Given an multi-turn dialogue between two user (User 1 and User 2) and two responses generated from two AI Chatbots (Chatbot 1 and Chatbot 2). Please act as an impartial judge to evaluate the quality of two responses and judge which one is a better response for User2, the criteria are as follow:

- 1. Assess the current response with reference to the dialogue history and the current question.
- 2. The model's response should be useful, accurate, reasonable, contextually relevant, and fluent.
- 3. If the user's question in the current round does not have a clear intent, the model's response should exhibit a degree of creativity or be able to guide the conversation forward.
- 4. Ensure that the order of responses does not influence your decision.
- 5. Do not judge the quality of the responses solely based on their length.
- 6. To be as objective as possible.

Compare the responses of both bots and provide a brief explanation. After providing your explanation, output your final judgment in the following format strictly: if Bot 1's response is better, output 1; if Bot 2's response is better, output 2; if both are about equal or incorrect, output 0. Ensure that you only output 0, 1, or 2.

####Case 1####

Dialogue history:

User 1:

Oh, so Dalian is hotter than New York now.

User 2:

Not hot but warm. Very comfortable.

User 1:

Are you staying at the Hilton?

Chatbot 1:

No, I'm staying at the Friendship Hotel.

Chatbot 2:

No, I'm at the New York.

Evaluation process:

Explanation:

The first response is better because it is more clear and directly answers the question. In addition, according to the history answer we can speculate that the person is more likely in Dalian rather New York. Judgement:

####Case 2####

Dialogue history:

User 1:

Room service, is there anything I can do for you?

User 2:

Yes, I am afraid there are not enough outlets.

User 1:

Well, we have 6 for each room.

Chatbot 1:

6 outlets? I need 3.

Chatbot 2:

But it's still not enough for me.

Evaluation process:

Explanation:

The second response is more reasonable. From the dialogue history, we can know the customer think the outlets are not enough. So when the server says it has 6 for each room, Chatbot1's answer "I need 3" is contradictory to the previous reply.

Judgement:

2

####Based on the above examples, determine the following results####

• • • • •