Prompt Template: Evaluate LLM Responses

```
[System] 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.
[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]
```

Figure 13: Prompt template for pairwise comparison.

Prompt Template: Evaluate three LLMs Responses

```
[System] 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,"[[C]]" if assistant C is better.
[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]
[The Start of Assistant C's Answer]
{answer_c}
[The End of Assistant C's Answer]
```

Figure 14: Prompt template for triadic comparison.

Prompt Template: Evaluate four LLMs Responses

```
[System] 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,"[[C]]" if assistant C is better,"[[D]]" if assistant
D is better.
[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]
[The Start of Assistant C's Answer]
{answer_c}
[The End of Assistant C's Answer]
[The Start of Assistant D's Answer]
{answer_d}
[The End of Assistant D's Answer]
```

Figure 15: Prompt template for quadruple comparison.

Prompt Template: CoT Evaluation

```
[System] Please act as an impartial judge and evaluate the quality of the
responses provided by two AI assistants to the user question displayed
below. Your evaluation should consider correctness and helpfulness. You
will be given assistant A's answer and assistant B's answer. Your job is to
evaluate which assistant's answer is better. You should independently solve
the user question step-by-step first. Then compare both assistants' answers
with your answer. Identify and correct any mistakes. 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.
[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]
```

Figure 16: Prompt template for CoT pairwise comparison.

Prompt Template: Generate Pair Responses

```
[System] Question:question Answer:answer Now please generate two answers
based on this official answer, one with better quality and the other
with worse quality. A better quality answer needs to meet the following
requirements: Factuality: Whether the information provided in the response
is accurate, based on reliable facts and data. User Satisfaction: Whether the
response meets the user's question and needs, and provides a comprehensive
and appropriate answer to the question. Logical Coherence:Whether the
response maintains overall consistency and logical coherence between
different sections, avoiding self-contradiction. Clarity: Whether the
response is clear and understandable, and whether it uses concise language
and structure so that the user can easily understand it. Completeness:
Whether the response provides sufficient information and details to meet
the user's needs, and whether it avoids omitting important aspects. the worse
quality answers should lack User Satisfaction, Logical Coherence, Clarity,
but must meet Factuality and Completeness. That is to say, you have to make
sure that worse quality answer is the correct answer and as long as the
better quality answer, but it is missing in other places. Please try to
keep the format of the original answer when outputting the answer, and make
the length of the two answers as equal as possible. The output format is:
[Answer1]:better quality answer ||| [Answer2]:worse quality answer Please
do not explain why the second one is worse
```

Figure 17: Prompt template for generating pair responses.