shot approach, accurately identified incorrect trainee responses across all three lessons we examined. We then used the GPT-4 model to rephrase incorrect responses into the desired responses. Our results demonstrated that the quality of rephrased responses provided by GPT-4, using a few-shot approach, achieved performance comparable to that of human experts. These results indicate that our proposed automatic explanatory feedback system shows promise in providing real-time feedback. Our study sheds light on the development of feedback provision for learners. By integrating our feedback system, we expect it can facilitate the tutor training process and further alleviate the the challenges associated with recruiting qualified tutors.

Declarations

Ethics Approval The study presented in this paper obtained the Institutional Review Boards (IRB) approval from Carnegie Mellon University.

Conflicts of Interest The authors have no relevant financial or non-financial interests to disclose, nor conflicting interests nor competing interests.

References

Attali, Y., & Powers, D. (2010). Immediate feedback and opportunity to revise answers to open-ended questions. *Educational and Psychological Measurement, 70*(1), 22–35.

Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Gianinazzi, L., Gajda, J., Lehmann, T., Podstawski, M., Niewiadomski, H., Nyczyk, P., et al. (2023). Graph of thoughts: Solving elaborate problems with large language models. arXiv:2308.09687

Butler, A. C., Godbole, N., & Marsh, E. J. (2013). Explanation feedback is better than correct answer feedback for promoting transfer of learning. *Journal of Educational Psychology, 105*(2), 290.

Cao, T., Law, M. T., & Fidler, S. (2019). A theoretical analysis of the number of shots in few-shot learning. In *International Conference on Learning Representations*

Carvalho, P. F., & Goldstone, R. L. (2022). A computational model of context-dependent encodings during category learning. *Cognitive Science, 46*(4), 13128.

Chhabra, P., Chine, D., Adeniran, A., Gupta, S., & Koedinger, K. (2022). An evaluation of perceptions regarding mentor competencies for technology-based personalized learning. In *Society for Information*

Technology & Teacher Education International Conference (pp. 1620–1625). Association for the Advancement of Computing in Education (AACE)

Dai, W., Lin, J., Jin, F., Li, T., Tsai, Y.-S., Gasevic, D., & Chen, G. (2023). Can large language models provide feedback to students? a case study on chatgpt

Demszky, D., Liu, J., Hill, H. C., Jurafsky, D., & Piech, C. (2021). Can automated feedback improve teachers' uptake of student ideas? evidence from a randomized controlled trial in a large-scale online course. *edworkingpaper* (no. 21-483). Annenberg Institute for School Reform at Brown University.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Henderson, M., Ajjawi, R., Boud, D., & Molloy, E. (2019). *The Impact of Feedback in Higher Education: Improving Assessment Outcomes for Learners*. Switzerland AG: Springer.

Hirunyasiri, D., Thomas, D. R., Lin, J., Koedinger, K. R., & Aleven, V. (2023). Comparative analysis of gpt-4 and human graders in evaluating praise given to students in synthetic dialogues. arXiv:2307.02018

Koedinger, K. R., Carvalho, P. F., Liu, R., & McLaughlin, E. A. (2023). An astonishing regularity in student learning rate. *Proceedings of the National Academy of Sciences, 120*(13), 2221311120.

Kraft, M. A., & Falken, G. T. (2021). A blueprint for scaling tutoring and mentoring across public schools. *AERA Open, 7*, 23328584211042856.

Levonian, Z., Li, C., Zhu, W., Gade, A., Henkel, O., Postle, M.-E., & Xing, W. (2023). Retrieval-augmented generation to improve math question-answering: Trade-offs between groundedness and human preference. arXiv:2310.03184

Lin, J., Dai, W., Lim, L.-A., Tsai, Y.-S., Mello, R. F., Khosravi, H., Gasevic, D., & Chen, G. (2023). Learner-centred analytics of feedback content in higher education. In *LAK23: 13th International Learning Analytics and Knowledge Conference* (pp. 100–110)

Lin, J., Rakovic, M., Lang, D., Gasevic, D., & Chen, G. (2022). Exploring the politeness of instructional strategies from human-human online tutoring dialogues. In *LAK22: 12th International Learning Analytics and Knowledge Conference* (pp. 282–293)

Lin, J., Raković, M., Xie, H., Lang, D., Gašević, D., Chen, G., & Li, Y. (2023). On the role of politeness in online human–human tutoring. *British Journal of Educational Technology*

Lin, J., Thomas, D. R., Han, F., Gupta, S., Tan, W., Nguyen, N. D., & Koedinger, K. R. (2023). Using large language models to provide explanatory feedback to human tutors. arXiv:2306.15498

Lin, J., Singh, S., Sha, L., Tan, W., Lang, D., Gašević, D., & Chen, G. (2022). Is it a good move? mining effective tutoring strategies from human-human tutorial dialogues. *Future Generation Computer Systems, 127*, 194–207.

Li, Y., Sha, L., Yan, L., Lin, J., Raković, M., Galbraith, K., Lyons, K., Gašević, D., & Chen, G. (2023). Can large language models write reflectively. *Computers and Education: Artificial Intelligence, 4*, 100140.

MacNeil, S., Tran, A., Mogil, D., Bernstein, S., Ross, E., & Huang, Z. (2022). Generating diverse code explanations using the gpt-3 large language model. In *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 2* (pp. 37–39).

McNichols, H., Feng, W., Lee, J., Scarlatos, A., Smith, D., Woodhead, S., & Lan, A. (2023). Exploring automated distractor and feedback generation for math multiple-choice questions via in-context learning. arXiv:2308.03234

Neuendorf, K. A. (2017). *The Content Analysis Guidebook*. Los Angeles, CA: SAGE Publications.

Nickow, A., Oreopoulos, P., & Quan, V. (2020). *The impressive effects of tutoring on prek-12 learning: A systematic review and meta-analysis of the experimental evidence*.

OpenAI (2023). *GPT-4 Technical Report*.

Pardo, A., Bartimote, K., Shum, S. B., Dawson, S., Gao, J., Gašević, D., Leichtweis, S., Liu, D., Martínez-Maldonado, R., Mirriahi, N., et al. (2018). Ontask: Delivering data-informed, personalized learning support actions. *Journal of Learning Analytics, 5*(3), 235–249.

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., & Launay, J. (2023). *The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only*.

Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C. P., Wang, X.-Z., & Wu, Q. J. (2022). A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence*.

Thomas, D., Yang, X., Gupta, S., Adeniran, A., Mclaughlin, E., & Koedinger, K. (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* (pp. 250–261).

Thompson, M., Owho-Ovuakporie, K., Robinson, K., Kim, Y. J., Slama, R., & Reich, J. (2019). Teacher moments: A digital simulation for preservice teachers to approximate parent-teacher conversations. *Journal of Digital Learning in Teacher Education, 35*(3), 144–164.

Torres, J. (2022). Feedback as open-ended conversation: Inviting students to coregulate and metacognitively reflect during assessment. *Journal of the Scholarship of Teaching and Learning, 22*(1), 81–94.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models*.

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (csur), 53*(3), 1–34.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. arXiv:2305.10601

Yu, X., Peng, B., Galley, M., Gao, J., & Yu, Z. (2023). Teaching language models to self-improve through interactive demonstrations. arXiv:2310.13522

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Jionghao Lin[1] [iD] **· Zifei Han[1] · Danielle R. Thomas[1] · Ashish Gurung[1] · Shivang Gupta[1] · Vincent Aleven[1] · Kenneth R. Koedinger[1]**

✉ Jionghao Lin
    jionghao@cmu.edu; jiongh.lin@gmail.com

    Zifei Han
    hanzifeifei@gmail.com

    Danielle R. Thomas
    drthomas@cmu.edu

    Ashish Gurung
    agurung@andrew.cmu.edu

    Shivang Gupta
    shivang@cmu.edu

    Vincent Aleven
    aleven@cs.cmu.edu

    Kenneth R. Koedinger
    koedinger@cmu.edu

[1]   Human-Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA