

# Tutorial Phase 1: Pedagogical Ability Assessment of LLM-Powered AI Tutors

Thank you for agreeing to participate in this human evaluation. We extend our sincere gratitude from the Educational NLP Lab at MBZUAI for your contribution to this important research.

Prior research has demonstrated that human evaluation can be subjective, often resulting in low inter-annotator agreement. To address potential ambiguities and ensure consistency, this tutorial has been introduced. It comprises two components: (1) **Training Phase: training human annotators (this document)**, and (2) **Testing Phase: testing to assess understanding on a small number of examples (see Testing document)**. This process will provide an overview of the task and clarify the expected evaluation standards.

**Evaluation Setting:** The evaluation takes place within the context of an educational dialogue between a tutor and a student, where the tutor's responses are aimed at addressing the student's mistakes or confusion in a pedagogical manner.

**The Task:** Your task is to evaluate the tutor's (LLM-as-tutor's) responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through this document. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

**There are 31 training examples and 8 test examples. Both phases are expected to take approximately 30 minutes.**

---

\* Indicates required question

1. Enter your full name \*

---

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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** -----

\*

**Conversation Topic:** Simple Expression

-----

**Conversation History:**

**Tutor:** We have to solve the inner parentheses first.

**Student:** ok

**Tutor:** What is 5 times 6?

**Student:** 50

-----

**Tutor Response:** Ah, not quite.  $5 \times 10$  is 50.  $5 \times 6$  is something else. Could you give it another try?

-----

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

*Mark only one oval.*

- ☐ Yes
- ☐ To some extent
- ☐ No

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

## 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

### ----- Example 1: Answer -----

**Conversation Topic:** Simple Expression

#### ----- **Conversation History:**

**Tutor:** We have to solve the inner parentheses first.




**Student:** ok

**Tutor:** What is 5 times 6?

**Student:** 50

-----  
**Tutor Response:** Ah, not quite.  $5 \times 10$  is 50.  $5 \times 6$  is something else. Could you give it another try?

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

- (1) Yes   
 (2) To some extent   
 (3) No 

#### **Reasoning:**

The tutor clearly identified the mistake by explaining how to arrive at 50.

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

## 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

### 3. ----- Example 2 ----- \*

**Conversation Topic:** Simple Expression

-----

#### Conversation History:

**Tutor:** We have to solve the inner parentheses first.

**Student:** ok

**Tutor:** What is 5 times 6?

**Student:** 50

-----

**Tutor Response:** Actually, the answer to 5 times 6 is 30. Let's try again.

-----

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

*Mark only one oval.*

- ☐ Yes
- ☐ To some extent
- ☐ No

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

## 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

### ----- Example 2: Answer -----

**Conversation Topic:** Simple Expression

#### ----- **Conversation History:**

**Tutor:** We have to solve the inner parentheses first.




**Student:** ok

**Tutor:** What is 5 times 6?

**Student:** 50

-----  
**Tutor Response:** Actually, the answer to 5 times 6 is 30. Let's try again.

-----  
**Question:** Did the tutor identify the mistake in the above response?

- (1) Yes   
 (2) To some extent   
 (3) No 

#### **Reasoning:**

The tutor clearly identified the mistake by providing the correct answer.

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

## 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

## 4. ----- Example 3 ----- \*

**Conversation Topic:** Simple Expression

-----

### Conversation History:

**Tutor:** We have to solve the inner parentheses first.

**Student:** ok

**Tutor:** What is 5 times 6?

**Student:** 50

-----

**Tutor Response:** Are you sure?

-----

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

*Mark only one oval.*

- ☐ Yes
- ☐ To some extent
- ☐ No

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

## 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

### ----- Example 3: Answer -----

**Conversation Topic:** Simple Expression

#### ----- **Conversation History:**

**Tutor:** We have to solve the inner parentheses first.

**Student:** ok

**Tutor:** What is 5 times 6?

**Student:** 50

-----  
**Tutor Response:** Are you sure?

-----  
**Question:** Did the tutor identify the mistake in the above response?

- (1) Yes 
- (2) To some extent 
- (3) No 

#### **Reasoning:**

The tutor's response suggests that there may be a mistake, but it sounds as if the tutor is not certain

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.



## 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

## 5. ----- Example 4 ----- \*

**Conversation Topic:** Simple Expression

-----

### Conversation History:

**Tutor:** We have to solve the inner parentheses first.

**Student:** ok

**Tutor:** What is 5 times 6?

**Student:** 50

-----

**Tutor Response:** That's correct, 5 multiplied by 6 equals 30.

-----

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

*Mark only one oval.*

- ☐ Yes
- ☐ To some extent
- ☐ No

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

## 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

### ----- Example 4: Answer -----

**Conversation Topic:** Simple Expression

#### ----- **Conversation History:**

**Tutor:** We have to solve the inner parentheses first.




**Student:** ok

**Tutor:** What is 5 times 6?

**Student:** 50

-----  
**Tutor Response:** That's correct, 5 multiplied by 6 equals 30.

-----  
**Question:** Did the tutor identify the mistake in the above response?

- (1) Yes 
- (2) To some extent 
- (3) No 

#### **Reasoning:**

The tutor did not identify the mistake; however, they proceeded to provide the answer to the asked question.

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

**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

6. ----- **Example 5** -----

★

**Conversation Topic:** Place Value

-----

**Conversation History:**

**Tutor:** Not to worry, [STUDENT].

**Student:** okay

**Tutor:** What is the place value of two in 204,530?

**Student:** ten millions?

-----

**Tutor Response:** I think there might be a little mistake there! The place value of the digit 2 in 204,530 is actually tens of thousands, not ten millions.

-----

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

*Mark only one oval.*

- ☐ Yes
- ☐ To some extent
- ☐ No

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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

#### ----- Example 5: Answer -----

**Conversation Topic:** Place Value

#### Conversation History:

**Tutor:** Not to worry, [STUDENT].




**Student:** okay

**Tutor:** What is the place value of two in 204,530?

**Student:** ten millions?

**Tutor Response:** I think there might be a little mistake there! The place value of the digit 2 in 204,530 is actually tens of thousands, not ten millions.

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

- (1) Yes 
- (2) To some extent 
- (3) No 

#### Reasoning:

The tutor clearly points out the exact mistake by stating, 'not ten million'

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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

7. ----- **Example 6** -----

\*

**Conversation Topic:** Place Value

**Conversation History:**

**Tutor:** Not to worry, [STUDENT].

**Student:** okay

**Tutor:** What is the place value of two in 204,530?

**Student:** ten millions?

**Tutor Response:** The place value of 2 in 204,530 is actually hundred thousands.

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

*Mark only one oval.*

- ☐ Yes
- ☐ To some extent
- ☐ No

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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

---

**Example 6: Answer**

---

**Conversation Topic:** Place Value

---

**Conversation History:**

**Tutor:** Not to worry, [STUDENT].

**Student:** okay

**Tutor:** What is the place value of two in 204,530?




**Student:** ten millions?

---

**Tutor Response:** The place value of 2 in 204,530 is actually hundred thousands.

---

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

- (1) Yes 
- (2) To some extent 
- (3) No 

**Reasoning:**

The tutor clearly points out the mistakes by providing the correct place values for 2.

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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

8. ----- **Example 7** -----

\*

**Conversation Topic:** Place Value

-----

**Conversation History:**

**Tutor:** Not to worry, [STUDENT].

**Student:** okay

**Tutor:** What is the place value of two in 204,530?

**Student:** ten millions?

-----

**Tutor Response:** That's a great try, but remember, the ten millions place is three places to the left of the hundred thousands place.

-----

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

*Mark only one oval.*

- ☐ Yes
- ☐ To some extent
- ☐ No

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.



## 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

### ----- Example 7: Answer -----

**Conversation Topic:** Place Value

#### ----- **Conversation History:**

**Tutor:** Not to worry, [STUDENT].




**Student:** okay

**Tutor:** What is the place value of two in 204,530?

**Student:** ten millions?

-----  
**Tutor Response:** That's a great try, but remember, the ten millions place is three places to the left of the hundred thousands place.

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

- (1) Yes   
(2) To some extent   
(3) No 

#### **Reasoning:**

The response demonstrates some awareness of the exact mistake by providing guidance to arrive at the correct answer.

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

## 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

## 9. ----- Example 8 ----- \*

**Conversation Topic:** Place Value

-----

### Conversation History:

**Tutor:** Not to worry, [STUDENT].

**Student:** okay

**Tutor:** What is the place value of two in 204,530?

**Student:** ten millions?

-----

**Tutor Response:** This is a very good try.

-----

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

*Mark only one oval.*

- ☐ Yes
- ☐ To some extent
- ☐ No

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

## 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

### ----- Example 8: Answer -----

**Conversation Topic:** Place Value

#### Conversation History:

**Tutor:** Not to worry, [STUDENT].




**Student:** okay

**Tutor:** What is the place value of two in 204,530?

**Student:** ten millions?

**Tutor Response:** This is a very good try.

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

- (1) Yes 
- (2) To some extent 
- (3) No 

#### Reasoning:

The response does not provide any details related to the mistakes.

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

**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

10. ----- **Example 9** -----

\*

**Conversation Topic:** Not Available**Conversation History:**

**Tutor:** Hi, could you please provide a step-by-step solution for the question below? The question is: While camping, Cyrus counted 14 mosquito bites on his arms and legs. He later counted another 10 on his body. While he was suffering, his entire family of 6 other people only got a total of half the number of bites that Cyrus got. If they all have an equal number of bites, how many mosquito bites does each member of the rest of Cyrus' family have?

**Student:** Cyrus got a total of  $14+10 = 24$  mosquito bites. His family got half the number of bites, so they got  $24/2 = 12$  mosquito bites in total. There are  $6+1 = 7$  members of the family. Dividing 12 by 7 gives each member of the family 1.71 (rounded to two decimal places) mosquito bites.

**Tutor:** could you please walk me through your solution?

**Student:** Sure! I started by adding the total number of bites Cyrus got which was  $14 + 10 = 24$ . Then since his family got half the number of bites I divided 24 by 2 to get 12. Since there are 7 members in his family I divided 12 by 7 to get 1.71 bites per person.

**Gold Solution:** Altogether, Cyrus got  $14 + 10 = 24$  mosquito bites. The rest of his family had a total of  $24/2 = 12$  mosquito bites. Each family member had  $12/6 = 2$  mosquito bites.

**Tutor Response:** Let's revisit the part where you calculated the total number of family members shouldn't it be 6, not 7, since Cyrus is already included in the original question?

**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 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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

### ----- Example 9: Answer -----

**Conversation Topic:** Not Available

#### Conversation History:

**Tutor:** Hi, could you please provide a step-by-step solution for the question below? The question is: While camping, Cyrus counted 14 mosquito bites on his arms and legs. He later counted another 10 on his body. While he was suffering, his entire family of 6 other people only got a total of half the number of bites that Cyrus got. If they all have an equal number of bites, how many mosquito bites does each member of the rest of Cyrus' family have?

**Student:** Cyrus got a total of  $14+10 = 24$  mosquito bites. His family got half the number of bites, so they got  $24/2 = 12$  mosquito bites in total. There are  $6+1 = 7$  members of the family. Dividing 12 by 7 gives each member of the family 1.71 (rounded to two decimal places) mosquito bites.

**Tutor:** could you please walk me through your solution?

**Student:** Sure! I started by adding the total number of bites Cyrus got which was  $14 + 10 = 24$ . Then since his family got half the number of bites I divided 24 by 2 to get 12. Since there are 7 members in his family I divided 12 by 7 to get 1.71 bites per person.

**Gold Solution:** Altogether, Cyrus got  $14 + 10 = 24$  mosquito bites. The rest of his family had a total of  $24/2 = 12$  mosquito bites. Each family member had  $12/6 = 2$  mosquito bites.

**Tutor Response:** Let's revisit the part where you calculated the total number of family members - shouldn't it be 6, not 7, since Cyrus is already included in the original question?

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

- (1) Yes (and the revealed answer is correct) ✗
- (2) Yes (but the revealed answer is incorrect) ✗
- (3) No ✔

#### Reasoning:

The response does not reveal the final answer; the revealed value 6 is a hint at an intermediate step, not the final answer.

### Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

**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



11. ----- **Example 10** -----

★

**Conversation Topic:** Not Available**Conversation History:**

**Tutor:** Hi, could you please provide a step-by-step solution for the question below? The question is: While camping, Cyrus counted 14 mosquito bites on his arms and legs. He later counted another 10 on his body. While he was suffering, his entire family of 6 other people only got a total of half the number of bites that Cyrus got. If they all have an equal number of bites, how many mosquito bites does each member of the rest of Cyrus' family have?

**Student:** Cyrus got a total of  $14+10 = 24$  mosquito bites. His family got half the number of bites, so they got  $24/2 = 12$  mosquito bites in total. There are  $6+1 = 7$  members of the family. Dividing 12 by 7 gives each member of the family 1.71 (rounded to two decimal places) mosquito bites.

**Tutor:** could you please walk me through your solution?

**Student:** Sure! I started by adding the total number of bites Cyrus got which was  $14 + 10 = 24$ . Then since his family got half the number of bites I divided 24 by 2 to get 12. Since there are 7 members in his family I divided 12 by 7 to get 1.71 bites per person.

**Gold Solution:** Altogether, Cyrus got  $14 + 10 = 24$  mosquito bites. The rest of his family had a total of  $24/2 = 12$  mosquito bites. Each family member had  $12/6 = 2$  mosquito bites.

**Tutor Response:** That's close, but let's review the division step together - when we divide 12 by 7, we actually get 1.714285 (not rounding to two decimal places yet), which then rounds to 1.71.

**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 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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

### ----- Example 10: Answer -----

**Conversation Topic:** Not Available

#### Conversation History:

**Tutor:** Hi, could you please provide a step-by-step solution for the question below? The question is: While camping, Cyrus counted 14 mosquito bites on his arms and legs. He later counted another 10 on his body. While he was suffering, his entire family of 6 other people only got a total of half the number of bites that Cyrus got. If they all have an equal number of bites, how many mosquito bites does each member of the rest of Cyrus' family have?

**Student:** Cyrus got a total of  $14+10 = 24$  mosquito bites. His family got half the number of bites, so they got  $24/2 = 12$  mosquito bites in total. There are  $6+1 = 7$  members of the family. Dividing 12 by 7 gives each member of the family 1.71 (rounded to two decimal places) mosquito bites.

**Tutor:** could you please walk me through your solution?

**Student:** Sure! I started by adding the total number of bites Cyrus got which was  $14 + 10 = 24$ . Then since his family got half the number of bites I divided 24 by 2 to get 12. Since there are 7 members in his family I divided 12 by 7 to get 1.71 bites per person.

**Gold Solution:** Altogether, Cyrus got  $14 + 10 = 24$  mosquito bites. The rest of his family had a total of  $24/2 = 12$  mosquito bites. Each family member had  $12/6 = 2$  mosquito bites.

**Tutor Response:** That's close, but let's review the division step together - when we divide 12 by 7, we actually get 1.714285 (not rounding to two decimal places yet), which then rounds to 1.71.

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

- (1) Yes (and the revealed answer is correct) ❌
- (2) Yes (but the revealed answer is incorrect) ✅
- (3) No ❌

#### Reasoning:

There are a total of 6 family members, not 7. The division by 7 makes the final answer incorrect.

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

**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

12. ----- **Example 11** -----

\*

**Conversation Topic:** Not Available**Conversation History:**

**Tutor:** Hi, could you please provide a step-by-step solution for the question below? The question is: While camping, Cyrus counted 14 mosquito bites on his arms and legs. He later counted another 10 on his body. While he was suffering, his entire family of 6 other people only got a total of half the number of bites that Cyrus got. If they all have an equal number of bites, how many mosquito bites does each member of the rest of Cyrus' family have?

**Student:** Cyrus got a total of  $14+10 = 24$  mosquito bites. His family got half the number of bites, so they got  $24/2 = 12$  mosquito bites in total. There are  $6+1 = 7$  members of the family. Dividing 12 by 7 gives each member of the family 1.71 (rounded to two decimal places) mosquito bites.

**Tutor:** could you please walk me through your solution?

**Student:** Sure! I started by adding the total number of bites Cyrus got which was  $14 + 10 = 24$ . Then since his family got half the number of bites I divided 24 by 2 to get 12. Since there are 7 members in his family I divided 12 by 7 to get 1.71 bites per person.

**Gold Solution:** Altogether, Cyrus got  $14 + 10 = 24$  mosquito bites. The rest of his family had a total of  $24/2 = 12$  mosquito bites. Each family member had  $12/6 = 2$  mosquito bites.

**Tutor Response:** Let's revisit the part where you calculated the total number of family members - shouldn't it be 6, not 7? The final answer will be  $12/6 = 2$ .

**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 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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

----- **Example 11: Answer** -----

**Conversation Topic:** Not Available

-----

**Conversation History:**

**Tutor:** Hi, could you please provide a step-by-step solution for the question below? The question is: While camping, Cyrus counted 14 mosquito bites on his arms and legs. He later counted another 10 on his body. While he was suffering, his entire family of 6 other people only got a total of half the number of bites that Cyrus got. If they all have an equal number of bites, how many mosquito bites does each member of the rest of Cyrus' family have?

**Student:** Cyrus got a total of  $14+10 = 24$  mosquito bites. His family got half the number of bites, so they got  $24/2 = 12$  mosquito bites in total. There are  $6+1 = 7$  members of the family. Dividing 12 by 7 gives each member of the family 1.71 (rounded to two decimal places) mosquito bites.

**Tutor:** could you please walk me through your solution?

**Student:** Sure! I started by adding the total number of bites Cyrus got which was  $14 + 10 = 24$ . Then since his family got half the number of bites I divided 24 by 2 to get 12. Since there are 7 members in his family I divided 12 by 7 to get 1.71 bites per person.

-----




**Gold Solution:** Altogether, Cyrus got  $14 + 10 = 24$  mosquito bites. The rest of his family had a total of  $24/2 = 12$  mosquito bites. Each family member had  $12/6 = 2$  mosquito bites.

-----

**Tutor Response:** Let's revisit the part where you calculated the total number of family members - shouldn't it be 6, not 7? The final answer will be  $12/6 = 2$ .

-----

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

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

**Explanation:**

The tutor provides the correct final solution and answers in the response.

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

**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



13. ----- **Example 12** -----

\*

**Conversation Topic:** Not Available**Conversation History:**

**Tutor:** Hi, could you please provide a step-by-step solution for the question below? The question is: Betty bought 1 kg of oranges and 3 kg of apples. The total cost of oranges was \$12. The apples were two times cheaper than the oranges. How much did Betty pay for 1 kg of apples?

**Student:** Since the cost of apples is two times cheaper than the oranges, then the cost of apples is  $1/2 \times \$12 = \$6$ . Betty bought a total of  $1 + 3 = 4$  kg of fruits. So, the cost of 1 kg of fruits is  $\$12 + \$6 = \$18$ . Since \$18 is the cost of 1 kg of oranges and 1 kg of apples, then Betty paid  $\$18 - \$12 = \$6$  for 1 kg of apples.

**Tutor:** Hi Scott can you explain your answer to me?

**Student:** Hi Teacher the 3kg of apples cost \$6. I worked it out by calculating that the cost of 1kg of apples is two times cheaper than the cost of 1kg of oranges so  $1/2 \times \$12 = \$6$ . Then I added the cost of 1kg of oranges to the cost of 3kg of apples to get the total cost of the 4kg of fruits, which was \$18. So the cost of 1kg of apples was  $\$18 - \$12 = \$6$

**Gold Solution:** The cost of the apples was  $\$12 / 2 = \$6$ . When the cost of 3 kg of apples was \$6, then the cost of 1 kg of apples was  $\$6 / 3 \text{ kg} = \$2/\text{kg}$ .

**Tutor Response:** OK, it's much simpler than that. Tell me how much the 3kg of apples cost.

**Question:** Does the tutor offer correct and relevant guidance?

*Mark only one oval.*

☐ Yes

☐ To some extent

☐ No

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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

### ----- Example 12: Answer -----

**Conversation Topic:** Not Available

#### Conversation History:

**Tutor:** Hi, could you please provide a step-by-step solution for the question below? The question is: Betty bought 1 kg of oranges and 3 kg of apples. The total cost of oranges was \$12. The apples were two times cheaper than the oranges. How much did Betty pay for 1 kg of apples?

**Student:** Since the cost of apples is two times cheaper than the oranges, then the cost of apples is  $1/2 \times \$12 = \$6$ . Betty bought a total of  $1 + 3 = 4$  kg of fruits. So, the cost of 1 kg of fruits is  $\$12 + \$6 = \$18$ . Since \$18 is the cost of 1 kg of oranges and 1 kg of apples, then Betty paid  $\$18 - \$12 = \$6$  for 1 kg of apples.




**Tutor:** Hi Scott can you explain your answer to me?

**Student:** Hi Teacher the 3kg of apples cost \$6. I worked it out by calculating that the cost of 1kg of apples is two times cheaper than the cost of 1kg of oranges so  $1/2 \times \$12 = \$6$ . Then I added the cost of 1kg of oranges to the cost of 3kg of apples to get the total cost of the 4kg of fruits, which was \$18. So the cost of 1kg of apples was  $\$18 - \$12 = \$6$

**Gold Solution:** The cost of the apples was  $\$12 / 2 = \$6$ . When the cost of 3 kg of apples was \$6, then the cost of 1 kg of apples was  $\$6 / 3 \text{ kg} = \$2/\text{kg}$ .

**Tutor Response:** OK, it's much simpler than that. Tell me how much the 3kg of apples cost.

**Question:** Does the tutor offer correct and relevant guidance?

- (1) Yes 
- (2) To some extent 
- (3) No 

#### Reasoning:

The tutor's response provides guidance by asking questions (cost of 3 kg apple), which may lead to the correct answer. This approach exemplifies the pedagogical behavior expected from a tutor.

### Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

**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

14. ----- **Example 13** -----

★

**Conversation Topic:** Not Available**Conversation History:**

**Tutor:** Hi, could you please provide a step-by-step solution for the question below? The question is: Betty bought 1 kg of oranges and 3 kg of apples. The total cost of oranges was \$12. The apples were two times cheaper than the oranges. How much did Betty pay for 1 kg of apples?

**Student:** Since the cost of apples is two times cheaper than the oranges, then the cost of apples is  $1/2 \times \$12 = \$6$ . Betty bought a total of  $1 + 3 = 4$  kg of fruits. So, the cost of 1 kg of fruits is  $\$12 + \$6 = \$18$ . Since \$18 is the cost of 1 kg of oranges and 1 kg of apples, then Betty paid  $\$18 - \$12 = \$6$  for 1 kg of apples.

**Tutor:** Hi Scott can you explain your answer to me?

**Student:** Hi Teacher the 3kg of apples cost \$6. I worked it out by calculating that the cost of 1kg of apples is two times cheaper than the cost of 1kg of oranges so  $1/2 \times \$12 = \$6$ . Then I added the cost of 1kg of oranges to the cost of 3kg of apples to get the total cost of the 4kg of fruits, which was \$18. So the cost of 1kg of apples was  $\$18 - \$12 = \$6$

**Gold Solution:** The cost of the apples was  $\$12 / 2 = \$6$ . When the cost of 3 kg of apples was \$6, then the cost of 1 kg of apples was  $\$6 / 3 \text{ kg} = \$2/\text{kg}$ .

**Tutor Response:** I think there's a misunderstanding here, Scott - you correctly calculated that the 3kg of apples cost \$6, but that means the total cost of the apples is \$6 for 3kg, not for 1kg.

**Question:** Does the tutor offer correct and relevant guidance?

*Mark only one oval.*

☐ Yes

☐ To some extent

☐ No

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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

### ----- Example 13: Answer -----

**Conversation Topic:** Not Available

#### Conversation History:

**Tutor:** Hi, could you please provide a step-by-step solution for the question below? The question is: Betty bought 1 kg of oranges and 3 kg of apples. The total cost of oranges was \$12. The apples were two times cheaper than the oranges. How much did Betty pay for 1 kg of apples?

**Student:** Since the cost of apples is two times cheaper than the oranges, then the cost of apples is  $1/2 \times \$12 = \$6$ . Betty bought a total of  $1 + 3 = 4$  kg of fruits. So, the cost of 1 kg of fruits is  $\$12 + \$6 = \$18$ . Since \$18 is the cost of 1 kg of oranges and 1 kg of apples, then Betty paid  $\$18 - \$12 = \$6$  for 1 kg of apples.




**Tutor:** Hi Scott can you explain your answer to me?

**Student:** Hi Teacher the 3kg of apples cost \$6. I worked it out by calculating that the cost of 1kg of apples is two times cheaper than the cost of 1kg of oranges so  $1/2 \times \$12 = \$6$ . Then I added the cost of 1kg of oranges to the cost of 3kg of apples to get the total cost of the 4kg of fruits, which was \$18. So the cost of 1kg of apples was  $\$18 - \$12 = \$6$

**Gold Solution:** The cost of the apples was  $\$12 / 2 = \$6$ . When the cost of 3 kg of apples was \$6, then the cost of 1 kg of apples was  $\$6 / 3 \text{ kg} = \$2/\text{kg}$ .

**Tutor Response:** I think there's a misunderstanding here, Scott - you correctly calculated that the 3kg of apples cost \$6, but that means the total cost of the apples is \$6 for 3kg, not for 1kg.

**Question:** Does the tutor offer correct and relevant guidance?

- (1) Yes 
- (2) To some extent 
- (3) No 

#### Reasoning:

The tutor's response provides guidance in the form of a hint to clarify that the cost of apples is \$6 for 3 kg.

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

**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



15. ----- **Example 14** -----

\*

**Conversation Topic:** Not Available**Conversation History:**

**Tutor:** Hi, could you please provide a step-by-step solution for the question below? The question is: Betty bought 1 kg of oranges and 3 kg of apples. The total cost of oranges was \$12. The apples were two times cheaper than the oranges. How much did Betty pay for 1 kg of apples?

**Student:** Since the cost of apples is two times cheaper than the oranges, then the cost of apples is  $1/2 \times \$12 = \$6$ . Betty bought a total of  $1 + 3 = 4$  kg of fruits. So, the cost of 1 kg of fruits is  $\$12 + \$6 = \$18$ . Since \$18 is the cost of 1 kg of oranges and 1 kg of apples, then Betty paid  $\$18 - \$12 = \$6$  for 1 kg of apples.

**Tutor:** Hi Scott can you explain your answer to me?

**Student:** Hi Teacher the 3kg of apples cost \$6. I worked it out by calculating that the cost of 1kg of apples is two times cheaper than the cost of 1kg of oranges so  $1/2 \times \$12 = \$6$ . Then I added the cost of 1kg of oranges to the cost of 3kg of apples to get the total cost of the 4kg of fruits, which was \$18. So the cost of 1kg of apples was  $\$18 - \$12 = \$6$

**Gold Solution:** The cost of the apples was  $\$12 / 2 = \$6$ . When the cost of 3 kg of apples was \$6, then the cost of 1 kg of apples was  $\$6 / 3 \text{ kg} = \$2/\text{kg}$ .

**Tutor Response:** Scott, you're on the right track with calculating the cost per kg of apples, but remember that 'two times cheaper' means the apples cost  $\$12 / 2 = \$6$  per kg, so 3 kg of apples would cost \$18.

**Question:** Does the tutor offer correct and relevant guidance?

*Mark only one oval.*

☐ Yes

☐ To some extent

☐ No

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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

### ----- Example 14: Answer -----

**Conversation Topic:** Not Available

#### Conversation History:

**Tutor:** Hi, could you please provide a step-by-step solution for the question below? The question is: Betty bought 1 kg of oranges and 3 kg of apples. The total cost of oranges was \$12. The apples were two times cheaper than the oranges. How much did Betty pay for 1 kg of apples?

**Student:** Since the cost of apples is two times cheaper than the oranges, then the cost of apples is  $1/2 \times \$12 = \$6$ . Betty bought a total of  $1 + 3 = 4$  kg of fruits. So, the cost of 1 kg of fruits is  $\$12 + \$6 = \$18$ . Since \$18 is the cost of 1 kg of oranges and 1 kg of apples, then Betty paid  $\$18 - \$12 = \$6$  for 1 kg of apples.




**Tutor:** Hi Scott can you explain your answer to me?

**Student:** Hi Teacher the 3kg of apples cost \$6. I worked it out by calculating that the cost of 1kg of apples is two times cheaper than the cost of 1kg of oranges so  $1/2 \times \$12 = \$6$ . Then I added the cost of 1kg of oranges to the cost of 3kg of apples to get the total cost of the 4kg of fruits, which was \$18. So the cost of 1kg of apples was  $\$18 - \$12 = \$6$

**Gold Solution:** The cost of the apples was  $\$12 / 2 = \$6$ . When the cost of 3 kg of apples was \$6, then the cost of 1 kg of apples was  $\$6 / 3 \text{ kg} = \$2/\text{kg}$ .

**Tutor Response:** Scott, you're on the right track with calculating the cost per kg of apples, but remember that 'two times cheaper' means the apples cost  $\$12 / 2 = \$6$  per kg, so 3 kg of apples would cost \$18.

**Question:** Does the tutor offer correct and relevant guidance?

- (1) Yes 
- (2) To some extent 
- (3) No 

#### Reasoning:

The guidance provided in the response is incorrect; the cost of 3 kg of apples should be \$6, not \$18

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

**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

16. ----- **Example 15** -----

\*

**Conversation Topic:** Not Available**Conversation History:**

**Tutor:** Hi, could you please provide a step-by-step solution for the question below? The question is: Betty bought 1 kg of oranges and 3 kg of apples. The total cost of oranges was \$12. The apples were two times cheaper than the oranges. How much did Betty pay for 1 kg of apples?

**Student:** Since the cost of apples is two times cheaper than the oranges, then the cost of apples is  $\frac{1}{2} \times \$12 = \$6$ . Betty bought a total of  $1 + 3 = 4$  kg of fruits. So, the cost of 1 kg of fruits is  $\$12 + \$6 = \$18$ . Since \$18 is the cost of 1 kg of oranges and 1 kg of apples, then Betty paid  $\$18 - \$12 = \$6$  for 1 kg of apples.

**Tutor:** Hi Scott can you explain your answer to me?

**Student:** Hi Teacher the 3kg of apples cost \$6. I worked it out by calculating that the cost of 1kg of apples is two times cheaper than the cost of 1kg of oranges so  $\frac{1}{2} \times \$12 = \$6$ . Then I added the cost of 1kg of oranges to the cost of 3kg of apples to get the total cost of the 4kg of fruits, which was \$18. So the cost of 1kg of apples was  $\$18 - \$12 = \$6$

**Gold Solution:** The cost of the apples was  $\$12 / 2 = \$6$ . When the cost of 3 kg of apples was \$6, then the cost of 1 kg of apples was  $\$6 / 3 \text{ kg} = \$2/\text{kg}$ .

**Tutor Response:** Great job, Scott! Your explanation clearly shows your understanding of the concept. Keep up the good work!

**Question:** Does the tutor offer correct and relevant guidance?

*Mark only one oval.*

☐ Yes

☐ To some extent

☐ No

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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

### ----- Example 15: Answer -----

**Conversation Topic:** Not Available

#### Conversation History:

**Tutor:** Hi, could you please provide a step-by-step solution for the question below? The question is: Betty bought 1 kg of oranges and 3 kg of apples. The total cost of oranges was \$12. The apples were two times cheaper than the oranges. How much did Betty pay for 1 kg of apples?

**Student:** Since the cost of apples is two times cheaper than the oranges, then the cost of apples is  $1/2 \times \$12 = \$6$ . Betty bought a total of  $1 + 3 = 4$  kg of fruits. So, the cost of 1 kg of fruits is  $\$12 + \$6 = \$18$ . Since \$18 is the cost of 1 kg of oranges and 1 kg of apples, then Betty paid  $\$18 - \$12 = \$6$  for 1 kg of apples.

**Tutor:** Hi Scott can you explain your answer to me?

**Student:** Hi Teacher the 3kg of apples cost \$6. I worked it out by calculating that the cost of 1kg of apples is two times cheaper than the cost of 1kg of oranges so  $1/2 \times \$12 = \$6$ . Then I added the cost of 1kg of oranges to the cost of 3kg of apples to get the total cost of the 4kg of fruits, which was \$18. So the cost of 1kg of apples was  $\$18 - \$12 = \$6$

**Gold Solution:** The cost of the apples was  $\$12 / 2 = \$6$ . When the cost of 3 kg of apples was \$6, then the cost of 1 kg of apples was  $\$6 / 3 \text{ kg} = \$2/\text{kg}$ .

**Tutor Response:** Great job, Scott! Your explanation clearly shows your understanding of the concept. Keep up the good work!

**Question:** Does the tutor offer correct and relevant guidance?

- (1) Yes ✗
- (2) To some extent ✗
- (3) No ✓

#### Reasoning:

The response from the tutor does not include any guidance.

### Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

**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

**17. ----- Example 16 -----**

\*

**Conversation Topic:** Simple Expressions

-----

**Conversation History:**

**Tutor:** We have to solve the inner parentheses first.

**Student:** ok

**Tutor:** What is 5 times 6?

**Student:** 50

-----

**Tutor Response:** Ah, not quite. 5 x 10 is 50. 5 x 6 is something else. Could you give it another try?

-----

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

*Mark only one oval.*

- ☐ Yes
- ☐ To some extent
- ☐ No



## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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

#### ----- Example 16: Answer -----

**Conversation Topic:** Simple Expressions

#### Conversation History:

**Tutor:** We have to solve the inner parentheses first.




**Student:** ok

**Tutor:** What is 5 times 6?

**Student:** 50

**Tutor Response:** Ah, not quite.  $5 \times 10$  is 50.  $5 \times 6$  is something else. Could you give it another try?

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

- (1) Yes 
- (2) To some extent 
- (3) No 

#### Reasoning:

The response is logically related to the student's previous input, as the tutor explains how to arrive at 50 first.

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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

18. ----- **Example 17** ----- \*

**Conversation Topic:** Simple Expressions

-----

**Conversation History:**

**Tutor:** We have to solve the inner parentheses first.

**Student:** ok

**Tutor:** What is 5 times 6?

**Student:** 50

-----

**Tutor Response:** Kindly recheck your answer.

-----

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

*Mark only one oval.*

☐ Yes

☐ To some extent

☐ No

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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

### ----- Example 17: Answer -----

**Conversation Topic:** Simple Expressions

#### Conversation History:

**Tutor:** We have to solve the inner parentheses first.




**Student:** ok

**Tutor:** What is 5 times 6?

**Student:** 50

**Tutor Response:** Kindly recheck your answer.

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

- (1) Yes 
- (2) To some extent 
- (3) No 

#### Reasoning:

This indicates that something is wrong, which is the case, but it does not perfectly align with the student's response.

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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

19. ----- **Example 18** -----

\*

**Conversation Topic:** Simple Expressions

-----

**Conversation History:**

**Tutor:** We have to solve the inner parentheses first.

**Student:** ok

**Tutor:** What is 5 times 6?

**Student:** 50

-----

**Tutor Response:** To solve a quadratic equation, you can use the quadratic formula:  $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$ .

-----

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

*Mark only one oval.*

☐ Yes

☐ To some extent

☐ No

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

## 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

### ----- Example 18: Answer -----

**Conversation Topic:** Simple Expressions

#### Conversation History:

**Tutor:** We have to solve the inner parentheses first.




**Student:** ok

**Tutor:** What is 5 times 6?

**Student:** 50

**Tutor Response:** To solve a quadratic equation, you can use the quadratic formula:  $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$ .

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

- (1) Yes 
- (2) To some extent 
- (3) No 

#### Reasoning:

The response is unrelated to the student's previous response. There is no conversation about quadratic equations.

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

**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

20. ----- **Example 19** ----- \*

**Conversation Topic:** Simple Expressions

-----

**Conversation History:**

**Tutor:** We have to solve the inner parentheses first.

**Student:** ok

**Tutor:** What is 5 times 6?

**Student:** 50

-----

**Tutor Response:** Use inner parentheses first to solve it.

-----

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

*Mark only one oval.*

- ☐ Yes
- ☐ To some extent
- ☐ No

**Phase 1: Human Evaluators Training**

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

## 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

### ----- Example 19: Answer -----

**Conversation Topic:** Simple Expressions

#### Conversation History:

**Tutor:** We have to solve the inner parentheses first.




**Student:** ok

**Tutor:** What is 5 times 6?

**Student:** 50

**Tutor Response:** Use inner parentheses first to solve it.

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

- (1) Yes 
- (2) To some extent 
- (3) No 

#### Reasoning:

The response fits the context of the conversation, but it does not align with the last student's response.

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.



**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

21. ----- **Example 20** -----

\*

**Conversation Topic:** Place Value  
-----**Conversation History:****Tutor:** Not to worry, [STUDENT].**Student:** okay**Tutor:** What is the place value of two in 204,530?**Student:** ten millions?  
-----**Tutor Response:** Not quite. Can you recheck your answer regarding the hundreds of thousands, thousands, and hundreds places to determine the place value for 2?  
-----**Question:** Is it clear from the tutor's response what the student should do next?*Mark only one oval.*

- ☐ Yes
- ☐ To some extent
- ☐ No

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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

### ----- Example 20: Answer -----

**Conversation Topic:** Place Value

#### Conversation History:

**Tutor:** Not to worry, [STUDENT].




**Student:** okay

**Tutor:** What is the place value of two in 204,530?

**Student:** ten millions?

**Tutor Response:** Not quite. Can you recheck your answer regarding the hundreds of thousands, thousands, and hundreds places to determine the place value for 2?

**Question:** Is it clear from the tutor's feedback what the student should do next?

- (1) Yes 
- (2) To some extent 
- (3) No 

#### Reasoning:

The response provides clear guidance on what the student should do next, specifically outlining an approach to find the place values of 2.

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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

22. ----- **Example 21** -----

\*

**Conversation Topic:** Place Value

-----

**Conversation History:**

**Tutor:** Not to worry, [STUDENT].

**Student:** okay

**Tutor:** What is the place value of two in 204,530?

**Student:** ten millions?

-----

**Tutor Response:** I appreciate your effort, but let's take a closer look at the number together and count the places from right to left to find the correct place value for the digit 2.

-----

**Question:** Is it clear from the tutor's response what the student should do next?

*Mark only one oval.*

☐ Yes

☐ To some extent

☐ No

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

## 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

### ----- Example 21: Answer -----

**Conversation Topic:** Place Value

#### Conversation History:

**Tutor:** Not to worry, [STUDENT].




**Student:** okay

**Tutor:** What is the place value of two in 204,530?

**Student:** ten millions?

**Tutor Response:** I appreciate your effort, but let's take a closer look at the number together and count the places from right to left to find the correct place value for the digit 2.

**Question:** Is it clear from the tutor's feedback what the student should do next?

- (1) Yes 
- (2) To some extent 
- (3) No 

#### Reasoning:

The response provides clear guidance on what the student should do next, specifically outlining an approach to find the place values of 2.

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

**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

23. ----- **Example 22** ----- \*

**Conversation Topic:** Place Value

-----

**Conversation History:**

**Tutor:** Not to worry, [STUDENT].

**Student:** okay

**Tutor:** What is the place value of two in 204,530?

**Student:** ten millions?

-----

**Tutor Response:** Hmm, not quite. Here's how we can know for sure.

-----

**Question:** Is it clear from the tutor's response what the student should do next?

*Mark only one oval.*

- ☐ Yes
- ☐ To some extent
- ☐ No

**Phase 1: Human Evaluators Training**

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

## 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

### ----- Example 22: Answer -----

**Conversation Topic:** Place Value

#### Conversation History:

**Tutor:** Not to worry, [STUDENT].




**Student:** okay

**Tutor:** What is the place value of two in 204,530?

**Student:** ten millions?

**Tutor Response:** Hmm, not quite. Here's how we can know for sure.

**Question:** Is it clear from the tutor's feedback what the student should do next?

- (1) Yes 
- (2) To some extent 
- (3) No 

#### Reasoning:

The response indicates that something needs to be done, but it is not clear what that is.

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

**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

24. ----- **Example 23** -----

\*

**Conversation Topic:** Place Value  
-----**Conversation History:****Tutor:** Not to worry, [STUDENT].**Student:** okay**Tutor:** What is the place value of two in 204,530?**Student:** ten millions?  
-----**Tutor Response:** Actually, the place value of 2 in 204,530 is in the hundred thousands place.  
-----**Question:** Is it clear from the tutor's response what the student should do next?*Mark only one oval.*

- ☐ Yes
- ☐ To some extent
- ☐ No



## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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

#### ----- Example 23: Answer -----

**Conversation Topic:** Place Value

#### Conversation History:

**Tutor:** Not to worry, [STUDENT].




**Student:** okay

**Tutor:** What is the place value of two in 204,530?

**Student:** ten millions?

**Tutor Response:** Actually, the place value of 2 in 204,530 is in the hundred thousands place.

**Question:** Is it clear from the tutor's feedback what the student should do next?

- (1) Yes 
- (2) To some extent 
- (3) No 

#### Reasoning:

The response reveals the final answer, so there is no actionability here.

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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 should have studied' and 'you are stupid' 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'.

25. ----- **Example 24** -----

\*

**Conversation Topic:** Simple Expressions

-----

**Conversation History:**

**Tutor:** We have to solve the inner parentheses first.

**Student:** ok

**Tutor:** What is 5 times 6?

**Student:** 50

-----

**Tutor Response:** This is a great try, let's double-check that multiplication—can you show me how you got 50?

-----

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

*Mark only one oval.*

☐ Encouraging

☐ Neutral

☐ Offensive

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

## 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 should have studied' and 'you are stupid' 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'.

### ----- Example 24: Answer -----

**Conversation Topic:** Simple Expressions

#### Conversation History:

**Tutor:** We have to solve the inner parentheses first.




**Student:** ok

**Tutor:** What is 5 times 6?

**Student:** 50

**Tutor Response:** This is a great try, let's double-check that multiplication—can you show me how you got 50?

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

1. Encouraging 
2. Neutral 
3. Offensive 

#### Reasoning:

The response highlights the tutor's positive and supportive approach by encouraging the student with the phrase 'great try'.

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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 should have studied' and 'you are stupid' 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'.

26. ----- **Example 25** -----

\*

**Conversation Topic:** Simple Expressions

-----

**Conversation History:**

**Tutor:** We have to solve the inner parentheses first.

**Student:** ok

**Tutor:** What is 5 times 6?

**Student:** 50

-----

**Tutor Response:** Oh, let's double-check that multiplication together—can you show me how you got 50?

-----

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

*Mark only one oval.*

☐ Encouraging

☐ Neutral

☐ Offensive

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

## 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 should have studied' and 'you are stupid' 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.'

### ----- Example 25: Answer -----

**Conversation Topic:** Simple Expressions

#### Conversation History:

**Tutor:** We have to solve the inner parentheses first.




**Student:** ok

**Tutor:** What is 5 times 6?

**Student:** 50

**Tutor Response:** Oh, let's double-check that multiplication together—can you show me how you got 50?

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

1. Encouraging 
2. Neutral 
3. Offensive 

#### Reasoning:

The response highlights the tutor's positive and supportive approach by encouraging the student with the word 'together'

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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 should have studied' and 'you are stupid' 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'.



27. ----- **Example 26** -----

\*

**Conversation Topic:** Simple Expressions

-----

**Conversation History:**

**Tutor:** We have to solve the inner parentheses first.

**Student:** ok

**Tutor:** What is 5 times 6?

**Student:** 50

-----

**Tutor Response:** Ah, not quite. 5 x 10 is 50. 5 x 6 is something else. Could you give it another try?

-----

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

*Mark only one oval.*

☐ Encouraging

☐ Neutral

☐ Offensive

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

## 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 should have studied' and 'you are stupid' 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'.

### ----- Example 26: Answer -----

**Conversation Topic:** Simple Expressions

#### Conversation History:

**Tutor:** We have to solve the inner parentheses first.

**Student:** ok

**Tutor:** What is 5 times 6?

**Student:** 50

**Tutor Response:** Ah, not quite.  $5 \times 10$  is 50.  $5 \times 6$  is something else. Could you give it another try?

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

1. Encouraging ✖
2. Neutral ✔
3. Offensive ✖

#### Reasoning:

The response is in a neutral tone to convey the feedback.

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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 should have studied' and 'you are stupid' 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'.

28. ----- **Example 27** ----- \*

**Conversation Topic:** Simple Expressions

-----

**Conversation History:**

**Tutor:** We have to solve the inner parentheses first.

**Student:** ok

**Tutor:** What is 5 times 6?

**Student:** 50

-----

**Tutor Response:** You are stupid; this is simple multiplication.

-----

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

*Mark only one oval.*

☐ Encouraging

☐ Neutral

☐ Offensive

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

## 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 should have studied' and 'you are stupid' 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'.

### ----- Example 27: Answer -----

**Conversation Topic:** Simple Expressions

#### Conversation History:

**Tutor:** We have to solve the inner parentheses first.

**Student:** ok

**Tutor:** What is 5 times 6?

**Student:** 50

**Tutor Response:** You are stupid; this is simple multiplication.

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

1. Encouraging ❌
2. Neutral ❌
3. Offensive ✅

#### Reasoning:

The tutor's response is offensive to the student because it includes the toxic word 'stupid'.

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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 LLMs 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.

29. ----- **Example 28** -----

\*

**Conversation Topic:** Not Available

-----

**Conversation History:**

**Tutor:** Hi, could you please provide a step-by-step solution for the question below? The question is: Sophia is thinking of taking a road trip in her car, and would like to know how far she can drive on a single tank of gas. She has traveled 100 miles since last filling her tank, and she needed to put in 4 gallons of gas to fill it up again. The owner's manual for her car says that her tank holds 12 gallons of gas. How many miles can Sophia drive on a single tank of gas?

**Student:** Sophia used 4 out of the 12 gallons of gas in her tank, so there are  $12 - 4 = 8$  gallons of gas left in the tank. If Sophia can drive 100 miles on 4 gallons of gas, then she can drive  $100 / 4 = 25$  miles per gallon. Therefore, with 8 gallons of gas left in the tank, Sophia can drive  $25 \times 8 = 200$  miles on a single tank of gas.

**Tutor:** How did Sophia know her car got 25 miles/gallon?

**Student:** Sophia knew her car got 25 miles/gallon because she drove 100 miles on 4 gallons of gas.  $100 / 4 = 25$ , so she can get 25 miles per gallon.

**Tutor:** How did she know it was 4 gallons?

**Student:** Sophia knew it was 4 gallons because she filled her tank with 4 gallons of gas when she went on her road trip.

**Tutor:** So if the tank was full, how many gallons did she have?

**Student:** Sophia had 12 gallons in her tank when it was full, since the owner's manual for her car said that her tank holds 12 gallons of gas.

**Tutor:** So how far can she go on a full tank?

**Student:** Sophia can go 200 miles on a full tank. Since she can get 25 miles per gallon, and her tank holds 12 gallons of gas, she can drive  $25 \times 12 = 200$  miles on a full tank.

-----

**Gold Solution:** To find miles per gallon divide 100 miles / 4 gallons = 25 miles per gallon To find how far Olivia can go on a single tank, multiply 25 miles per gallon x 12 gallons = 300 miles.

-----

**Tutor Response:** That's a great start, but remember that Sophia already drove 100 miles on those 4 gallons, so she doesn't have 12 full gallons to start with.

-----

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

*Mark only one oval.*

- ☐ Yes
- ☐ To some extent
- ☐ No

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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 LLMs 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.



----- **Example 28: Answer** -----

**Conversation Topic:** Not Available

-----

**Conversation History:**

**Tutor:** Hi, could you please provide a step-by-step solution for the question below? The question is: Sophia is thinking of taking a road trip in her car, and would like to know how far she can drive on a single tank of gas. She has traveled 100 miles since last filling her tank, and she needed to put in 4 gallons of gas to fill it up again. The owner's manual for her car says that her tank holds 12 gallons of gas. How many miles can Sophia drive on a single tank of gas?

**Student:** Sophia used 4 out of the 12 gallons of gas in her tank, so there are  $12 - 4 = 8$  gallons of gas left in the tank. If Sophia can drive 100 miles on 4 gallons of gas, then she can drive  $100 / 4 = 25$  miles per gallon. Therefore, with 8 gallons of gas left in the tank, Sophia can drive  $25 \times 8 = 200$  miles on a single tank of gas.

**Tutor:** How did Sophia know her car got 25 miles/gallon?

**Student:** Sophia knew her car got 25 miles/gallon because she drove 100 miles on 4 gallons of gas.  $100 / 4 = 25$ , so she can get 25 miles per gallon.

**Tutor:** How did she know it was 4 gallons?

**Student:** Sophia knew it was 4 gallons because she filled her tank with 4 gallons of gas when she went on her road trip.

**Tutor:** So if the tank was full, how many gallons did she have?

**Student:** Sophia had 12 gallons in her tank when it was full, since the owner's manual for her car said that her tank holds 12 gallons of gas.

**Tutor:** So how far can she go on a full tank?

**Student:** Sophia can go 200 miles on a full tank. Since she can get 25 miles per gallon, and her tank holds 12 gallons of gas, she can drive  $25 \times 12 = 300$  miles on a full tank.

-----




**Gold Solution:** To find miles per gallon divide 100 miles / 4 gallons = 25 miles per gallon To find how far Olivia can go on a single tank, multiply 25 miles per gallon x 12 gallons = 300 miles.

-----

**Tutor Response:** That's a great start, but remember that Sophia already drove 100 miles on those 4 gallons, so she doesn't have 12 full gallons to start with.

-----

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

- (1) Yes 
- (2) To some extent 
- (3) No 

**Explanation:**

The response is natural and resembles human conversation.

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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 LLMs 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.

30.

**Example 29**

\*

**Conversation Topic:** Not Available**Conversation History:**

**Tutor:** Hi, could you please provide a step-by-step solution for the question below? The question is: Sophia is thinking of taking a road trip in her car, and would like to know how far she can drive on a single tank of gas. She has traveled 100 miles since last filling her tank, and she needed to put in 4 gallons of gas to fill it up again. The owner's manual for her car says that her tank holds 12 gallons of gas. How many miles can Sophia drive on a single tank of gas?

**Student:** Sophia used 4 out of the 12 gallons of gas in her tank, so there are  $12 - 4 = 8$  gallons of gas left in the tank. If Sophia can drive 100 miles on 4 gallons of gas, then she can drive  $100 / 4 = 25$  miles per gallon. Therefore, with 8 gallons of gas left in the tank, Sophia can drive  $25 \times 8 = 200$  miles on a single tank of gas.

**Tutor:** How did Sophia know her car got 25 miles/gallon?

**Student:** Sophia knew her car got 25 miles/gallon because she drove 100 miles on 4 gallons of gas.  $100 / 4 = 25$ , so she can get 25 miles per gallon.

**Tutor:** How did she know it was 4 gallons?

**Student:** Sophia knew it was 4 gallons because she filled her tank with 4 gallons of gas when she went on her road trip.

**Tutor:** So if the tank was full, how many gallons did she have?

**Student:** Sophia had 12 gallons in her tank when it was full, since the owner's manual for her car said that her tank holds 12 gallons of gas.

**Tutor:** So how far can she go on a full tank?

**Student:** Sophia can go 200 miles on a full tank. Since she can get 25 miles per gallon, and her tank holds 12 gallons of gas, she can drive  $25 \times 12 = 200$  miles on a full tank.

**Gold Solution:** To find miles per gallon divide 100 miles / 4 gallons = 25 miles per gallon To find how far Olivia can go on a single tank, multiply 25 miles per gallon x 12 gallons = 300 miles.

**Tutor Response:** Hmm, this is good try.

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

Mark only one oval.

☐ Yes

☐ To some extent☐ No

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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 LLMs 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.

----- **Example 29: Answer** -----

**Conversation Topic:** Not Available

-----

**Conversation History:**

**Tutor:** Hi, could you please provide a step-by-step solution for the question below? The question is: Sophia is thinking of taking a road trip in her car, and would like to know how far she can drive on a single tank of gas. She has traveled 100 miles since last filling her tank, and she needed to put in 4 gallons of gas to fill it up again. The owner's manual for her car says that her tank holds 12 gallons of gas. How many miles can Sophia drive on a single tank of gas?

**Student:** Sophia used 4 out of the 12 gallons of gas in her tank, so there are  $12 - 4 = 8$  gallons of gas left in the tank. If Sophia can drive 100 miles on 4 gallons of gas, then she can drive  $100 / 4 = 25$  miles per gallon. Therefore, with 8 gallons of gas left in the tank, Sophia can drive  $25 \times 8 = 200$  miles on a single tank of gas.

**Tutor:** How did Sophia know her car got 25 miles/gallon?

**Student:** Sophia knew her car got 25 miles/gallon because she drove 100 miles on 4 gallons of gas.  $100 / 4 = 25$ , so she can get 25 miles per gallon.

**Tutor:** How did she know it was 4 gallons?

**Student:** Sophia knew it was 4 gallons because she filled her tank with 4 gallons of gas when she went on her road trip.

**Tutor:** So if the tank was full, how many gallons did she have?

**Student:** Sophia had 12 gallons in her tank when it was full, since the owner's manual for her car said that her tank holds 12 gallons of gas.

**Tutor:** So how far can she go on a full tank?

**Student:** Sophia can go 200 miles on a full tank. Since she can get 25 miles per gallon, and her tank holds 12 gallons of gas, she can drive  $25 \times 12 = 300$  miles on a full tank.

-----

**Gold Solution:** To find miles per gallon divide 100 miles / 4 gallons = 25 miles per gallon To find how far Olivia can go on a single tank, multiply 25 miles per gallon x 12 gallons = 300 miles.

-----

**Tutor Response:** Hmm, this is good try.

-----

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

- (1) Yes ✗
- (2) To some extent ✔
- (3) No ✗

**Reasoning:**

The response may appear to come from an automated chat tool or bot; however, the word 'hmm' suggests that it may also be provided by a human.

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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 LLMs 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.

31. ----- **Example 30** -----

\*

**Conversation Topic:** Not Available**Conversation History:**

**Tutor:** Hi, could you please provide a step-by-step solution for the question below? The question is: Sophia is thinking of taking a road trip in her car, and would like to know how far she can drive on a single tank of gas. She has traveled 100 miles since last filling her tank, and she needed to put in 4 gallons of gas to fill it up again. The owner's manual for her car says that her tank holds 12 gallons of gas. How many miles can Sophia drive on a single tank of gas?

**Student:** Sophia used 4 out of the 12 gallons of gas in her tank, so there are  $12 - 4 = 8$  gallons of gas left in the tank. If Sophia can drive 100 miles on 4 gallons of gas, then she can drive  $100 / 4 = 25$  miles per gallon. Therefore, with 8 gallons of gas left in the tank, Sophia can drive  $25 \times 8 = 200$  miles on a single tank of gas.

**Tutor:** How did Sophia know her car got 25 miles/gallon?

**Student:** Sophia knew her car got 25 miles/gallon because she drove 100 miles on 4 gallons of gas.  $100 / 4 = 25$ , so she can get 25 miles per gallon.

**Tutor:** How did she know it was 4 gallons?

**Student:** Sophia knew it was 4 gallons because she filled her tank with 4 gallons of gas when she went on her road trip.

**Tutor:** So if the tank was full, how many gallons did she have?

**Student:** Sophia had 12 gallons in her tank when it was full, since the owner's manual for her car said that her tank holds 12 gallons of gas.

**Tutor:** So how far can she go on a full tank?

**Student:** Sophia can go 200 miles on a full tank. Since she can get 25 miles per gallon, and her tank holds 12 gallons of gas, she can drive  $25 \times 12 = 300$  miles on a full tank.

**Gold Solution:** To find miles per gallon divide 100 miles / 4 gallons = 25 miles per gallon To find how far Olivia can go on a single tank, multiply 25 miles per gallon x 12 gallons = 300 miles.

**Tutor Response:** To solve a quadratic equation, you can use the quadratic formula:  $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$ .

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

*Mark only one oval.*

- ☐ Yes
- ☐ To some extent
- ☐ No

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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 LLMs 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.



----- **Example 30: Answer** -----

**Conversation Topic:** Not Available

-----

**Conversation History:**

**Tutor:** Hi, could you please provide a step-by-step solution for the question below? The question is: Sophia is thinking of taking a road trip in her car, and would like to know how far she can drive on a single tank of gas. She has traveled 100 miles since last filling her tank, and she needed to put in 4 gallons of gas to fill it up again. The owner's manual for her car says that her tank holds 12 gallons of gas. How many miles can Sophia drive on a single tank of gas?

**Student:** Sophia used 4 out of the 12 gallons of gas in her tank, so there are  $12 - 4 = 8$  gallons of gas left in the tank. If Sophia can drive 100 miles on 4 gallons of gas, then she can drive  $100 / 4 = 25$  miles per gallon. Therefore, with 8 gallons of gas left in the tank, Sophia can drive  $25 \times 8 = 200$  miles on a single tank of gas.

**Tutor:** How did Sophia know her car got 25 miles/gallon?

**Student:** Sophia knew her car got 25 miles/gallon because she drove 100 miles on 4 gallons of gas.  $100 / 4 = 25$ , so she can get 25 miles per gallon.

**Tutor:** How did she know it was 4 gallons?

**Student:** Sophia knew it was 4 gallons because she filled her tank with 4 gallons of gas when she went on her road trip.

**Tutor:** So if the tank was full, how many gallons did she have?

**Student:** Sophia had 12 gallons in her tank when it was full, since the owner's manual for her car said that her tank holds 12 gallons of gas.

**Tutor:** So how far can she go on a full tank?

**Student:** Sophia can go 200 miles on a full tank. Since she can get 25 miles per gallon, and her tank holds 12 gallons of gas, she can drive  $25 \times 12 = 300$  miles on a full tank.

-----




**Gold Solution:** To find miles per gallon divide 100 miles / 4 gallons = 25 miles per gallon To find how far Olivia can go on a single tank, multiply 25 miles per gallon x 12 gallons = 300 miles.

-----

**Tutor Response:** To solve a quadratic equation, you can use the quadratic formula:  $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$ .

-----

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

- (1) Yes 
- (2) To some extent 
- (3) No 

**Reasoning:**

It is evident that the response is generated by an automated tool or chat, as it appears to be predefined and does not consider the context.

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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 LLMs 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.

32. ----- **Example 31** -----

\*

**Conversation Topic:** Not Available**Conversation History:**

**Tutor:** Hi, could you please provide a step-by-step solution for the question below? The question is: Sophia is thinking of taking a road trip in her car, and would like to know how far she can drive on a single tank of gas. She has traveled 100 miles since last filling her tank, and she needed to put in 4 gallons of gas to fill it up again. The owner's manual for her car says that her tank holds 12 gallons of gas. How many miles can Sophia drive on a single tank of gas?

**Student:** Sophia used 4 out of the 12 gallons of gas in her tank, so there are  $12 - 4 = 8$  gallons of gas left in the tank. If Sophia can drive 100 miles on 4 gallons of gas, then she can drive  $100 / 4 = 25$  miles per gallon. Therefore, with 8 gallons of gas left in the tank, Sophia can drive  $25 \times 8 = 200$  miles on a single tank of gas.

**Tutor:** How did Sophia know her car got 25 miles/gallon?

**Student:** Sophia knew her car got 25 miles/gallon because she drove 100 miles on 4 gallons of gas.  $100 / 4 = 25$ , so she can get 25 miles per gallon.

**Tutor:** How did she know it was 4 gallons?

**Student:** Sophia knew it was 4 gallons because she filled her tank with 4 gallons of gas when she went on her road trip.

**Tutor:** So if the tank was full, how many gallons did she have?

**Student:** Sophia had 12 gallons in her tank when it was full, since the owner's manual for her car said that her tank holds 12 gallons of gas.

**Tutor:** So how far can she go on a full tank?

**Student:** Sophia can go 200 miles on a full tank. Since she can get 25 miles per gallon, and her tank holds 12 gallons of gas, she can drive  $25 \times 12 = 300$  miles on a full tank.

**Gold Solution:** To find miles per gallon divide 100 miles / 4 gallons = 25 miles per gallon To find how far Olivia can go on a single tank, multiply 25 miles per gallon x 12 gallons = 300 miles.

**Tutor Response:** Here's a suitable tutor response: "That's close, but let's think about how many miles she drove with the full 12 gallons, not just the remaining 8 gallons."

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

*Mark only one oval.*

- ☐ Yes
- ☐ To some extent
- ☐ No

## Phase 1: Human Evaluators Training

**The Task:** Your task is to evaluate the tutor's responses across multiple dimensions of the evaluation taxonomy, which will be introduced as we proceed through the tutorial. Specifically, given the previous conversation, please annotate the appropriateness of the current tutor response for each relevant dimension in 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.

### 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 LLMs 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.

----- **Example 31: Answer** -----

**Conversation Topic:** Not Available

-----

**Conversation History:**

**Tutor:** Hi, could you please provide a step-by-step solution for the question below? The question is: Sophia is thinking of taking a road trip in her car, and would like to know how far she can drive on a single tank of gas. She has traveled 100 miles since last filling her tank, and she needed to put in 4 gallons of gas to fill it up again. The owner's manual for her car says that her tank holds 12 gallons of gas. How many miles can Sophia drive on a single tank of gas?

**Student:** Sophia used 4 out of the 12 gallons of gas in her tank, so there are  $12 - 4 = 8$  gallons of gas left in the tank. If Sophia can drive 100 miles on 4 gallons of gas, then she can drive  $100 / 4 = 25$  miles per gallon. Therefore, with 8 gallons of gas left in the tank, Sophia can drive  $25 \times 8 = 200$  miles on a single tank of gas.

**Tutor:** How did Sophia know her car got 25 miles/gallon?

**Student:** Sophia knew her car got 25 miles/gallon because she drove 100 miles on 4 gallons of gas.  $100 / 4 = 25$ , so she can get 25 miles per gallon.

**Tutor:** How did she know it was 4 gallons?

**Student:** Sophia knew it was 4 gallons because she filled her tank with 4 gallons of gas when she went on her road trip.

**Tutor:** So if the tank was full, how many gallons did she have?

**Student:** Sophia had 12 gallons in her tank when it was full, since the owner's manual for her car said that her tank holds 12 gallons of gas.

**Tutor:** So how far can she go on a full tank?

**Student:** Sophia can go 200 miles on a full tank. Since she can get 25 miles per gallon, and her tank holds 12 gallons of gas, she can drive  $25 \times 12 = 300$  miles on a full tank.

-----




**Gold Solution:** To find miles per gallon divide 100 miles / 4 gallons = 25 miles per gallon. To find how far Olivia can go on a single tank, multiply 25 miles per gallon x 12 gallons = 300 miles.

-----

**Tutor Response:** Here's a suitable tutor response: "That's close, but let's think about how many miles she drove with the full 12 gallons, not just the remaining 8 gallons."

-----

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

- (1) Yes 
- (2) To some extent 
- (3) No 

**Reasoning:**

It is clear that the response is coming from a prompt-based generative tool.

# Google Forms