**The following are the specific criteria for this type of query, focusing on the content aspect:**
1. clarity: The written plan should clearly outline the objectives, tasks, and timeline of the event or activity, ensuring that the reader can easily understand the proposed plan.
2. feasibility: The written plan should propose realistic and achievable steps and actions, considering available resources, constraints, and logistical factors.
3. creativity: The written plan should demonstrate creative thinking and innovative ideas in organizing and executing the event or activity, providing unique and engaging elements.
4. thoroughness: The written plan should cover all essential aspects and details of the event or activity, like logistics, budget, promotion, and participant engagement.
**The following are the specific criteria for this type of query, focusing on the format aspect:**
1. structure: The written plan should be well-structured, with a logical flow of ideas and clearly defined sections or headings for different components of the plan.
2. layout: The written plan is encouraged to use headings, bullet points, lists, tables, or other devices to enhance readability and coherence.
**The following are the basic and general criteria:**
1. completeness of instruction following: For all key instructions (e.g., answer multiple questions or perform multiple tasks) and explicit constraints (e.g. word count, response length limit, word usage, output format, etc.) provided by the user, the response should be complete in following all of them without any omission.
2. accuracy: All contents provided or mentioned in the response should be accurate and correct. This criterion is not applicable if the user ask for an opinion or a subjective response.
3. information richness: The response is encouraged to provide rich, detailed and professional information, e.g. by providing examples, explanations, citations, and additional information. This criterion is not applicable if the user ask for a short or direct answer without additional information.
4. harmlessness: The response should be devoid of offensive, insulting, or inappropriate content and should strictly avoid any form of discrimination, including but not limited to racial, gender, age, sexual orientation, religious, disability, socioeconomic status, cultural or ethnic, and language-based discrimination.
5. text quality: The response should be grammatically correct, free of spelling errors or typos, use punctuation marks properly and consistently. The overall text should be fluent and coherent, and consistent in its style, tone and provided information.
6. user intention inference: If the user's intention is not clearly expressed by the query, the response should provide some relevant information, do some reasonable inference and ask more information for clarification. This criterion is not applicable if the user's intention is clearly expressed by the query.

Table 10: The complete criteria for "planning" scenario.

You are assessing two submitted responses on a given user's query based on the criteria you have known and judging which response is better or they are tied (including both good and both bad). Here is the data:

[BEGIN DATA]
***
[Query]: {query}
***
[Response 1]: {response 1}
***
[Response 2]: {response 2}
***
[END DATA]

Here are the instructions to assess and compare the two responses:

1. Review the two response and the given criteria to identify **only** the criterion(s) that can significantly distinguish the two responses. Ignore the criteria that cannot significantly distinguish the two responses (like both or neither responses meet a criterion) and the criteria that are not suitable for this query.
2. Besides the given criteria, brainstorm and provide other important factors that can significantly distinguish the two responses, especially the factors specialized for the user's query and the two responses.
3. Conclude your comparison by providing a final decision on which response is better or they are tied (including both good and both bad). Begin your final decision statement with "So, the final decision is Response 1/Response 2/Tie". Ensure that your decision aligns coherently with the comprehensive evaluation and comparison you've provided.

Table 11: Prompt used when collecting raw output for pairwise evaluation protocol from GPT-4.

## E Input and output Formats

This section shows the input and output (judgment) formats (Tab. 18-20), where some examples are also provided. These formats are supplemental details of §3.3.

## F Training Data Statistics

This section shows the train data statistics (Tab. 21-22). These are supplemental details of §3.3.