# Leveraging a human-in-the-loop, chain-of-thought prompting approach to generate feedback on middle school short-answer responses in science

ANONYMOUS AUTHOR(S)*

This research explores a novel human-in-the-loop approach that goes beyond traditional prompt engineering approaches to harness Large Language Models (LLMs) with chain-of-thought prompting for grading middle school students' short answer formative assessments in science and generating useful feedback. While recent efforts have successfully applied LLMs and generative AI to automatically grade assignments in secondary classrooms, the focus has primarily been on providing scores for mathematical and programming problems with little work targeting the generation of actionable insight from the student responses. This paper addresses these limitations by exploring a human-in-the-loop approach to make the process more intuitive and more effective. By incorporating the expertise of educators, this approach seeks to bridge the gap between automated assessment and meaningful educational support in the context of science education for middle school students. We have conducted a preliminary user study, which suggests that (1) co-created models improve the performance of formative feedback generation, and (2) educator insight can be integrated at multiple steps in the process to inform what goes into the model and what comes out. Our findings suggest that in-context learning and human-in-the-loop approaches may provide a scalable approach to automated grading, where the performance of the automated LLM-based grader continually improves over time, while also providing actionable feedback that can support students' open-ended science learning.

CCS Concepts: • **Human-centered computing** → **HCI design and evaluation methods**; • **Applied computing** → **Education**.

Additional Key Words and Phrases: LLM, formative feedback, STEM learning

**ACM Reference Format:**

Anonymous Author(s). 2023. Leveraging a human-in-the-loop, chain-of-thought prompting approach to generate feedback on middle school short-answer responses in science. In . ACM, New York, NY, USA, 22 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

Recent advances in Large Language Models (LLMs) raise a plethora of new research inquiries that are centered around the intricate relationship between stakeholders and these AI models [17, 28, 44]. From a design perspective, this includes the optimization of user interactions and intelligent systems that provide more accurate results, are ethical and equitable, and leverage the human-AI partnership. From an application perspective, this involves meeting instructors' specific needs and reducing their effort in their day-to-day grading of assignments, but at the same time producing results that are interpretable and useful to students. In line with this research landscape, our paper embarks on a journey to create an 'educator-in-the-loop' tool that empowers educators by providing them with mechanisms to use LLMs for automatically grading and delivering feedback on short answer responses for middle school students in science.

Recently, there has been a push to shift science learning from rote, fact-based instruction towards one that promotes a deeper understanding of concepts and processes, and links the science to real-world, problem-based approaches. Therefore, instructors have to be more involved in facilitating and orchestrating students' science knowledge construction and



Manuscript submitted to ACM

problem-solving skill development [38]. Educators must monitor, evaluate, and respond to copious amounts of differing student data (e.g., classroom discourse, written responses, problem solutions, etc) in ways that give students' agency in their learning — all while aligning and adhering to school, state, and national standards and learning objectives for each target concept [16]. While these problem-based learning approaches have supported students' integrated learning of Science, Technology, Engineering, and Mathematics (STEM) domains, particularly through the leveraging of technology-enhanced learning environments [31, 38], they also warrant innovative ways to evaluate and provide timely feedback to support all students as they construct new knowledge, ideas, and problem-solving skills [41].

Formative assessments emerge as a pivotal tool in this endeavor, aiding students in cultivating self-evaluation skills and providing timely feedback and guidance when they encounter challenges [4]. However, the labor-intensive nature of grading and generating personalized feedback for formative assessments, especially when conducted at frequent intervals, presents a significant burden for educators and is susceptible to errors [14, 30].

This is where Large Language Models (LLMs) can come into play, offering the potential for automating the scoring of short answer responses [12, 45], providing students with feedback to identify their successes and overcome challenges [20], and assisting educators in identifying opportunities for engaging in and supporting student learning, monitoring student difficulties, and enacting enhanced learning experiences [46]. However, there exists limited prior research that (1) specifically addresses the automation of formative assessment grading and feedback provision in science domains, (2) identifies pathways for actionable insights that can be leveraged by educators, and (3) incorporates educator insight into the technical level implementation of these technologies through human-in-the-loop approaches.

In this paper, we present a user study on an educator-researcher partnership for developing a methodology to support formative assessment short answer scoring with explanations for the assigned scores. In particular, we focus on *human-in-the-loop* learning combined with in-context learning and chain-of-thought reasoning with the goal of developing a methodology that can be customized for individual educator needs. We demonstrate the effectiveness of our approach using a dataset of formative question answers collected in a middle school classroom for a water runoff science curricular unit. In the discussion section, we consider issues in interface design to support educator usability and the use of our approach as an 'educator-in-the-loop' feedback generation tool to support student learning. We also discuss the limitations of our work and future directions that may help us address these limitations.

## 2 BACKGROUND

Advances in Natural Language Processing (NLP) have ignited interest in enhancing and automating assessment scoring [46]. These approaches have produced methodologies such as next sentence prediction strategies [45], prototypical neural networks [47], cross-prompt fine-tuning [12], and reinforcement learning from human feedback (RLHF) [26] methods. These methods have highlighted the potential of leveraging LLMs for automated scoring but fall short of providing actionable feedback separate from a performance score. In addition, they focus mainly on structured tasks in disciplines such as mathematics. The complexities of analyzing open-ended responses in science may limit or convolute the application of such approaches for scoring and generating feedback in this domain.

A key limiting factor of automated evaluation and feedback in such contexts is data impoverishment. These educational datasets tend to be characterized by limited data volumes, imbalanced response representations, and non-standardized syntax and semantics [7]. For example, LLM fine-tuning methods require (1) a substantial amount of annotated data and accompanying computational resources to accomplish short-answer scoring tasks and (2) ungeneralizable approaches that utilize a unique model for each computational task [12]. More application research is needed to understand how to best leverage LLM advances to support this important educational need.

In science, teachers need to actively evaluate students' developing ideas and reasoning skills as they construct knowledge of the key science concepts and practices [16]. Evidence-based approaches have been used to develop formative assessments that align with science standards and learning science theory [41, 42]. But, given the need to assess students' critical thinking skills, the assessments are often structured to allow students to provide open-ended responses [13]. At the same time, it is important for the teachers to assess students' evolving knowledge and provide timely feedback that helps them achieve their learning objectives [41]. Continual grading of short answer formative assessments can unduly burden teachers who also have to focus on problem-based instruction of difficult science material. Therefore, automating formative assessment grading with teacher input can provide large benefits to teachers and support student learning. However, more work is needed to automate the evaluation and feedback generation process in ways that maintain alignment with educational goals and consider the educational needs of all students.

Recent approaches to grading and feedback generation have leveraged human-the-loop methodologies to address data impoverishment and rubric-alignment concerns with success (e.g., 19). *Active learning* [29, 35] is a human-in-the-loop approach that improves model training by consulting an "oracle" (in this case, the human) to label additional training instances. In our case, active learning allows the educator to examine the model's incorrectly scored instances to identify LLM errors that caused the model to incorrectly score multiple instances. Incorrectly predicted instances containing these patterns can then be reinserted into the prompt as few-shot exemplars, where chain-of-thought reasoning is used to address the model's reasoning errors that caused the scoring misalignment with the human.

In the past, automated assessment evaluations have been developed for well-structured tasks (e.g., in introductory computer science courses) with human-in-the-loop training conducted by researchers familiar with the intricacies of LLMs. In educational technology design, efforts to integrate educator insight in the design of learning technologies have highlighted the need to ensure proper training (particularly in AI literacy) and AI explainability to encourage educator insight and direct contribution to the application development process. This allows for better accommodation of the needs, concerns, and preferences of educators in developing automated assessment schemes for their classrooms [16, 21]. In this work, our goals are to extend these approaches to develop an "*educator-in-the-loop*" interface that allows educators to partner with LLM models to enrich the grading schemes while also generating meaningful feedback that provides them with actionable information and their students with explanations that help them overcome their difficulties and misconceptions. To do so, we take a first step in our educator-in-the-loop approach by developing a methodology where researchers, familiar with both LLMs and educators' needs, develop a methodology for classroom teachers to fine-tune their LLM assessment models for support their own instruction and student feedback.

## 3 FRAMING OUR EDUCATOR-IN-THE-LOOP LLM APPROACH

Fig. 1 illustrates our framework that combines human-in-the-loop learning with learning sciences theory to design tools for grading and feedback for formative assessments. From a technical perspective, the previous section highlights the need to better understand how LLM advances can support open-ended, short-answer assessment evaluations, given the limitations of educational datasets, and the need for interfaces and a methodology to support human-in-the-loop tuning and customization. The Literature Review box emphasizes that the design and development processes need to use the technology appropriately to identify methods that ensure alignment with learning objectives and analyze student responses in ways that are equitable for all learners.

From the Human-In-The-Loop Input perspective, the design and development processes need to consider how to solicit and incorporate educator perspectives, which allows them to tailor the grading scheme and feedback to their

prior and current classroom experiences. It is also important that the educator be supported in developing some level of AI literacy.

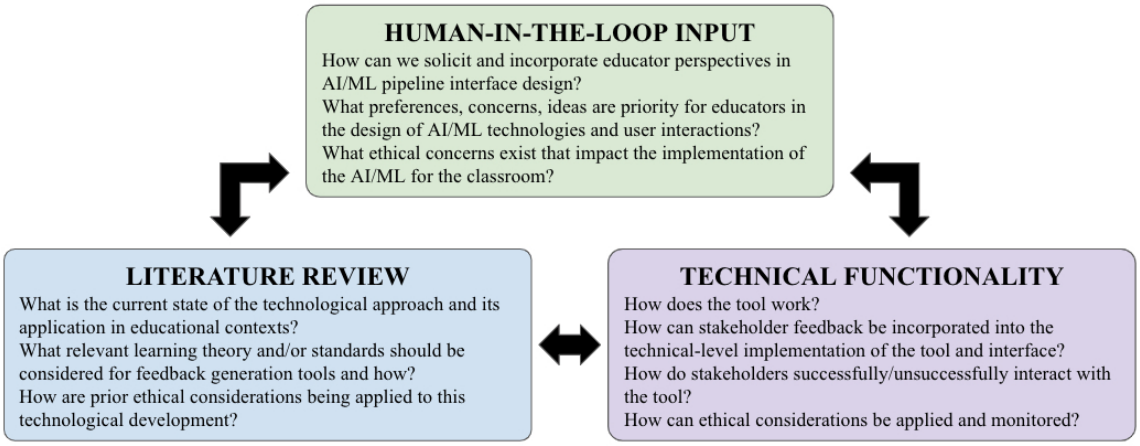



Fig. 1. Human-in-the-loop and learning theory considerations for AI/ML technology development

From the Technical Functionality perspective, our method extends previous work [1]. The key idea for initiating LLM prompt engineering is the use of an educator-designed rubric to initially inform the LLM-based grading scheme. The rubric design incorporates educator insight into the scoring processes, thereby creating a partnership between the LLM and the educator. To further extend the human-in-the-loop approach, we leverage an emergent behavior in LLMs known as *in-context learning* (ICL) [5], where the LLM uses a few labeled instances in the prompt (few-shot) to inform its grading and explanation generation without traditional training (i.e., training or tuning to make a lot of parameter updates). This allows educators to use the same foundation model across different formative assessments and domains simply by changing the prompt to match their desires for a particular assessment. We utilize inter-rater reliability (IRR) as a method to not only achieve consensus among human scorers but also to identify "sticking points" where the differences between human scorers may similarly make it difficult for the LLM to score in a consistent manner.

We include a subset of these sticking point instances in the initial prompt and augment each of the few-shot instances in the prompt with a form of ICL called *chain-of-thought* (CoT) reasoning [39] to align the model with the human scoring consensus achieved during IRR. While traditional ICL instances are comprised of question-and-answer pairs, CoT answers are accompanied by reasoning chains that explain the rationale behind the correct answers. CoT reasoning has been shown to improve model performance over traditional ICL [39], and these reasoning chains are of further benefit to educators who can use the model's scoring explanations to provide feedback to students, for example, why an assessment point was or was not awarded. Furthermore, the model's reasoning chains can be used to highlight specific causes of scoring misalignment between the LLM and the educator, which can then be further addressed via CoT reasoning in the prompt to help correct the LLM's misalignment and improve its scoring and reasoning capabilities.

This process can also inform rubric refinement, as it alerts the educator to ambiguities in the rubric and formative assessment questions that are confusing for the LLM and possibly the students. Further, combining CoT and active learning can help the educator identify human errors made when generating the initial scoring procedure. In such cases, the educator (human scorer) may side with the LLM over his or her initial score.

## 4 METHOD

Our human-in-the-loop approach to fostering a partnership between an educator and LLM to score and explain students' short answer formative assessment responses in the Earth Science domain is summarized in Figure 2. It summarizes a researcher interface for open-ended, short-answer formative feedback generation. In the application of this approach, a computer science researcher and learning scientist with teaching experience collaborated in using GPT-4 with few-shot, in-context learning with chain-of-thought reasoning, and active learning to design and develop a scoring and feedback mechanism for middle school formative assessments that covered applications of conservation of matter principles in Earth Science. The overall curriculum integrates earth sciences, computing, and engineering to teach students about water runoff by developing conceptual models; then construction, debugging, and testing of computational models; and finally, the use of those models to design a schoolyard that minimizes water runoff while meeting cost and accessibility constraints.

We describe each step of this formative feedback generation application in this section. Specific details for applying our method (the blue diamonds inside the green box in Figure 2) to each formative assessment question can be found in the Supplemental Materials.

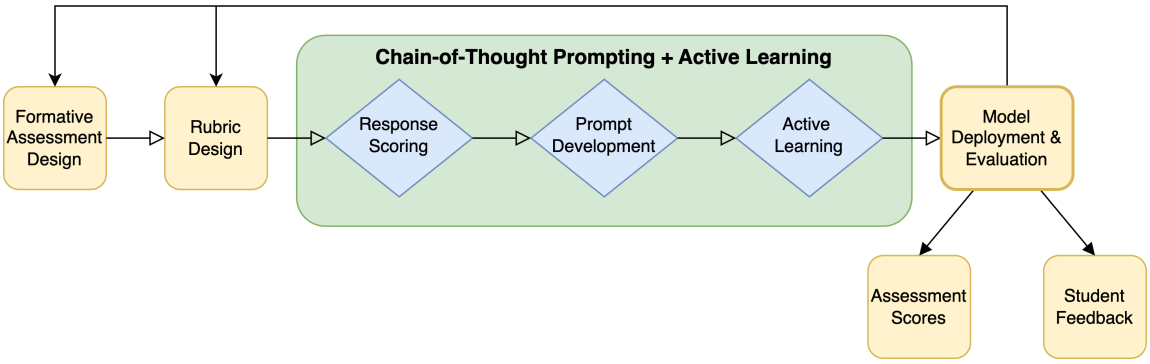



Fig. 2. Chain-of-Thought Prompting + Active Learning. The blue diamonds in the green box are specific steps in the method, while the yellow boxes correspond to the process's classroom applications and areas of interest to the educator.

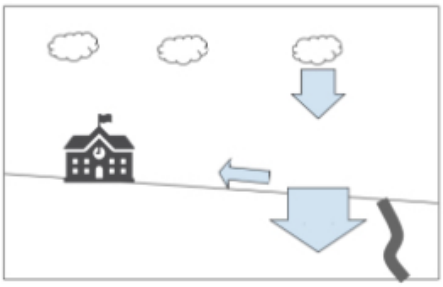
### 4.1 Formative Assessment and Rubric Design

We leverage an evidence-centered design (ECD; [22]) approach for assessment development. This process supports developing assessments that focus on science knowledge concepts and practices as well as problem-solving skill development. ECD ensures that systematic links are established between components of the curriculum and assessments that provide evidence of students' proficiency with the target knowledge and skills [2]. including engaging in argument from evidence, and developing and using models [24].

In-time analysis of these assessments provides us with opportunities to provide evidence-based, formative feedback to better support students' learning by construction, debugging, and evaluation of their developing conceptual models. In this paper, we analyze three formative questions that focus on students' conceptual knowledge development. The three questions and the rubrics for grading each question are provided in Fig. 3. The design of the rubrics supported the Prompt Development phase, discussed in more detail below.

**Abbreviations**
**SCI**: Application of Science Domain Knowledge
**MOD**: Application of Scientific Modeling Knowledge
**SEP**: Science and Engineering Practice (from NGSS)
**DCI**: Disciplinary Core Idea
**CCC**: Crosscutting Concept (from NGSS)

**Libby's Model**



**FA1-Q1: What do you think the different sized arrows in Libby's model could mean?**

| Point | Description | Domain |
|---|---|---|
| 1 | Evidence of: different sized arrows illustrate amount of water (SEP) | SCI, MOD |
| 0 | Other: Implies importance of action, describes where/how water goes | SCI, MOD |

**FA1-Q2. What are two things that Libby's model does a good job of explaining? Explain your answer.**

| Item | Point | Description | Domain |
|---|---|---|---|
| A | 1 | Demonstrating how elements in the model represent the target science concepts (student can describe any and/or all): rainfall, absorbed and/or runoff (DCI) | SCI |
| | 1 | Demonstrates scientific reasoning in their explanation; *e.g., the student discusses how the model uses the arrow from the cloud to demonstrate that it is raining* | SCI |
| B | 1 | Using arrow size to indicate water amounts; the student can describe the rainfall or runoff arrows or a general description of the symbol usage (SEP) | MOD |
| | 1 | Demonstrates scientific reasoning in their explanation; *e.g., the student discusses how to use symbol size to indicate different amounts of water - student does not get a point if they state that the absorption size is correct because it is depicted as larger than the rainfall* | MOD |
| | 0 | Showing direction of runoff* | SCI |
| | 0 | Illustrating the amount of absorption based on rainfall** | SCI |

* demonstrates conceptual knowledge error w/runoff (DCI)
**demonstrates misunderstanding/misapplication of conservation of matter principle (CCC)

**FA1-Q3. What are two things that you would change about Libby's model to explain where the water goes? Explain your answer.**

| Item | Point | Description | Domain |
|---|---|---|---|
| A | 1 | Identified that the direction of the runoff arrow is incorrect (SEP) | SCI |
| | 1 | Demonstrates scientific reasoning in their explanation *e.g., the student describes that the runoff arrow should be facing the other direction as the school is on a hill and water that is not absorbed/on surface would move downhill.* | SCI |
| B | 1 | The size of the arrows must change to correctly adhere to conservation of matter (SEP, CCC) | SCI |
| | 1 | Demonstrates scientific reasoning in their explanation *e.g., the student describes that the absorption arrow size cannot be larger than the rainfall since total rainfall should be equal to the amount of absorption and the amount of runoff* | SCI |
| | 0 | Model representation: more arrows, more words* | MOD |
| | 0 | Science representation: adding an arrow to stream, flooding** | SCI |

* demonstrates difficulties with model representation scheme/semantics (SEP)
**demonstrates misunderstanding/misapplication of conservation of matter principle (DCI, CCC)

Fig. 3. The formative assessment questions and their accompanying ECD-based rubrics that we analyze in this paper.

### 4.2 Response Scoring

Two of this paper's authors (education researcher + computer science researcher, henceforth referred to as the "raters") independently scored a random subset (20%) of the student responses for each of the three questions using the rubric developed by the education researcher in consultation with classroom teachers who are intimately familiar with the curriculum. While conducting IRR, the raters noted instances where they agreed and disagreed on the scores they assigned to students' answers.

The raters set themselves the following goals: (1) reach an agreement on rubric definitions that would be used to describe the grading and explanation generation schemes to the LLM; and (2) support human-scoring of all assessment items to compare performance. This task aligns with general educator goals of succinct rubric generation so that learners can leverage rubrics as a resource for completing tasks. Particular attention was paid to disagreements that caused multiple instances to be scored differently. Instances that were agreed upon by both before the consensus-generation activity served as "ground truth" exemplars, providing the LLM with an initial alignment to the human scorers. "Sticking point" instances (i.e., disagreements that caused *multiple* instances to be scored differently by the raters) were included in the prompt to highlight specific reasons for misalignment between humans that the model was likely to have problems with for similar reasons.

We repeated this process for each (of the three) formative assessment questions until an inter-rater Cohen's $k > 0.7$ was achieved across all subscores. After this, the educator scored the full set of student responses. We split the dataset into training and testing sets (80% and 20%, respectively) prior to developing the initial prompt. The training set instances that were not used as few-shot examples in the prompt were used as a validation set for the active learning process.

### 4.3 Prompt Development

The first component of the prompt introduced the LLM to its task via the *persona pattern* [40]. We informed the LLM that its job was to play the role of a middle school teacher to align the model goals with the pedagogical objectives of the educator when evaluating students' formative assessment responses. In the following portion of the prompt, we introduced the rubric, which served three purposes: (1) to provide the framework that the model should use for its scoring decisions, (2) to provide the format for improving the readability of the model's generated responses, and (3) to enable programmatic response parsing.

The next part of the prompt included ground truth and sticking point instances to further refine the LLM scoring and explanations. Ground truth examples included CoT reasoning to explain why the student should or should not receive a point for each allocated subscore according to the rubric (i.e., as further support for "ground truth" instance scoring). Sticking point instances contained similar CoT reasoning chains but were directed at aligning the model with the IRR consensus. These were the instances where the human scorers struggled with reaching an agreement during IRR. As discussed, we believed that these responses would also prove difficult for the model to score. Each few-shot instance in the prompt adhered to the following CoT template: *evidence in the student response + reference to the rubric + score* (as shown below).

> The student says X. The rubric states Y. Based on the rubric, the student earned a score of Z.

This process involved citing quotations from the student's response as evidence, tying that evidence to the rubric, and returning a score and explanation to the model to guide its reasoning during inference. Our approach mirrors the original CoT publication [39], which prompted the LLM to break down algebraic word problems step-by-step to guide the model toward generating correct solutions to the problem.

Additional few-shot examples were added to the prompt to balance the individual subscores to be as equally distributed as possible. While we did not use data augmentation in this work, prior work demonstrated augmenting underrepresented classes to balance the dataset can improve model performance considerably [7–9]. In our case, we simply selected additional instances from the training set and added them to the list of labeled instances in the prompt.

However, the small and imbalanced nature of the dataset was a constraint. Because we were balancing across four subscores per student, i.e., assigning multiple labels for responses to Q2 and Q3, it was not always possible to achieve an exact balance. This is because adding one additional instance to the prompt to augment a single subscore naturally affects the balance across all other subscores. In some cases, it was simply not possible to achieve a perfect balance given the training set data, but we included at least one positive and one negative instance across all subscores in each of the three formative assessment question prompts. The initial prompts for each of the three questions are included in the Supplemental Materials.

### 4.4 Active Learning

During active learning, the validation set instances (i.e., those instances in the training set not used as few-shot examples in the initial prompt) were fed through the LLM. As a next step, the raters conducted error analysis to identify patterns where the LLM still generated incorrect responses (scores and/or explanations). We paid particular attention to the *reasoning* provided by the system to each of its incorrect scoring predictions and took note of any reasoning errors that caused the model to mislabel multiple instances (similar to "sticking points" during *Response Scoring*). The raters determined that these responses would be the candidates for generating new few-shot instances to be inserted into the prompt. The prompt would use CoT reasoning to address the LLM errors and better align the model with the human scorers. This enabled us to potentially correct several mis-scored instances. For the educator, this was analogous to identifying misunderstandings and providing new learning opportunities to overcome student difficulties.

The number of student responses that were mis-scored was used to prioritize candidate selection to maximize the impact of CoT prompting. For Q1, there were only a handful of incorrectly predicted scores in the validation set, so we added them all back into the prompt during active learning. For Q2 and Q3, the researcher identified the $n$ most useful instances to add back into the prompt, where $n$ was chosen based on the number of model reasoning errors, context length, and API call and token limits.

For all questions, CoT reasoning was provided to the LLM with each additional labeled instance to correct the model's faulty reasoning found during active learning. We again rebalanced the few-shot instances in the prompt across subscores as needed. In general, active learning can be performed until one of several stopping conditions is met:

- The model no longer produces any incorrect validation generation scores (i.e., it achieves convergence);
- The model predicts more validation scores incorrectly than in previous iterations (i.e., it overfits);
- There are not enough instances remaining in the validation set to achieve an acceptable data balance in the prompt; and
- Other real-world constraints, such as API call limits, API token limits, context-length limits, cost limits, and so on prevent additional analyses.

While testing our method, we performed one active learning iteration for each of the three formative assessment questions. For each subscore, we first identified trends in the model's scoring errors by answering the following question: *are model scoring errors primarily false negatives (underscoring) or false positives (overscoring)*? This informed us of the "direction" we needed to guide the model to accurately present the human scorers' consensus. Once we identified the

model's scoring error trend, we examined those instances to ascertain the common causes for the model's incorrectly predicted scores in the validation set. We identified the most prevalent model reasoning error (i.e., the error resulting in the greatest number of mis-scored validation set instances) and reinserted one of these instances back into the prompt with CoT reasoning to correct the error. As an example, we found the ratio of false positives to false negatives was 5:2 for the Q3 *Runoff Direction* subscore. We also discovered that more than half of these false positives were the result of the model awarding points to students who mentioned the arrows in the diagram needed to change direction. This was actually incorrect, as only the runoff arrow needed to change direction (the other arrows were already pointing in the correct direction). To address this problem, we chose one of these incorrectly predicted validation instances, inserted it into the prompt, and used CoT reasoning to help correct the model's misconception going forward.

### 4.5 Model Evaluation

To evaluate our approach, we chose GPT-4[1] [25] as our LLM given its wide recognition as the current state-of-the-art in language models [25, 36, 48]. We evaluated our short answer assessment approach by comparing its performance on the test set to three incremental baselines: (1) *Zero-Shot*, (2) *Few-Shot*, and (3) *Few-Shot CoT*. The Zero-Shot baseline included the rubric in the prompt but no labeled examples. The Few-Shot baseline added labeled instances to the prompt, but those instances did not include CoT reasoning in the answers (i.e., only numerical scores were provided as labels). The Few-Shot CoT baseline added CoT reasoning to the labeled instances from the Few-Shot baseline. Active learning was incorporated as a last step, and all three baselines were then compared to our full Chain-of-Thought Prompting + Active Learning method. By evaluating our method across incremental baselines, we examined the effects of adding specific components of the method and ascertained the degree to which each component affected the model's performance in terms of both scoring and providing informative explanations to the educator.

Going by prevalence in the literature, we chose Macro-F1 and Cohen's Quadratic Weighted Kappa (QWK) [11] as our performance metrics for evaluating the overall model performance [32, 34]. Macro-F1 as opposed to Micro-F1 was more useful because of our dataset's imbalance across subscores. In particular, the scientific reasoning subscores are heavily weighted in favor of the negative class, as students often do not demonstrate scientific reasoning in their formative assessment responses. Cohen's QWK was chosen over the traditional Cohen's $k$ [10] because it accounts for the degree of disagreement between reviewers, making it a very useful metric for our ordinal data. The mathematical notation for both metrics is discussed below. In addition, we computed the accuracy of the results for the different methods, but this metric is not used in the performance comparisons. Model performance comparisons for each of the three formative assessment questions are presented in Sections 5.1, 5.2, and 5.3. We provide code for both the LLM response generation and our analysis in the Supplemental Materials.

Equation 1 shows the computation of the F1-score from a confusion matrix consisting of true and false positives (TP and FP) and true and false negatives (TN and FN):

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

Macro-F1 is simply the arithmetic mean of the F1-scores across all $n$ classes. This is shown in Equation 2.

$$Macro\text{-}F1 = \frac{\sum_{i=1}^{n} F1}{n} \tag{2}$$

[1]https://openai.com/research/gpt-4

Equation 3 shows the Cohen's QWK calculation. $w_{ij}$, $f_{o_{ij}}$, $f_{e_{ij}}$ denote the quadratic weights, observed frequencies, and expected frequencies, respectively. $f_{o_{ij}}$ is the matrix of observed ratings by the reviewers. $f_{e_{ij}}$ (expected frequencies) are calculated by first multiplying the row- and column-wise sums of $f_{o_{ij}}$ divided by the number of total ratings to get the expected probabilities. The expected probabilities are then multiplied by the total number of ratings to get the expected frequencies matrix, $f_{e_{ij}}$.

$$QWK = 1 - \frac{\sum w_{ij} f_{o_{ij}}}{\sum w_{ij} f_{e_{ij}}} \tag{3}$$

The quadratic weights are represented by the two-dimensional matrix $w$, where each dimension of the matrix ($i$ and $j$) corresponds to a different rater and $w_{ij}$ is a specific weight in the matrix (see Equation 4). $N$ is the number of available (ordinal) categories available to raters.

$$w_{ij} = \frac{(i-j)^2}{(N-1)^2} \tag{4}$$

QWK scores typically range from [0,1]. Scores less than 0 indicate "less than chance" agreement between raters. 0 indicates chance agreement, and 1 indicates perfect agreement [33].

Finally, to inform our evaluation of the model output, we first took memos [15] on the differences in model output and human scores that inform prompt engineering needs as well as curriculum and rubric refinements. Utilizing these memos, the raters provide key themes identified from our analytical approach and provide vignettes to demonstrate theme examples.

## 5 RESULTS

We analyze the performance of our methodology for the three questions that we presented in Figure 3. For each question, we (1) evaluate the performance of the co-created model and (2) provide vignettes of the educator-in-the-loop process that our educator+researcher team applied to improve the model.

### 5.1 Evaluating Performance for FA1-Q1

FA1-Q1 asked students to state what the *different-sized arrows* in the model meant. The question was scored for one point (science concepts), and students received this point if they correctly identified that the size of the arrows in the diagram corresponded to the quantity of water flow. Table 1 presents our method's performance for the three baselines.

| Q1 Arrow Size | n | Acc | F1 | QWK |
|---|---|---|---|---|
| Zero-Shot | 0 | 0.87 | 0.84 | 0.68 |
| Few-Shot | 4 | 1.00 | **1.00** | **1.00** |
| Few-Shot, CoT | 4 | 0.96 | 0.95 | 0.89 |
| CoT + AL | 12 | 0.98 | 0.97 | 0.95 |

Table 1. Chain-of-Thought Prompting + Active Learning performance results for the Q1: *Arrow Size* subscore in comparison to the three incremental baselines. $n$ is the number of few-shot instances included in the prompt. For all questions, each subscore and metric is in bold for the best-performing implementation.

Even in a zero-shot setting, GPT-4 aligned with the human scorer to a *moderate* degree, i.e., QWK $\geq$ 0.6. Adding the few-shot instances enabled the model to achieve perfect alignment with the human on the test set. When CoT was subsequently used, performance dropped slightly for both Macro-F1 and QWK. Presumably, the model was already

perfectly aligned with the human after the few-shot instances were added to the prompt, and the CoT reasoning chains caused the model to overfit because it created misalignments between the human scorer and the LLM. However, the goal of our approach was not just accurate scoring but proper explanations for the assigned scores. We discuss this issue in greater detail below.

When active learning was implemented, with the addition of other few-shot instances that used CoT reasoning to correct the model's reasoning errors from the validation set, the model realigned itself with the human scorer to a large degree. In this case, the model predicted only a handful of instances incorrectly in the validation set, and we reinserted all of them back into the prompt with corrective CoT reasoning chains within the constraints of the GPT 4 API calls and token limitations.

For Q1, the educator and researcher achieved consensus after two rounds of IRR. Q1 initially appeared to be the most straightforward of the three questions. The score was based on one science concept (*Arrow Size*) and no scores were allocated for scientific reasoning. The IRR process highlighted a major sticking point, where the researchers disagreed in their scoring of multiple instances. In several student answers, there were only general statements about the meaning of size (e.g., *more of something* or *less of something*) without specific mention of arrow size or amount of water. In these instances, one reviewer initially scored these answers as correct, while the other researcher scored them as incorrect. After discussion, the educator-researcher team agreed that the primary purpose of this question was to recognize that *arrow size indicates the amount*. Therefore, the *more of something* responses also deserved the point. To help align the LLM with this consensus, one of these *more of something* instances was selected as a few-shot instance in the initial prompt, and our rationale for awarding the point was explained via CoT reasoning.

Typically, when training neural networks, an *early stopping* criterion is often employed to halt the training process once the model converges (i.e., validation loss stops decreasing) [27] precisely to avoid overfitting the data. In our case, this would mean stopping after the *Few-Shot* implementation, and using this model for subsequent grading as it was the best-performing. However, this is not as useful to educators and their students, as numerical scores alone do not provide explanations for the model's scoring decisions or drive rubric and formative assessment question refinement during training. For Q1, eliciting CoT reasoning from the model came at the cost of scoring, but active learning helped bridge this performance gap. This is an important finding, as it demonstrates our method's ability to both provide the educator with useful insight into the model's scoring and minimize the performance cost of eliciting scoring explanations from the LLM.

### 5.2 Evaluating Performance for FA2-Q2

Q2 asked students to *select two things that the diagram did a good job of explaining*. The scoring rubric assigned four possible points: two for answering science concepts correctly, and two for providing the correct scientific reasoning. Science concept subscores include: (1) *Arrow Direction*. Students were awarded a point if they identified that the diagram correctly demonstrated that the water originated from the sky in the form of rain, that the water was absorbed into the ground, or that the water became runoff; and (2) *Arrow Size*. Students were awarded a point if they identified that the diagram used the arrow size correctly to represent the amount of water. Associated with each science concept was an additional point for correct scientific reasoning by the student. Model performance comparisons for Q2 are presented in Table 2.

Scoring of Q2 science concepts subscores (*Arrow Direction* and *Arrow Size*) achieved their best performance (or tied for it) only when the full Chain-of-Thought + Active Learning method was used. The scientific reasoning subscores, however, saw performance decrease as additional components of the method were added. While the total score achieved

| **Q2 Arrow Direction** | n | Acc | F1 | QWK |
|---|---|---|---|---|
| Zero-Shot | 0 | 0.91 | 0.89 | 0.78 |
| Few-Shot | 5 | 0.87 | 0.79 | 0.60 |
| Few-Shot, CoT | 5 | 0.98 | **0.98** | **0.95** |
| CoT + AL | 10 | 0.98 | **0.98** | **0.95** |
| **Q2 Arr. Dir., Reasoning** | n | Acc | F1 | QWK |
| Zero-Shot | 0 | 0.92 | **0.73** | **0.47** |
| Few-Shot | 5 | 0.89 | 0.67 | 0.36 |
| Few-Shot, CoT | 5 | 0.91 | 0.70 | 0.41 |
| CoT + AL | 10 | 0.92 | 0.65 | 0.3 |
| **Q2 Arrow Size** | n | Acc | F1 | QWK |
| Zero-Shot | 0 | 0.77 | 0.69 | 0.39 |
| Few-Shot | 5 | 0.77 | 0.69 | 0.39 |
| Few-Shot, CoT | 5 | 0.91 | 0.88 | 0.77 |
| CoT + AL | 10 | 0.94 | **0.92** | **0.83** |
| **Q2 Arr. Sz., Reasoning** | n | Acc | F1 | QWK |
| Zero-Shot | 0 | 0.96 | 0.82 | 0.65 |
| Few-Shot | 5 | 0.98 | **0.90** | **0.79** |
| Few-Shot, CoT | 5 | 0.94 | 0.77 | 0.55 |
| CoT + AL | 10 | 0.96 | 0.82 | 0.65 |
| **Q2 Total Score** | n | Acc | F1 | QWK |
| Zero-Shot | 0 | 0.60 | 0.59 | 0.65 |
| Few-Shot | 5 | 0.53 | 0.52 | 0.55 |
| Few-Shot, CoT | 5 | 0.75 | **0.80** | 0.80 |
| CoT + AL | 10 | 0.85 | 0.79 | **0.87** |

Table 2. Performance comparisons for Question 2.

its highest QWK with the complete method, this assessment question highlighted a considerable misalignment between the LLM and human scorer for scientific reasoning subscores and provided an opportunity for the educator to refine both the rubric and the formative assessment question to help students provide better answers for scientific reasoning.

Q2 required three rounds of IRR and was the most difficult for the human scorers to achieve consensus. This was because many student answers were vague or ambiguous. Several students listed both correct and incorrect attributes of the model in their responses, which was a major sticking point during IRR. Much of the difficulty in human scoring could be attributed to the open-ended nature of the question. The question asked students to list two things the model did a good job of demonstrating. However, there are several things the model does well. Many of these the researchers did not realize until they saw the student responses on paper.

Consider the following student response: "*A good job Taylor explained was which cloud rained the most, they made the cloud dark.*" While the student received a point (for *Arrow Direction*) because he or she mentioned rainfall, the student did not receive the additional point for reasoning. The mention that the diagram does a good job of showing that the cloud with more rain appears darker may demonstrate correct scientific understanding and that the model may illustrate this scientific concept. However, there was no specific subscore in the rubric to capture this point, and the student was credited with only one of the four possible points despite having correctly identified an additional attribute of the model that was scientifically correct (more dark clouds means more rain). In total, six different students mentioned the darker rain cloud as a good example of what the diagram did a good job of showing; however, none of them received a

point for doing so because it was not linked to a relevant subscore in the rubric. This example highlights how both the IRR and active learning (i.e., human-in-the-loop) processes can help educators improve rubrics and formulate formative assessment questions that are more precise or to expand the scope of the rubric to accommodate answers that are scientifically relevant given the wording of the question.

For this question, students were simply instructed to *explain your answer*. One student remarked, *I think that the big arrow that is pointing at the ground is the water that is being absorb* (sic) *into the grass because it shows that the water is going down to the grass*. While the student received a point for *Arrow Direction* and provided an explanation, he or she did not demonstrate scientific reasoning with respect to the scientific process itself and, therefore, did not earn a point for *Arrow Direction Reasoning*. However, the LLM disagreed and awarded the scientific reasoning point to the student with the following justification:

> ...the student says "because it shows that the water is going down to the grass". This demonstrates that the student used [scientific] reasoning to justify his or her response with regard to [Arrow Direction]. Based on the rubric, the student earned a score of 1.[2]

It was common for the LLM to relate words like *because* with scientific reasoning (when this was often not the case). In the future, the educator should revise the rubric and reword the formative assessment question to better define what constitutes scientific reasoning for both the student and the LLM, and continue to add labeled instances in the prompt with CoT reasoning to address additional model misconceptions.

The model often had difficulty scoring student responses for many of the same reasons the educator and researcher found it difficult to achieve consensus during IRR – even in cases where CoT reasoning was used in the prompt to address these same issues. One student's response was, "*The amount of absorption and runoff.*" One could argue that the student understands arrow size represents water amount; however, the absorption arrow in the model is actually incorrect, as it is bigger than the rainfall arrow, so it violates the law of conservation of matter. Because Q2 asks for "good" (i.e., correct) examples in the diagram, and the absorption arrow is incorrect, both reviewers agreed that this and similar responses should not receive a point for *Arrow Size* even though the student may have understood that the arrow size represented water amount.

In addition, the student mentioned *the amount of runoff*, which is actually correct (i.e., the runoff arrow is smaller than the rainfall arrow, so it does not violate the law of conservation of matter). In this case, the student listed one correct example and one incorrect example, and the student also demonstrated an understanding that arrow size corresponds to water amount. All of this was linked to the same subscore, *Arrow Size*. Issues like these illustrate why it is difficult for an LLM to correctly predict and explain scores for formative assessment questions, rubrics, and student responses that are not sufficiently precise or otherwise ambiguous. During active learning, the model erroneously awarded points to several of these types of responses. We tried using CoT reasoning to address these issues, but the model began to mislabel other instances it had previously scored correctly because of overfitting. The question of achieving accuracy in scoring with proper explanation versus overfitting is probably best resolved by making the wording of the question precise, as we discuss below.

Q2 is a good example of how this educator-researcher partnership and the LLM inform the educator to revise the wording of the question and the rubric for formative assessment questions to provide clearer guidance to students. For Q2, this is exactly what we decided to do going forward, and future work with Q2 will involve rewording the question

[2]Items in brackets refer to terms that differed slightly in the actual prompt and were later renamed for readability in the manuscript. The original, raw prompts can be found in the Supplemental Materials in the *Initial Prompts* section.

to better explain to students the nature of the scientific reasoning they are being asked to provide. In addition, the educator can also reformulate the rubric to better describe the expected scientific reasoning answers for the question.

### 5.3 Evaluating Performance for FA2-Q3

Q3 asked students to *identify two things they would change to correct the diagram.* Similar to Q2, Q3 was assigned a total of 4 possible points: 2 for science concepts and 2 for science reasoning. For science concepts, students were awarded a point for *Runoff Direction* if they mentioned the runoff arrow in the diagram was pointing in the wrong direction and needed to be changed (i.e., the runoff arrow had to be changed from pointing uphill to pointing downhill). Students were awarded a point for *Arrow Size* if they indicated that the arrow sizes in the diagram needed to be changed (i.e., the absorption arrow was larger than the rainfall arrow, and this violated the law of conservation of matter). Each science concept subscore also had an additional point if students provided a correct scientific reason for their response. Performance comparisons for Q3 are presented in Table 3.

| **Q3 Runoff Direction** | n | Acc | F1 | QWK |
|---|---|---|---|---|
| Zero-Shot | 0 | 0.89 | 0.88 | 0.77 |
| Few-Shot | 5 | 0.91 | 0.90 | 0.80 |
| Few-Shot, CoT | 5 | 0.92 | **0.92** | **0.84** |
| CoT + AL | 9 | 0.89 | 0.88 | 0.75 |
| **Q3 Run. Dir., Reasoning** | n | Acc | F1 | QWK |
| Zero-Shot | 0 | 0.94 | 0.89 | 0.79 |
| Few-Shot | 5 | 0.94 | 0.91 | 0.82 |
| Few-Shot, CoT | 5 | 0.94 | 0.92 | 0.83 |
| CoT + AL | 9 | 0.98 | **0.97** | **0.94** |
| **Q3 Arrow Size** | n | Acc | F1 | QWK |
| Zero-Shot | 0 | 0.87 | 0.83 | 0.67 |
| Few-Shot | 5 | 0.89 | 0.87 | 0.73 |
| Few-Shot, CoT | 5 | 0.85 | 0.83 | 0.65 |
| CoT + AL | 9 | 0.92 | **0.92** | **0.83** |
| **Q3 Arr. Sz., Reasoning** | n | Acc | F1 | QWK |
| Zero-Shot | 0 | 0.98 | 0.90 | 0.79 |
| Few-Shot | 5 | 1.00 | **1.00** | **1.00** |
| Few-Shot, CoT | 5 | 0.94 | 0.82 | 0.64 |
| CoT + AL | 9 | 1.00 | **1.00** | **1.00** |
| **Q3 Total Score** | n | Acc | F1 | QWK |
| Zero-Shot | 0 | 0.74 | **0.80** | 0.85 |
| Few-Shot | 5 | 0.75 | 0.73 | 0.87 |
| Few-Shot, CoT | 5 | 0.75 | 0.71 | 0.79 |
| CoT + AL | 9 | 0.81 | **0.80** | **0.90** |

Table 3. Performance comparisons for Question 3.

With the exception of the Macro-F1 for total score, all Q3 subscores (science concepts and scientific reasoning) improved for Macro-F1 and QWK metrics once the few-shot examples were added to the prompt. When CoT reasoning was added to those instances, performance increased for both *Runoff Direction* subscores but decreased considerably for both *Arrow Size* subscores. *Arrow Size Reasoning*, in particular, saw a sizeable drop in performance. This was similar to what happened in Q1 with the *Arrow Size* subscore, where the addition of the CoT reasoning chains caused the

model to overfit and become misaligned with the human scorer. Once active learning was introduced, however, all Q3 subscores (and total score) except *Runoff Direction* achieved their best performance (or tied for it) for both Macro-F1 and QWK. *Runoff Direction* performed best after the CoT reasoning chains were added but overfit when active learning was introduced. Unlike Q2, where the best-performing subscores were the science concepts subscores, the best-performing subscores for Q3 were the science reasoning subscores – both of which outperformed their science concepts counterparts for the Macro-F1 and QWK metrics.

For Q3, the human scorers reached a scoring consensus after only one round of IRR, and there were very few sticking points that caused scoring disagreements between researchers. Similarly, the model's own misconceptions for the scientific reasoning subscores were easily (relative to Q2) corrected using CoT reasoning chains in the prompt. One common issue the model encountered with Q3 had nothing to do with the formative assessment content. The LLM had a tendency to cite the same piece of evidence to justify awarding points across different subscores (i.e., the model would use the same snippet of a student's response as evidence for awarding both science concepts and science reasoning points). This led to overscoring. The reviewers had agreed beforehand that, to receive a point for any one subscore, the student must provide *new* information in his or her response that was not used to earn a point for any other subscore. We did not, however, think to include this in the initial prompt or the few-shot CoT reasoning chains.

This issue was largely resolved by active learning, and performance increased for both metrics for 3 out of the 4 subscores, and the total score (*Runoff Direction* was the only exception). This example highlights another instance where the human-in-the-loop, working in tandem with the LLM during active learning, was able to align the model to the human scorers' consensus and improve model performance.

The considerable performance gap between Q2 and Q3 illustrates how effective the partnership between humans and the LLM can be as they work together to refine prompts to maximize scorer-LLM alignment. The more the model and the human are aligned, the more informative the LLM's scoring explanations are. The following is an example of a student's formative assessment response and corresponding model generation from the test set after the active learning process was implemented:

> **Student**: *The first thing I would change is, the direction the arrow is for runoff. I would change this because the run off goes into the stream. The second thing I would change is how much absorption there is*
>
> **LLM**: [Runoff Direction]: *The student says "the direction the arrow is for runoff". This indicates that the student understands the runoff arrow should be pointing in the opposite direction (i.e., downhill). Based on the rubric, the student earned a score of 1.*
>
> [Runoff Direction Reasoning]: *The student received a point for [Runoff Direction], and the student justified his or her response by saying "because the run off goes into the stream". This demonstrates [scientific] reasoning with regard to the [Runoff Direction]. Based on the rubric, the student earned a score of 1.*
>
> [Arrow Size]: *The student says "how much absorption there is". This demonstrates that the student understands the arrow sizes need to be changed. Based on the rubric, the student earned a score of 1.*
>
> [Arrow Size Reasoning]: *The student received a point for [Arrow Size], but the student did not provide an explanation that demonstrated [scientific] reasoning with respect to the [Arrow Size]. Based on the rubric,*

*the student earned a score of 0.*

Total Score: *The student received 1 point for [Runoff Direction], 1 point for [Runoff Direction Reasoning], and 1 point for [Arrow Size], so the student's total score is 3 points.*

In this example, the model generated the correct subscore for each component of the answer (and thus the total score, as well) in accordance with the human scorer. In addition, the model was able to identify the exact component of the student's response (i.e., evidence) to be used in its decision-making about student reasoning. This includes describing positive student work examples that warranted points and, as the educator described, serves as positive feedback to support future applications of student reasoning. This is the same evidence that educators would likely use when providing their scoring decisions. Because the researcher was able to use the active learning process (by adding CoT reasoning chains in the prompt) to address the LLM's tendency to cite the same piece of evidence as justification for awarding points across multiple subscores, the model correctly identified each piece of relevant evidence for each subscore using completely distinct pieces of the student's response. In its generation, the model demonstrated that the student saying "*the direction the arrow is for runoff*" constitutes a *Runoff Direction* point, and the student's justification "*because the run off goes into the stream*" demonstrated scientific reasoning (i.e., water should flow down the hill and not uphill) and, therefore, was awarded a point for *Arrow Direction Reasoning*.

In this case, an additional, positive observation is that the model understood that "run off" and "runoff" represented the same concept. This is not always the case, as we discussed in previous work [1]. Overall, this example shows that LLMs are capable of providing informative, insightful formative assessment feedback when working alongside an educator to iteratively correct model misalignments.

### 5.4 Evaluation Summary

For all questions, the model was generally aligned with the human scorers. Out of 11 subscores and total scores, 9 of them had "strong" agreement (QWK >= 0.8) or better between the human and the LLM at some point in the process (i.e., during one of the three baselines or the full Chain-of-Thought + Active Learning approach). Four subscores even achieved "almost perfect" (QWK > 0.9) agreement between the human and the LLM. With the exception of Q2 *Arrow Direction Reasoning*, all subscores saw a Macro-F1 of 0.90 or greater at some point in the process.

We also demonstrated that the LLM can overfit when the CoT reasoning and active learning components are added to the pipeline. This is particularly true for the less complex science concept subscores (Q1 *Arrow Size* and Q3 *Runoff Direction*) and the more ambiguous scientific reasoning subscores (Q2 *Arrow Direction Reasoning* and Q2 *Arrow Size Reasoning*).

Often, the model was confronted by the same issues the humans encountered during the IRR process, and we noted that the degree of agreement between the human raters during IRR also provided some expectations of the model's scoring abilities. We found assessment questions that were easier for the human scorers to achieve consensus on were similarly easier for the model to align with the humans. The opposite was also true: when the human scorers had difficulty reaching a scoring consensus (requiring several rounds of IRR, e.g., Q2), the model also found it difficult to accurately score the student responses and align with the human consensus. However, more research needs to be done to evaluate this issue in more systematic ways.

We provide two additional examples from our analysis to highlight the LLM's abilities and limitations, both of which were part of the same model generation (i.e., the model's scoring of a single student's formative assessment question

response) for Q3. For Q3 *Runoff Direction*, the model remarks: '*The student says "Show a bit of runoff going towards the stream". This indicates that the student understands the runoff arrow should be pointing in the opposite direction (i.e., downhill). Based on the rubric, the student earned a score of 1.*' This was an example of a false positive instance, as the model awarded a point for *Runoff Direction* when it should not have. One of the raters' Q3 sticking points during *Response Scoring* was whether to award students a point for *Runoff Direction* if they mentioned adding another arrow to the diagram to indicate where the runoff should be flowing but failed to correct (i.e., remove) the original arrow in the diagram that was pointing the wrong direction. The raters ultimately agreed that, since the model would still be incorrect if a new (correct) arrow was added but the old (incorrect) arrow remained, the student would not receive a point for *Runoff Direction*. However, even though this issue was addressed in the initial prompt via CoT reasoning, the model still erroneously awarded the point. This demonstrates that, even when CoT reasoning chains are added to address specific scoring issues, the LLM still may not be able to effectively use them to output the correct scores. Addressing these issues serves as future directions of our work.

An example of a correct model score and explanation for Q3 *Arrow Size Reasoning* from the same generation as the above example is shown below:

> The student received a point for [Arrow Size], and the student explains his or her response by saying "the absorption is not usually bigger than the amount of rainfall". This illustrates that the student understands conservation of matter, in that the size of the runoff arrow and the size of the absorption arrow must sum to the size of the rainfall arrow. This demonstrates [scientific] reasoning with regard to Size of Arrows. Based on the rubric, the student earned a score of 1.

In this example, the LLM realizes the student was (correctly) awarded a point for the *Arrow Size* subscore, so the student is eligible to receive a point for *Arrow Size Reasoning*. The LLM also cites the student, highlighting the portion of the student's response that should be used when deciding to award the student the *Arrow Size Reasoning* point. The LLM expounds upon this, explaining that the student's saying the absorption arrow should not be bigger than the rainfall arrow illustrates his or her understanding of the law of conservation of matter. Finally, the LLM is able to tie this reasoning back to the rubric and correctly award the student a point for *Arrow Size Reasoning*.

## 6 DISCUSSION

Our results indicate two key findings: (1) co-created models in which stakeholders, i.e., researchers and educators, interact to systematically inform the model can improve the performance of formative feedback generation; and (2) educator insight can be integrated through a well-designed interface that supports the kinds of interactions we have reported at multiple steps in the process to create the educator-LLM partnership for grading science formative assessments.

### 6.1 Performance Considerations

The results of our approach demonstrate the effectiveness of our methods in improving model performance and provide clear future directions for improving issues such as overfitting. We demonstrated the advantages of chain-of-thought reasoning from an educator's perspective. This kind of interaction is more natural for educators than a prompt generation and prompt refinement approach that can be difficult and tedious.

Chain-of-thought reasoning also allowed directly leveraging the educator to better inform the model's learning needs (what information it needed to know) as well as how to define the output (what information it needed to generate useful

explanations). In addition, the aligning of formative feedback generation with the learning objectives and curricular design of the educator allowed us to systematically design prompts, inform prompt refinement needs, and improve the partnership of the educator, researcher, and model.

We believe this approach can be generalized and applied to short-answer response grading in other science topics and domains, as it is dependent on the rubrics generated by educators followed by systematic additional inputs to subsequent refinement steps in tuning the LLM performance. As we continue to leverage advances of LLMs to ease the burden on our educators, we believe this approach provides an opportunity to not only ease grading constraints but also to inform assessment improvements as a team.

### 6.2 Educator Considerations

Our approach provides systematic methods for integrating educator insight at each step of the LLM response refinement process. Based on our experience, we identified three key themes for integrating educator insight:

- **Explain LLM Process Leveraging Educator Background**: we identified that steps in the model process mitigated AI literacy concerns by comparing them to educator background experience. For instance, reaching an agreement on rubric items to ensure that the rubric contains enough insight for a model to leverage the rubric in the task aligns with educators' experience in interactively refining rubrics to support their students equitably.
- **Interactions Should Align with Current Pedagogical Processes**: In the process of improving model output, interactions between educators and researchers and AI models can help capture and include educators' existing teaching practices into the grading of formative assessments. This is particularly apparent through CoT and active learning. For example, an educator may provide personalized feedback on students' misconceptions to help their students overcome their difficulties. To do so, the teacher must understand the knowledge state, compare student material to other examples of how misunderstandings manifested, and refine their feedback accordingly. In the future, we will design formative feedback interfaces for students using teacher input to help students overcome their difficulties. This will also help us to systematically improve the active learning process.
- **Leverage Existing, Standard-Aligned Rubrics and Learning Objectives to Support Model Training**: A key advantage in our approach was the use of evidence-centered design for rubric and curriculum creation as it supported a more systematic prompt engineering process for initial preparation of the model. Having an interface for taking educator-developed rubrics as input to the model directly connected the learning objectives to a resource the model would use for its feedback generation task.

With these considerations in mind, there is an opportunity for future research to dive more deeply into how teachers interact with feedback technologies (a current limitation in user experience and learning analytics research [6]) and better align interactions with their lived classroom experiences.

### 6.3 Limitations

Using any LLM carries with it innate risks such as ethical concerns relating to privacy and bias, as well as hallucinations [49]. GPT-4, in particular, raises additional concerns due to its opacity and private ownership. In fact, even its underlying architecture remains undisclosed. Because it is not openly available to the public (i.e., we cannot create our own local implementations), we do not know how OpenAI uses the data it accumulates through its LLMs (e.g., to train and

improve its models). This means GPT-4 is not an option for researchers handling sensitive data or who are otherwise obligated to maintain complete control over their data pipeline.

In addition, not all researchers and educators have access to GPT-4 (or its API) due to the model's exclusivity, API call and token limits, and cost[3]. For these reasons (and others), open-source LLMs are preferred; however, there is still a considerable gap between open and closed models performance-wise [36], and model size plays a large role in customizing LLMs [18]. Our initial attempts at applying our method using a smaller, open source LLM (Falcon-7B-Instruct[4]) did not yield LLM responses with sufficient accuracy to conduct further informative analyses that we have presented in this paper.

While CoT reasoning chains have been shown to improve model performance over traditional ICL, the degree to which this reasoning actually guides the model's decision-making processes (if at all) is still an open problem [37]. Trying to align LLM responses with human intentions often comes at the expense of decreased model performance and can have other unintended consequences [43]. This is known as the *alignment problem*[5], and it remains an active area of LLM research. Another drawback to our Chain-of-Thought Prompting + Active Learning approach (and ICL, in general) is that prompts can become long. During inference, each instance must be accompanied by the task description, rubric, few-shot examples, and CoT reasoning chains, which can drive up API costs and create context-length issues [3]. Increasing context-length in LLMs is another currently active field of research.

Last, our results indicate both CoT reasoning and active learning can cause overfitting, particularly when applied to simpler and easier-to-define subproblems, as well as those subproblems whose rubrics are more ambiguous. In the case of the former, LLM-based methods may be overkill. This was demonstrated by researchers who recently found rule-based approaches outperformed GPT-4 for detecting common item-writing flaws in student-generated multiple-choice questions [23].

## 7 CONCLUSIONS AND FUTURE DIRECTIONS

Our findings offer insight into the potential of educator-AI partnerships for actionable, learning-aligned feedback to support student learning and teacher understanding of and engagement in their students' developing science ideas. We offer precedent knowledge on how these human-in-the-loop interactions may be applied to support similar tasks, in particular for learning domains where responses are less structured than in math problem-solving or evaluating programming assignments.

As we move forward in developing actionable LLM tools for teachers, we aim to eliminate the need for the researcher in the loop through additional human-computer interaction analysis on how to generate a teacher user-interface that enhances the educator-AI partnership, and explores additional question types, such as the evaluation of causal reasoning in science. Features of this tool can support summary feedback to teachers on individual, group, and class work as well as individual student feedback creation that is designed by the teacher. We will also explore how to utilize smaller, customized LLMs for effectively grading formative assessments with less computational and cost expenses. While the potential for better supporting our teachers has become possible because of the advances in LLMs and other generative AI, it is critical that we consider the educator in the loop and provide interaction mechanisms that give educators agency in their interactions and how they engage with their students.

[3]For reference, we spent $91.62 using the OpenAI API for testing, refining, and evaluating our method.
[4]https://huggingface.co/tiiuae/falcon-7b-instruct
[5]https://openai.com/blog/our-approach-to-alignment-research

## ACKNOWLEDGEMENTS

This work is supported by National Science Foundation Awards XXX-XXXXXXX and XXX-XXXXXXX.

## REFERENCES

[1] Anonymous. 1234. Anonymous. (1234). Anonymous.

[2] Anonymous. 1234. Anonymous. In *Anonymous*. Anoymous, Anonymous, Anonymous, 00–00.

[3] Anonymous. 1234. Anonymous. (1234). Anonymous.

[4] Benjamin Bloom, George Madaus, and J. Hastings. 1971. *Handbook on Formative and Summative Evaluation of Student Learning*. McGraw-Hill, New York.

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv e-prints* n/a, n/a, Article arXiv:2005.14165 (May 2020), 75 pages. https://doi.org/10.48550/arXiv.2005.14165 arXiv:2005.14165 [cs.CL]

[6] Fabio Campos, June Ahn, Daniela K. DiGiacomo, Ha Nguyen, and Maria Hays. 2021. Making Sense of Sensemaking: Understanding How K–12 Teachers and Coaches React to Visual Analytics. *Journal of Learning Analytics* 8, 3 (2021), 60–80. https://doi.org/10.18608/jla.2021.7113

[7] Keith Cochran, Clayton Cohn, and Peter M. Hastings. 2023. Improving NLP Model Performance on Small Educational Data Sets Using Self-Augmentation. In *Proceedings of the 15th International Conference on Computer Supported Education, CSEDU 2023, Volume 1, Prague - Czech Republic, April 21-23, 2023*, Jelena Jovanovic, Irene-Angelica Chounta, James Uhomoibhi, and Bruce M. McLaren (Eds.). SCITEPRESS, Prague, Czech Republic, 70–78. https://doi.org/10.5220/0011857200003470

[8] Keith Cochran, Clayton Cohn, Nicole Hutchins, Gautam Biswas, and Peter Hastings. 2022. Improving Automated Evaluation of Formative Assessments with Text Data Augmentation. In *Artificial Intelligence in Education*, Maria Mercedes Rodrigo, Noburu Matsuda, Alexandra I. Cristea, and Vania Dimitrova (Eds.). Springer International Publishing, Cham, 390–401.

[9] Keith Cochran, Clayton Cohn, Jean Francois Rouet, and Peter Hastings. 2023. Improving Automated Evaluation of Student Text Responses Using GPT-3.5 for Text Data Augmentation. In *Artificial Intelligence in Education: 24th International Conference, AIED 2023, Tokyo, Japan, July 3–7, 2023, Proceedings* (Tokyo, Japan). Springer-Verlag, Berlin, Heidelberg, 217–228. https://doi.org/10.1007/978-3-031-36272-9_18

[10] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[11] Jacob Cohen. 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin* 70, 4 (1968), 213.

[12] Hiroaki Funayama, Yuya Asazuma, Yuichiroh Matsubayashi, Tomoya Mizumoto, and Kentaro Inui. 2023. Reducing the Cost: Cross-Prompt Pre-finetuning for Short Answer Scoring. In *Artificial Intelligence in Education*, Ning Wang, Genaro Rebolledo-Mendez, Noboru Matsuda, Olga C. Santos, and Vania Dimitrova (Eds.). Springer Nature Switzerland, Cham, 78–89.

[13] Shuchi Grover. 2017. Assessing algorithmic and computational thinking in K-12: Lessons from a middle school classroom. *Emerging research, practice, and policy on computational thinking* n/a, n/a (2017), 269–288.

[14] Kevin C Haudek, Jennifer J Kaplan, Jennifer Knight, Tammy Long, John Merrill, Alan Munn, Ross Nehm, Michelle Smith, and Mark Urban-Lurain. 2011. Harnessing technology to improve formative assessment of student conceptions in STEM: forging a national network. *CBE—Life Sciences Education* 10, 2 (2011), 149–155.

[15] Christopher Hoadley. 2002. Creating context: Design-based research in creating and understanding CSCL. In *Proceedings of the Conference on Computer Support for Collaborative Learning: Foundations for a CSCL Community*. Taylor & Francis Group, n/a, 453–462. https://doi.org/10.3115/1658616.1658679

[16] Nicole M. Hutchins and Gautam Biswas. 2023. Co-designing teacher support technology for problem-based learning in middle school science. *British Journal of Educational Technology* n/a, n/a (2023). https://doi.org/10.1111/bjet.13363 arXiv:https://bera-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.13363

[17] Jaeho Jeon and Seongyong Lee. 2023. Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies* n/a, n/a (2023), 1–20.

[18] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv e-prints* n/a, n/a, Article arXiv:2001.08361 (Jan. 2020), 30 pages. https://doi.org/10.48550/arXiv.2001.08361 arXiv:2001.08361 [cs.LG]

[19] Evan Liu, Moritz Stephan, Allen Nie, Chris Piech, Emma Brunskill, and Chelsea Finn. 2022. Giving Feedback on Interactive Student Programs with Meta-Exploration. *Advances in Neural Information Processing Systems* 35 (2022), 36282–36294.

[20] Stephen MacNeil, Andrew Tran, Arto Hellas, Joanne Kim, Sami Sarsa, Paul Denny, Seth Bernstein, and Juho Leinonen. 2023. Experiences from Using Code Explanations Generated by Large Language Models in a Web Software Development E-Book. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1* (Toronto ON, Canada) *(SIGCSE 2023)*. Association for Computing Machinery, New York, NY, USA, 931–937. https://doi.org/10.1145/3545945.3569785

[21] Roberto Martinez-Maldonado, Abelardo Pardo, Negin Mirriahi, Kalina Yacef, Judy Kay, and Andrew Clayphan. 2016. LATUX: an Iterative Workflow for Designing, Validating and Deploying Learning Analytics Visualisations. *Journal of Learning Analytics* 2, 3 (Feb. 2016), 9–39. https://doi.org/10.18608/jla.2015.23.3

[22] Robert J. Mislevy and Geneva D. Haertel. 2006. Implications of Evidence-Centered Design for Educational Testing. *Educational Measurement: Issues and Practice* 25, 4 (2006), 6–20. https://doi.org/10.1111/j.1745-3992.2006.00075.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1745-3992.2006.00075.x

[23] Steven Moore, Huy A. Nguyen, Tianying Chen, and John Stamper. 2023. Assessing the Quality of Multiple-Choice Questions Using GPT-4 and Rule-Based Methods. *arXiv e-prints* n/a, n/a, Article arXiv:2307.08161 (July 2023), 15 pages. https://doi.org/10.48550/arXiv.2307.08161 arXiv:2307.08161 [cs.CL]

[24] NGSS. 2013. *Next Generation Science Standards: For States, By States.* The National Academies Press, Washington, DC, USA.

[25] OpenAI. 2023. GPT-4 Technical Report. *arXiv e-prints* n/a, n/a, Article arXiv:2303.08774 (March 2023), 100 pages. https://doi.org/10.48550/arXiv.2303.08774 arXiv:2303.08774 [cs.CL]

[26] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[27] Lutz Prechelt. 1998. Automatic early stopping using cross validation: quantifying the criteria. *Neural Networks* 11, 4 (1998), 761–767. https://doi.org/10.1016/S0893-6080(98)00010-0

[28] Charvi Rastogi, Marco Tulio Ribeiro, Nicholas King, Harsha Nori, and Saleema Amershi. 2023. Supporting Human-AI Collaboration in Auditing LLMs with LLMs. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Montréal, QC, Canada) *(AIES '23)*. Association for Computing Machinery, New York, NY, USA, 913–926. https://doi.org/10.1145/3600211.3604712

[29] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. 2021. A survey of deep active learning. *ACM computing surveys (CSUR)* 54, 9 (2021), 1–40.

[30] Fátima Rodrigues and Paulo Oliveira. 2014. A system for formative assessment and monitoring of students' progress. *Computers & Education* 76 (2014), 30–41. https://doi.org/10.1016/j.compedu.2014.03.001

[31] Pratim Sengupta, John S Kinnebrew, Satabdi Basu, Gautam Biswas, and Douglas Clark. 2013. Integrating computational thinking with K-12 science education using agent-based computation: A theoretical framework. *Education and Information Technologies* 18, 2 (2013), 351–380.

[32] Shubhankar Singh, Anirudh Pupneja, Shivaansh Mital, Cheril Shah, Manish Bawkar, Lakshman Prasad Gupta, Ajit Kumar, Yaman Kumar, Rushali Gupta, and Rajiv Ratn Shah. 2023. H-AES: Towards Automated Essay Scoring for Hindi. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 13 (Sep. 2023), 15955–15963. https://doi.org/10.1609/aaai.v37i13.26894

[33] Shubhankar Singh, Anirudh Pupneja, Shivaansh Mital, Cheril Shah, Manish Bawkar, Lakshman Prasad Gupta, Ajit Kumar, Yaman Kumar, Rushali Gupta, and Rajiv Ratn Shah. 2023. H-AES: Towards Automated Essay Scoring for Hindi. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 13 (Sep. 2023), 15955–15963. https://doi.org/10.1609/aaai.v37i13.26894

[34] Yaman Kumar Singla, Sriram Krishna, Rajiv Ratn Shah, and Changyou Chen. 2022. Using Sampling to Estimate and Improve Performance of Automated Scoring Systems with Guarantees. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 11 (Jun. 2022), 12835–12843. https://doi.org/10.1609/aaai.v36i11.21563

[35] Wei Tan, Jionghao Lin, David Lang, Guanliang Chen, Dragan Gašević, Lan Du, and Wray Buntine. 2023. Does informativeness matter? Active learning for educational dialogue act classification. In *International Conference on Artificial Intelligence in Education*. Springer, Springer Nature Switzerland, Cham, 176–188.

[36] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv e-prints* n/a, n/a (2023), arXiv:2307.09288.

[37] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *arXiv e-prints* n/a, n/a, Article arXiv:2305.04388 (May 2023), 29 pages. https://doi.org/10.48550/arXiv.2305.04388 arXiv:2305.04388 [cs.CL]

[38] Janet Walkoe, Michelle Wilkerson, and Andrew Elby. 2017. Technology-Mediated Teacher Noticing: A Goal for Classroom Practice, Tool Design, and Professional Development. In *Proceedings of the 12th International Conference on Computer Supported Collaborative Learning (CSCL) 2017* (Philadelphia, PA, USA). International Society of the Learning Sciences, n/a, n/a.

[39] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv e-prints* n/a, n/a, Article arXiv:2201.11903 (Jan. 2022), 43 pages. https://doi.org/10.48550/arXiv.2201.11903 arXiv:2201.11903 [cs.CL]

[40] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. *arXiv e-prints* n/a, n/a, Article arXiv:2302.11382 (Feb. 2023), 19 pages. https://doi.org/10.48550/arXiv.2302.11382 arXiv:2302.11382 [cs.SE]

[41] Korah J. Wiley, Yannis Dimitriadis, Allison Bradford, and Marica C. Linn. 2020. From Theory to Action: Developing and Evaluating Learning Analytics for Learning Design. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (Frankfurt, Germany) *(LAK '20)*. Association for Computing Machinery, New York, NY, USA, 569–578. https://doi.org/10.1145/3375462.3375540

[42] Alyssa Friend Wise and David Williamson Shaffer. 2015. Why Theory Matters More than Ever in the Age of Big Data. *Journal of Learning Analytics* 2, 2 (Dec. 2015), 5–13. https://doi.org/10.18608/jla.2015.22.2

[43] Yotam Wolf, Noam Wies, Oshri Avnery, Yoav Levine, and Amnon Shashua. 2023. Fundamental Limitations of Alignment in Large Language Models. *arXiv e-prints* n/a, n/a, Article arXiv:2304.11082 (April 2023), 29 pages. https://doi.org/10.48550/arXiv.2304.11082 arXiv:2304.11082 [cs.CL]

[44] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 385, 22 pages. https://doi.org/10.1145/3491102.3517582

[45] Xuansheng Wu, Xinyu He, Tianming Liu, Ninghao Liu, and Xiaoming Zhai. 2023. Matching Exemplar as Next Sentence Prediction (MeNSP): Zero-Shot Prompt Learning for Automatic Scoring in Science Education. In *Artificial Intelligence in Education*, Ning Wang, Genaro Rebolledo-Mendez, Noboru Matsuda, Olga C. Santos, and Vania Dimitrova (Eds.). Springer Nature Switzerland, Cham, 401–413.

[46] Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2023. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology* n/a, n/a (2023), 1–23. https://doi.org/10.1111/bjet.13370 arXiv:https://bera-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.13370

[47] Zijie Zeng, Lin Li, Quanlong Guan, Dragan Gašević, and Guanliang Chen. 2023. Generalizable Automatic Short Answer Scoring via Prototypical Neural Network. In *Artificial Intelligence in Education*, Ning Wang, Genaro Rebolledo-Mendez, Noboru Matsuda, Olga C. Santos, and Vania Dimitrova (Eds.). Springer Nature Switzerland, Cham, 438–449.

[48] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv e-prints* n/a, n/a (2023), arXiv:2303.18223.

[49] Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Red teaming ChatGPT via Jailbreaking: Bias, Robustness, Reliability and Toxicity. *arXiv e-prints* n/a, n/a, Article arXiv:2301.12867 (Jan. 2023), 17 pages. https://doi.org/10.48550/arXiv.2301.12867 arXiv:2301.12867 [cs.CL]

Received 09 October 2023; revised n/a; accepted n/a