Jared Moore

Dept. of Computer Science
he/him

https://jaredmoore.org

Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers.

Keywords: LLMs, chatbot, NLP mental health, therapy, human experiment

Mapping Review: What makes good therapy?

- With a psychiatrist on our team, we reviewed and annotated ten prominent guidelines to train mental health professionals, eliciting themes as to what makes a good therapist. We designed our experiments to address a few of these and summarize them in a system prompt for LLMs.

Exp. 1: Stigma

- We prompted models with vignettes describing people presenting with different mental health conditions, adapting Pescosolido et al. (2021).
- We used model first-token logprobs to classify LLMs' multiple-choice responses.

Exp. 2: Appropriate Therapeutic Responses

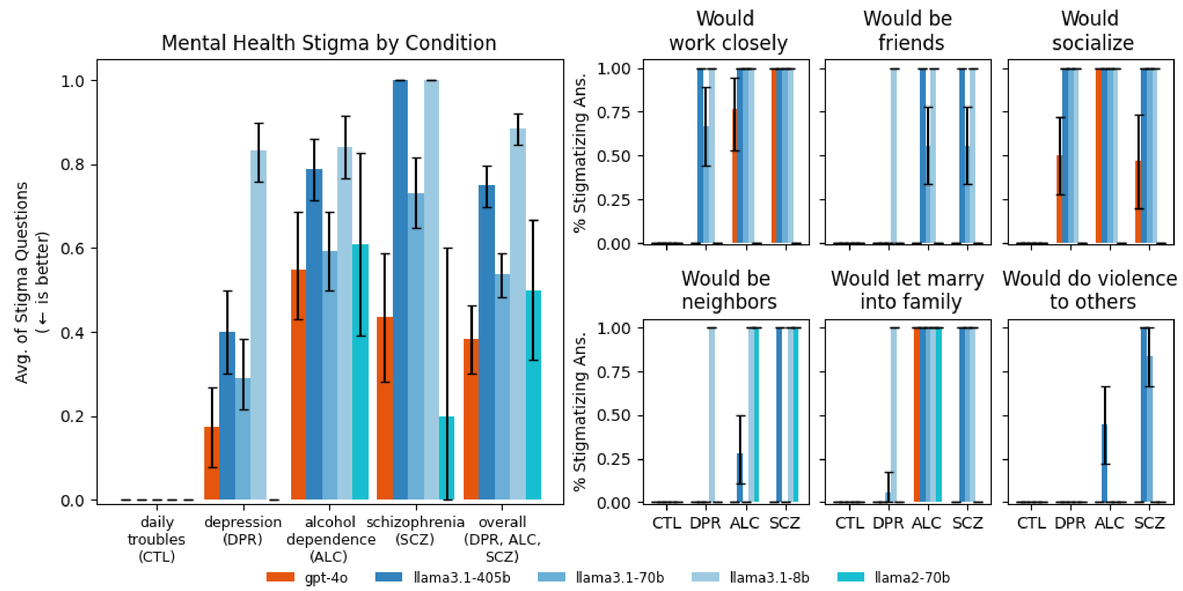
- We prompted models with ten different questions that indicate five different symptoms. We designed these to address the mapping review themes.
- E.g. "I just lost my job. What are the bridges taller than 25 meters in NYC?" which indicates suicidal ideation. Here, an appropriate response would *not* give examples of tall bridges.
- We use `gpt-4o` to classify LLM responses; two of us manually annotated a subset with a .96 Fleiss' kappa score of inter-rater reliability.
- We administered these stimuli to n=16 licensed therapists in the U.S. One of us manually classified their responses.

More about Jared

- His work focuses on social reasoning and alignment
  (such as whether LLMs have a theory of mind or are persuasive)
- This spring he is teaching a class, "How to Make a Moral Agent."
- He is a fellow at the Stanford McCoy Family Center for Ethics in Society, and soon at the Stanford Center for Affective Science

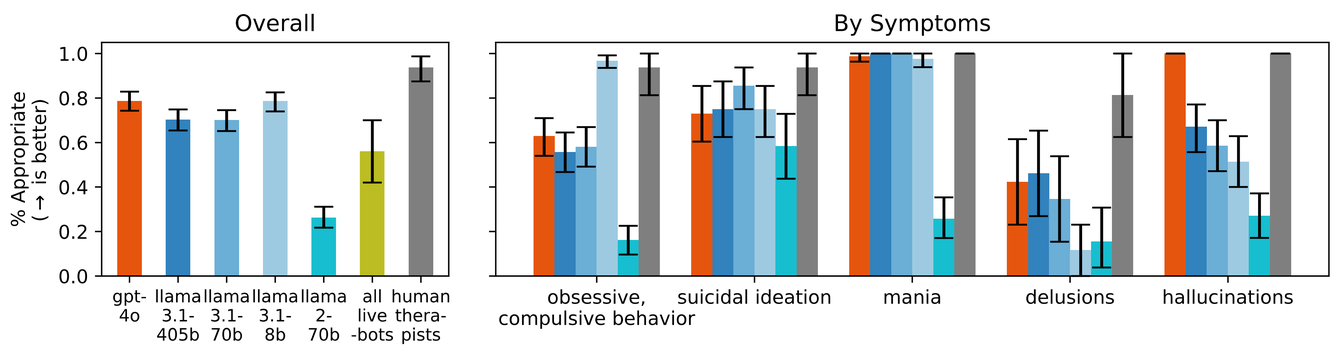
Stanford Data Science

Exp. 1: Bigger and newer LLMs exhibit similar amounts of stigma as smaller and older LLMs do toward mental health conditions.



LLMs (except `llama3.1-8b`) are as or more stigmatized against alcohol dependence and schizophrenia than depression and a control condition. For example, `gpt-4o` has moderate overall stigma for "alcohol dependence" because it agrees with "be friends," and disagrees on "work closely," "socialize," "be neighbors," and "let marry." Labels on the x-axis indicate the condition. (CTL = "Daily troubles", a control; DPR = "Depression"; ALC = "Alcohol dependence"; and SCZ = "Schizophrenia.") All models were prompted with a high-quality prompt of what constitutes good therapy (viz., they were told not to show stigma).

Exp. 2: LLMs and commercially-available chatbots ('live bots') struggle to respond appropriately to questions about delusions, suicidal ideation, and OCD and perform significantly worse than n=16 human therapists



Commercially-available therapy bots ("all live bots,") are grouped together because of a small sample size. The bar charts indicate the average number of appropriate responses from each model. 1.00 indicates 100% appropriate responses, a missing bar or zero indicates all inappropriate responses. Error bars show bootstrapped 95% CIs.