A Chain-of-Thought Prompting Approach with LLMs for Evaluating Students' Formative Assessment Responses in Science

Method Application Details

This section includes our step-by-step method application details for each of the three formative assessment questions, broken down by the three components in the method's pipeline: Response Scoring, Prompt Development, and Active Learning.

Q1 Method Application Details

Response Scoring

1. IRR Round 1
    a. Each of the raters (researcher and educator) manually scored a randomly sampled 20% of the Q1 formative assessment responses.
    b. Initial Cohen's *k* for the lone subscore was 0.56, which was lower than our 0.70 threshold required for consensus.
    c. During IRR Round 1, the raters identified one sticking point, which was whether or not to award a point for Q1 *Arrow Size* if the student made a vague statement relating to quantity (e.g., "more of something" or "less of something") without explicitly mentioning water.
    d. The researcher and educator discussed the disagreement and ultimately decided these types of responses *would* receive a point due to the student understanding arrow size represents quantity.
2. IRR Round 2
    a. Each rater manually scored another randomly sampled 20% of the Q1 formative assessment responses.
    b. Cohen's *k* was a perfect 1.0, which was greater than our 0.70 threshold (i.e., we achieved consensus)
    c. Based on this consensus, the educator manually scored the remaining 80% of the Q1 responses.
3. Train/Test Split. The labeled data was partitioned into training (80%) and testing (20%) sets.

Prompt Development

1. *Persona Pattern*. The prompt begins by employing the *persona pattern* (White et al., 2023) to inform the LLM of its role and task. This pattern was chosen due to its success in previous works (both anonymized for peer review).
2. Formative Assessment Question. The *persona pattern* is followed by the formative assessment question.
3. Additional Instructions. An additional instruction was provided to the LLM to ignore differences between students referencing "Libby's model" or "Taylor's model," as we felt the LLM may interpret multiple student names as multiple diagrams (when the two diagrams are, in fact, one and the same).
4. Rubric. The rubric is then inserted into the prompt (along with an explanation and response formatting information), and the model is instructed to score students' responses pursuant to that rubric.
5. Few-Shot Example Selection. Four few-shot examples were then selected for prompt inclusion. The number 4 was chosen due to it being the minimum number of instances we could include in the prompt while ensuring: (1) few-shot examples included at least one positive and one negative instance, (2) few-shot examples included both ground truth and sticking point instances, and (3) few-shot examples were balanced in the

prompt. The ordering of the few-shot instances was deliberately engineered to ensure an even label distribution within the prompt (i.e., to avoid labeled instances being "clumped" by score). Few-shot instances were chosen for the following reasons:

   a. Few-Shot Example 1. Ground truth instance in the positive class (i.e., score=1).
   b. Few-Shot Example 2. Ground truth instance in the negative class (i.e., score=0).
   c. Few-Shot Example 3. Sticking point instance in the positive class.
   d. Few-Shot Example 4. Negative instance added for data balance.

6. Chain-of-Thought Reasoning. CoT reasoning chains were added to each of the labeled examples in the prompt to alert the model as to why or why not the LLM should award the student a point for his or her Q1 response. For instances with sticking points, the CoT reasoning also addresses the raters' reasons for disagreement during IRR to help the model better align with the human consensus.
7. Few-Shot Example Inclusion. The selected few-shot examples with CoT reasoning were appended to the prompt, at which point the prompt was ready to be tested on the validation set.

Active Learning

1. Validation Set Inference. Model inference was performed on the validation set (i.e., all instances in the training set except for the four few-shot instances in the prompt) using the prompt we created during *Prompt Development*.
2. Scoring Trend Determination. We looked at the model's predictions and noted that the model made six incorrect predictions: 4 FNs, and 2 FPs.
3. Additional Few-Shot Example Selection. Due to the model only incorrectly predicting six student responses, all six were appended to the prompt. Two additional (randomly selected) instances from the negative class (score=0) were added for data balance. In total, 8 additional instances were added to the prompt during active learning.
4. Test Set Inference. The updated prompt (with the few-shot instances added during active learning) was fed through the LLM with the test set instances to conduct our analysis. For analysis, we performed: (1) quantitative analysis comparing the model's scores to the human's for both Macro-F1 and QWK (presented in this work), (2) a thematic analysis via inductive coding to identify trends in the LLM's scoring errors and patterns in its explanations, and (3) a researcher comparison between the educator's scoring and the LLM's scoring for all test set instances where the educator and LLM disagreed. 2 and 3 are outside the scope of this work but were presented in a previous work (anonymized for peer review). For reference, for Q1, the researcher sided with the educator over the LLM in each instance of a scoring disagreement.

Q2 Method Application Details

Response Scoring

1. IRR Round 1
    a. Each of the raters (researcher and educator) manually scored a randomly sampled 20% of the Q2 formative assessment responses.
    b. Initial Cohen's *k* across all Q2 subscores was 0.43, which was lower than our 0.70 threshold required for consensus.
    c. During IRR Round 1, the raters identified a sticking point having to do with the absorption arrow. In the diagram, the absorption arrow is incorrect, as it is larger than the rainfall arrow, so it violates the law of conservation of matter. However, several students mentioned the absorption arrow as a "good" example in that its larger size indicates more water, even though the arrow's size is actually incorrect. The question arose, "should students receive a point for Q2 *Arrow Size* if they indicate the diagram does a good job of using arrow size to represent the amount of water even when they use the incorrect absorption arrow as their example?"
    d. The raters discussed the disagreement and agreed that if the student only used the absorption arrow as his or her "good" example, then the student should not receive a point for Q2 *Arrow Size* because the absorption arrow in the diagram is incorrect. However, if the student mentioned other arrows (in addition to the absorption arrow) to highlight that the model did a good job of using arrow size to represent the amount of water, then the student would be awarded a point for Q2 *Arrow Size*.
2. IRR Round 2
    a. Each rater manually scored another randomly sampled 20% of the Q2 formative assessment responses.
    b. Cohen's *k* for the Q2 subscores was 0.68, which was again below the 0.70 threshold required for consensus.
    c. During IRR Round 2, the raters identified two additional sticking points. The first was a question about scoring "general" responses that lacked detail relating to the diagram (e.g., "What goes where and how much of it"), which was similar to the disagreement during IRR Round 1 for Q1. This sticking point applied to both Q2 science concepts subscores (*Arrow Direction Size* and *Arrow Size*). The other sticking point had to do with scientific reasoning (i.e., what, specifically, constitutes scientific reasoning?)
    d. The raters discussed both disagreements and came to a consensus. For the disagreement regarding how to treat "general" responses, the raters agreed that these types of responses were enough to earn a point for *Arrow Size* (e.g., "how much of it" in the example above) but not enough to earn a point for *Arrow Direction*. For *Arrow Direction*, the raters agreed that the students needed to use the scientific terminology to receive credit for that subscore, which required that students explicitly mention the science concepts (i.e., rainfall, absorption, and runoff). For the scientific reasoning disagreement, the raters agreed that points

would only be awarded for scientific reasoning subscores if students demonstrated reasoning with respect to the scientific process itself and that no scientific reasoning points would be awarded for simply using observations in the diagram (e.g., “I also like how he shaded in the clouds”) as justification for one’s response.

3. IRR Round 3
    a. Each rater (researcher and educator) manually scored another randomly sampled 20% of the Q2 formative assessment responses.
    b. Cohen’s *k* was 0.87, indicating the two raters had achieved consensus.
    c. Based on this consensus, the educator manually scored the remaining 80% of the Q2 responses.
4. Train/Test Split. The labeled data was partitioned into training (80%) and testing (20%) sets.

Prompt Development

1. *Persona Pattern*. The prompt begins by employing the *persona pattern* (White et al., 2023) to inform the LLM of its role and task. This is the same as for Q1.
2. Formative Assessment Question. The *persona pattern* is followed by the formative assessment question. This is the same as for Q1, but with a different formative assessment question (Q2 instead of Q1).
3. Additional Instructions. An additional instruction was provided to the LLM to ignore differences between students referencing “Libby’s model” or “Taylor’s model,” as we felt the LLM may interpret multiple student names as multiple diagrams (when the two diagrams are, in fact, one and the same).
4. Rubric. The rubric is then inserted into the prompt (along with an explanation and response formatting information), and the model is instructed to score students’ responses pursuant to that rubric. This is the same as for Q1, but with a different rubric specific for Q2.
5. Few-Shot Example Selection. For Q2, five few-shot examples were then selected for prompt inclusion for the same reasons as Q1, except Q2 required one additional instance to ensure: (1) at least one positive and one negative instance of each subscore was included in the prompt, and (2) data balance was maintained. Like with Q1, the ordering of the few-shot instances was deliberately engineered to ensure an even label distribution within the prompt (i.e., to avoid labeled instances being “clumped” by score). One major difference between Q1 and Q2 is that Q2 contains multiple binary subscores summing to a total score, which makes it a multilabel task. Because of this, balancing the subscores was not able to be done perfectly. The Q2 few-shot instances were chosen for the following reasons:
    a. Few-Shot Example 1. Ground truth instance containing positive *Arrow Direction* and *Arrow Direction Reasoning* subscores.
    b. Few-Shot Example 2. Ground truth instance containing positive *Arrow Direction*, *Arrow Size,* and *Arrow Size Reasoning* subscores.
    c. Few-Shot Example 3. Sticking point instance containing a positive *Arrow Size* subscore.

    d. Few-Shot Example 4. Sticking point instance containing positive *Arrow Direction* and *Arrow Size* subscores.
    e. Few-Shot Example 5. Instance with no positive subscores that was added for data balance, as positive subscores were overrepresented relative to the distribution of the dataset.
6. Chain-of-Thought Reasoning. CoT reasoning chains were added to each of the labeled examples in the prompt to alert the model as to why or why not the LLM should award the student a point for his or her Q2 response. For sticking point instances, the CoT reasoning also addresses the raters' reasons for disagreement during IRR to help the model better align with the human consensus.
7. Few-Shot Example Inclusion. The selected few-shot examples with CoT reasoning were appended to the prompt, at which point the prompt was ready to be tested on the validation set.

Active Learning

1. Validation Set Inference. Model inference was performed on the validation set (i.e., all instances in the training set except for the five few-shot instances in the prompt) using the prompt we created during *Prompt Development*.
2. Scoring Trend Determination. We looked at the model's predictions and noted that the ratio of FPs to FNs was 22:1 (i.e., the model overscored), which meant that we needed to use additional labeled instances with CoT reasoning chains to help the model better identify negative instances. Across all four subscores, FPs outweighed FNs by a considerable margin.
3. Additional Few-Shot Example Selection. In total, the model predicted 70 validation set subscores incorrectly (67 FPs, 3 FNs) across all instances, so adding every incorrectly predicted instance back into the prompt simply was not feasible (or recommended, as we encountered overfitting issues even when adding just a few labeled examples). Instead, we added 5 additional instances back into the prompt. When selecting additional instances during active learning, we prioritized instances whose reasons for being mis-scored caused several instances to be mis-scored, and those with multiple incorrectly predicted subscores in the same instance (i.e., the model and human total scores differed by a large degree). This allowed us to address the model's reasoning errors across multiple subscores in as few instances as possible.

   Four out of the five instances selected during Q2 active learning had scoring differentials of 3 between the model and the human (i.e., the model predicted a total score of 4, awarding one point for each of the four subscores, while the human awarded only one point for a single subscore for a total score of 1). During our analysis, we discovered that the model still had difficulty deciding when to award points for Q2 *Arrow Size* if the student used the absorption arrow as a "good" example. These four additional instances each addressed this issue to help the model better align with the human consensus. For all instances added to the prompt during active learning, corrective CoT reasoning chains were provided for each subscore to help guide the model toward the human consensus. Additionally, we noticed that the model had a tendency to be more generous

in its scoring with longer formative assessment question responses, so we selected some of these longer responses to be put back into the prompt during active learning. The fifth instance had a total score of 0 (i.e., the student did not receive a point from the human rater for any of the four subscores) and was added for data balance, as we again had an overrepresentation of positive subscores in the prompt relative to the dataset's label distribution.

4. Test Set Inference. The updated prompt (with the few-shot instances added during active learning) was fed through the LLM with the test set instances to conduct our analysis. Like Q1, our analysis included: (1) quantitative analysis comparing the model's scores to the human's for both Macro-F1 and QWK (presented in this work), (2) a thematic analysis via inductive coding to identify trends in the LLM's scoring errors and patterns in its explanations, and (3) a researcher comparison between the educator's scoring and the LLM's scoring for all test set instances where the educator and LLM disagreed. 2 and 3 are outside the scope of this work but were presented in a previous work (anonymized for peer review). For reference, for Q2, the researcher sided with the LLM over the human in 2/10 instances of scoring disagreement in the test set (both of which were scientific reasoning subscores).

Q3 Method Application Details

Response Scoring

1. IRR Round 1
    a. Each of the raters (researcher and educator) manually scored a randomly sampled 20% of the Q3 formative assessment responses.
    b. Initial Cohen's *k* across all Q3 subscores was 0.88, which was higher than our 0.70 threshold, so consensus was achieved.
    c. During IRR, the raters identified one sticking point having to do with the runoff arrow for the *Q3 Runoff Direction* subscore. Several students mentioned adding an additional runoff arrow pointing downhill, as the current runoff arrow was incorrect (i.e., pointing uphill). While this indicated the students understood the runoff direction needed to change, adding an extra runoff arrow pointing in the correct direction would not fix the diagram, as there would still be a runoff arrow in the diagram (i.e., the original runoff arrow) pointing in the incorrect direction.
    d. The raters discussed the disagreement and agreed that, if a student mentioned adding another runoff arrow pointing downhill, the student would only be awarded the point for Q3 *Runoff Direction* if he or she simultaneously mentioned that the old (incorrect) runoff arrow needed to be removed. Otherwise, the student would not be awarded the point, as adding a new, correct runoff arrow without removing the old, incorrect runoff arrow would not result in a correct diagram.
2. Train/Test Split. The labeled data was partitioned into training (80%) and testing (20%) sets.

Prompt Development

1. *Persona Pattern*. The prompt begins by employing the *persona pattern* (White et al., 2023) to inform the LLM of its role and task. This is the same as for Q1 and Q2.
2. Formative Assessment Question. The *persona pattern* is followed by the formative assessment question. This is the same as for Q1 and Q2, but with a different formative assessment question.
3. Additional Instructions. An additional instruction was provided to the LLM to ignore differences between students referencing "Libby's model" or "Taylor's model," as we felt the LLM may interpret multiple student names as multiple diagrams (when the two diagrams are, in fact, one and the same).
4. Rubric. The rubric is then inserted into the prompt (along with an explanation and response formatting information), and the model is instructed to score students' responses pursuant to that rubric. This is the same as for Q1 and Q2, but with a different rubric specific for Q3.
5. Few-Shot Example Selection. For Q3, five few-shot examples were then selected for prompt inclusion for the same reasons as Q1 and Q2. The ordering of the few-shot instances was deliberately engineered (as always, in this work) to ensure an even label distribution within the prompt (i.e., to avoid labeled instances being "clumped" by score). Just as in Q2, Q3 contains multiple binary subscores summing to a total score, which

makes it a multilabel task. Because of this, balancing the subscores was not able to be done perfectly, although we do ensure there is at least one positive and one negative instance for each subscore. The Q3 few-shot instances were chosen for the following reasons:

   a. Few-Shot Example 1. Ground truth instance where all four subscores were positive (total score of 4).
   b. Few-Shot Example 2. Ground truth instance where none of the four subscores were positive (total score of 0).
   c. Few-Shot Example 3. Instance added for data balance with positive subscores for science concepts only (*Runoff Direction* and *Arrow Size)*.
   d. Few-Shot Example 4. Sticking point instance with no positive subscores.
   e. Few-Shot Example 5. Instance added for data balance with a positive subscore for *Runoff Direction*.

6. Chain-of-Thought Reasoning. CoT reasoning chains were added to each of the labeled examples in the prompt to alert the model as to why or why not the LLM should award the student a point for his or her Q3 response. For the sticking point instance, the CoT reasoning also addresses the raters' reason for disagreement during IRR to help the model better align with the human consensus.
7. Few-Shot Example Inclusion. The selected few-shot examples with CoT reasoning were appended to the prompt, at which point the prompt was ready to be tested on the validation set.

Active Learning

1. Validation Set Inference. Model inference was performed on the validation set (i.e., all instances in the training set except for the five few-shot instances in the prompt) using the prompt we created during *Prompt Development*.
2. Scoring Trend Determination. We looked at the model's predictions and noted that the ratio of FPs to FNs was 9:4 (i.e., the model overscored), which meant that we needed to use additional labeled instances with CoT reasoning chains to help the model better identify negative instances. For three out of the four subscores, FPs outweighed FNs by a considerable margin. The one exception was Q3 *Arrow Size*, whose FN instances outweighed FP instances by a ratio of 7:4. This informed us that, in addition to accounting for the model's tendency to generally overscore, we had to address the model's underscoring for the *Arrow Size* subscore, specifically.
3. Additional Few-Shot Example Selection. In total, the model predicted 65 validation set subscores incorrectly (45 FPs, 20 FNs) across all instances. We added 4 additional instances back into the prompt. Two FP instances were added to the prompt to correct false positives corresponding to Q3 *Runoff Direction* and Q3 *Arrow Size Reasoning*. The model had a problem with overscoring (i.e., generating false positives), so these instances included CoT reasoning chains to address this for both science concepts and scientific reasoning. One FN instance was added with CoT reasoning to address the model's tendency to underscore Q3 *Arrow Size*. The fourth instance was added for data balance.

4. Test Set Inference. The updated prompt (with the few-shot instances added during active learning) was fed through the LLM with the test set instances to conduct our analysis. Like Qs 1 and 2, our analysis included: (1) quantitative analysis comparing the model's scores to the human's for both Macro-F1 and QWK (presented in this work), (2) a thematic analysis via inductive coding to identify trends in the LLM's scoring errors and patterns in its explanations, and (3) a researcher comparison between the educator's scoring and the LLM's scoring for all test set instances where the educator and LLM disagreed. 2 and 3 are outside the scope of this work but were presented in a previous work (anonymized for peer review). For reference, for Q3, the researcher sided with the LLM over the human in 1/10 instances of scoring disagreement in the test set, and that instance was a disagreement about a science concepts subscore.