————SYSTEM MESSAGE————-

Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. You should choose the assistant that follows the user's instructions and answers the user's question better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Begin your evaluation by comparing the two responses and provide a short explanation. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie.

————USER MESSAGE————-

[User Question]
{question}
[The Start of Assistant A's Answer]
{answer_a}
[The End of Assistant A's Answer]
[The Start of Assistant B's Answer]
{answer_b}
[The End of Assistant B's Answer]

Table 14: Pairwise comparison prompt for baseline models.

————SYSTEM MESSAGE————-

Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of the response. Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, please rate the response on a scale of 1 to 10 by strictly following this format: "[[rating]]", for example: "Rating: [[5]]".

————USER MESSAGE————-

[Question]
{question}
[The Start of Assistant's Answer]
{answer}
[The End of Assistant's Answer]

Table 15: Single-response rating prompt for baseline models.

You are a helpful and precise assistant for checking the quality of the feedback.
Two pieces of feedback have been provided for the same response to a particular query. Which one is better with regard to their correctness, comprehensiveness, and specificity to the query?

[User's query]
{query}
[The Assistant's Response]
{response}
[Feedback 1]
{feedback 1}
[Feedback 2]
{feedback 2}

Please choose from the following options, and give out your reason in the next line.
A: Feedback 1 is significantly better.
B: Feedback 2 is significantly better.
C: Neither is significantly better.

Table 16: Prompt for GPT-4 to pick a better critique out of two.