How Can I Improve? Using GPT to Highlight the Desired and Undesired Parts of Open-ended Responses

Presenter: Jionghao Lin

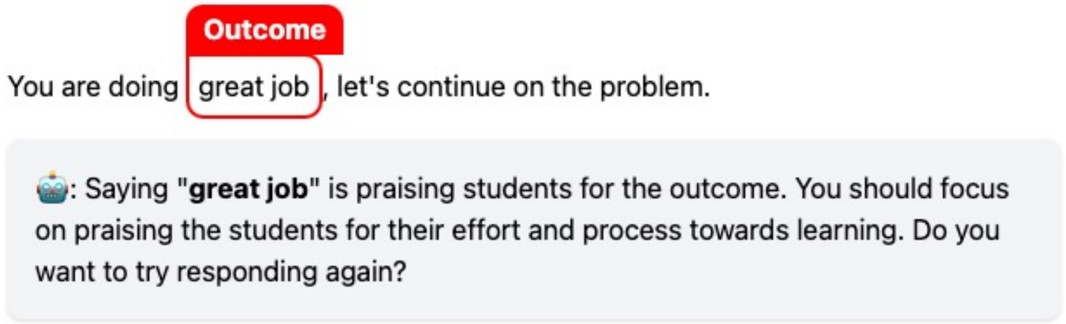
Eason Chen
Zifei Han
Ashish Gurung
Danielle R. Thomas
Wei Tan
Ngoc Dang Nguyen
Kenneth R. Koedinger

PLUS Personalized Learning²

Carnegie Mellon University

Research Questions (RQ):

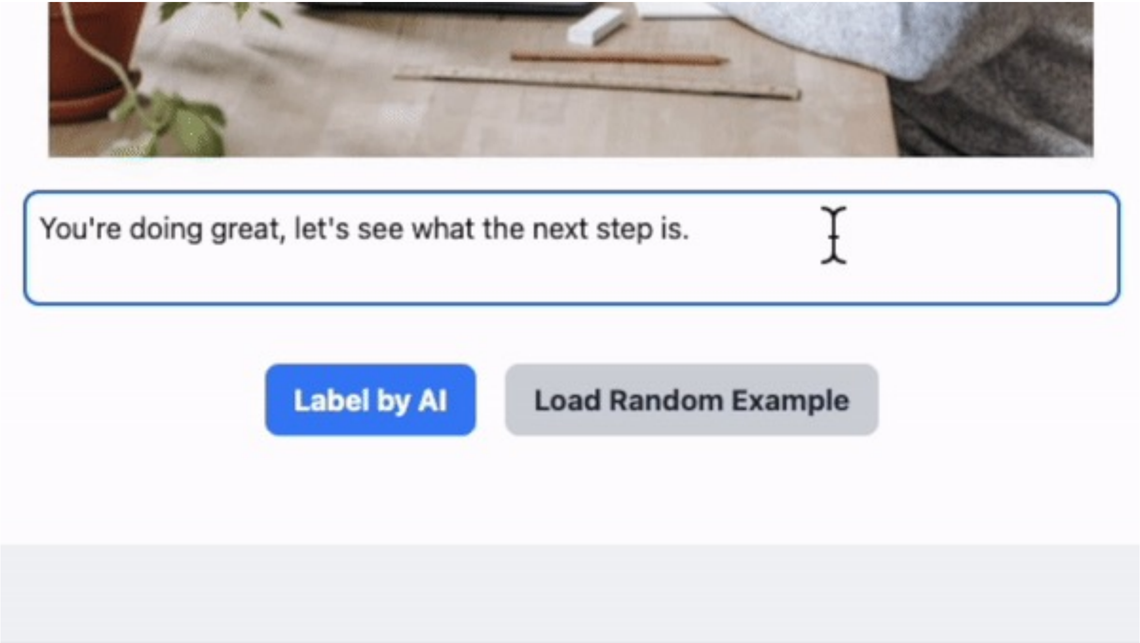
RQ1: Does GPT provide accurate explanatory feedback?
RQ2: Does fine-tuning improve the accuracy?



Demo of a tutor receiving explanatory feedback within the lesson on giving effective praise

Provide feedback on trainee tutor responses

Fine-tuned GPT-3.5 for highlighting key components on novice tutors' responses



Try our demo here

Background - Challenge

Tutoring is one of the most effective instructional methods which can help student improve their learning achievements, and yet:

- Finding qualified tutors is difficult
- Training tutors is more difficult
- Providing tutoring to ALL students is the most difficult

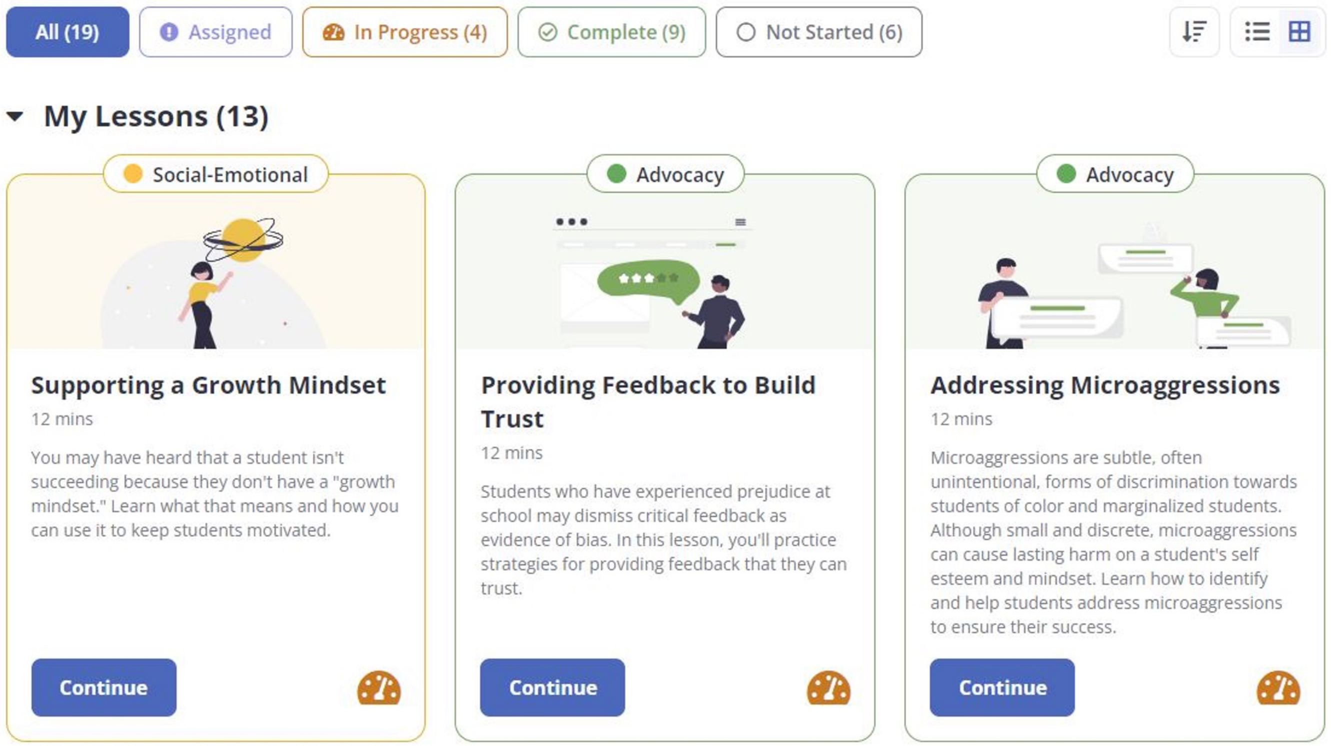
Personalized Learning Squared (PLUS)

PLUS

PLUS is a platform designed to optimize tutoring

Try our platform: https://tutors.plus

Background - PLUS Training Lessons



Scenario-based Question

Question scenario

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?

You are doing great job, let's continue on the problem.

975 chars remaining

Powered by OpenAI

PLUS Personalized Learning²

Review 1

Scenario-based Question

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 for trainees

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?

You are doing great job, let's continue on the problem.

975 chars remaining

Powered by OpenAI

PLUS Personalized Learning²

Review 1

Scenario-based Question

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?

Trainee's attempt → You are doing great job, let's continue on the problem.

975 chars remaining

Powered by OpenAI

PLUS Personalized Learning²

Review 1

How can we use generative AI to provide feedback to tutor trainees?

Research Questions

RQ1: To what extent can we prompt the GPT models to enhance the prediction accuracy of providing explanatory feedback?

RQ2: To what extent can the fine-tuned GPT models enhance the prediction accuracy of providing explanatory feedback?

Research Questions

RQ1: To what extent can we prompt the GPT models to enhance the prediction accuracy of providing explanatory feedback?

Does prompting GPT provide accurate explanatory feedback?

RQ2: To what extent can the fine-tuned GPT models enhance the prediction accuracy of providing explanatory feedback?

Does fine-tuning improve the accuracy?

Focus: Giving Effective Praise

Train novice tutors (adult learners/trainees) to give effective praise for middle school students.

Effective Praise should be:

- perceived as sincere and truthful.
- detailed what the student did well, not generic (e.g., *"great job"*).
- focused on the student learning process

Thomas, D., Yang, X., Gupta, S., Adeniran, A., Mclaughlin, E., & Koedinger, K. (2023, March). 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).

Praise Types

Effort-based praise: focus on the learning process (e.g., *"I like how you worked hard to…"*).

Outcome-based praise: focus on student's achievements, such as getting a problem correct, and is often, but not always, associated with unproductive praise (e.g., *"Great job!"*).

Thomas, D., Yang, X., Gupta, S., Adeniran, A., Mclaughlin, E., & Koedinger, K. (2023, March). 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).

Provide feedback on trainee tutor responses

Effort-based praise

Outcome-based praise

Trainee tutor: *"Good effort working through this, let's move to the next problem."*

Feedback for the trainee tutor:

Saying *"Good effort working through this"* is a nice example of effort-based praise, which praises students for their effort.

Provide feedback on trainee tutor responses

Effort-based praise

Outcome-based praise

Trainee tutor: *"Good effort working through this, let's move to the next problem."*

Feedback for the trainee tutor:

Saying "*Good effort working through this*" is a nice example of effort-based praise, which praises students for their effort.

Provide feedback on trainee tutor responses

Effort-based praise

Outcome-based praise

Trainee tutor: *"Well done, Kevin! Let's work on the next one"*

Feedback for the trainee tutor:

Saying "*Well done*" is praising students for the outcome. You should focus on praising the students for their effort and process towards learning. Do you want to try responding again?

Provide feedback on trainee tutor responses

- **RQ1:** Does GPT provide accurate explanatory feedback?
- **RQ2:** Does fine-tuning improve the accuracy?

Trainee tutor responses
(N = 129)

Founding source: Richard King Mellon Foundation, USA

Lin, J., Chen, E., Han, F., Gurung, A., Thomas, D. R., Tan, W., Ngoc, D. N., & Koedinger, K. R. (2024). How Can I Improve? Using GPT to Highlight the Desired and Undesired Parts of Open-ended Responses. *17th International Conference on Educational Data Mining*

Aleven, V., Popescu, O., & Koedinger, K. R. (2001). Towards tutorial dialog to support self-explanation: Adding natural language understanding to a cognitive tutor. In *Proceedings of Artificial Intelligence in Education* (pp. 246-255).

Provide feedback on trainee tutor responses

- **RQ1:** Does GPT provide accurate explanatory feedback?
- **RQ2:** Does fine-tuning improve the accuracy?

Trainee tutor responses (N = 129) → Expert labeling

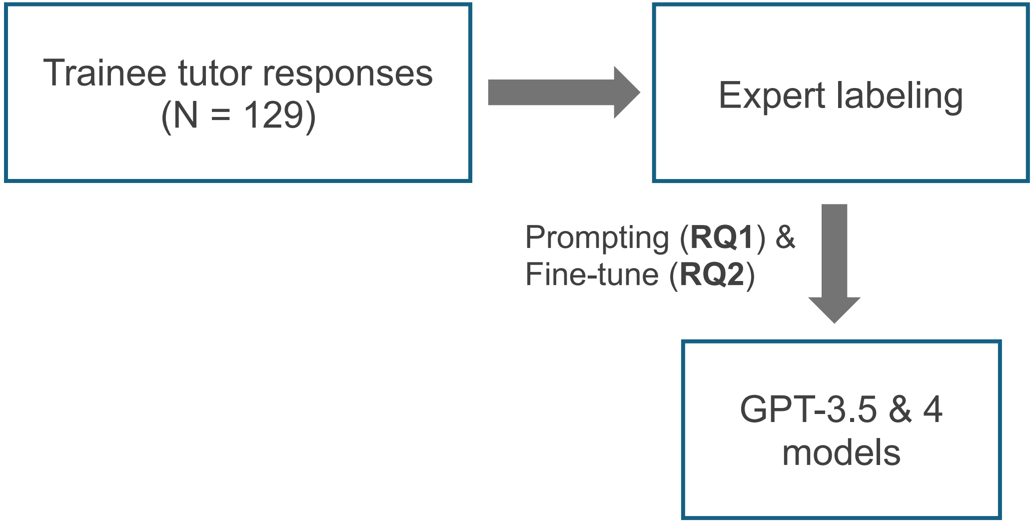
Founding source: Richard King Mellon Foundation, USA

Lin, J., Chen, E., Han, F., Gurung, A., Thomas, D. R., Tan, W., Ngoc, D. N., & Koedinger, K. R. (2024). How Can I Improve? Using GPT to Highlight the Desired and Undesired Parts of Open-ended Responses. *17th International Conference on Educational Data Mining*

Aleven, V., Popescu, O., & Koedinger, K. R. (2001). Towards tutorial dialog to support self-explanation: Adding natural language understanding to a cognitive tutor. In *Proceedings of Artificial Intelligence in Education* (pp. 246-255).

Provide feedback on trainee tutor responses

- **RQ1:** Does GPT provide accurate explanatory feedback?
- **RQ2:** Does fine-tuning improve the accuracy?



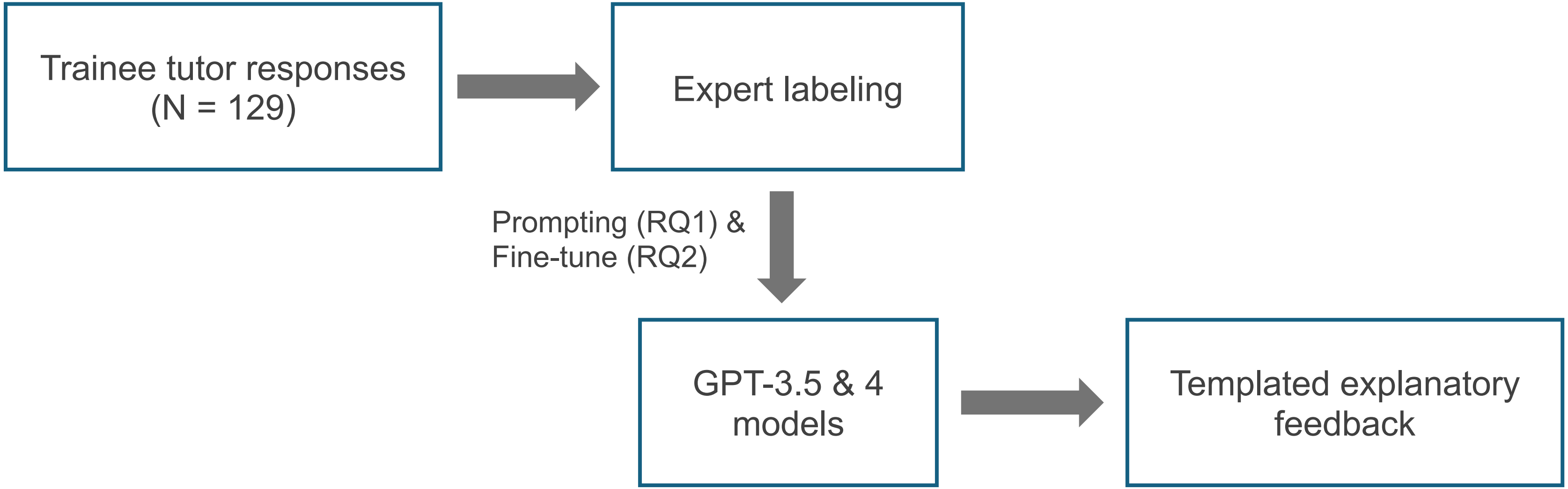
Founding source: Richard King Mellon Foundation, USA

Lin, J., Chen, E., Han, F., Gurung, A., Thomas, D. R., Tan, W., Ngoc, D. N., & Koedinger, K. R. (2024). How Can I Improve? Using GPT to Highlight the Desired and Undesired Parts of Open-ended Responses. *17th International Conference on Educational Data Mining*

Aleven, V., Popescu, O., & Koedinger, K. R. (2001). Towards tutorial dialog to support self-explanation: Adding natural language understanding to a cognitive tutor. In *Proceedings of Artificial Intelligence in Education* (pp. 246-255).

Provide feedback on trainee tutor responses

- **RQ1:** Does GPT provide accurate explanatory feedback?
- **RQ2:** Does fine-tuning improve the accuracy?



Founding source: Richard King Mellon Foundation, USA

Lin, J., Chen, E., Han, F., Gurung, A., Thomas, D. R., Tan, W., Ngoc, D. N., & Koedinger, K. R. (2024). How Can I Improve? Using GPT to Highlight the Desired and Undesired Parts of Open-ended Responses. *17th International Conference on Educational Data Mining*

Aleven, V., Popescu, O., & Koedinger, K. R. (2001). Towards tutorial dialog to support self-explanation: Adding natural language understanding to a cognitive tutor. In *Proceedings of Artificial Intelligence in Education* (pp. 246-255).

RQ1: Does GPT provide accurate explanatory feedback?

Highlights:

Proposed and validated a **metric (M-IoU)** to measure the quality of highlighted components identified by GPT models

Prompted GPT models to **highlight** the desired and undesired components from trainees’ responses

RQ1: Does GPT provide accurate explanatory feedback?

- Quality of sequence labeling

Table 3: Original training instance and different types of augmented instances. We highlighted outcome-based praise using yellow color and effort-based praise using blue.

	Instance	Label	Score
1	*John, you are making a really great effort.*	True	**1**

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. *AIED 2023 Workshop - Towards the Future of AI-augmented Human Tutoring in Math Learning*

RQ1: Does GPT provide accurate explanatory feedback?

- Quality of sequence labeling

Table 3: Original training instance and different types of augmented instances. We highlighted outcome-based praise using yellow color and effort-based praise using blue.

	Instance	Label	Score
1	*John, you are making a really great effort.*	True	**1**
2	*John, you are making a really great effort.*	Pred	**0.84**
3	*John, you are making a really great effort.*	Pred	**0.68**
4	*John, you are making a really great effort.*	Pred	**0.52**

Smaller penalty for extra predictions

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. *AIED 2023 Workshop - Towards the Future of AI-augmented Human Tutoring in Math Learning*

RQ1: Does GPT provide accurate explanatory feedback?
- Quality of sequence labeling

Table 3: Original training instance and different types of augmented instances. We highlighted outcome-based praise using yellow color and effort-based praise using blue.

	Instance	Label	Score
1	*John, you are making a really great effort.*	True	**1**
2	*John, you are making a really great effort.*	Pred	**0.84**
3	*John, you are making a really great effort.*	Pred	**0.68**
4	*John, you are making a really great effort.*	Pred	**0.52**
5	*John, you are making a really great effort.*	Pred	**0.2**

Larger penalty for missing predictions

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. *AIED 2023 Workshop - Towards the Future of AI-augmented Human Tutoring in Math Learning*

RQ1: Does GPT provide accurate explanatory feedback?

- Modified Intersection over Union (M-IOU)

TP – True Positive FP – False Positive FN – False Negative

$$F1\ score = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{TP}{TP + FP + FN}$$

Smaller penalty for extra predictions, larger penalty for missing predictions

RQ1: Does GPT provide accurate explanatory feedback?
- Modified Intersection over Union (M-IOU)

TP – True Positive FP – False Positive FN – False Negative

$$F1\ score = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{TP}{TP + FP + FN}$$

Smaller penalty for extra predictions, larger penalty for missing predictions

$$\text{M-IOU} = \frac{TP}{TP + \alpha \times FP + FN} \qquad (\alpha = 0.2)$$

M-IoU is used to measure the quality of highlighted components of tutor praise responses

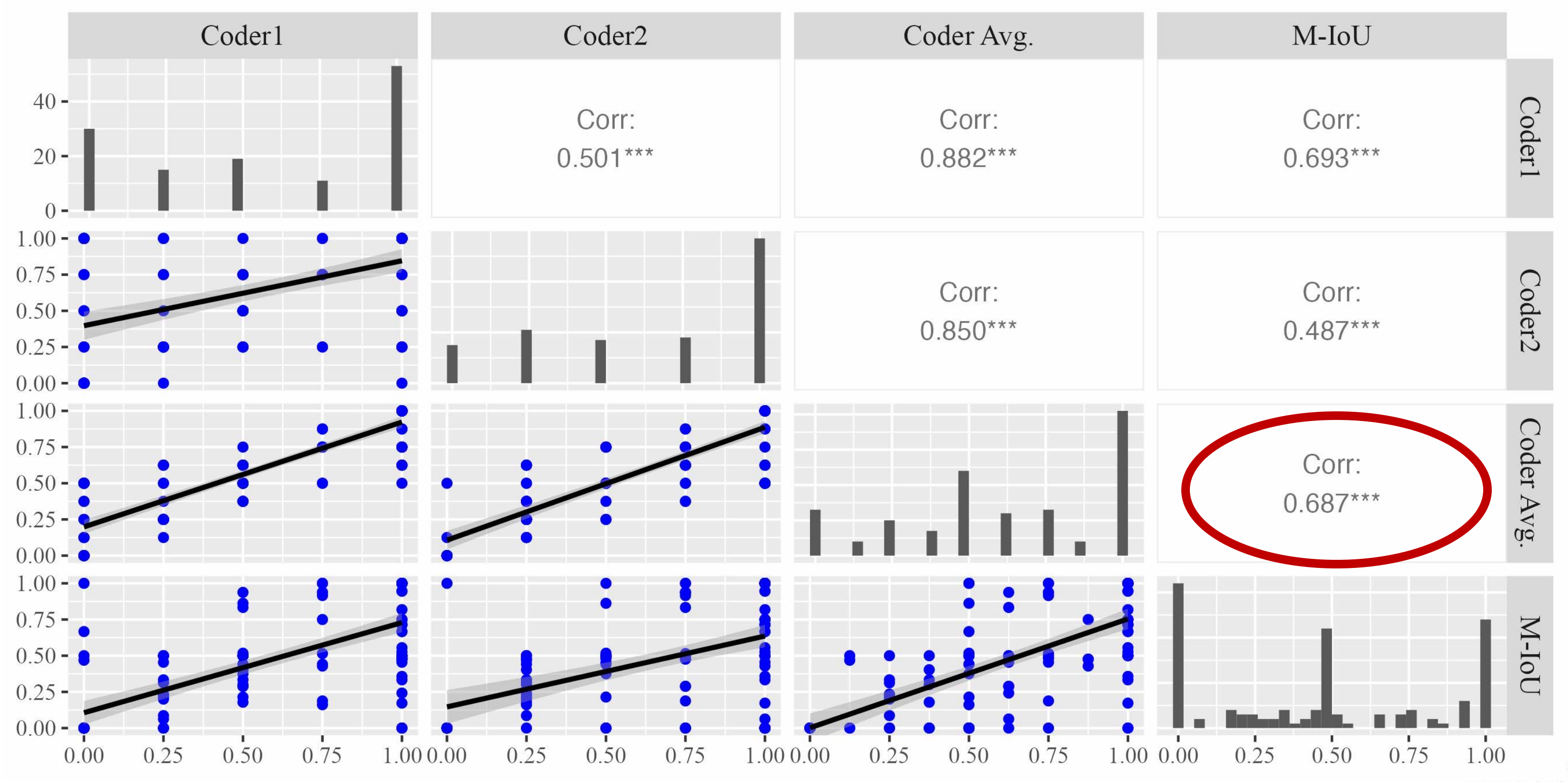
RQ1: Does GPT provide accurate explanatory feedback?
- Modified Intersection over Union (M-IOU)

Validate M-IOU score

Two human coders were assigned to evaluate the quality (five-point Likert scale) of shuffled highlighted responses, specifically focusing on the components of effort and outcome praise

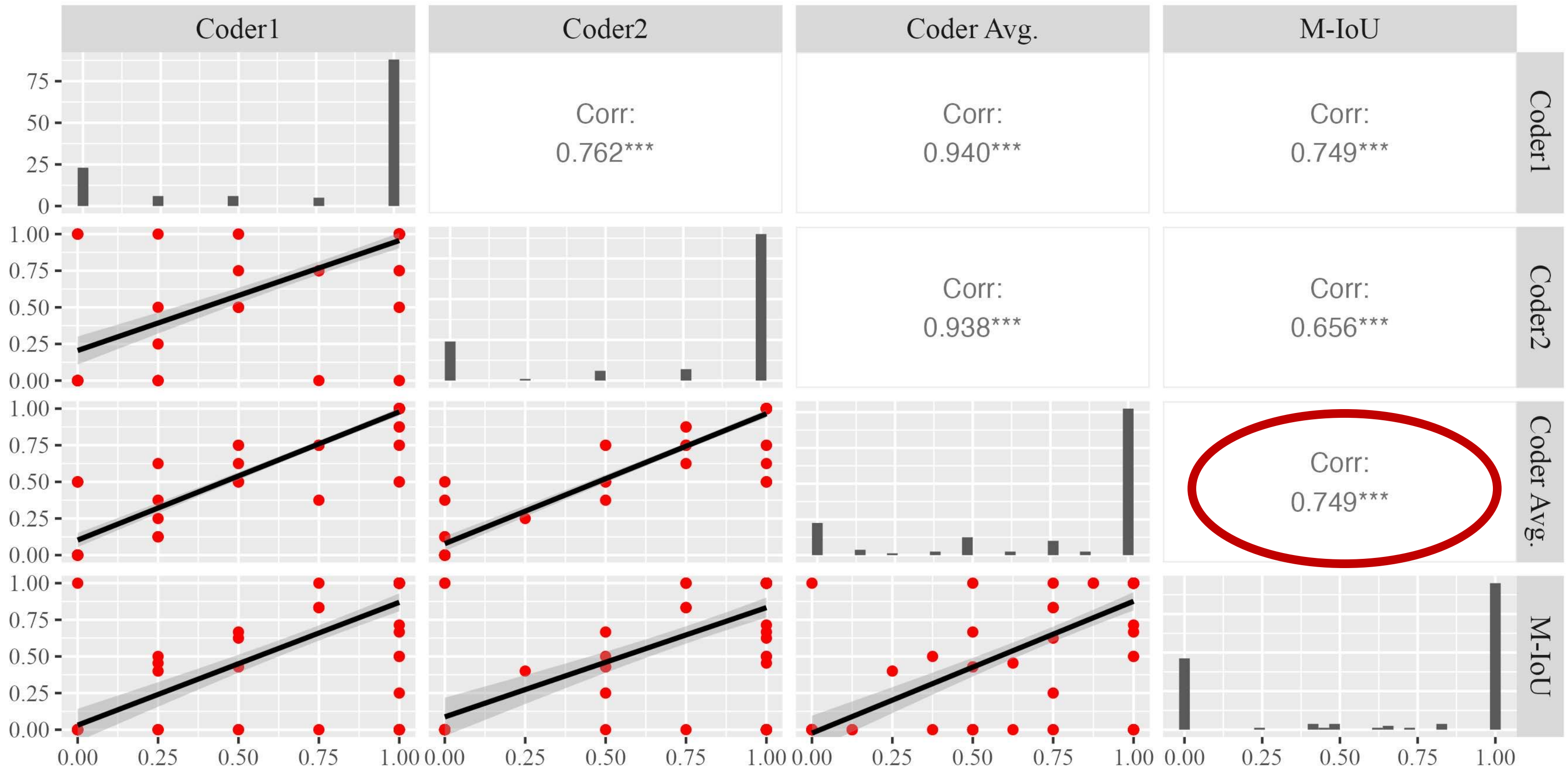
RQ1: Does GPT provide accurate explanatory feedback?

- Validate M-IOU score on effort-based praise



RQ1: Does GPT provide accurate explanatory feedback?

- Validate M-IOU score on outcome-based praise



RQ1: Does GPT provide accurate explanatory feedback?

- Analysis of prompting GPT models for highlighting key components

Table 4: Descriptive statistics of the scores rated by two human coders and measured by M-IoU scores

	GPT-3.5 turbo		GPT 4 turbo		Expert Annotation	
	Effort	Outcome	Effort	Outcome	Effort	Outcome
Comparison between the normalized human ratings and M-IoU scores						
Coder 1	$0.68_{0.38}$	$0.79_{0.40}$	$0.63_{0.36}$	$0.75_{0.39}$	$0.77_{0.35}$	$0.89_{0.30}$
Coder 2	$0.60_{0.43}$	$0.76_{0.40}$	$0.57_{0.40}$	$0.74_{0.40}$	$0.77_{0.35}$	$0.84_{0.35}$
Avg.	$0.64_{0.35}$	$0.77_{0.37}$	$0.60_{0.33}$	$0.75_{0.38}$	$0.77_{0.29}$	$0.87_{0.29}$
M-IoU	$0.46_{0.36}$	$0.68_{0.44}$	$0.47_{0.38}$	$0.64_{0.46}$	N/A	N/A
Proportion of human rating '*Agree*' or higher on our scale*						
Coder 1	64.06%	76.56%	53.13%	75.00%	73.44%	89.06%
Coder2	53.13%	73.44%	46.88%	71.88%	75.00%	84.38%
Avg.	56.25%	73.44%	53.13%	73.44%	75.00%	85.94%

Human coders rate **expert's highlight as accurate about 80% of the time** and they rate **GPT-3.5 as accurate about 65% of the time**.

RQ1: Does GPT provide accurate explanatory feedback?

- M-IoU scores for expert annotation on effort-based praise

Table 4: Descriptive statistics of the scores rated by two human coders and measured by M-IoU scores

	GPT-3.5 turbo		GPT 4 turbo		Expert Annotation	
	Effort	Outcome	Effort	Outcome	Effort	Outcome
Comparison between the normalized human ratings and M-IoU scores						
Coder 1	$0.68_{0.38}$	$0.79_{0.40}$	$0.63_{0.36}$	$0.75_{0.39}$	$0.77_{0.35}$	$0.89_{0.30}$
Coder 2	$0.60_{0.43}$	$0.76_{0.40}$	$0.57_{0.40}$	$0.74_{0.40}$	$0.77_{0.35}$	$0.84_{0.35}$
Avg.	$0.64_{0.35}$	$0.77_{0.37}$	$0.60_{0.33}$	$0.75_{0.38}$	$0.77_{0.29}$	$0.87_{0.29}$
M-IoU	$0.46_{0.36}$	$0.68_{0.44}$	$0.47_{0.38}$	$0.64_{0.46}$	N/A	N/A
Proportion of human rating '*Agree*' or higher on our scale*						
Coder 1	64.06%	76.56%	53.13%	75.00%	73.44%	89.06%
Coder2	53.13%	73.44%	46.88%	71.88%	75.00%	84.38%
Avg.	56.25%	73.44%	53.13%	73.44%	75.00%	85.94%

RQ1: Does GPT provide accurate explanatory feedback?

- M-IoU scores for expert annotation on effort-based praise

Table 4: Descriptive statistics of the scores rated by two human coders and measured by M-IoU scores

	GPT-3.5 turbo		GPT 4 turbo		Expert Annotation	
	Effort	Outcome	Effort	Outcome	Effort	Outcome
Comparison between the normalized human ratings and M-IoU scores						
Coder 1	$0.68_{0.38}$	$0.79_{0.40}$	$0.63_{0.36}$	$0.75_{0.39}$	$0.77_{0.35}$	$0.89_{0.30}$
Coder 2	$0.60_{0.43}$	$0.76_{0.40}$	$0.57_{0.40}$	$0.74_{0.40}$	$0.77_{0.35}$	$0.84_{0.35}$
Avg.	$0.64_{0.35}$	$0.77_{0.37}$	$0.60_{0.33}$	$0.75_{0.38}$	$0.77_{0.29}$	$0.87_{0.29}$
M-IoU	$0.46_{0.36}$	$0.68_{0.44}$	$0.47_{0.38}$	$0.64_{0.46}$	N/A	N/A
Proportion of human rating '*Agree*' or higher on our scale*						
Coder 1	64.06%	76.56%	53.13%	75.00%	73.44%	89.06%
Coder2	53.13%	73.44%	46.88%	71.88%	75.00%	84.38%
Avg.	56.25%	73.44%	53.13%	73.44%	75.00%	85.94%

RQ1: Does GPT provide accurate explanatory feedback?

- M-IoU scores for expert annotation on effort-based praise

Table 4: Descriptive statistics of the scores rated by two human coders and measured by M-IoU scores

	GPT-3.5 turbo		GPT 4 turbo		Expert Annotation	
	Effort	Outcome	Effort	Outcome	Effort	Outcome
Comparison between the normalized human ratings and M-IoU scores						
Coder 1	$0.68_{0.38}$	$0.79_{0.40}$	$0.63_{0.36}$	$0.75_{0.39}$	$0.77_{0.35}$	$0.89_{0.30}$
Coder 2	$0.60_{0.43}$	$0.76_{0.40}$	$0.57_{0.40}$	$0.74_{0.40}$	$0.77_{0.35}$	$0.84_{0.35}$
Avg.	$0.64_{0.35}$	$0.77_{0.37}$	$0.60_{0.33}$	$0.75_{0.38}$	$0.77_{0.29}$	$0.87_{0.29}$
M-IoU	$0.46_{0.36}$	$0.68_{0.44}$	$0.47_{0.38}$	$0.64_{0.46}$	N/A	N/A
Proportion of human rating '*Agree*' or higher on our scale*						
Coder 1	64.06%	76.56%	53.13%	75.00%	73.44%	89.06%
Coder2	53.13%	73.44%	46.88%	71.88%	75.00%	84.38%
Avg.	56.25%	73.44%	53.13%	73.44%	75.00%	85.94%

$$ratio = \frac{Avg(Coder1, Coder2)}{MIoU}$$

$$\widehat{MIoU}_{Effort-expert} = \frac{Avg(Coder1, Coder2)}{ratio}$$

RQ1: Does GPT provide accurate explanatory feedback?

- M-IoU scores for expert annotation on effort-based praise

Table 4: Descriptive statistics of the scores rated by two human coders and measured by M-IoU scores

	GPT-3.5 turbo		GPT 4 turbo		Expert Annotation	
	Effort	Outcome	Effort	Outcome	Effort	Outcome
Comparison between the normalized human ratings and M-IoU scores						
Coder 1	$0.68_{0.38}$	$0.79_{0.40}$	$0.63_{0.36}$	$0.75_{0.39}$	$0.77_{0.35}$	$0.89_{0.30}$
Coder 2	$0.60_{0.43}$	$0.76_{0.40}$	$0.57_{0.40}$	$0.74_{0.40}$	$0.77_{0.35}$	$0.84_{0.35}$
Avg.	$\boxed{0.64_{0.35}}$	$0.77_{0.37}$	$\boxed{0.60_{0.33}}$	$0.75_{0.38}$	$0.77_{0.29}$	$0.87_{0.29}$
M-IoU	$\boxed{0.46_{0.36}}$	$0.68_{0.44}$	$\boxed{0.47_{0.38}}$	$0.64_{0.46}$	N/A	N/A
Proportion of human rating '*Agree*' or higher on our scale*						
Coder 1	64.06%	76.56%	53.13%	75.00%	73.44%	89.06%
Coder2	53.13%	73.44%	46.88%	71.88%	75.00%	84.38%
Avg.	56.25%	73.44%	53.13%	73.44%	75.00%	85.94%

$$ratio = \frac{Avg(Coder1, Coder2)}{MIoU}$$

$$ratio_{Effort-GPT3.5} = \frac{0.64}{0.46} = 1.39$$

$$ratio_{Effort-GPT4} = \frac{0.60}{0.47} = 1.28$$

$$ratio = \frac{1.39 + 1.28}{2} = 1.34$$

RQ1: Does GPT provide accurate explanatory feedback?

- M-IoU scores for expert annotation on effort-based praise

Table 4: Descriptive statistics of the scores rated by two human coders and measured by M-IoU scores

	GPT-3.5 turbo		GPT 4 turbo		Expert Annotation	
	Effort	Outcome	Effort	Outcome	Effort	Outcome
Comparison between the normalized human ratings and M-IoU scores						
Coder 1	$0.68_{0.38}$	$0.79_{0.40}$	$0.63_{0.36}$	$0.75_{0.39}$	$0.77_{0.35}$	$0.89_{0.30}$
Coder 2	$0.60_{0.43}$	$0.76_{0.40}$	$0.57_{0.40}$	$0.74_{0.40}$	$0.77_{0.35}$	$0.84_{0.35}$
Avg.	$0.64_{0.35}$	$0.77_{0.37}$	$0.60_{0.33}$	$0.75_{0.38}$	$0.77_{0.29}$	$0.87_{0.29}$
M-IoU	$0.46_{0.36}$	$0.68_{0.44}$	$0.47_{0.38}$	$0.64_{0.46}$	**0.57**	N/A
Proportion of human rating '*Agree*' or higher on our scale*						
Coder 1	64.06%	76.56%	53.13%	75.00%	73.44%	89.06%
Coder2	53.13%	73.44%	46.88%	71.88%	75.00%	84.38%
Avg.	56.25%	73.44%	53.13%	73.44%	75.00%	85.94%

$$ratio = \frac{Avg(Coder1, Coder2)}{MIoU}$$

$$\widehat{MIoU}_{\ Effort-expert} = \frac{0.77}{1.34} = 0.57$$

If the prediction on effort-based praise can achieve an **M-IoU of 0.57**, the quality is comparable to expert annotation.

RQ1: Does GPT provide accurate explanatory feedback?

- M-IoU scores for expert annotation on outcome-based praise

Table 4: Descriptive statistics of the scores rated by two human coders and measured by M-IoU scores

	GPT-3.5 turbo		GPT 4 turbo		Expert Annotation	
	Effort	Outcome	Effort	Outcome	Effort	Outcome
Comparison between the normalized human ratings and M-IoU scores						
Coder 1	$0.68_{0.38}$	$0.79_{0.40}$	$0.63_{0.36}$	$0.75_{0.39}$	$0.77_{0.35}$	$0.89_{0.30}$
Coder 2	$0.60_{0.43}$	$0.76_{0.40}$	$0.57_{0.40}$	$0.74_{0.40}$	$0.77_{0.35}$	$0.84_{0.35}$
Avg.	$0.64_{0.35}$	$0.77_{0.37}$	$0.60_{0.33}$	$0.75_{0.38}$	$0.77_{0.29}$	$0.87_{0.29}$
M-IoU	$0.46_{0.36}$	$0.68_{0.44}$	$0.47_{0.38}$	$0.64_{0.46}$	**0.57**	**0.76**
Proportion of human rating '*Agree*' or higher on our scale*						
Coder 1	64.06%	76.56%	53.13%	75.00%	73.44%	89.06%
Coder2	53.13%	73.44%	46.88%	71.88%	75.00%	84.38%
Avg.	56.25%	73.44%	53.13%	73.44%	75.00%	85.94%

$$ratio = \frac{Avg(Coder1, Coder2)}{MIoU}$$

$$\widehat{MIoU}_{\ Outcome-expert} = \frac{0.87}{1.15} = 0.76$$

If the prediction on outcome-based praise can achieve an **M-IoU of 0.76**, the quality is comparable to expert annotation.

RQ2: Does fine-tuning improve the accuracy?

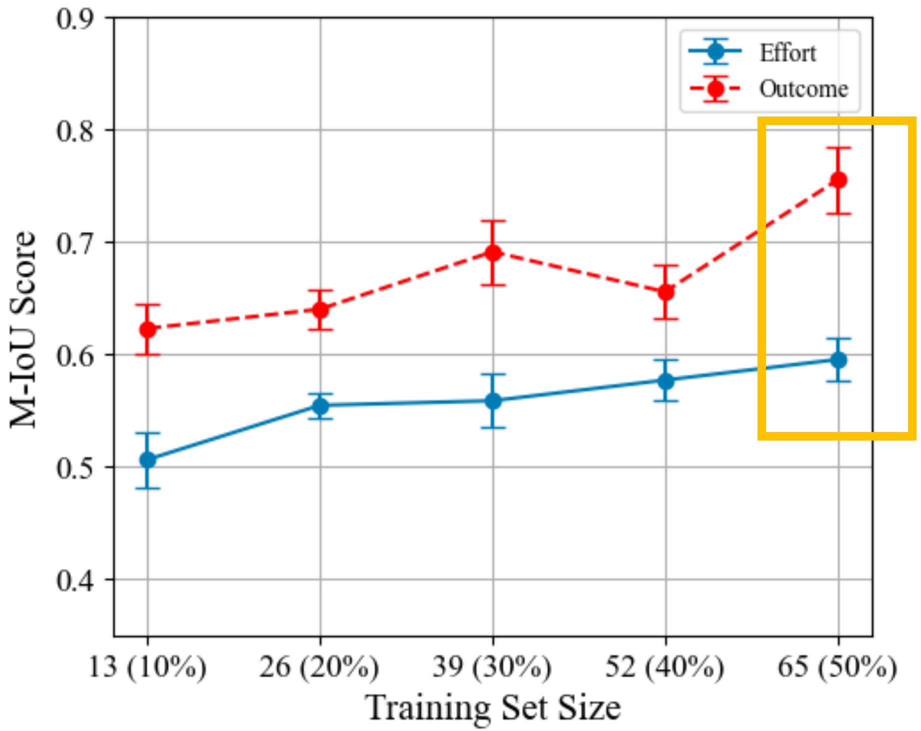
Highlights:

Fine-tuning GPT-3.5 can improve the accuracy of **highlighting** the desired and undesired components in trainees' responses

Examine the influence of different ratios of fine-tuning set on model performance.

RQ2: Does fine-tuning improve the accuracy?

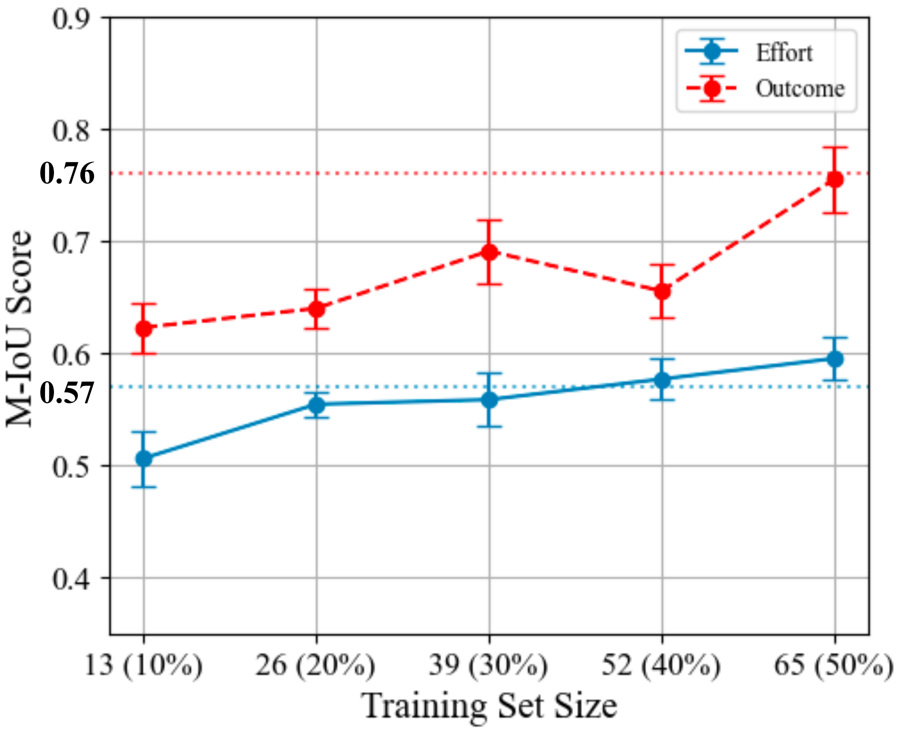
- Analysis of finetuning GPT models for highlighting key components



Praise type	Min (M-IoU)	Mean (M-IoU)	Max (M-IoU)
Effort	0.54	0.60	0.64
Outcome	0.66	0.76	0.84

RQ2: Does fine-tuning improve the accuracy?

- Analysis of finetuning GPT models for highlighting key components



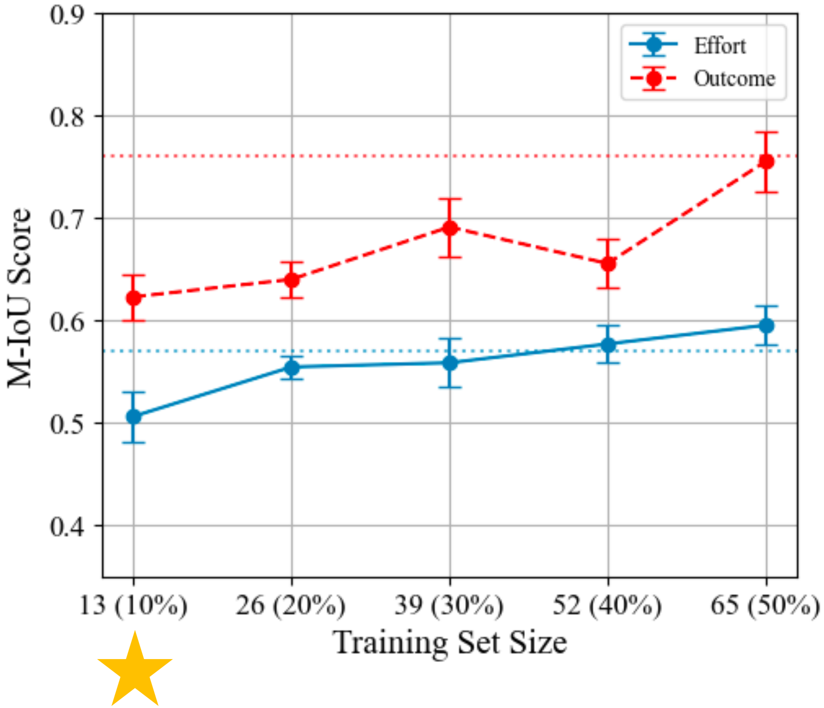
Praise type	Min (M-IoU)	Mean (M-IoU)	Max (M-IoU)
Effort	0.54	0.60	0.64
Outcome	0.66	0.76	0.84

Expected model performance

$$\widehat{MIoU}_{Effort-expert} = 0.57$$
$$\widehat{MIoU}_{Outcome-expert} = 0.76$$

Ongoing and Future Works

Data Augmentation



Our mission is to double math learning for 10,000 middle school students by 2026.

PLUS
Personalized Learning²

Carnegie Mellon University