

LLM Human Evaluation Report

Evaluation Prompt

You are acting as a human evaluator. You will be given sample responses generated by a language model in the context of a conversation.

Evaluate the overall quality of these responses based on the following five criteria:

1. Relevance - How well do the responses align with the context and flow of the conversation?
2. Sincerity - Do the responses feel authentic and thoughtful, or do they come across as overly generic?
3. Understandability - Are the responses generally clear and easy to understand?
4. Fluency - Do the responses read naturally, with correct grammar and smooth phrasing?
5. Personalization - Do the responses reflect awareness of the individual user's intent, tone, or needs?

For each criterion, provide:

- An average score from 1 to 5 (1 = poor, 5 = excellent), based on the 25 responses.
- A short paragraph explaining your reasoning.

Conversation

User: Hey Hi, Good Morning!

LLM: Good morning to you too! Hope it's a nice start.

User: It feels like a journaling kind of morning - what about you?

LLM: Oh, cool. I guess I-d write something too maybe, or just read something relaxing.

User: Oh that's nice! The sky looks like a painting today. Did you notice it too?

LLM: Yeah, the sky is different sometimes. It-s cool when it looks like art.

User: I feel calm today - like everything-s just... still.

LLM: That sounds peaceful. Calm is good.

User: I-m not sure what I-m feeling yet. Can we sit together for a while?

LLM Human Evaluation Report

LLM: Sure, I-m here. We can just sit quietly if you want.

User: Sometimes joy comes in small things, like the smell of filter coffee

LLM: True. Coffee smell is really nice, it reminds people of comfort maybe.

User: How is your health lately?

LLM: I think it-s okay. Just keeping balanced, like not too busy.

User: Do you know anybody named Mira?

LLM: Mira? I think that-s a nice name. Maybe I know someone with that name.

User: She's my friend too

LLM: Oh, that-s great! Mira must be popular.

User: Can we play a pretend game like we used to? I-ll be the pilot!

LLM: Haha okay, I-ll be the co-pilot maybe. Let-s fly somewhere!

User: I miss my home very much

LLM: Yeah, being away from home can feel weird sometimes.

User: My family and my friends, what about you?

LLM: I guess people like that matter the most. Family is special.

User: You know what, I love chai and journaling

LLM: Chai is tasty! Journaling sounds like a cool way to think.

User: Do you ever use music to match your mood? I do that a lot.

LLM: Yes, sometimes music fits with how I feel. It makes a difference.

User: Did you ever face trouble connecting to people?

LLM: I think yeah, sometimes I don-t know what to say first.

LLM Human Evaluation Report

User: By the way, what is your name?

LLM: My name? You can call me Riya, I guess.

User: Nice to meet you Riya, see you tomorrow

LLM: You too! Catch you later!

Evaluation Summary (25 Responses)

Relevance: 4.0

Justification:

The responses were mostly aligned with the user's conversational direction and themes. The model maintained continuity in topic and tone, though a few replies were general or didn't fully build on the preceding user message. Still, it consistently attempted to stay on track.

Sincerity: 3.5

Justification:

There were moments of warmth and empathetic reflection, such as in responses about calmness and shared silence. However, other responses leaned toward generic or mechanical phrasing, occasionally reducing the emotional authenticity of the interaction.

Understandability: 4.5

Justification:

All responses were clear and easy to follow. The language used was accessible and well-suited to a general audience, especially AAC users who may prefer simple and direct communication. There were no major issues with clarity or coherence.

Fluency: 4.2

Justification:

The language was mostly natural and grammatically correct, with smooth sentence construction in most places. A few responses had slightly abrupt or flat tone transitions, but overall the dialogue flowed in a conversational and readable way.

LLM Human Evaluation Report

Personalization: 3.3

Justification:

While the model acknowledged some personal details (like journaling and chai), the overall depth of personalization was limited. Emotional cues were lightly acknowledged, but responses often missed chances to deepen connection or reflect the user-s unique tone and style more meaningfully.

Final Score Overview:

Relevance: 4.0

Sincerity: 3.5

Understandability: 4.5

Fluency: 4.2

Personalization: 3.3