

Prompts

Below are all of our initial CoT prompts before Active Learning. Student responses appear as-is and were not corrected for spelling and grammar errors. For each question's prompt, we explain why we chose each few-shot instance. The prompt also instructs the LLM to ignore whether the student referred to "Libby's" model or "Taylor's" model. We use "&&&" as a message delimiter for string parsing purposes and ":" as a delimiter to separate the API message *role* from the *content*.

Additionally, the names of the subscores in the manuscript differ slightly from the names of the subscores in the prompts. In the manuscript, the subscore names were altered for readability and to avoid confusion. We provide a mapping below. For Q1, *Arrow Size* is simply referred to as *Score* in the prompt because it is the only subscore in Q1. The last two lines of each prompt are placeholders to let the user know where to insert the instance during inference.

<u>FA Question</u>	<u>Manuscript Subscore Name</u>	<u>Prompt Subscore Name</u>
1	Arrow Size	Score
2	Arrow Direction	Representing Science Concepts
	Arrow Direction Reasoning	Reasoning About Science Concepts
	Arrow Size	Size of Arrows Representing Amounts
	Arrow Size Reasoning	Reasoning About Size of Arrows
3	Runoff Direction	Direction of Runoff Arrow
	Runoff Direction Reasoning	Reasoning About Direction of Runoff Arrow
	Arrow Size	Size of Arrows
	Arrow Size Reasoning	Reasoning About Size of Arrows

Q1 Prompt

system: You are a teacher whose job it is to score middle school student short answer formative assessment question responses in the Earth Science domain.

Students are asked the following question: What do you think the different sized arrows in the model could mean?

Importantly, it does not matter if the student refers to Libby's model or Taylor's model because both models are identical.

You are to score the responses based on the following rubric:

Score [0 or 1]: 1 point if the student correctly identifies that the different sized arrows correspond to how much water (i.e., the size, amount, or quantity), otherwise 0 points.

&&&

user: The amount of water used. Rainfall, runoff, absorption.

&&&

assistant: The student mentions "The amount of water used." This indicates that the student understands the size of the arrows corresponds to the quantity of water. Based on the rubric, the student earned a score of 1.

&&&

user: because the grass absorbs most of the water as it goes down the stream

&&&

assistant: The student says "the grass absorbs most of the water", but he or she does not give any indication that arrow size as a whole corresponds to the amount of water. Based on the rubric, the student earned a score of 0.

&&&

user: more of something or less of something.

&&&

assistant: The student says "more of something or less of something." Even though the student doesn't mention water specifically, it is clear that the student understands that the arrow sizes correspond to the quantity of water. Based on the rubric, the student earned a score of 1.

&&&

user: How much important it is in Water Cycle

&&&

assistant: The student mentions "How much important" but gives no indication that he or she understands arrow size is indicative of water amount. Based on the rubric, the student earned a score of 0.

&&&

user: [INSERT INSTANCE HERE]

&&&

assistant:

The prompt for Q1 contained 4 few-shot examples. The first two are ground truth instances both humans agreed on during *Response Scoring*. The third labeled instance is one where the two

researchers disagreed during IRR. In this case, there was debate over whether a student should receive a point if he or she just provided a general reference to quantity (e.g., "more of something") without explicitly mentioning water. With this example, the humans eventually agreed the student should earn a point, as the student is able to tie arrow size to quantity. The CoT reasoning in this instance addresses this sticking point pursuant to the consensus. The fourth instance was added for data balance.

Q2 Prompt

system: You are a teacher whose job it is to score middle school student short answer formative assessment question responses in the Earth Science domain.

Students are asked the following question: What are two things that the model does a good job of explaining?

Importantly, it does not matter if the student refers to Libby's model or Taylor's model because both models are identical.

You are to score the responses based on the following rubric:

Representing Science Concepts [0 or 1]: 1 point if the student correctly identifies that the model does a good job of showing that water is absorbed, water runs off, or water originates from the sky or clouds in the form of rain; otherwise, 0 points.

Reasoning About Science Concepts [0 or 1]: 1 point if the student received a point for Representing Science Concepts AND provides an explanation or justification that demonstrates mechanistic reasoning with regard to water being absorbed, water running off, or water falling from the sky in the form of rain. 0 points if the student did not receive a point for Representing Science Concepts, or if the student did receive a point for Representing Science Concepts but failed to demonstrate mechanistic reasoning with regard to Representing Science Concepts in his or her response.

Size of Arrows Representing Amounts [0 or 1]: 1 point if the student correctly identifies that the model does a good job of using arrow size to correspond to the amount, size, or quantity of water; otherwise, 0 points.

Reasoning About Size of Arrows [0 or 1]: 1 point if the student received a point for Size of Arrows Representing Amounts AND provides an explanation or justification that demonstrates mechanistic reasoning with regard to arrow size corresponding to amount, size, or quantity of water. 0 points if the student did not receive a point for Size of Arrows Representing Amounts, or if the student did receive a point for Size of Arrows Representing Amounts but failed to demonstrate mechanistic reasoning with regard to Size of Arrows Representing Amounts in his or her response.

Total Score [0 to 4]: Sum of scores from Representing Science Concepts, Reasoning About Science Concepts, Size of Arrows Representing Amounts, and Reasoning About Size of Arrows scores.

&&&

user: libby did great on the rainfall and absorption because it shows where it goes or points

&&&

assistant: Representing Science Concepts: The student explicitly mentions both "rainfall" and "absorption." This demonstrates that the student correctly identified that the model does a good job of indicating the water originates from the sky in the form of rain and that the water is absorbed. Based on the rubric, the student earned a score of 1.

Reasoning About Science Concepts: The student received 1 point for Representing Science Concepts, and the student says "because it shows where

it goes or points." This demonstrates that the student used mechanistic reasoning to justify his or response with regard to Representing Science Concepts. Based on the rubric, the student earned a score of 1.

Size of Arrows Representing Amounts: The student did not mention arrow size or amount of water, so the student did not indicate that the model does a good job of using arrow size to correspond to quantity of water. Based on the rubric, the student earned a score of 0.

Reasoning About Size of Arrows: The student did not receive a point for Size of Arrows Representing Amounts, so the student is ineligible to receive a point for Reasoning About Size of Arrows. Based on the rubric, the student earned a score of 0.

Total Score: The student received 1 point for Representing Science Concepts and 1 point for Reasoning About Science Concepts, so the student's total score is 2 points.

&&&

user: Two things that Taylor's model does a good job of explaining is how much water is runoff or could be potentially, and how deep the rainfall could be absorbed into the surface material, and or how heavily the absorption of the rainfall could be, form the thickness or the thinness of the arrows show in the model.

&&&

assistant: Representing Science Concepts: The student says "water is runoff", "the rainfall", and "could be absorbed". This demonstrates that the student correctly identified that the model does a good job of showing water becomes runoff, water originates from the sky in the form of rain, and water gets absorbed. Based on the rubric, the student earned a score of 1.

Reasoning About Science Concepts: The student received a point for Representing Science Concepts, but he or she did not provide an explanation or justification demonstrating mechanistic reasoning with respect to Representing Science Concepts in his or her response. Based on the rubric, the student earned a score of 0.

Size of Arrows Representing Amounts: The student says "how much water". This indicates that the student understands the model does a good job of using arrow size to correspond to the amount of water. Based on the rubric, the student earned a score of 1.

Reasoning About Size of Arrows: The student received a point for Size of Arrows Representing Amounts, and the student mentions "the thickness or the thinness of the arrows" as a justification for his or her response. This explanation displays mechanistic reasoning with respect to Size of Arrows Representing Amounts. Based on the rubric, the student earned a score of 1.

Total Score: The student received 1 point for Representing Science Concepts, 1 point for Size of Arrows Representing Amounts, and 1 point for Reasoning About Size of Arrows, so the student's total score is 3 points.

&&&

user: What goes where and how much of it

&&&

assistant: Representing Science Concepts: The student mentions "What goes where", which indicates arrow direction, but the student does not explicitly mention that water get absorbed, runs off, or originates from the sky or

clouds in the form of rain. Based on the rubric, the student earned a score of 0.

Reasoning About Science Concepts: The student did not receive a point for Representing Science Concepts, so the student is ineligible to receive a point for Reasoning About Science Concepts. Based on the rubric, the student earned a score of 0.

Size of Arrows Representing Amounts: The student says "how much of it". This demonstrates that the student understands the model does a good job explaining that arrow sizes correspond to quantity of water. Based on the rubric, the student earned a score of 1.

Reasoning About Size of Arrows: The student received a point for Size of Arrows Representing Amounts, but the student did not provide an explanation in his or her response, so no mechanistic reasoning was used with regard to Size of Arrows Representing Amounts. Based on the rubric, the student earned a score of 0.

Total Score: The student received 1 point for Size of Arrows Representing Amounts, so the student's total score is 1 point.

&&&

user: It does a good job of showing a soft surfaces because there is a lot water being absorbed and barely any runoff

&&&

assistant: Representing Science Concepts: The student explicitly mentions "water being absorbed" and "runoff". This indicates that the student understands the model does a good job of explaining that water is absorbed or becomes runoff. Based on the rubric, the student earned a score of 1.

Reasoning About Science Concepts: The student received a point for Representing Science Concepts, but the student does not provide an explanation of his or her response with respect to Representing Science Concepts that demonstrates mechanistic reasoning. Based on the rubric, the student earned a score of 0.

Size of Arrows Representing Amounts: The student says "a lot water being absorbed" and "barely any runoff". This indicates that the student understands the model does a good job of explaining that the different arrow sizes correspond to different amounts of water. Based on the rubric, the student earned a score of 1.

Reasoning About Size of Arrows: The student received a point for Size of Arrows Representing Amounts, but he or she does not provide an explanation or justification in the response that demonstrates mechanistic reasoning with respect to Size of Arrows Representing Amounts. The student says "because there is a lot water", but this reasoning is applied to the model's surfaces and not to Size of Arrows Representing Amounts. Based on the rubric, the student earned a score of 0.

Total Score: The student received 1 point for Representing Science Concepts and 1 point for Size of Arrows Representing Amounts, so the student's total score is 2 points.

&&&

user: Arrows are good, kind of show what the words mean?!

&&&

assistant: Representing Science Concepts: The student does not explicitly mention that water gets absorbed, runs off, or originates from the sky in the form of rain. Based on the rubric, the student earned a score of 0.

Reasoning About Science Concepts: The student did not earn a point for Representing Science Concepts, so the student is ineligible to receive a point for Reasoning About Science Concepts. Based on the rubric, the student earned a score of 0.

Size of Arrows Representing Amounts: The student mentions that "Arrows are good" but gives no indication that he or she understands arrow size corresponds to the size or quantity of water. Based on the rubric, the student earned a score of 0.

Reasoning About Size of Arrows: The student did not receive a point for Size of Arrows Representing Amount, so the student is ineligible to receive a point for Reasoning About Size of Arrows. Based on the rubric, the student earned a score of 0.

Total Score: The student did not receive any points for Representing Science Concepts, Reasoning About Science Concepts, Size of Arrows Representing Amounts, or Reasoning About Size of Arrows, so the student's total score is 0 points.

&&&

user: [INSERT INSTANCE HERE]

&&&

assistant:

The prompt for Q2 contained five few-shot examples. The first two are ground truth instances both researchers agreed on during IRR. The third and fourth instances are those the researchers did not agree on, where we used CoT to address the specific sticking points. In the third few-shot example, the researchers did not initially agree whether to award a point for *Arrow Direction* if the student mentioned where the water was going or where it came from without explicitly mentioning rain, runoff, or absorption. The researchers eventually agreed that the student needed to use the science concepts (i.e., rain, rainfall, absorb, absorption, runoff, etc.) to receive credit. The fourth example was mentioned in the manuscript: does the student receive credit for *Arrow Size* if he or she uses the absorption arrow as a “good” example, even though the absorption arrow in the diagram is wrong? In this case, the researchers agreed to award the student the point because the student mentioned “a lot water being absorbed” and “barely any runoff,” both of which are true statements and indicate the student understands the diagram does a good job of showing arrow size corresponds to quantity of water. If the student were to have simply said, “the amount of the absorption arrow,” however, the point would not have been awarded. The fifth few-shot instance was added for data balance.

Q3 Prompt

system: You are a teacher whose job it is to score middle school student short answer formative assessment question responses in the Earth Science domain.

Students are asked the following question: What are two things that you would change about the model to explain where the water goes?

Importantly, it does not matter if the student refers to Libby's model or Taylor's model because both models are identical.

You are to score the responses based on the following rubric:

Direction of Runoff Arrow [0 or 1]: 1 point if the student mentions the water runoff arrow should be pointing in the opposite direction (it currently points uphill, but it should point downhill). 0 points if the student does not identify that the runoff arrow needs to change directions.

Reasoning About Direction of Runoff Arrow [0 or 1]: 1 point if the student received a point for Direction of Runoff Arrow AND provides an explanation or justification for why the runoff arrow should point in the opposite direction that demonstrates mechanistic reasoning with respect to Direction of Runoff Arrow. 0 points if the student did not receive a point for Direction of Runoff Arrow, or if the student did receive a point for Direction of Runoff Arrow but failed to demonstrate mechanistic reasoning with regard to Direction of Runoff Arrow.

Size of Arrows [0 or 1]: 1 point if the student mentions that the arrow sizes need to change (in the model, the absorption and runoff arrows sum to a value larger than the rainfall arrow, which violates conservation of matter). 0 points if the student does not mention that the arrow sizes need to be changed.

Reasoning About Size of Arrows [0 or 1]: 1 point if the student received a point for Size of Arrows AND provides an explanation or justification for why the arrow sizes need to be changed that demonstrates mechanistic reasoning with respect to Size of Arrows. 0 points if the student did not receive a point for Size of Arrows, or if the student did receive a point for Size of Arrows but did not demonstrate mechanistic reasoning with regard to Size of Arrows.

Total Score [0 to 4]: Sum of scores from Direction of Runoff Arrow, Reasoning About Direction of Runoff Arrow, Size of Arrows, and Reasoning About Size of Arrows scores.

&&&

user: I think that Taylor needs to make the rainfall the biggest arrow because the absorption and runoff is derived from the rainfall. I also think that the runoff should point down because water can't move uphill.

&&&

assistant: Direction of Runoff Arrow: The student says "the runoff should point down". 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.

Reasoning About Direction of Runoff Arrow: The student received a point for Direction of Runoff Arrow, and the student justifies his or her response by saying "because water can't move uphill". This demonstrates mechanistic reasoning with regard to Direction of Runoff Arrow. Based on the rubric, the student earned a score of 1.

Size of Arrows: The student says "Taylor needs to make the rainfall the biggest arrow". This demonstrates that the student understands the arrow sizes need to be changed. Based on the rubric, the student earned a score of 1.

Reasoning About Size of Arrows: The student received a point for Reasoning About Direction of Runoff Arrow, and the student explains his or her response by saying "because the absorption and runoff is derived from the 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 mechanistic reasoning with regard to Size of Arrows. Based on the rubric, the student earned a score of 1.

Total Score: The student received 1 point for Direction of Runoff Arrow, 1 point for Reasoning About Direction of Runoff Arrow, 1 point for Size of Arrows, and 1 point for Reasoning About Size of Arrows, so the student's total score is 4 points.

&&&

user: She could have been able to explain what the arrow that was going up means. It would be clearer if it was labeled.

&&&

assistant: Direction of Runoff Arrow: The student made no mention of the runoff arrow needing to change to point in the opposite direction. Based on the rubric, the student earned a score of 0.

Reasoning About Direction of Runoff Arrow: The student did not receive a point for Direction of Runoff Arrow, so the student is ineligible to receive a point for Reasoning About Direction of Runoff Arrow. Based on the rubric, the student earned a score of 0.

Size of Arrows: The student did not discuss the size of the arrows or that the arrow sizes need to change. Based on the rubric, the student earned a score of 0.

Reasoning About Size of Arrows: The student did not receive a point for Size of Arrows, so the student is ineligible to receive a point for Reasoning About Size of Arrows. Based on the rubric, the student earned a score of 0.

Total Score: The student did not receive any points for Direction of Runoff Arrow, Reasoning About Direction of Runoff Arrow, Size of Arrows, or Reasoning About Size of Arrows, so the student's total score is 0 points.

&&&

user: The runoff should go down hill and less water should be absorbed.

&&&

assistant: Direction of Runoff Arrow: The student states "The runoff should go down hill". This indicates the student understands that the runoff arrow needs to change to point downhill in the opposite direction. Based on the rubric, the student earned a score of 1.

Reasoning About Direction of Runoff Arrow: The student received a point for Direction of Runoff Arrow, but the student did not justify or explain his or her response, so the student does not demonstrate mechanistic reasoning with respect to Direction of Runoff Arrow. Based on the rubric, the student earned a score of 0.

Size of Arrows: The student says "less water should be absorbed". This shows that the student understands the absorption arrow should be smaller and that the size of the arrows needs to change. Based on the rubric, the student earned a score of 1.

Reasoning About Size of Arrows: The student received a point for Size of Arrows, but the student failed to demonstrate mechanistic reasoning with respect to Size of Arrows because the student did not explain or justify his or her response. Based on the rubric, the student earned a score of 0.

Total Score: The student received 1 point for Direction of Runoff Arrow and 1 point for Size of Arrows, so the student's total score is 2 points.

&&&

user: I would put an arrow going towards the Stream because some water goes there.

&&&

assistant: Direction of Runoff Arrow: The student says "I would put an arrow going towards the Stream". At first glance, it seems that the student understands the direction of the runoff arrow needs to change to point downhill towards the stream. However, if an arrow pointing towards the stream is added to the model, but the original arrow is not removed, the model will still be wrong because there will now be two runoff arrows (one pointing downhill and one pointing uphill), when there should only be one runoff arrow that points downhill towards the stream. Because of this, if the student says that an arrow should be added to the model that points downhill in the correct direction, but the student fails to mention removing the previous incorrect runoff arrow pointing uphill, the student does not earn a point for Direction of Runoff Arrow. Based on the rubric, the student earned a score of 0.

Reasoning About Direction of Runoff Arrow: The student did not receive a point for Direction of Runoff Arrow, so the student is ineligible to receive a point for Reasoning About Direction of Runoff Arrow. Based on the rubric, the student earned a score of 0.

Size of Arrows: The student does not mention arrow size or demonstrate that he or she understands that the size of the arrows need to be changed. Based on the rubric, the student earned a score of 0.

Reasoning About Size of Arrows: The student did not receive a point for Size of Arrows, so the student is ineligible to receive a point for

Reasoning About Size of Arrows. Based on the rubric, the student earned a score of 0.

Total Score: The student did not receive any points for Direction of Runoff Arrow, Reasoning About Direction of Runoff Arrow, Size of Arrows, or Reasoning About Size of Arrows, so the student's total score is 0 points.

&&&

user: I would change the direction of the runoff. I would also put what the different sized arrows mean and why they are there

&&&

assistant: Direction of Runoff Arrow: The student mentions he or she "would change the direction of the runoff". This shows that the student understands that the runoff arrow is pointing in the wrong direction and needs to be changed to point downhill towards the stream. Based on the rubric, the student earned a score of 1.

Reasoning About Direction of Runoff Arrow: The student received a point for Direction of Runoff Arrow, but the student did not exhibit mechanistic reasoning with regard to Direction of Runoff Arrow because the student did not justify or explain his or her response. Based on the rubric, the student earned a score of 0.

Size of Arrows: The student mentions "the different sized arrows" but does not specifically mention that the arrow sizes need to be changed. Based on the rubric, the student earned a score of 0.

Reasoning About Size of Arrows: The student did not receive a point for Size of Arrows, so the student is ineligible to receive a point for Reasoning About Size of Arrows. Based on the rubric, the student earned a score of 0.

Total Score: The student received 1 point for Direction of Runoff Arrow, so the student's total score is 1 point.

&&&

user: [INSERT INSTANCE HERE]

&&&

assistant:

The prompt for Q3 contained five few-shot examples. Like Qs 1 and 2, the first two instances are ground truth examples the researchers agreed on during IRR. The fourth instance represents a sticking point encountered during IRR. In this case, the student mentioned he or she would add an arrow pointing downhill towards the stream. While this indicates the student understands the runoff direction needs to be changed, this is still technically incorrect, as adding a new runoff arrow pointing in the correct direction does not fix the arrow still pointing in the incorrect direction. As such, the reviewers agreed that this student should not receive a point for *Runoff Direction* for Q3. This rationale is explained in the example's CoT reasoning in the prompt. The third and fifth instances were added for balance.