GPT-4o Lacks Core Features of Theory of Mind

John Muchovej
Yale University, New Haven, Connecticut, United States

Amanda Royka
Yale University, New Haven, Connecticut, United States

Shane Lee
Yale University, New Haven, Connecticut, United States

Julian Jara-Ettinger
Yale University, New Haven, Connecticut, United States

Abstract

Do Large Language Models (LLMs) possess a Theory of Mind (ToM)? Research into this question has found that LLMs succeed on a range of benchmark tasks. However, these evaluations do not test for the actual representations posited by ToM: namely, a causal model of mental states and behavior. Here, we use a cognitively-grounded definition of ToM to develop and test a new evaluation framework. Specifically, our approach probes whether LLMs have a coherent, abstract, and consistent model of how mental states cause behavior—regardless of whether that model matches a human-like ToM. We test our evaluation against GPT-4o and find that even though it succeeds in approximating human judgments in a simple ToM paradigm, GPT-4o fails at a logically-equivalent task and exhibits low consistency between its action predictions and corresponding mental state inferences. As such, these findings suggest that GPT-4o's social proficiency is not the result of a ToM.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*