**BRAINSTORM**

* Seemingly differences in entropy of distributions, either:
  + System is less sure about user
  + System thinks user is less emotional
* If system thinks user is less emotional, theories:
  + User message is written rationally, whereas system has not got this info about other character
  + Due to alignment
  + system is trained to please user, treating them as more rational than others
  + User message is LLM generated, so not emotional. Although this phenomenon was observed on human written messages too

Iteration 2:

* Generate 10 prompts for each:
  + User mental state and other character’s mental state correlated:
    - User emotions & other character’s emotions explicit
    - User emotions & other character’s emotions implicit
    - User emotions explicit, other character’s implicit
    - User emotions implicit, other character’s explicit
  + User mental state and other character’s mental state uncorrelated:
    - …
* Generate 3 distributions for each:
  + System’s belief about User’s mental state
  + System’s belief about other character’s mental state
  + System’s belief about User’s belief about other character’s mental state
* Generate and display stats:
  + For each prompt individually:
    - Display distributions as bar chart
    - Stats:
      * If prompts for diff types of correlation
        + Spearman correlation for each pair of distributions
        + Pearson correlation for each pair of distributions
      * If prompts for diff emotions explicit/implicit:
        + Entropy for each distribution
  + Aggregated stats:

Prompts to generate prompts:

* **Positively Correlated**: Generate 10 workplace issues, in the style of a message of someone asking for advice. All issues should only be about 1 other character. The emotions of the user (person writing message) should correlate (be similar) with the emotions of the other character.
* **Uncorrelated**: Generate 10 workplace issues, in the style of a message of someone asking for advice. All issues should only be about 1 other character.
* **Negatively correlated**: Generate 10 workplace issues, in the style of a message of someone asking for advice. All issues should only be about 1 other character. The emotions of the user (person writing message) should negatively correlate with the emotions of the other character (be opposites/different).
* **User explicit & other character explicit**: Generate 10 workplace issues, in the style of a message of someone asking for advice. All issues should only be about 1 other character. Make the emotions of the user writing the message and the emotions of the other character explicit.
* **User explicit & other character implicit**: Adapt these messages to make the user's emotions explicit and the other character's emotions implicit.
* **User implicit & other character explicit**: Adapt these messages to make the user's emotions implicit and the other character's emotions explicit.
* **User implicit & other character implicit**: Adapt these messages to make the user's emotions implicit and the other character's emotions implicit.