

Tutorial Phase 2: Pedagogical Ability Assessment of LLM-Powered AI Tutors

Phase 2: Testing Phase

For the 8 questions in this phase, we ask you to assess your understanding based on what you learned during the training phase. **Once you've completed the test, you can view feedback on your responses by clicking 'View Score'.**

* Indicates required question

1. Enter your full name *

Phase 2: Testing Phase

The Task: Your task is to evaluate the tutor's responses across various dimensions of the evaluation taxonomy introduced during the training phase. Specifically, given the previous conversation between the tutor and student, annotate the appropriateness of the current tutor response for each dimension of the taxonomy.

Note: Please assume that the evaluation dimensions are *orthogonal* to each other; that is, the evaluation of one dimension should not influence the evaluation of other dimensions. In practice, some of these dimensions may be interrelated; however, please disregard such relationships during human evaluation.

Once you've completed the test, you can view feedback on your responses by clicking 'View Score'.

Evaluation Dimension 1: Mistake Identification

Definition: Has the tutor *identified* a mistake in a student's response?

Annotation Labels:

1. Yes
2. To some extent
3. No

2. ----- **Example 1** ----- * 1 point

Conversation Topic: Converting Units of Measure

Conversation History:

Tutor: I will help you solve this question.

Tutor: One pound is equal to 16 ounces.

Tutor: So, how many ounces are equal to 12 pounds?

Student: 4

Tutor Response: Oh, I see where you might have gotten confused - let's work through this together to find the correct answer.

Question: Has the tutor identified the mistake in the above response?

Mark only one oval.

- ☐ Yes
- ☐ To some extent
- ☐ No

Phase 2: Testing Phase

The Task: Your task is to evaluate the tutor's responses across various dimensions of the evaluation taxonomy introduced during the training phase. Specifically, given the previous conversation between the tutor and student, annotate the appropriateness of the current tutor response for each dimension of the taxonomy.

Note: Please assume that the evaluation dimensions are *orthogonal* to each other; that is, the evaluation of one dimension should not influence the evaluation of other dimensions. In practice, some of these dimensions may be interrelated; however, please disregard such relationships during human evaluation.

Once you've completed the test, you can view feedback on your responses by clicking 'View Score'.

Evaluation Dimension 2: Mistake Location

Definition: Does the tutor's response *accurately point* to a *genuine* mistake and its location?

Annotation Labels:

1. Yes
2. To some extent
3. No

3. ----- **Example 2** ----- * 1 point

Conversation Topic: Converting Units of Measure

Conversation History:

Tutor: I will help you solve this question.

Tutor: One pound is equal to 16 ounces.

Tutor: So, how many ounces are equal to 12 pounds?

Student: 4

Tutor Response: Oh, I see where you might have gotten confused - let's work through this together to find the correct answer.

Question: Did the tutor indicate the exact mistake or location in the above response?

Mark only one oval.

- ☐ Yes
- ☐ To some extent
- ☐ No

Phase 2: Testing Phase

The Task: Your task is to evaluate the tutor's responses across various dimensions of the evaluation taxonomy introduced during the training phase. Specifically, given the previous conversation between the tutor and student, annotate the appropriateness of the current tutor response for each dimension of the taxonomy.

Note: Please assume that the evaluation dimensions are *orthogonal* to each other; that is, the evaluation of one dimension should not influence the evaluation of other dimensions. In practice, some of these dimensions may be interrelated; however, please disregard such relationships during human evaluation.

Once you've completed the test, you can view feedback on your responses by clicking 'View Score'.

Evaluation Dimension 3: Revealing of the Answer

Definition: Does the tutor reveal the *final* answer (whether correct or not)?

Annotation Labels:

1. Yes (and the revealed answer is *correct*)
2. Yes (but the revealed answer is *incorrect*)
3. No

4. ----- **Example 3** ----- * 1 point

Conversation Topic: Converting Units of Measure

Conversation History:

Tutor: I will help you solve this question.

Tutor: One pound is equal to 16 ounces.

Tutor: So, how many ounces are equal to 12 pounds?

Student: 4

Tutor Response: It looks like you divided 12 by 3, but we need to multiply by 16 since there are 16 ounces in each pound.

Question: Did the above tutor's response reveal the final answer?

Mark only one oval.

- ☐ Yes (and the revealed answer is correct)
- ☐ Yes (but the revealed answer is incorrect)
- ☐ No

Phase 2: Testing Phase

The Task: Your task is to evaluate the tutor's responses across various dimensions of the evaluation taxonomy introduced during the training phase. Specifically, given the previous conversation between the tutor and student, annotate the appropriateness of the current tutor response for each dimension of the taxonomy.

Note: Please assume that the evaluation dimensions are *orthogonal* to each other; that is, the evaluation of one dimension should not influence the evaluation of other dimensions. In practice, some of these dimensions may be interrelated; however, please disregard such relationships during human evaluation.

Once you've completed the test, you can view feedback on your responses by clicking 'View Score'.

Evaluation Dimension 4: Providing Guidance

Definition: Does the tutor offer *correct* and *relevant* guidance, such as an explanation, elaboration, hint, examples, and so on?

Annotation Labels:

1. Yes (guidance is *correct* and *relevant* to the mistake)
2. To some extent (guidance is provided but it is fully or partially *incorrect* or *incomplete*)
3. No

5. ----- **Example 4** ----- * 1 point

Conversation Topic: Converting Units of Measure

Conversation History:

Tutor: I will help you solve this question.

Tutor: One pound is equal to 16 ounces.

Tutor: So, how many ounces are equal to 12 pounds?

Student: 4

Tutor Response: That was a good try. What is the product of 6 and 2?

Question: Does the tutor offer correct and relevant guidance?

Mark only one oval.

☐ Yes

☐ To some extent

☐ No

Phase 2: Testing Phase

The Task: Your task is to evaluate the tutor's responses across various dimensions of the evaluation taxonomy introduced during the training phase. Specifically, given the previous conversation between the tutor and student, annotate the appropriateness of the current tutor response for each dimension of the taxonomy.

Note: Please assume that the evaluation dimensions are *orthogonal* to each other; that is, the evaluation of one dimension should not influence the evaluation of other dimensions. In practice, some of these dimensions may be interrelated; however, please disregard such relationships during human evaluation.

Once you've completed the test, you can view feedback on your responses by clicking 'View Score'.

Evaluation Dimension 5: Coherent

Definition: Is the tutor's response *logically consistent* with the student's previous response?

Annotation Labels:

1. Yes
2. To some extent
3. No

6. ----- **Example 5** ----- * 1 point

Conversation Topic: Converting Units of Measure

Conversation History:

Tutor: I will help you solve this question.

Tutor: One pound is equal to 16 ounces.

Tutor: So, how many ounces are equal to 12 pounds?

Student: 4

Tutor Response: That was a good try. What is the product of 6 and 2?

Question: Is the tutor's response logically consistent with the student's previous response?

Mark only one oval.

- ☐ Yes
- ☐ To some extent
- ☐ No

Phase 2: Testing Phase

The Task: Your task is to evaluate the tutor's responses across various dimensions of the evaluation taxonomy introduced during the training phase. Specifically, given the previous conversation between the tutor and student, annotate the appropriateness of the current tutor response for each dimension of the taxonomy.

Note: Please assume that the evaluation dimensions are *orthogonal* to each other; that is, the evaluation of one dimension should not influence the evaluation of other dimensions. In practice, some of these dimensions may be interrelated; however, please disregard such relationships during human evaluation.

Once you've completed the test, you can view feedback on your responses by clicking 'View Score'.

Evaluation Dimension 6: Actionability**Definition:** Is it clear from the tutor's feedback *what the student should do next?***Annotation Labels:**

1. Yes
2. To some extent
3. No

7. ----- **Example 6** ----- * 1 point

Conversation Topic: Converting Units of Measure
-----**Conversation History:****Tutor:** I will help you solve this question.**Tutor:** One pound is equal to 16 ounces.**Tutor:** So, how many ounces are equal to 12 pounds?**Student:** 4
-----**Tutor Response:** : It looks like you divided 12 by 3, but we need to multiply by 16 since there are 16 ounces in each pound.
-----**Question:** Is it clear from the tutor's feedback what the student should do next?*Mark only one oval.*

- ☐ Yes
- ☐ To some extent
- ☐ No

Phase 2: Testing Phase

The Task: Your task is to evaluate the tutor's responses across various dimensions of the evaluation taxonomy introduced during the training phase. Specifically, given the previous conversation between the tutor and student, annotate the appropriateness of the current tutor response for each dimension of the taxonomy.

Note: Please assume that the evaluation dimensions are *orthogonal* to each other; that is, the evaluation of one dimension should not influence the evaluation of other dimensions. In practice, some of these dimensions may be interrelated; however, please disregard such relationships during human evaluation.

Once you've completed the test, you can view feedback on your responses by clicking 'View Score'.

Evaluation Dimension 7: Tutor Tone

Definition: Is the tutor's response *encouraging*, *neutral*, or *offensive*?

Annotation Labels:

1. Encouraging
2. Neutral
3. Offensive

Note: The 'Encouraging' label encompasses a range of positive characteristics typically exhibited by a tutor, such as support, care, understanding, moral encouragement, and empathy. Responses like 'good try,' 'let's work together,' or 'you are close' fall under this label. Conversely, responses that contain offensive or toxic content, such as 'you are stupid' and "you should have studied," are categorized under the 'Offensive' label. Finally, responses that sound neutral or natural, but do not fit into either of the above categories, are labeled as 'Neutral.'

8. ----- **Example 7** ----- * 1 point

Conversation Topic: Converting Units of Measure

Conversation History:

Tutor: I will help you solve this question.

Tutor: One pound is equal to 16 ounces.

Tutor: So, how many ounces are equal to 12 pounds?

Student: 4

Tutor Response: It looks like you divided 12 by 3, but we need to multiply by 16 since there are 16 ounces in each pound.

Question: Is the above tutor's response encouraging, neutral, or offensive?

Mark only one oval.

☐ Encouraging

☐ Neutral

☐ Offensive

Phase 2: Testing Phase

The Task: Your task is to evaluate the tutor's responses across various dimensions of the evaluation taxonomy introduced during the training phase. Specifically, given the previous conversation between the tutor and student, annotate the appropriateness of the current tutor response for each dimension of the taxonomy.

Note: Please assume that the evaluation dimensions are *orthogonal* to each other; that is, the evaluation of one dimension should not influence the evaluation of other dimensions. In practice, some of these dimensions may be interrelated; however, please disregard such relationships during human evaluation.

Once you've completed the test, you can view feedback on your responses by clicking 'View Score'.

Evaluation Dimension 8: Humanness

Definition: Does the tutor's response sound *natural*, rather than *robotic* or *artificial*?

Annotation Labels:

1. Yes
2. To some extent
3. No

Note: Based on our observations, most of the outputs generated by the LLM resemble human-like responses. There are only a few instances where the outputs do not exhibit human-like qualities, such as: (a) automated responses like 'try again,' (b) cases where the model includes prompt instructions in the response, and so on.

9. ----- **Example 8** ----- * 1 point

Conversation Topic: Converting Units of Measure

Conversation History:

Tutor: I will help you solve this question.

Tutor: One pound is equal to 16 ounces.

Tutor: So, how many ounces are equal to 12 pounds?

Student: 4

Tutor Response: Actually, 12 pounds would be equal to 192 ounces, since you multiply 12 by 16

Question: Does the tutor's response sound natural and not robotic or artificial?

Mark only one oval.

- ☐ Yes
- ☐ To some extent
- ☐ No

