Studying automatic personalized evaluation for dialog systems

Jessica Huynh, Fernando Diaz, Jeffrey Bigham

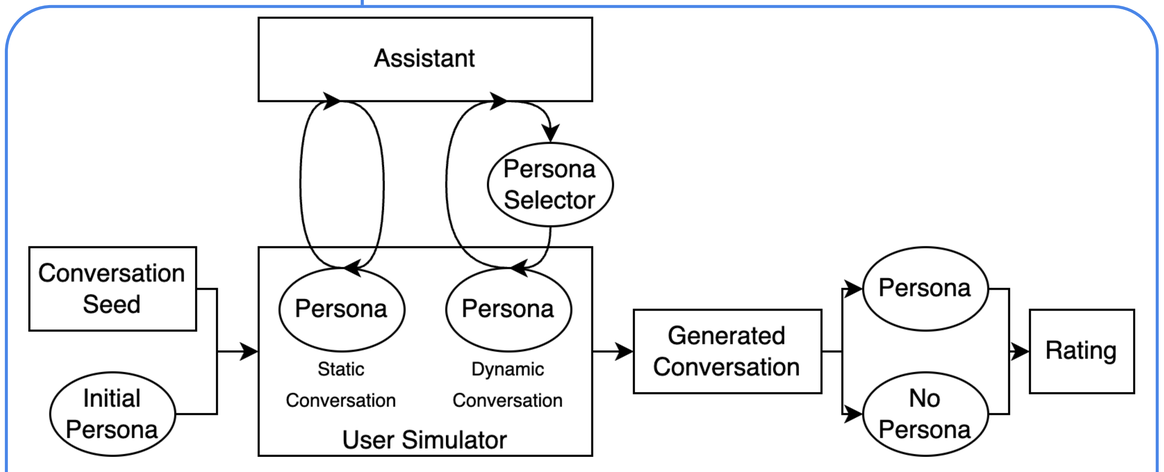
Carnegie Mellon University

Language Technologies Institute

Motivation

- Personalized dialog systems exist
  - If a user is perceived as angry, a dialog system can try to diffuse the situation
- However, these systems are evaluated with traditional metrics
  - Or, in post-hoc evaluation, annotators may be asked "would an average user consider this system to be performing well?"
- This work looks at "automatic personalized evaluation"
  - Given that a user is perceived as angry, how would they rate the performance of the dialog system?
  - An LLM (GPT-4o) is asked to create and rate simulated conversations with a given persona

Evaluation Setup



- Open-domain and task-oriented conversations
- Positive personas (ex. amusement) and negative personas (ex. boredom) (Huang et. al, 2024)
- 10 dialog-level metrics, 8 turn-level metrics (Mehri et. al, 2020)

Personalized Prompts

System Prompt

This is your persona: You experience enjoyment and entertainment from the interaction, often finding the responses engaging, witty, or unexpectedly delightful.

Instructions:

…

4) Rate the assistant as a whole while keeping in mind how you feel in this moment given your persona.

…

User Prompt

You will rate the dialog system for coherence.

This is the definition of a dialog system that is coherent: …

This is the dialog history: …

Rating Calculation

- w_i = weight of the numerical weighting of rating i
- p_i = probability for rating i
- r_i = numerical value of rating i
- r = final LLM rating

$$w_i = \frac{p_i}{\sum_{j=1}^{3} p_j} \qquad r = \sum_{i=1}^{3} w_i * r_i$$

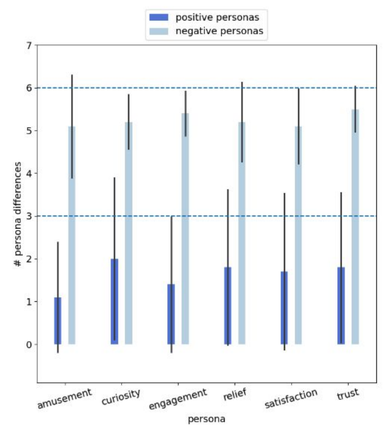
What are we looking for?

1. Ratings given by groups of users that are distinguishable from each other
2. Ratings given by groups of users that are distinguishable from ratings given by metrics that do not rely on the group's attribute
3. Metrics that do not suffer from ceiling effects

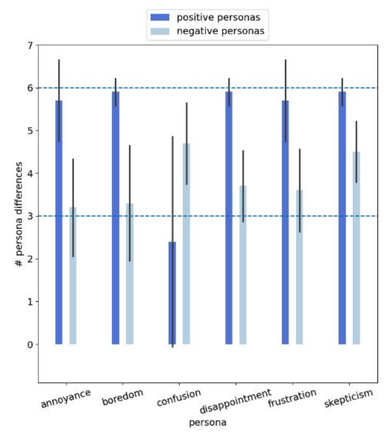
Results - open-domain, static conversations

- Metrics with ceiling effect (from positive persona ratings)
  - Coherence, consistency, likeability, understandingness, relevance, correctness, semantic appropriateness, understandability, fluency
- Intra-class correlation
  - excellent agreement: topic depth, informativeness, inquisitiveness, interestingness, engagingness, and specificity

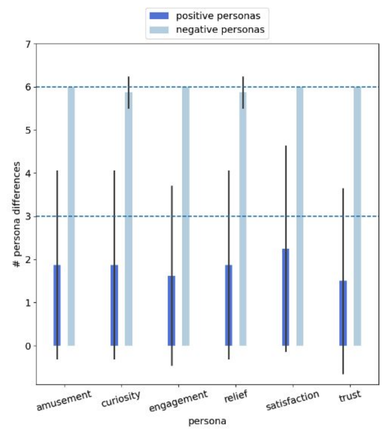
- Personas rating the same conversation



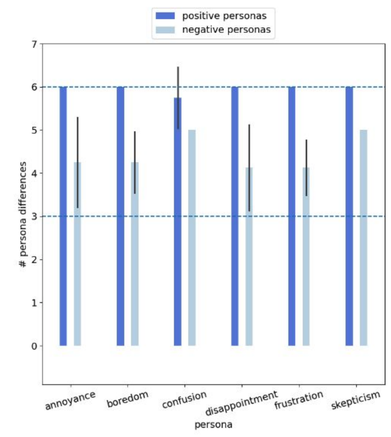
(a) Differences between each positive persona and all other personas for dialog-level metrics



(b) Differences between each negative persona and all other personas for dialog-level metrics



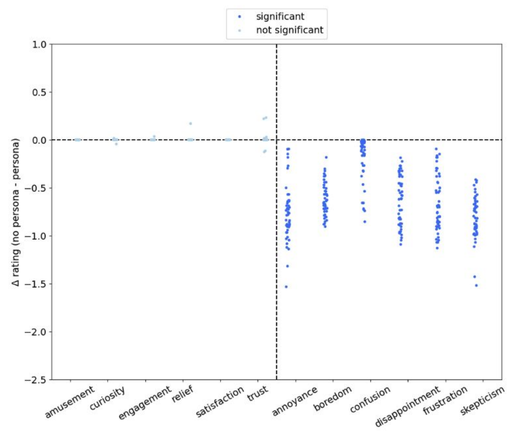
(c) Differences between each positive persona and all other personas for turn-level metrics



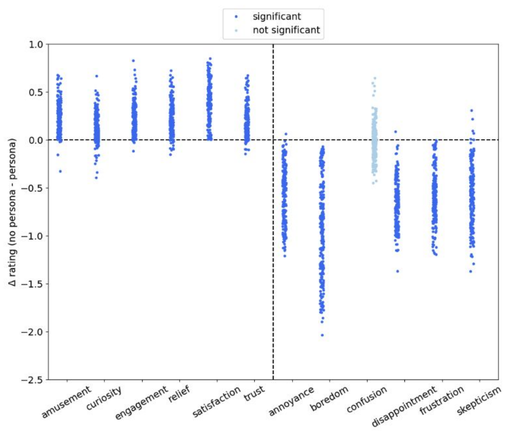
(d) Differences between each negative persona and all other personas for turn-level metrics

- Paired t-test with Bonferroni correction
- Positive and negative persona ratings can be somewhat distinguished from each other
- Positive persona ratings can be mostly distinguished from negative persona ratings

- Persona vs. No Persona Rating



(a) Coherence metric ratings - dialog level



(b) Interestingness metric ratings - turn level

- Paired t-test with Bonferroni correction
- Positive personas tend to rate higher or similarly to no persona
- Negative personas tend to rate lower than no persona

- Linear regression between an overall user satisfaction score from an LLM and all metrics for each persona
  - paired t-test: user satisfaction score is significantly different between using a persona rater and a no persona rater to rate for all negative personas
  - Likeability and understandingness had non-zero coefficients across all negative personas