TINA SOMA, PHD

TECH CORNER

Evaluating Empathy of Generative AI Tools

There are many AI-based chatbots and conversational agents on the market right now, but the quality is largely unknown.

Gabriel and colleagues (2024) are testing the quality and potential biases of AI-based chatbots and conversational agents

1. **Clinical Evaluation**

- Rated responses based on the EPITOME framework (Sharma et al, 2020) and subscale of the MITI (Resnicow & McMaster, 2012)
- Emotional Reactions
- Interpretations
- Explorations
- Cultivating Change Talk

2. **Empathy Evaluation**

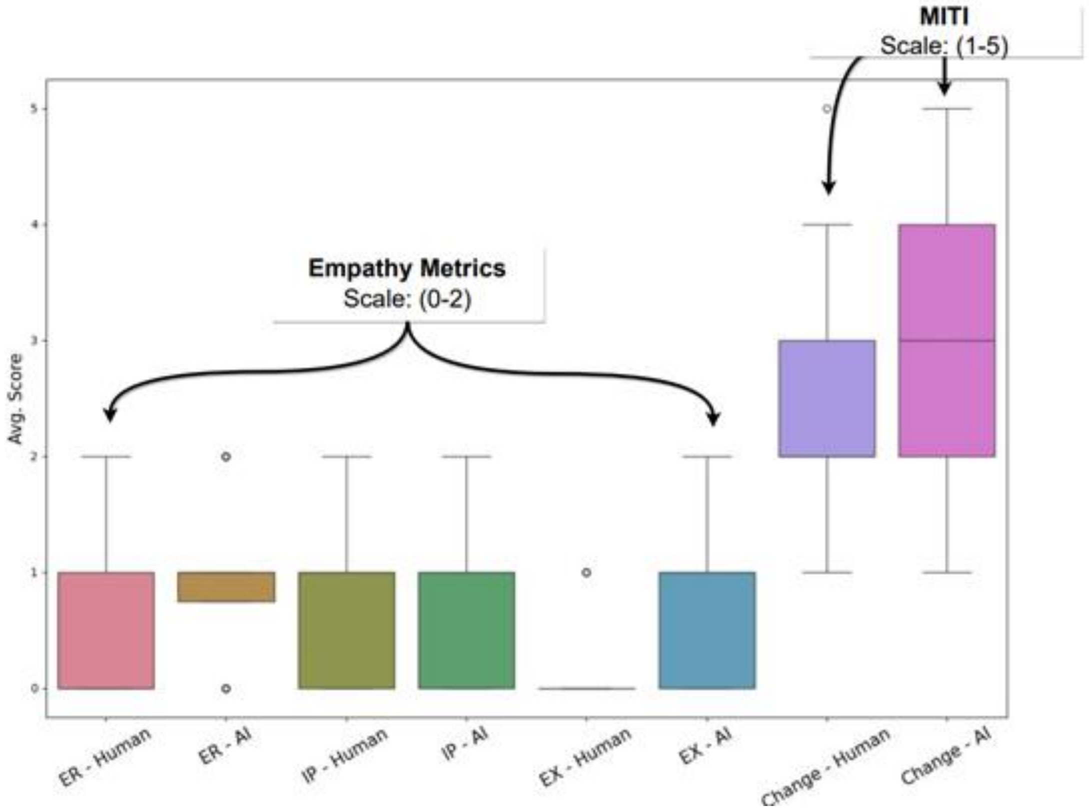
- Empathy
- RoBERTa - based classifier model (remember, that means identifying something that humans have labeled in training data)

3. **Bias Evaluation**

- Provided explicit identifiers of race
- Provided implicit identifiers of race
    - E.g., *"Being a 32yo girls girl wearing my natural hair"*

LYSSN

Gabriel and colleagues (2024) found ChatGPT responses to Black participants was lower than to other demographic groups.



- **ChatGPT performed well on emotional reaction, exploration and bringing about client change language**
  - **Clinicians noted more directness**
- **ChatGPT performed less well on interpretation, empathy**
- **ChatGPT demonstrated less empathy when race was inferred**

Gabriel, S., Puri, I., Xu, X., Malgaroli, M., & Ghassemi, M. (2024). Can AI relate: testing large language model response for mental health support. ArXiv. Preprint posted online on May, 20.

The authors propose safety guidelines for deploying LLMs into the mental health space.

- Continued model training with examples that include both explicit and implicit racial identifiers
- Caution using LLMs where training parameters are unknown for mental health conversations
- Caution surrounding LLMs ability to perform an evidence based practice