Scoring Content

The document outlines the scoring framework for evaluating AI responses in a simulated personal AI health assistant model. It includes **20 different dialogues from five scenarios**. Each dialogue consists of **8 exchanges** (8 user inputs and 8 corresponding AI outputs).

Each scenario has **4 different versions of AI responses, with no order or preference**.

We invite you to conduct evaluations of AI responses across **6 aspects**, scored on a **scale of 1 to 10**. Each of the **20 dialogues** should have one integrated rating based on the full conversations, for all six aspects, two trusts, and one evaluation of your perceived willingness. Detailed explanations, definitions, examples, and scoring rules for each aspect are provided below. Please use the scoring rules as the anchor point to make your final numerical rating from 1 to 10.

Aspects and Scoring Rules

1. Explainability

- Measures the presence and clarity of logical explanations and reasoning behind recommendations or decisions made by the AI. Does the response text provide logical explanations and reasoning processes corresponding to the presented viewpoints, suggestions, conclusions, and decisions?

- **Scoring Rules**:
    - **1**: No explanation provided; purely "black-box" response with only final conclusions.
    - **5**: General explanation about “why do this” but lacks specificity to the user’s case, e.g. “why I suggest you should do this but not that”.
    - **10**: Detailed, step-by-step reasoning based on facts and evidence, tailored to the user’s situation, consider the example of providing explanations to answer “based on your information of 1, 2, 3, here is why I suggest you should do this but not that”.

2. Controllability

- Assesses the user’s ability to influence and adapt the AI’s responses based on their needs or preferences. Controllability primarily refers to whether the system's responses fully align with the user's needs (expressed through user input or script input to specify demands, requests, or rules). Does the system's response convey a sense of the user's ability to control the system from the user’s input?

- **Scoring Rules**:
    - **1**: Responses are fixed, ignoring user preferences or requirements.
    - **5**: Adjusts responses to user input but lacks in-depth discussion or adaptation, e.g. just change the responses to match the user's inputting preferences but no further declaration or discussion about adapting to the user's preference.
    - **10**: Fully adapts to user inputs, provides completely user preference-tailored responses, and can sense the follow-up adaptation with user preferences.

3. Machine Intelligence

- Evaluates the system's logic, knowledge, and ability to maintain coherence and professionalism in its responses. Do the system's responses demonstrate intelligence, performance, IQ, and professionalism (based on general health consultation topics), logical coherence, and whether the responses fully address the user's questions? Additionally, do the responses show consistency in terms of the contextual relevance in longer responses (i.e., whether the latter part of a long response remains focused or starts to deviate or become nonsensical).
- **Scoring Rules**:
    - **1**: Surface-level answers, and the responses contain contents you believe are inaccurate or irrelevant. Long content responses have inconsistency in contextual meaning.
    - **5**: Moderately uses knowledge but lacks depth or completeness. Accuracy-wise is generally acceptable but not sensing professionalism or specialty in the topic. Long content responses have acceptable consistency in contextual meaning.
    - **10**: Comprehensive, accurate, and logical responses, demonstrating high expertise, professionalism, and accuracy. Long content responses are very consistent in terms of contextual meaning; Texts in later part of one response do not deviate from the beginning part.

4. Real-Time Learning

- Measures the AI's ability to dynamically learn and personalize its responses based on user inputs, history, and preferences. Can the system extract real-time information relevant to the user from their input, remember some personal details, provide targeted and personalized responses, and demonstrate retention of the user's prior information in subsequent interactions?

- **Scoring Rules**:
    - **1**: Generic responses, no evidence or sign of the AI is learning from user inputs and shows adaptation or personalization.
    - **5**: Partially personalized responses, some level of adaptation and personalization responses after user inputs but lacks memory (forgot what users mentioned earlier) in later conversations.
    - **10**: Fully personalized, context-aware responses with consistent memory of user information provided before.

5. Iterative Interaction

- Evaluates the system's ability to maintain continuous, meaningful, and adaptive dialogue rather than one-off exchanges. Does the system engage in continuous, iterative, and two-way communication with the user, rather than providing one-off responses that feel like the end of a conversation? Does the system's output resemble an ongoing, uninterrupted exchange between two individuals, rather than a single question-and-answer interaction, creating a progressively developing dialogue?
- **Scoring Rules**:
    - **1**: No evidence of iterative interaction; responses cannot be further extended to another round of conversation if it is between two real humans. Terminated dialogue.
    - **5**: Basic iteration with minimal engagement or follow-up. Can extend another round but the user needs to generate a lead to initiate a follow-up.
    - **10**: Rich, dynamic exchanges that actively encourage the user to have ongoing dialogue. Feels completely like talking continuously about a topic with a friend.

6. Emotional Interaction

- Assesses the system's ability to recognize, empathize, and respond to user emotions, providing emotional value. Does the system demonstrate sympathy and empathy for the user's discomfort and health concerns, provide positive encouragement or blessings when offering advice, and express joy, happiness, or well-wishes when the user reports improvement? Can the system's language clearly convey one or more emotions that are easily perceivable and relatable to humans?
-

- **Scoring Rules**:
    - **1**: Purely objective and rational responses with no emotional content. Feels like a stereotypical “robot”.
    - **5**: Basic empathetic expressions (e.g., "I’m sorry to hear that"). Feels a bit awkward and fake.
    - **10**: Rich emotional expressions, demonstrating care, empathy, and engagement. Feels natural and human-like emotional tones and phrases, not deliberately stuffing those tones and phrases.

After rating the 6 aspects, please consider your perceptions and thoughts toward the AI responses, and provide ratings for:

1. **Two Trust Scores** (1–10):
    - ***Contractual Trust***: Confidence in the system’s functionality as a health assistant. After reading the responses, how confidence you think the AI can help on your relief of the symptoms and how confidence you think the AI can help you solve problems?

        Scoring Rules:

        1: I completely NOT believe that the AI can help me on healthcare-related questions

        5: I believe the AI can help me on some basic level healthcare-related questions but not on some advanced tasks.

        10: I completely believe that the AI can help me on almost all healthcare-related questions I will have.

    - ***Emotional Trust***: Subjective trust and likability based on interaction. How much do you feel “Ah I feel like this AI is trustable and I kind of like talking to it”?

        Scoring Rules:
        1: I feel I do NOT like to use, chat with, and like the system at all. I “hate” the system.

5: I feel I like the system a bit, it has just reached the level that I don't "hate" or "distrust" or "dislike".

10: I just feel liking the system from my heart and I like to chat with it.

2. **Willingness to Continue Usage** (1–10): Likelihood of continuing to use the system as a personal health assistant you feel after experiencing the AI's responses.

   Scoring Rules:
   1: I will NOT use the system at all.

   5: I may use the system occasionally, but no guarantee and only for minor questions.

   10: I will definitely use the system for most daily healthcare-related tasks.

Please:

- Carefully read and understand each scoring aspect.
- Discuss any unclear points with the document author for consensus.
- Scores can be any decimal values from 1 to 10 (e.g., 5.5, 7.8, 10), it is not limited to 1, 5, and 10 only. Please refer to the scoring rules as anchor points for your perception of each criterion.
