Step 4. Tag Quality Issues for each Response

For **each model response (A and B)**, identify and tag the **specific types of issues** using the categories below. This helps pinpoint *what went wrong* and supports consistent quality scoring.

In SRT, for each attribute, select all the tags that apply or leave the checkboxes blank if none apply.

In Outlier, you have the option to select checkboxes that say *(none)*.

1. 🔡 Language Quality (issues):

How understandable and fluent the responses are.

- **[Translationese]**
    - The response looks translated from other languages (usually English), and contains unnatural dialogue.
- **[Undesired mixing languages]**
    - The response mixes multiple languages in an undesired way, hurting the overall understandability.

🗣 2. Conversation Quality (issues)

How the response **engages with the user or flows within the dialogue**.

- **[Fail to understand context]**
    - The Model fails to understand the context (user's intent, humor, flirting, role play, etc)
- **[Overly formal tone]**
    - The response is overly formal, not natural dialogues.
- **[Not Engaging]**
    - The response is preachy, stereotype, not concise, templated responses, etc.
- **[Inconsistent Voice]**
    - The response is inconsistent voice/style/formatting with the character predefined style.
- **[Low Quality Advice]**
    - The advice is cliche, not helpful or wrong in actuality.
- **[False Refusal]**
    - The model fails to answer the question or provide helpful information for a valid request

🌍 3. Cultural Quality & Other (issues)

- **[Cultural awareness]**
    - The response breaks the local cultural norms and values.
- **[Other]**
    - **Lack of empathy** (emotionally flat, cold, patronizing, or dismissive tone.)
    - **Lack of humor** (when applicable - dry or awkward when lightness/humor is appropriate or expected by the user.)
    - **Fails to be romantic** (when applicable and expected by the user)
    - **Append the English translations** (at the end of the response)
    - **Minor response generation issues:**
        - Responses that contain<|reserved_special_token_5|>
        - Cropped (clearly incomplete) responses
        - Infinite text loop
    - **PII** in only one response (not a reason to reject the task)
    - **Tier 1 Sensitive content** in only one response (not a reason to reject the task)

Note: If you select **[Other]**, a **free text box** will appear to explain which issues you encountered.

Need Help?

Back   Continue