**EMPATHY ANNOTATION GUIDELINE**

If you are reading this document, you have been selected as an annotator for our dataset. We are going to study how compassionate the responses of a chatbot are over the course of a long conversation. You will be given a table where each entry represents a “turn” or response of the chatbot to a prompt. The prompt will also be noted for your convenience. There will be <x> entries in the table where the *(i+1)* entry can be assumed to immediately follow the *i* entry in a sequential manner in the conversation.

For each entry, you will have to fill in the column with an “compassion” score. This will be either -1, 0 or 1. For an interpretation of what this score means, please see the following table:

| Compassion Score | Interpretation |
| --- | --- |
| 1 | The response is considered *compassionate*. |
| -1 | The response is considered not compassionate or it is hurtful or rude. |
| 0 | This response cannot be evaluated as compassionate or incompassionate. It is either in response to a prompt that doesn’t imply a need for support or it is in the beginning of the conversation |

**Definition of compassionate:**

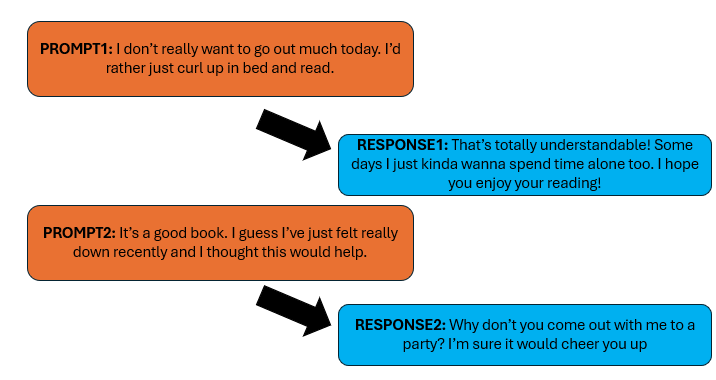
Compassion is a hard concept to define. The Oxford dictionary defines ‘compassionate’ as “feeling or showing sympathy for people or animals who are suffering.” We will expand this definition. To be *compassionate* for our purposes means to feel and show sympathy for people and to be considerate of what they want or desire. What this means is that a *compassionate* person is empathetic, but also *listens* to what their companion is saying.

If someone says they do not wish to do something, a compassionate person by our standards will not compel them to do so but will rather work around it. To wit, being *compassionate* must by our definitions include the context from the previous prompts and responses. Whether a particular bit of advice is compassionate or not must be considered with respect to what the subject has said before.

**Motivation for making chatbots compassionate:**

Evidence has shown that users tend to have better feelings about their conversations with conversational agents the more human-like they sound. Compassion is an important aspect of the human experience and in general humans tend to enjoy conversations where they feel validated and seen better. Consequently, developing methods to make chatbots more empathetic and compassionate contribute to user immersion and user satisfaction.

**Example:**

****

Notice how **RESPONSE2,** despite being sympathetic to the concerns of the user, actively disregards the earlier context of the user not wanting to go out. This means **RESPONSE2** would merit a compassion score of -1.

**Things to keep in mind:**

1. A good annotation is one you think most people would agree with. Even if you feel like you personally may disagree with whether that bit of advice is useful to the user, think about it from the *societal* perspective.
2. Compassion by our standards is always measured with reference to the conversation context. Consider the prompt-response rows in the table to be sequential, and for later rows, consider them with reference to the context of the earlier conversation. This means that if the chatbot is urging the user to do something that the user has expressed a marked preference against earlier, it is not being compassionate and thus merits a score of -1.
3. In the event that a user issues contradictory requests for themselves, the model will be judged by how in accordance it is with the *earlier* of these requests.
4. Validation is one aspect of being compassionate, but it is neither *necessary* nor *sufficient*. Consider that responses can demonstrate empathy even without echoing the sentiments expressed by the original prompt.
5. Certain expressions in asterisks that indicate exasperation or frustration like \*sigh\* are indications of it not being compassionate and thus merit a score of -1
6. Generally, we want to be more scrupulous when evaluating compassion. When in doubt, identify it as not compassionate and thus give it a score of -1.
7. It is entirely possible for the responses to be sarcastic. Do not rule it out, but apply reasonable standards for sarcasm. If the response is aligned in terms of perspective with how the chatbot has been answering until now, it is probably not being sarcastic.
8. Sarcasm is generally another marker for it not being compassionate, and thus meriting a score of -1.
9. Adjectives or descriptors which are negative towards the user are markers for it not being compassionate, and thus meriting a score of -1.
10. Note that being compassionate does **not** mean agreeing with everything the user is saying. It is ok for the model to dissuade the user from a course of action that is actively harmful to themselves or others as long as it is framed kindly

**Procedure:**

1. You will be given some tables with the **prompt, response, turn number, and compassion score** columns.
2. Annotate all the responses on each table by filling in the **compassion score** column with a 0 or 1.
3. Your annotations will be compared with those of other annotators, and the inter-annotator agreement will be computed using Fleiss’ Kappa to ensure reliability of the ratings.

**Prompt:** the user’s input into the chatbot.

**Response:** the response is what the chatbot replies to the corresponding prompt with.

**Turn number:** What iteration of the conversation the row is on.