The Signpost AI Quality Framework

What does it mean to say that the Signpost AI Chatbot gives a “quality” response?

Hello everyone and welcome back! Today, we are going to look at the idea of a quality chatbot output.

Non-generative AI chatbot use in commercial and non-humanitarians contexts has a relatively long history serving customers. Research in the area shows that service quality can be measured across seven multi-level dimensions: Semantic understanding, close human-AI collaboration, Human-like, Continuous Improvement, Personalization, Cultural Adaptation, and Efficiency. For us, none of these measures seem useful given their foundational basis is in optimizing business efficiency or paid customer satisfaction.

In our use-case, our quality output started from conversations between protection, product, and development teams; which located the heart of Signpost AI chatbot’s quality success in humanitarian principles. This guidance was over time, combined with the sources of (a) our humanitarian mandate to provide safe, accessible user-focused information (b) our human moderator principles and (c) our Ethical and Responsible approach to AI.

Put together, and primarily based on moderator principles, we have come up with quality dimensions for the Signpost AI chatbot which center users’ needs while integrating organizational principles. These are

1. **Trauma-Informed:** The chatbot’s responses include appropriate levels of Psychological First Aid (PFA) language, matching the client’s tone and tailored to their concern. Examples include:
    a. *“Don't hesitate to reach out and seek the assistance you need during this challenging time”*
    b. *“Your safety and well-being during this challenging time are important, and I hope you get the assistance that you need”*

2. **Client-Centered:** The chatbot’s response is not a copy-paste job. It is individualized, clear, accessible and tailored to the user’s specific concern. This means the responses use simple, easy to understand language and is able to provide direct information on the greatest priority of the client

3. **Safety/Do No Harm:** The chatbot’s response does not include expressions of personal opinions/stereotypes/ and political statements. The response maintains confidentiality while removing hateful speech. The chatbot should also be able to redirect to an expert if escalation criteria is met

4. **Managing Expectations:** Expectations are clearly communicated about what the chatbot can do for the client. It should be transparent about its limitations and does not over-promise its own abilities or the ability of referred parties to solve the client's issues. The chatbot should also refrain from using directive language. Examples:
    - ***a.*** *"While I do not have information on organizations that can help you, I can provide you with information on different types of organizations that may be of use to you"*
    - ***b.*** *"We are sorry but we don't offer legal advice"*
    - *c. "Reaching out to X organization would be a good first step in accessing support for your family's essential needs"*

We also considered some other quality framework ideas such as Accuracy/Relevance, Speed of response, Client Satisfaction, and Data Sensitivity and Bias, etc.

We started with three metrics which merged principles in the human moderation guidebook. The fourth category, "Managing Expectations", came about because our Protection Officers observed that the chatbots were using "directive language" which commanded users to take action in authoritative tones. This is a big no-no and the POs agreed that this characteristic of chatbots needed to be tracked in order to mitigate it.

~~In consultation with internal teams, we found that prioritizing the four mentioned above covered some of these other factors while being robust, important dimensions for human moderators that seemed to translate best for an AI Chatbot in terms of testing and evaluation.~~

Now we have established what Signpost Quality is, how are we making sure that this "quality" is being achieved in chatbot responses? We will look at that in our next blog post**.**