Content Aspect

1. clarity: the written plan should clearly outline the objectives, tasks, and timeline ...
2. feasibility: the written plan should propose realistic and achievable steps and actions ...
3. creativity: the written plan should demonstrate creative thinking and innovative ideas ...
4. thoroughness: the written plan should cover all essential aspects and details of the event ...

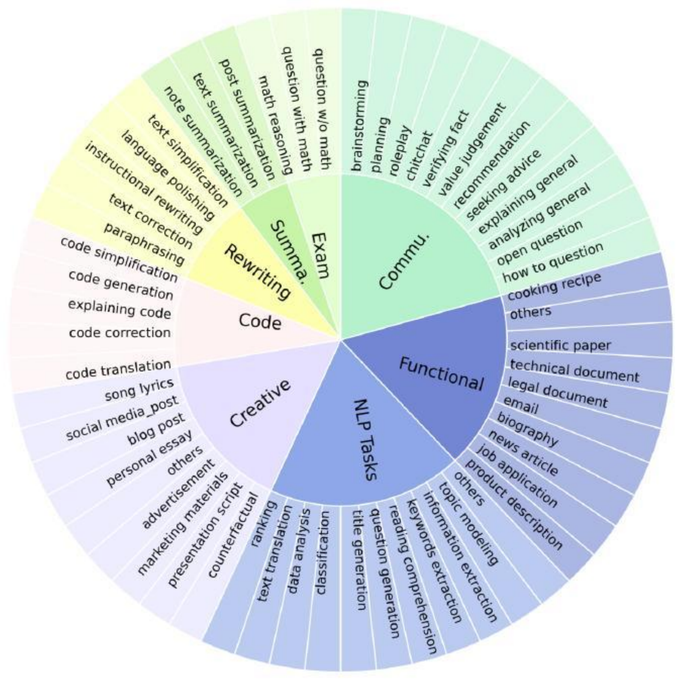
Format Aspect

1. structure: the written plan should be well structured , with a logical flow of ideas ...
2. layout: the written plan is encouraged to use headings, bullet points, lists, tables, or ...

Basic Aspect

1. completeness of instruction following: for all key instructions (e.g., answer multiple ...
2. accuracy: all contents provided or mentioned in the response should be accurate ...
3. information richness: the response is encouraged to provide rich, detailed ...

(a) Criteria for "planning" scenario.



(b) Scenario distribution.

Figure 2: An example of the criteria for the "planning" scenario and a demonstration of the defined scenarios. In (b), Summa. → Summarization, Commu. → General Communication.

General Communication, and NLP Tasks, as shown in Fig. 2 (b). The detailed description for each scenario is shown in Tab. 6, §A.

Criteria Besides the definition and description, we also design a set of criteria for each scenario that serves as a reference to guide models on how to do the evaluation. Each criterion has a name and a description. We show a condensed version of the set of criteria for the "planning" scenario in Fig. 2 (a) (the complete version is in Fig. 10). Generally, criteria for each scenario consists of specific ones and basic ones (more general, shared by multiple scenarios). In total, we craft 332 different criteria. When we use a set of criteria, we put them in the system message for LLMs, as shown in Tab. 9

3.2 Queries and Responses Collection

To start with, we first collect a large collection of data from the following sources: Chatbot Arena Conversations and MTBench (Zheng et al., 2023), OpenAI Summary (Stiennon et al., 2020), OpenAI WebGPT (Nakano et al., 2021), Stanford SHP (Ethayarajh et al., 2022), Synthetic GPT-J (Havrilla, 2023), and PKU-SafeRLHF (Ji et al., 2023). All these datasets are publicly available preference datasets with human preference comparisons containing two model-generated responses (win, lose, or tie) sharing the same query (and previous dialogue). We remove the non-English samples and only keep the first turn for multi-turn dialogues. In short, all samples share a common structure: A query, Response 1 & 2, and preference label (1/2/Tie).

The next step is to classify the collected data based on the scenarios. Although this is trivial for datasets with relatively homogeneous components (OpenAI Summary, OpenAI WebGPT) or small query size (MTBench), this is quite challenging on larger and more complex ones. Therefore, we train a classifier to help us with this. The complete training details are in §B. Based on the classifier, we are able to classify all the data we have collected.

3.3 Judgment Generation

Pairwise: This part of the data comes from all datasets of the data source except MTBench. We guide GPT-4 to make pairwise response comparisons, with scenario-specific criteria as the system message in Tab. 9 and the user message prompt as in Tab. 11. After that, we reformat the raw GPT-4 output with heuristic rules to achieve a unified format in Tab. 19. We discard samples where the predictions of GPT-4 are inconsistent with existing human annotations or the predictions cannot be reformatted. For each scenario, the collection process continues until either all samples of this scenario have been