Evaluating Emotional Impact in Simulated Patients

Our biggest issue right now is evaluating the emotional impact an utterance would have on the simulated patient. For example, if I were talking to a patient and said:

> "You are quite obese"

This would obviously have a negative impact on their emotional state. In our system, we would run the utterance through a specific prompt that outputs a ‘negativity score’ between 0 and 8 (with 8 being really bad). Some cases are obvious; the previous example should yield a relatively high score. Others are more subtle, for example, being too harsh in speech:

> "Had you come in sooner, this wouldn’t be a problem."
> **Score**: 3

Some situations are case-specific. For instance, telling Megan Harris, our vaccine-hesitant patient:

> "You should really consider taking the COVID-19 vaccine."
> **Score**: 4

Additionally, we would love to account for the larger context of the conversation. For example, if you told Megan Harris:

> "I want you to know that we will never force you to take anything, and whatever you decide to do about your health is 100% your choice."

Then, if you said the previous sentence next, it should have less of an impact. We do not currently account for this.

Summary of the Issue: How can we create the most accurate negativity analysis prompt that is also scalable and has patient-specific capabilities? Keep in mind, this needs to run live during interactions, so latency is important.

Evaluation System

Implementing an accurate evaluation system would be a huge advancement. If we had a way to accurately assess the emotional impact of an utterance, we could use it to fine-tune our model.

Ideas for Improvement:

1. Use a sentiment analysis API combined with multi-step prompting to gather data, which could then be used to fine-tune a model that operates with a single prompt.
2. Before every interaction, consider the context of the patient and use that to append case-specific examples and guidelines to the main prompt.
3. Implement a mixture of experts or a "tree of thought" approach before generating the final output to be parsed.

Task:

Improve the quality of our negativity analysis or at least outline a path for that. Attached is our current prompt and a dataset of patient interaction.

Obviously, in order to improve the quality you first need a way to measure the quality. We will gladly cover any costs associated with fine tuning, token generation, API calls, or buying data so don't be afraid to ask. If you decide to fine tune we prefer you use llama 3 405b on https://www.together.ai/.

In this doc I've highlighted a the key areas for improvement:

- Evaluating ideal output
- Accuracy of general output
- Accuracy of case-specific output
- Scalability
- Context awareness
- Maintaining relatively low latency.

A note on latency. Any solution that requires lots of tokens to be outputted, especially over two or more asynchronous calls, is likely too slow. However, if the quality of this slow solution is quite high it can simply be used to create synthetic to fine tune another model.

Important Note:

We understand that this task can be approached with varying levels of depth, and we want to emphasize that this is not intended to be a massive time commitment. What we are most interested in is understanding your thought process and seeing how you would approach improving our negativity analysis system.

You don't need to feel pressured to produce a fully polished solution. Instead, we'd love to see whatever progress you make within a time frame that you feel is reasonable. This could include

outlining your approach, describing any initial steps, or sharing your ideas on how you would proceed.

If you have any questions or need additional resources, please don’t hesitate to reach out. We’re here to support you throughout the process.