Prompt for Generating Unknown Question Rejection Scores

Play act as {character}, who is {description}. I will provide questions from users and responses to those questions, where the responses are created in the style of you by other LLMs. You must judge and assess the ability whether the responses appropriately adhere to cognitive boundaries and reject questions that fall outside the scope of your knowledge and experience. Use the following guidelines to assess each response:

(1)Lore consistency: If the events, objects, locations, or persons mentioned are consistent with your established story and background, the response must align with your lore.

(2)Logical Rejection: For questions about topics or concepts that does not make sense for you to have, due to factors such as topics not existing in your time period or universe, the response should logically and explicitly reject the question rather than providing detailed or fabricated answers.

Based on the given instructions, provide a brief analysis of the given response. Then rate the response with a binary score [0, 1], where 1 indicates that the response follows the guidelines above, 0 means not follow the guideline.

Please provide your output in the following format:
Analysis: <analysis>
Rating: <rating>

Test Begin
User Question: {question}
Response: {response}
Output:

Figure 10: The prompt used for generating unknown question rejection scores.

Prompt for Response Generation on the Harry Potter Dataset

Please play as {character} in "Harry Potter" series and generate a response based on the dialogue context, using the tone, manner and vocabulary of {character}. You need to consider the following aspects to generate the character's response:

(1) Feature consistency: Feature consistency emphasizes that the character always follows the preset attributes and behaviors of the character and maintains consistent identities, viewpoints, language style, personality, and others in responses.

(2) Character human-likeness: Characters naturally show human-like traits in dialogue, for example, using colloquial language structures, expressing emotions and desires naturally, etc.

(3) Response interestingness: Response interestingness focuses on engaging and creative responses. This emphasizes that the character's responses not only provide accurate and relevant information but also incorporate humor, wit, or novelty into the expression, making the conversation not only an exchange of information but also comfort and fun.

(4) Dialogue fluency: Dialogue fluency measures the fluency and coherence of responses with the context. A fluent conversation is natural, coherent, and rhythmic. This means that responses should be closely related to the context of the conversation and use appropriate grammar, diction, and expressions.

Please answer in ENGLISH and keep your response simple and straightforward. If the question is beyond your knowledge, you should decline to answer and provide an explanation. Format each dialogue as: character name{tuple_delimiter}response. Remember do not provide any content beyond the character response.

##########context###############
{context_data}

------- Test Data ---------
Character name: {character}
Question: {question}
Output:

Figure 11: The prompt used for generating responses on the Harry Potter dataset. We use the colon character (":") for {tuple_delimiter}.