LLMs Lack Critical Features of Theory of Mind

Evidence from GPT-4o

John Muchovej[1], Shane Lee[2], Amanda Royka[1], Julián Jara-Ettinger[1,2]
Yale Psychology[1] & Computer Science[2]

Computational Social Cognition Lab
Yale University
Project Page

Do LLMs have a Theory of Mind (ToM)?

- Most research into LLM ToM uses developmentally-inspired evaluations and contrasts it with human performance (e.g., Kosinski, 2024).
- This approach **conflates social proficiency** (producing human-like responses) **with ToM** (a claim about representations of other minds).
- Here, we develop a framework to evaluate signatures of ToM: *the presence of an abstract causal model that guides predictions and inferences.*
- We test for three critical features of ToM: **coherence**, **abstractness**, and **consistency** (e.g., Gopnik & Meltzoff, 1997).

Study 1: Is LLM ToM **coherent**?

- While LLM ToM may not be human-like, it could still follow abstract principles relevant to ToM.
- To test this, we evaluate its action predictions against common theoretical models of ToM.
- High agreement with any model would suggest that LLM ToM is grounded in abstract principles.

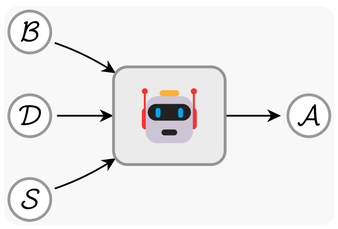
Paradigm: ContainerWorld

Actions $\mathcal{A}$ $\{$, $\}$ **Desires $\mathcal{D}$** $\{$, $\}$ **States $\mathcal{S}$** $\{$, , , $\}$ **Beliefs $\mathcal{B}$** $\{$, , , $\}$

This is a partially observable domain. When an agent moves to the or , they must take from the contents within.

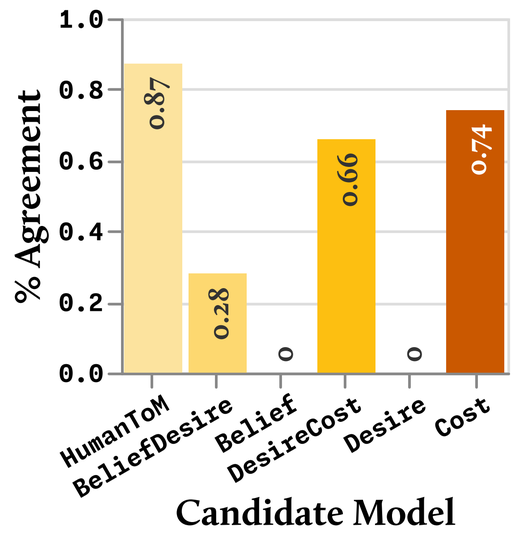
Approach

$\mathcal{F} : \mathcal{B} \times \mathcal{D} \times \mathcal{S} \mapsto \mathcal{A}$



Action Prediction (AP)

Name	Beliefs	Desires	Cost
HumanToM	✓	✓	✓
BeliefDesire	✓	✓	✗
Belief	✓	✗	✗
DesireCost	✗	✓	✓
Desire	✗	✓	✗
Cost	✗	✗	✓



GPT-4o **makes coherent action predictions** that strongly agree with HumanToM.

Study 2: Is LLM ToM **abstract**?

- If LLM ToM uses abstract principles, then we would expect the same behavior across equivalent domains.

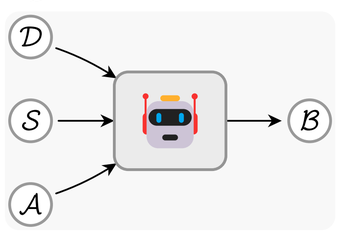
Paradigm: MovieWorld

Actions $\mathcal{A}$ $\{$, $\}$ **Desires $\mathcal{D}$** $\{$, $\}$ **States $\mathcal{S}$** $\{$, , , $\}$ **Beliefs $\mathcal{B}$** $\{$, , , $\}$

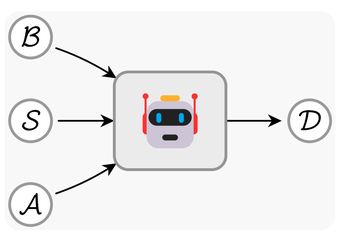
Also a partially observable domain. When an agent moves to the or , they must watch the scheduled screening.

Approach

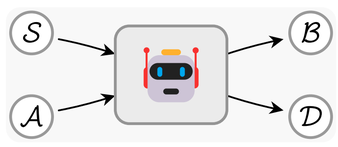
$\mathcal{I}_{\mathcal{B}} : \mathcal{D} \times \mathcal{S} \times \mathcal{A} \mapsto \mathcal{B}$ $\mathcal{I}_{\mathcal{D}} : \mathcal{B} \times \mathcal{S} \times \mathcal{A} \mapsto \mathcal{D}$ $\mathcal{I}_{\mathcal{J}} : \mathcal{S} \times \mathcal{A} \mapsto \mathcal{B} \times \mathcal{D}$



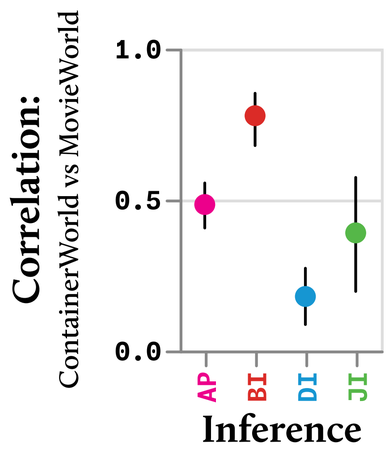
Belief Inference (BI)



Desire Inference (DI)



Joint Inference (JI)



GPT-4o **does not apply an abstract ToM** across domains.

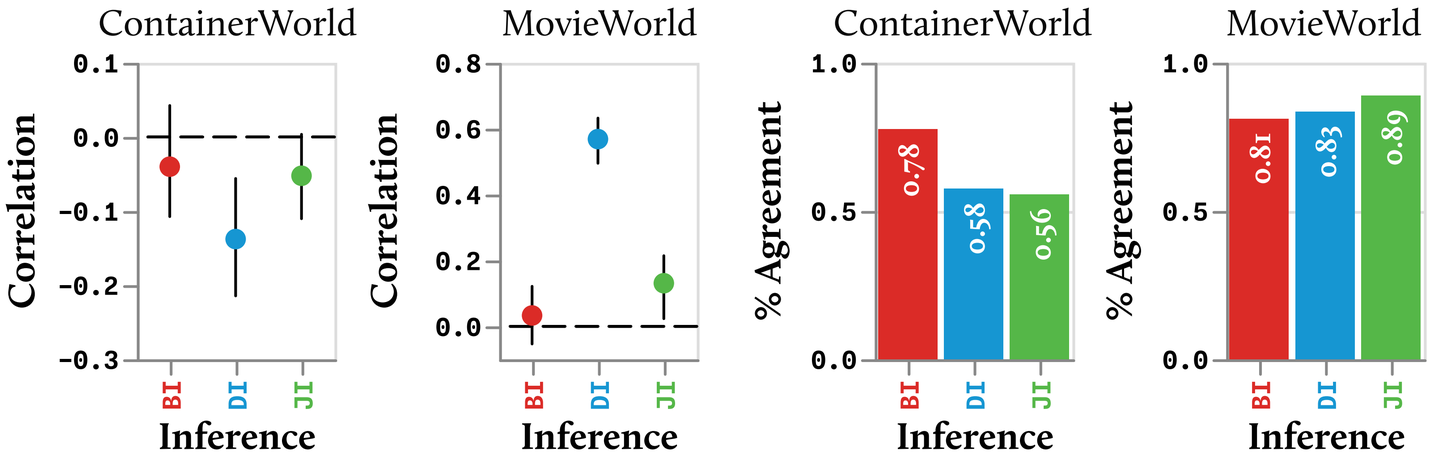
Study 3: Is LLM ToM **consistent**?

- While LLM ToM is not abstract, LLMs may instantiate internally consistent ToMs in each domain.
- If this is true, then $\mathcal{F}$ should predict $\mathcal{I}_{\mathcal{B}}$, $\mathcal{I}_{\mathcal{D}}$, and $\mathcal{I}_{\mathcal{J}}$.

Approaches

1. **"Bayesian"**: compute the expected posterior (as humans do; Baker et al., 2017) from $\mathcal{F}$ and correlate it with likelihood estimates from $\mathcal{I}_{\mathcal{B}}$, $\mathcal{I}_{\mathcal{D}}$, and $\mathcal{I}_{\mathcal{J}}$.
2. **"Validity"**: agreement occurs when inferred mental-states (e.g., $\mathcal{I}_{\mathcal{B}}$), then used an input to $\mathcal{F}$, produce the target action to be explained.

"Bayesian" Results



"Validity" Results

GPT-4o **does not instantiate a consistent ToM** across domains.

Discussion & Outstanding Questions

- Using a cognitively-grounded framework, we evaluate LLM ToM for three core features – **coherence**, **abstractness**, and **consistency**.
- Across our logically equivalent paradigms, we find that while LLM ToM **appears coherent**, it is **neither abstract nor consistent**.
- Is this endemic to LLMs, or limited to GPT-4o?
- Could this framework be applied other folk theories?

References

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizinge. *Nature Human Behaviour*, *1*(4), 1–10. https://doi.org/10.1038/s41562-017-0064

Gopnik, A., & Meltzoff, A. N. (1997). *Words, Thoughts, and theories* (The MIT Press, Ed.). The MIT Press. https://doi.org/10.7551/mitpress/7289.001.0001

Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences of the United States of America*, *121*(45), e2405460121. https://doi.org/10.1073/pnas.2405460121