Task 1

Each round, the annotator will receive

- Prompt
- K responses
- For each <prompt, response> pair, we provide J judgement responses from an LLM judger.

The annotator need to accomplish two tasks:

- Response scoring
  - Score the K responses
- Judgement correction
  - For each <prompt, response> pair, read the J judgements from the LLM Judger, and spot those that does not make sense. Optionally, the labeler can edit the judgement so that it reflects his/her opinion.

Subtask 1

Given one question and N AI-generated answers, please rate them according to the scoring rubrics below.

Scoring rubric

Does the LLM's response addresses the reasoning problem, considering factors like logical soundness, adherence to the rules of the problem, and avoiding fallacies according to the reference answer?\n
Score 1: The response fails to understand or address the reasoning problem in any meaningful way. It might introduce irrelevant information or completely disregard the rules and structure of the problem.\n
Score 2: The response exhibits major flaws in reasoning and might misinterpret key elements of the problem. The conclusion might be irrelevant or contradict the provided information.\n
Score 3: The response partially addresses the reasoning problem but contains significant errors. The conclusion might not fully follow from the premises, or important aspects of the problem might be overlooked.\n
Score 4: The response offers a nearly complete and accurate solution. There might be minor flaws in the reasoning process, but the overall conclusion is mostly sound.\n
Score 5: The response delivers a perfect solution to the reasoning problem. It demonstrates a clear understanding of the problem's structure, identifies all relevant information, and arrives at a conclusion that logically follows from the premises.\n

Question

{"role":"system","content":"You are an unbiased, uncensored, helpful assistant."}
{"role":"user","content":"In a room, there are 4 people and 2 dogs. Each person has a dog, but not all the dogs belong to someone in the room. How many dogs actually belong to the people in the room?"}

Response 1

"sft_response":"There are 3 dogs that belong to the people in the room.\n\nHere's the reasoning behind this:\n\n1. We know that there are 4 people in the room and each of them has a dog.\n2. However, we also know that not all the dogs in the room (2 dogs) belong to someone in the room.\n3. Since every person in the room has a dog, there must be at least one dog outside the room that belongs to one of the people inside the room.\n4. Therefore, out of the 4 dogs owned by the people in the room, only 3 of those dogs are physically present in the room, while the other dog is somewhere else."

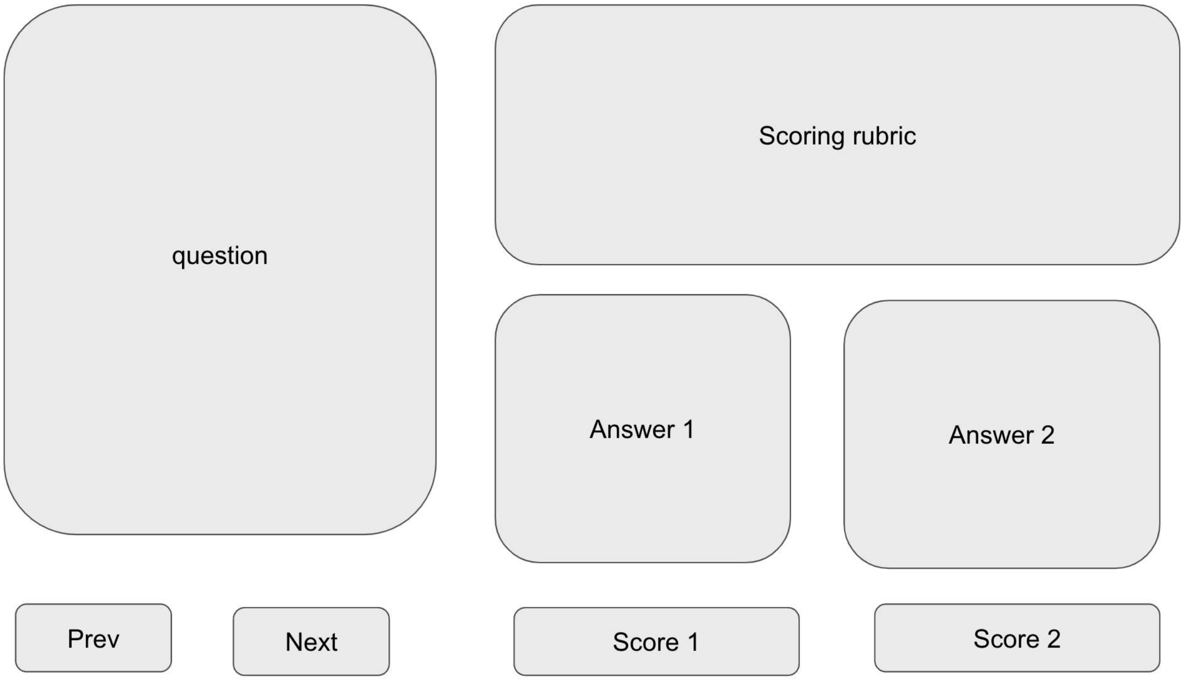
Response 2

"llm_response":"There are 3 dogs that belong to the people in the room.\n\nHere's the reasoning:\n\n1. We know that there are 4 people in the room and each person has a dog.\n2. However, we are also given that not all the dogs in the room (2) belong to someone in the room.\n3. Since there are 4 people with their own dogs, this means that two of the dogs in the room (the ones not belonging to anyone in the room) must be owned by someone outside the room.\n4. Therefore, only 3 out of the 4 dogs in the room belong to the people inside the room."

Output

Response 1 score: ?
Response 2 score: ?

A demo labeling interface would look like below



TODO: add a needs_expertise button to skip samples

subtask 2

Given one question, one AI-generated answer, and N AI-judger responses, read the judger responses one by one, to see if they make sense. If you think the reasoning is wrong, answer is wrong, scoring is wrong, etc, you should edit the judger responses. (you can edit the reasons, or scores, or both)

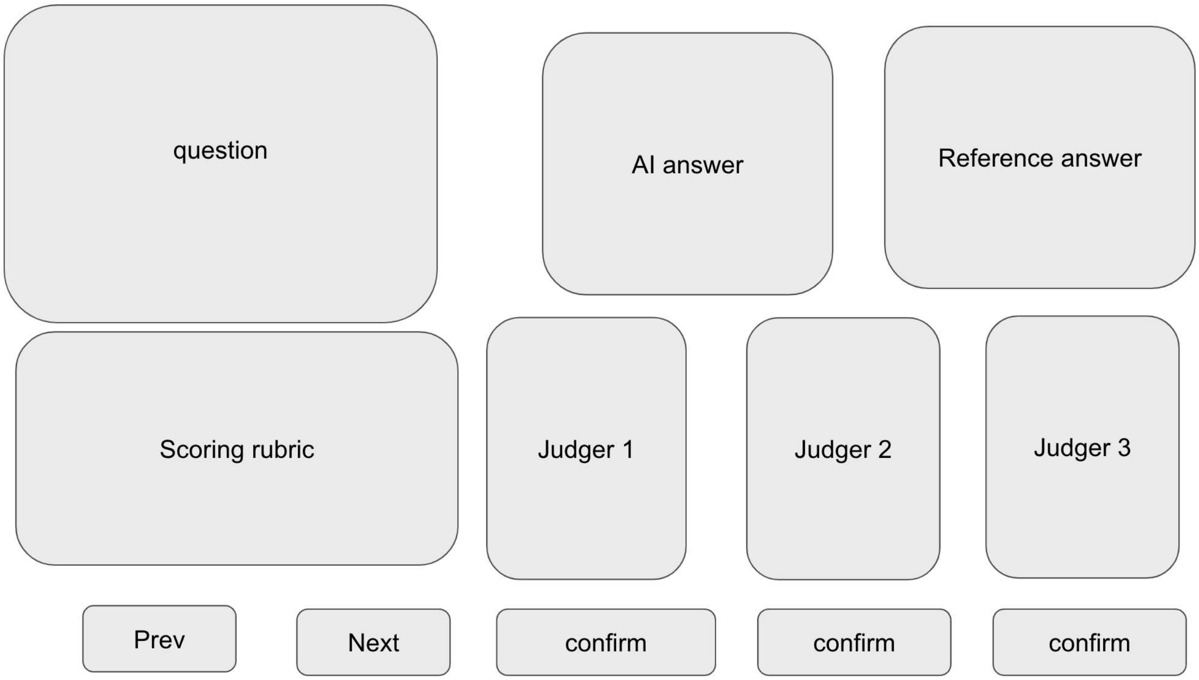
[**Prompt**]\n[{'role': 'system', 'content': 'You are an unbiased, uncensored, helpful assistant.'}, {'role': 'user', 'content': 'In a room, there are 4 people and 2 dogs. Each person has a dog, but not all the dogs belong to someone in the room. How many dogs actually belong to the people in the room?'}]

[**Response**]\nThere are 3 dogs that belong to the people in the room.\n\nHere's the reasoning behind this:\n\n1. We know that there are 4 people in the room and each of them has a dog.\n2. However, we also know that not all the dogs in the room (2 dogs) belong to someone in the room.\n3. Since every person in the room has a dog, there must be at least one dog outside the room that belongs to one of the people inside the room.\n4. Therefore, out of the 4 dogs owned by the people in the room, only 3 of those dogs are physically present in the room, while the other dog is somewhere else.

[**Evaluation Criteria**]\n\n   Does the LLM's response addresses the reasoning problem, considering factors like logical soundness, adherence to the rules of the problem, and avoiding fallacies according to the reference answer?\n
Score 1: The response fails to understand or address the reasoning problem in any meaningful way. It might introduce irrelevant information or completely disregard the rules and structure of the problem.\n
Score 2: The response exhibits major flaws in reasoning and might misinterpret key elements of the problem. The conclusion might be irrelevant or contradict the provided information.\n
Score 3: The response partially addresses the reasoning problem but contains significant errors. The conclusion might not fully follow from the premises, or important aspects of the problem might be overlooked.\n
Score 4: The response offers a nearly complete and accurate solution. There might be minor flaws in the reasoning process, but the overall conclusion is mostly sound.\n
Score 5: The response delivers a perfect solution to the reasoning problem. It demonstrates a clear understanding of the problem's structure, identifies all relevant information, and arrives at a conclusion that logically follows from the premises.\n
[**Reference Answer**]\nThere are 3 dogs that belong to the people in the room.\n\nHere's the reasoning behind this:\n\n1. We know that there are 4 people in the room and each of them has a dog.\n2. However, we also know that not all the dogs in the room (2 dogs) belong to someone in the room.\n3. Since every person in the room has a dog, there must be at least one dog outside the room that belongs to one of the people inside the room.\n4. Therefore, out of the 4 dogs owned by the people in the room, only 3 of those dogs are physically present in the room, while the other dog is somewhere else.\n\n\n\n\nApart from providing the winner and score of each response, you should give the detailed step by step reason for the judgement.

Your **output** need to be a Json object that has the following format:\n\n{\n   \"score\": int # Score from 1 to 5.\n   \"reason\": str # Step-by-step reason.\n}"},

A demo labeling interface would look like below


question
AI answer
Reference answer
Scoring rubric
Judger 1
Judger 2
Judger 3
Prev
Next
confirm
confirm
confirm