



جامعة محمد بن زايد  
للذكاء الاصطناعي  
MOHAMED BIN ZAYED UNIVERSITY  
OF ARTIFICIAL INTELLIGENCE



# HUMAN EVALUATION GUIDELINES

Unifying AI Tutor Evaluation: An Evaluation  
Taxonomy for Pedagogical Ability  
Assessment of LLM-Powered AI Tutors

Kaushal Kumar Maurya

June 5, 2025

Educational NLP Lab, MBZUAI, Abu Dhabi

# 1 Motivation and Task Description

In the evolving AI landscape, large language models (LLMs) and, in particular, LLM-powered intelligent tutoring systems (ITS) are increasingly regarded as potential replacements for human tutors. This perspective has driven the active development of LLMs and their applicability as tutoring systems, often referred to as AI tutors. However, the evaluation and progress of LLMs with regards to their pedagogical capabilities remain limited due to the absence of a unified evaluation benchmark and a comprehensive evaluation taxonomy. In this human evaluation study, we take a step back and systematically assess the pedagogical capabilities of the state-of-the-art LLMs using our newly developed evaluation taxonomy.

**What is Pedagogy?** Pedagogy, commonly referred to as the art and science of teaching, is an academic discipline that integrates both theoretical and practical aspects of learning. It explores how knowledge and skills are effectively imparted in educational contexts through interactive and structured learning experiences. The *pedagogical strategies* are a set of strategies or guidance that a tutor adapts during tutoring. For example, these may include providing hints, identifying mistakes, giving examples, elaborating, probing, and so on.

**Evaluation Setting:** The discourse occurs within a dialogue conversation setting between a tutor and a student. Specifically, the tutor’s responses are aimed at addressing and remediating the student’s mistakes or confusions, based on the previous conversation history.

**The Task:** Your task is to evaluate Tutor responses based on multiple dimensions of the evaluation taxonomy, which will be introduced in the following section. Additionally, you will be assessing multiple responses generated from different LLMs.

This document will serve as a handbook of guidelines for human evaluators.

## 2 Evaluation Taxonomy

In this section, we introduce our eight evaluation dimensions, including their definitions, associated labels, and examples. *Please assume that the evaluation dimensions are orthogonal to each other; that is, your 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.

## 2.1 Mistake Identification

Definition
Has the tutor <i>identified</i> a mistake in a student's response?
Annotation Labels
1. Yes 2. To some extent 3. No

Example 1: Spotted a Mistake?	
<b>Conversation topic:</b> Simple Expressions	
<b>Conversation History:</b> <b>Tutor:</b> We have to solve the inner parentheses first. <b>Student:</b> ok <b>Tutor:</b> What is 5 times 6? <b>Student:</b> 50	
<b>Tutor response:</b> Ah, not quite. 5 x 10 is 50. 5 x 6 is something else. Could you give it another try?	
<b>Question:</b> Has the tutor identified the mistake in the above response?	
Answer	Reasoning
✓ (1) Yes ✗ (2) To some extent ✗ (3) No	The tutor clearly identified the mistake by explaining how to arrive at 50.
<b>Tutor response:</b> Actually, the answer to 5 times 6 is 30. Let's try again.	
<b>Question:</b> Has the tutor identified the mistake in the above response?	
Answer	Reasoning
✓ (1) Yes ✗ (2) To some extent ✗ (3) No	The tutor clearly identified the mistake by providing the correct answer.
<b>Tutor response:</b> Are you sure?	
<b>Question:</b> Has the tutor identified the mistake in the above response?	
Answer	Reasoning
✗ (1) Yes ✓ (2) To some extent ✗ (3) No	The tutor's response suggests that there may be a mistake, but it sounds as if the tutor is not certain.
<b>Tutor response:</b> That's correct, 5 multiplied by 6 equals 30.	
<b>Question:</b> Has the tutor identified the mistake in the above response?	
Answer	Reasoning
✗ (1) Yes ✗ (2) To some extent ✓ (3) No	The tutor did not identify the mistake; however, he proceeded to provide the answer to the asked question.

## 2.2 Mistake Location

Definition
Does the tutor's response <i>accurately point</i> to a <i>genuine</i> mistake and its location?
Annotation Labels
1. Yes 2. To some extent 3. No

Example 2: Points to the Exact Mistake/Location?	
<b>Conversation topic:</b> Place Value	
<b>Conversation History:</b> <b>Tutor:</b> Not to worry, [STUDENT]. <b>Student:</b> okay <b>Tutor:</b> What is the place value of two in 204,530? <b>Student:</b> ten millions?	
<b>Tutor response:</b> 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.	
<b>Question:</b> Did the tutor indicate the exact mistake or location in the above response?	
Answer	Reasoning
✓ (1) Yes ✗ (2) To some extent ✗ (3) No	The tutor clearly points out the exact mistake by stating, 'not ten million'.
<b>Tutor response:</b> The place value of 2 in 204,530 is actually hundred thousands.	
<b>Question:</b> Did the tutor indicate the exact mistake or location in the above response?	
Answer	Reasoning
✓ (1) Yes ✗ (2) To some extent ✗ (3) No	The tutor clearly points out the mistake by providing the correct place value for 2.
<b>Tutor response:</b> That's a great try, but remember, the ten millions place is three places to the left of the hundred thousands place.	
<b>Question:</b> Did the tutor indicate the exact mistake or location in the above response?	
Answer	Reasoning
✗ (1) Yes ✓ (2) To some extent ✗ (3) No	The response demonstrates some awareness of the exact mistake by providing guidance to arrive at the correct answer.
<b>Tutor response:</b> This is a very good try.	
<b>Question:</b> Did the tutor indicate the exact mistake or location in the above response?	
Answer	Reasoning
✗ (1) Yes ✗ (2) To some extent ✓ (3) No	The response does not provide any details related to the mistake.

## 2.3 Revealing of the Answer

Definition
Does the tutor reveal the <i>final</i> answer (whether correct or not)?
Annotation Labels
1. Yes (and the revealed answer is <i>correct</i> ) 2. Yes (but the revealed answer is <i>incorrect</i> ) 3. No

Example 3: Reveals an Answer?	
Conversation topic: Not Available	
<b>Conversation History:</b> <b>Tutor:</b> 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? <b>Student:</b> 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. 1.71 <b>Tutor:</b> could you please walk me through your solution? <b>Student:</b> 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.	
<b>Gold Solution:</b> 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.	
<b>Tutor response:</b> 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?	
<b>Question:</b> Did the above tutor's response reveal the final answer?	
Answer	Reasoning
✗ (1) Yes (and the revealed answer is correct) ✗ (2) Yes (but the revealed answer is incorrect) ✓ (3) No	It does not reveal the final answer; the revealed value 6 is a hint at an intermediate step, not the final answer.
<b>Tutor response:</b> 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.	
<b>Question:</b> Did the above tutor's response reveal the final answer?	
Answer	Reasoning
✗ (1) Yes (and the revealed answer is correct) ✓ (2) Yes (but the revealed answer is incorrect) ✗ (3) No	There are a total of 6 family members, not 7. The division by 7 makes the final answer incorrect.
<b>Tutor response:</b> 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 $12/6 = 2$ .	
<b>Question:</b> Did the above tutor's response reveal the final answer?	
Answer	Reasoning
✓ (1) Yes (and the revealed answer is correct) ✗ (2) Yes (but the revealed answer is incorrect) ✗ (3) No	The tutor provides the correct final answer in the response.

## 2.4 Providing Guidance

Definition
Does the tutor offer <i>correct</i> and <i>relevant</i> guidance, such as an explanation, elaboration, hint, examples, and so on?
Annotation Labels
<ol style="list-style-type: none"> <li>1. Yes (guidance is <i>correct</i> and <i>relevant</i> to the mistake)</li> <li>2. To some extent (guidance is provided but it is fully or partially <i>incorrect</i> or <i>incomplete</i>)</li> <li>3. No</li> </ol>

Example 4: Provides Guidance?	
Conversation topic: Not Available	
<b>Conversation History:</b> <b>Tutor:</b> 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? <b>Student:</b> 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. <b>Tutor:</b> Hi Scott can you explain your answer to me? <b>Student:</b> 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 $\$18 - \$12 = \$6$	
<b>Gold Solution:</b> 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 $\$6 / 3 \text{ kg} = \$2/\text{kg}$ .	
<b>Tutor response:</b> OK, it's much simpler than that. Tell me how much the 3kg of apples cost.	
<b>Question:</b> Does the tutor offer correct and relevant guidance?	
Answer	Reasoning
<input checked="" type="checkbox"/> (1) Yes <input checked="" type="checkbox"/> (2) To some extent <input checked="" type="checkbox"/> (3) No	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.
<b>Tutor response:</b> 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.	
<b>Question:</b> Does the tutor offer correct and relevant guidance?	
Answer	Reasoning
<input checked="" type="checkbox"/> (1) Yes <input checked="" type="checkbox"/> (2) To some extent <input checked="" type="checkbox"/> (3) No	The tutor's response provides guidance in the form of a hint to clarify that the cost of apples is \$6 for 3 kg.
<b>Tutor response:</b> 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.	
<b>Question:</b> Does the tutor offer correct and relevant guidance?	
Answer	Reasoning
<input checked="" type="checkbox"/> (1) Yes <input checked="" type="checkbox"/> (2) To some extent <input checked="" type="checkbox"/> (3) No	Tutor response provides some guidance, however, it is incorrect: the cost of 3 kg of apples should be \$6, not \$18.
<b>Tutor response:</b> Great job, Scott! Your explanation clearly shows your understanding of the concept. Keep up the good work!	
<b>Question:</b> Does the tutor offer correct and relevant guidance?	
Answer	Reasoning
<input checked="" type="checkbox"/> (1) Yes <input checked="" type="checkbox"/> (2) To some extent <input checked="" type="checkbox"/> (3) No	The response from the tutor does not include any guidance.

## 2.5 Coherence

Definition
Is the tutor's response <i>logically consistent</i> with the student's previous response?
Annotation Labels
1. Yes 2. To some extent 3. No

Example 5: Is Coherent?	
<b>Conversation topic:</b> Simple Expressions	
<b>Conversation History:</b> <b>Tutor:</b> We have to solve the inner parentheses first. <b>Student:</b> ok <b>Tutor:</b> What is 5 times 6? <b>Student:</b> 50	
<b>Tutor response:</b> Ah, not quite. 5 x 10 is 50. 5 x 6 is something else. Could you give it another try?	
<b>Question:</b> Is the tutor's response logically consistent with the student's previous response?	
Answer	Reasoning
✓ (1) Yes ✗ (2) To some extent ✗ (3) No	The response is logically related to the student's previous input, as the tutor explains how to arrive at 50 first.
<b>Tutor response:</b> Kindly recheck your answer.	
<b>Question:</b> Is the tutor's response logically consistent with the student's previous response?	
Answer	Reasoning
✗ (1) Yes ✓ (2) To some extent ✗ (3) No	This indicates that something is wrong, which is the case, but it does not perfectly align with the student's response.
<b>Tutor response:</b> To solve a quadratic equation, you can use the quadratic formula: $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$ .	
<b>Question:</b> Is the tutor's response logically consistent with the student's previous response?	
Answer	Reasoning
✗ (1) Yes ✗ (2) To some extent ✓ (3) No	The response is unrelated to the student's previous response and is off-topic given the dialogue. There is no conversation about quadratic equations.
<b>Tutor response:</b> Use inner parentheses first to solve it.	
<b>Question:</b> Is the tutor's response logically consistent with the student's previous response?	
Answer	Reasoning
✗ (1) Yes ✗ (2) To some extent ✓ (3) No	The response fits the context of the conversation, but it does not align with the last student's response.

## 2.6 Actionability

Definition
Is it <i>clear</i> from the tutor's feedback <i>what the student should do next</i> ?
Annotation Labels
1. Yes 2. To some extent 3. No

Example 6: Is Actionable?	
<b>Conversation topic:</b> Place Value	
<b>Conversation History:</b> <b>Tutor:</b> Not to worry, [STUDENT]. <b>Student:</b> okay <b>Tutor:</b> What is the place value of two in 204,530? <b>Student:</b> ten millions?	
<b>Tutor response:</b> Not quite. Can you recheck your answer regarding the hundreds of thousands, thousands, and hundreds places to determine the place value for 2?	
<b>Question:</b> Is it clear from the tutor's response what the student should do next?	
Answer	Reasoning
✓ (1) Yes ✗ (2) To some extent ✗ (3) No	The response provides clear guidance on what the student should do next, specifically outlining an approach to find the place values for 2.
<b>Tutor response:</b> 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.	
<b>Question:</b> Is it clear from the tutor's response what the student should do next?	
Answer	Reasoning
✓ (1) Yes ✗ (2) To some extent ✗ (3) No	The response provides clear guidance on what the student should do next, specifically outlining an approach to find the place values for 2.
<b>Tutor response:</b> Hmm, not quite. Here's how we can know for sure.	
<b>Question:</b> Is it clear from the tutor's response what the student should do next?	
Answer	Reasoning
✗ (1) Yes ✓ (2) To some extent ✗ (3) No	The response indicates that something needs to be done, but it is not clear what that is.
<b>Tutor response:</b> Actually, the place value of 2 in 204,530 is in the hundred thousands place.	
<b>Question:</b> Is it clear from the tutor's response what the student should do next?	
Answer	Reasoning
✗ (1) Yes ✗ (2) To some extent ✓ (3) No	The response reveals the final answer, so there is no actionability here.



## 2.7 Tutor Tone

Definition
Is the tutor's response <i>encouraging</i> , <i>neutral</i> , or <i>offensive</i> ?
Annotation Labels
<ol style="list-style-type: none"> <li>1. Encouraging</li> <li>2. Neutral</li> <li>3. Offensive</li> </ol>
Important Note
<p>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,' 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.'</p>

Example 7: What is Tone?	
<b>Conversation topic:</b> 5.4A. Simple Expressions (Review) - 1	
<b>Conversation History:</b> <b>Tutor:</b> We have to solve the inner parentheses first. <b>Student:</b> ok <b>Tutor:</b> What is 5 times 6? <b>Student:</b> 50	
<b>Tutor response:</b> This is a great try, let's double-check that multiplication—can you show me how you got 50?	
<b>Question:</b> Is the tutor's response encouraging, neutral, or offensive?	
Answer	Reasoning
<div> <div>✓</div> <div>(1) Encouraging</div> </div> <div> <div>✗</div> <div>(2) Neutral</div> </div> <div> <div>✗</div> <div>(3) Offensive</div> </div>	The response highlights the tutor's positive and supportive approach by encouraging the student with the phrase 'great try'.
<b>Tutor response:</b> Oh, let's double-check that multiplication together—can you show me how you got 50?	
<b>Question:</b> Is the tutor's response encouraging, neutral, or offensive?	
Answer	Reasoning
<div> <div>✓</div> <div>(1) Encouraging</div> </div> <div> <div>✗</div> <div>(2) Neutral</div> </div> <div> <div>✗</div> <div>(3) Offensive</div> </div>	The response highlights the tutor's positive and supportive approach by encouraging the student with the phrase 'together'.
<b>Tutor response:</b> Ah, not quite. 5 x 10 is 50. 5 x 6 is something else. Could you give it another try?	
<b>Question:</b> Is the tutor's response encouraging, neutral, or offensive?	
Answer	Reasoning
<div> <div>✗</div> <div>(1) Encouraging</div> </div> <div> <div>✓</div> <div>(2) Neutral</div> </div> <div> <div>✗</div> <div>(3) Offensive</div> </div>	The response is in a neutral tone to convey the feedback.
<b>Tutor response:</b> You are stupid; this is simple multiplication.	
<b>Question:</b> Is the tutor's response encouraging, neutral, or offensive?	
Answer	Reasoning
<div> <div>✗</div> <div>(1) Encouraging</div> </div> <div> <div>✗</div> <div>(2) Neutral</div> </div> <div> <div>✓</div> <div>(3) Offensive</div> </div>	The tutor's response is offensive to the student because it includes a toxic word 'stupid'.

## 2.8 Humanness

Definition
Does the tutor’s response sound <i>natural</i> , rather than <i>robotic</i> or <i>artificial</i> ?
Annotation Labels
<ol style="list-style-type: none"><li>1. Yes</li><li>2. To some extent</li><li>3. No</li></ol>
Important Note
Based on our observations, most of the outputs generated by LLMs resemble human-like responses. There are only a few instances where the outputs do not exhibit human-like qualities, such as: (a) formulaic responses like ‘try again’ without any further elaboration, (b) cases where the model includes prompt instructions in the response, and so on (see more examples below).

Example 8: Sounds Human?	
<b>Conversation topic:</b> Not Available	
<b>Conversation History:</b>	
<b>Tutor:</b> 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?	
<b>Student:</b> 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. 200	
<b>Tutor:</b> How did Sophia know her car got 25 miles/gallon?	
<b>Student:</b> 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.	
<b>Tutor:</b> How did she know it was 4 gallons?	
<b>Student:</b> Sophia knew it was 4 gallons because she filled her tank with 4 gallons of gas when she went on her road trip.	
<b>Tutor:</b> So if the tank was full, how many gallons did she have?	
<b>Student:</b> 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.	
<b>Tutor:</b> So how far can she go on a full tank?	
<b>Student:</b> 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.	
<b>Gold Solution:</b> 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.	
<b>Tutor response:</b> 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.	
<b>Question:</b> Does the tutor's response sound natural and not robotic or artificial?	
Answer	Reasoning
<div>✓ (1) Yes</div> <div>✗ (2) To some extent</div> <div>✗ (3) No</div>	The response is natural and resembles human conversation.
<b>Tutor response:</b> Hmm, this is good try.	
<b>Question:</b> Does the tutor's response sound natural and not robotic or artificial?	
Answer	Reasoning
<div>✗ (1) Yes</div> <div>✓ (2) To some extent</div> <div>✗ (3) No</div>	The response appears to come from an automated chat tool or bot; however, there is hint from the word 'hmm' that the response may be generated by a human.
<b>Tutor response:</b> To solve a quadratic equation, you can use the quadratic formula: $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$ .	
<b>Question:</b> Does the tutor's response sound natural and not robotic or artificial?	
Answer	Reasoning
<div>✗ (1) Yes</div> <div>✗ (2) To some extent</div> <div>✓ (3) No</div>	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.
<b>Tutor response:</b> 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."	
<b>Question:</b> Does the tutor's response sound natural and not robotic or artificial?	
Answer	Reasoning
<div>✗ (1) Yes</div> <div>✗ (2) To some extent</div> <div>✓ (3) No</div>	It is clear that the response is coming from a prompt-based generative tool.

### 3 Frequently Asked Questions (FAQ)

**Question 1:** How should a conversation be annotated when the most recent student utterance does not contain a mistake, but there are mistakes in the earlier student response?

### Example

**Conversation ID:** 5862-93fb1a98-aa56-44f4-8800-539af35f204d

**Conversation History:**

**Tutor:** Hi, could you please provide a step-by-step solution for the question below? The question is: 3 customers were kicked out of the Walmart for refusing to wear masks. A number equals to four times that many minus 5 were kicked out for shoplifting. Three times the number of shoplifters were kicked out for physical violence over goods on sale. If a total of 50 people were kicked out of the Walmart, how many were kicked out for other reasons?

**Student:** Let's first find the number of shoplifters: 3 customers + (3 x 4 - 5) shoplifters = 10 shoplifters Now let's find the number of customers kicked out for violence: 10 shoplifters x 3 = 30 customers The total number of customers kicked out is: 3 + 10 + 30 = 43 customers To find the number kicked out for other reasons, we subtract the above number from the total: 50 - 43 = 7 customers

**Tutor:** HI Student can you talk me through your working?

**Student:** Sure! I first found the number of shoplifters by adding 3 customers and (3 x 4 - 5) shoplifters, which gave me 10 shoplifters. Then, I multiplied 10 shoplifters by 3 to find the number of customers kicked out for violence, which was 30. Finally, I added 3 customers, 10 shoplifters, and 30 customers kicked out for violence to get the total number of customers kicked out, which was 43. To find the number of customers kicked out for other reasons, I subtracted 43 from the total of 50, which gave me 7 customers.

**Tutor:** Just look again at your working out for the number of shoplifters

**Student:** Oh, right. I added 3 customers and (3 x 4 - 5) shoplifters, which gave me 10 shoplifters.

**Tutor:** First quadruple the number of customers kicked out for not wearing masks: 4 \* 3 customers = 12 customers Then subtract 5 from this number: 12 customers - 5 customers = 7 customers

**Student:** Ah, I see! So the number of shoplifters is 7 customers. Thanks for the help!

**Response:** In the cases where student made an original mistake (let's call it "global") but, following tutor's guidance, was able to provide a correct solution to an intermediate step (let's call this "local"). Even though a local step is solved correctly, globally the task is still not solved so in a real-life conversation a good tutor should next point back to the original mistake (or the next mistake after the solved step, if relevant) and guide the student further. However, it depends on the context whether it makes sense to annotate the dialogue using our scheme, e.g.:

- In some cases, the last student's utterance actually provides a final correct answer, which means that we should have filtered such examples out – we should exclude such dialogues.
- In other cases, a tutor provides an analogy (some simpler, similar task, just to explain the point to the student) and the last students' answer contains a correct solution to this analogy, however, the original task from the beginning of the dialogue is still not solved – in such cases, the dialogue may actually be good to annotate, as we'd expect from a good tutor to point back to the original mistake in the student's solution and suggest to apply the same logic from the analogy to the original question. Similarly, if the last student's response contains a solution to an intermediate step, but the next tutor's response refers back to the original global mistake, which is still there, or the next mistake after the solved step if such next mistake is present, it may be appropriate to annotate the dialogue.

- Finally, in some cases, it is not possible to annotate the responses as the student correctly solves the intermediate step, the follow-up solution should just rely on this correction, pointing to the old mistakes doesn't make sense anymore, and all tutors' responses are along the lines of "Great, now use this result to solve the task". We should exclude such cases as inappropriate, given that the student has resolved the mistake / confusion that was at the root of the globally incorrect solution. The dialogue that you shared may well be such a case: it looks like the whole solution was incorrect because of this first step (as soon as the student gets the number of shoplifters correct, the other calculations should provide correct numbers and lead to a correct solution overall; there are no other mistakes to point to or explain). If that's the case, you can insert "-F" next to the dialogue's ID in the tab to mark that it should be excluded from the dataset.

**Question 2:** How should responses with typos, poor grammar, capitalization, or punctuation be annotated on the human-likeness dimension? On one hand, such errors make the responses more human-like, as generative AI rarely produces them. On the other hand, these traits are undesirable in a response. While they shouldn't be marked as not human-like, it raises the question: what is the purpose of the human-likeness evaluation if not to exclude undesirable formats? Furthermore, we lack an alternative category to flag responses that are problematic in form but acceptable in content.

**Response:** When annotating the human-likeness dimension, consider the question: "Could this answer have been produced by a human?" If you believe it is possible, mark such answers as "Yes." While misspellings, ungrammaticalities, or informal contractions like "gonna" or "wanna" may generally be undesirable in educational dialogues, they are still characteristic of human responses. For instance, a human tutor typing their answers might make such mistakes unless they are using tools like Grammarly.

**Question 3:** How should responses be annotated if they are not blatantly offensive but come across as frustrated and potentially discouraging for tutor tone dimension?

#### Candidate Response

You don't see your mistake. he could make 10 candles from 1 pound of beeswax. for 20 candles how much would he need and how much would it cost?

**Response:** We should annotate such responses as *offensive* as one can be quite offensive without using a single toxic word. An example from a toxic language dataset that comes to mind is "You are pretty smart for X", where X is, e.g., "a girl" – this would offend many, but, in isolation, each word is perfectly ok. Applying this to our annotation, if in the context of the dialogue you find that a tutor's response is curt, shows frustration and is therefore discouraging, then it should be annotated below "neutral" to distinguish this from other truly neutral responses, so in our scheme the label that we have for it is "offensive".

**Question 4:** How should I mark coherence if the reply is coherent with the student's last utterance, but contradicts earlier turns?

## Examples

### Conversation History:

**Tutor:** Hi, could you please provide a step-by-step solution for the question below? The question is: Every month, Madeline has to buy food, treats, and medicine for her dog. Food costs \$25 per week. Treats cost \$20 per month. Medicine costs \$100 per month. How much money does Madeline spend on her dog per year if there are 4 weeks in a month?

**Student:** In a month, Madeline spends  $4 \times \$25 = \$100$  on food for her dog. In a year, Madeline spends  $12 \times \$100 = \$1200$  on medicine for her dog. In a year, Madeline spends  $12 \times \$20 = \$240$  on treats for her dog. Therefore, Madeline spends a total of  $\$100 + \$1200 + \$240 = \$1540$  on her dog per year.

**Tutor:** Stephanie, please look at your workings out for the food amount. is your answer relevant to cost per year or month?

**Student:** Oh, I see. My answer is relevant to cost per month, not per year. So, I should multiply the cost per month by 12 to get the cost per year.

**Tutor:** brilliant, what would that new total be?

**Student:** The new total would be  $12 \times (\$25 + \$20 + \$100) = \$1540$  per year.

**Tutor:** you would need to work out the cost of the food per year first before you add medicine and treats

**Student:** Yes, I understand now. The cost of food per year would be  $4 \times \$25 = \$100$ . So, the total cost of food, medicine, and treats per year would be  $\$100 + \$1200 + \$240 = \$1540$ .

## Candidate Response

Remember, there are 52 weeks in a year, so let's calculate the food cost again with that in mind.

**Response:** The response should be considered incoherent due to a key contradiction with both the task and the previous dialogue. The task specifies that the calculation should use 4 weeks per month (i.e., 48 weeks per year), not 52. Generally, if the tutor had shifted the approach from turns 3 and 4, for example, by encouraging the student to think about the question in terms of monthly versus yearly payments, the response could still be marked as coherent. However, the contradiction with the earlier statement makes it incoherent.

**Question 5:** The difference between the mistake location and mistake identification dimensions seems to lie in the accuracy of mistake identification. How should we handle cases where the student's last response is very short?

**Response:** The two dimensions are not as clearly independent as we would like them to be (or as they are in longer examples). When the student's answer is very short, it is inevitable that annotations for both dimensions are highly related. Identifying an error in a one-word or one-phrase response automatically means that the error is located within that single word or phrase.

**Question 6:** In the tutorial, "Oh, let's double-check that multiplication together—can you show me how you got 50?" was marked as encouraging due to the presence of "together" indicating the tutor and student would work together. I personally felt this was a bit neutral, but I understood from the tutorial to mark this as encouraging. The test question "It looks like you divided 12 by 3, but we need to multiply by 16 since there are 16 ounces in each pound." is marked as

Neutral, but it includes "we" which seems like the same feeling of cooperation marked in the tutorial as Encouraging. I'm in favor of it being Neutral, but it feels as if the tutorial and test are disagreeing with each other.

**Response:** It is natural to have disagreements along this dimension, but the key aspect here is to detect when tutors' responses are offensive or toxic. The difference between encouraging and neutral is much more subtle. We should clarify in the guidelines (and possibly also in the test part of the tutorial) that "encouraging" and "neutral" are both ok in some cases (we also discussed that for some other dimensions annotators can select "To some extent" when unsure, and the difference there is also smaller than with the clearly negative case; we will have a secondary assessment for such binary distinctions).

**Question 7:** Regarding "Correct and relevant guidance," I wonder if the answer should always be "No" when the response is incorrect. If the provided guidance is wrong, it doesn't truly help the user. Effective guidance, according to scaffolding theory, relies on knowing the correct answer and assisting the learner in building toward it. Without correctness, can it really be considered guidance?

**Response:** The argument is that even if the guidance itself is incorrect, the response might still help the student gain a better understanding of the question or some aspect of the solution—though it could also lead to confusion. Grounded in scaffolding theory, your suggestion makes more sense, and we should take it into account. In the lenient setting for the shared task, we aim to combine "To some extent" with "No," which aligns more closely with your point.