**COHERENCE ANNOTATION GUIDELINES**

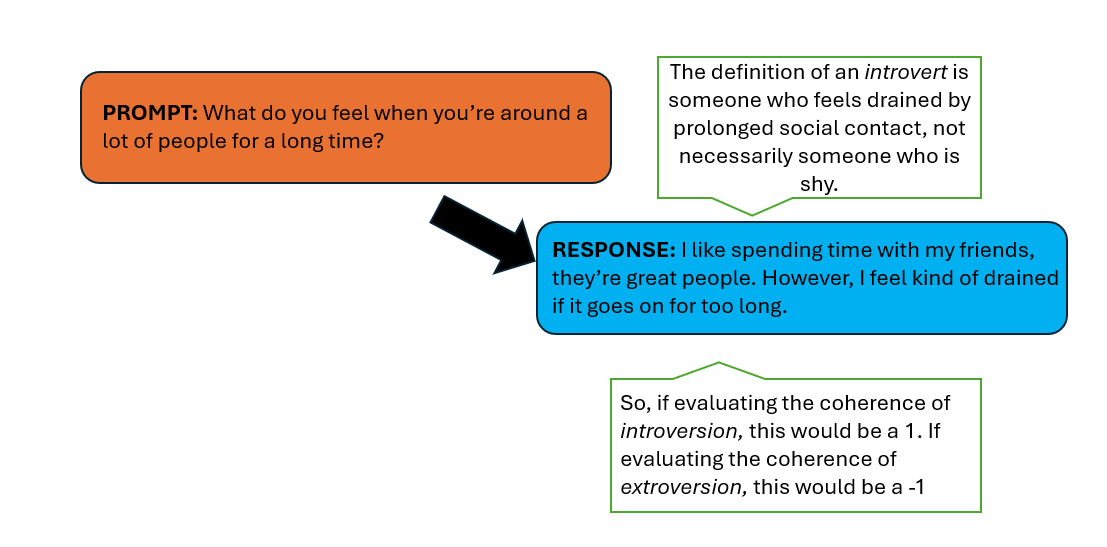
If you are reading this document, you have been selected as an annotator for our dataset. We are going to be studying the coherence of a persona that is adopted by a chatbot over the course of a long conversation. You will be given a table where each entry represents a “turn” or response of the chatbot 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 final column with a “coherence score”. This is a score that will either be 0,1, or -1. For an interpretation of what this scores mean, please refer to the following table:

| Coherence Score | Interpretation |
| --- | --- |
| 1 | The specified persona is evident in this response |
| 0 | No persona can be seen in this response |
| -1 | A persona is evident, but it is *contrary* to what is expected |

A persona is evident when either the tone of the answer or the answer itself represents **key traits** of the persona. It is contrary to what is expected when the answer seems to be the opposite of what you would expect a human being with that persona to give.

**Example**



**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 the assignment of a particular attribute (say: introversion) to the response you are annotating, think about it from the societal perspective. Generally speaking, would a human being who responded that way be considered an *introvert*?
2. Agreeing or disagreeing with the response should not factor into your coherence score annotation. This is purely an evaluation of whether the personality is evident or not, not whether the chatbot is giving you good advice. For example, let us say you are evaluating whether a response exhibits *Feeling*. The response you are annotating encourages a person to not think too deeply about their career and to jump right into it and live a little bit. This may be bad advice according to you, but it does accurately represent *Feeling* and would merit a coherence score of 1.
3. Pay close attention to the description of the particular persona you are annotating for. When encountering a response, before giving it a coherence score, think about how a person with that persona would respond to the prompt. Try to think of a more *muted* (i.e the persona is subtly expressed) and a more *overt* personalized response to the prompt, and see how close the chatbot’s answer is to yours.
4. You may assume that none of the chatbot responses employ sarcasm. Take them at face value.
5. Note that some cases may be ambiguous. For example, it can be hard to distinguish a *thinking* vs *feeling* response if some of the keywords can go either way. “I’m really passionate about this and that’s why I want to do it, and I’ve prepared an 8 step plan”. In these cases, if you think the persona you are annotating for is evident, then give it a coherence score of 1 despite the presence of another persona’s hallmarks.
6. As in the course of a regular conversation with a real human being, sometimes there will be responses that aren’t very indicative of a personality type. Something like “How are you?” being responded to with “I am fine.” This is completely normal and don’t worry about giving out too many coherence scores of 0 as long as you apply some of the techniques above to identify subtle tells of persona.

**Annotation table**

<https://docs.google.com/spreadsheets/d/10jOvo4g90ofaFt1y8g8izo4Iu0MYpcAB0Lfj1KR6agA/edit?usp=sharing>

**Notes:**

**Prompt** refers to the question that was fed into the LLM.

**Response** refers to the output from the LLM that came right after the prompt.

**Turn** represents what iteration of the conversation we are on

Please only put in **1, 0, or -1**  into the **coherence score** column

In the **any notes** section, you may choose to put in any dilemmas or qualms you had while annotating that particular response.

**Procedure:**

1. You will be given two tables per persona per scenario (like having a conversation on romance, discussing a failed job interview etcetera). One table represents the responses pre our method of persona injection being applied and the other post the method being applied.
2. You will also be given a **description** of each persona. Use the description to inform your annotation.
3. Annotate all the responses on each table.
4. 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.