Towards A Holistic Landscape of Situated Theory of Mind in Large Language Models

Ziqiao Ma, Jacob Sansom, Run Peng, Joyce Chai.
Computer Science and Engineering Division, University of Michigan.

Motivation

Theory of Mind (ToM), the ability to attribute mental states to oneself and others, is integral to human cognition and social reasoning. ToM has influenced AI, particularly in developing machine ToM to enhance AI agents' social intelligence. The emergence of large language models like ChatGPT and GPT-4 has intensified discussions about machine ToM, with debates centering on their capabilities and limitations in complex social reasoning. Current benchmarks for evaluating machine ToM are seen as limited and susceptible to data-related issues.

To advance this field, here are two critical questions:
1) how to comprehensively categorize the machine ToM;
2) what constitutes a more effective evaluation protocol for it.

LLM as Theory of Mind Agents

In this section, we first survey recent research presenting evidence and counter-evidence for the emergence of ToM in LLMs.

Q: Do Machine ToM Emerge in LLMs?

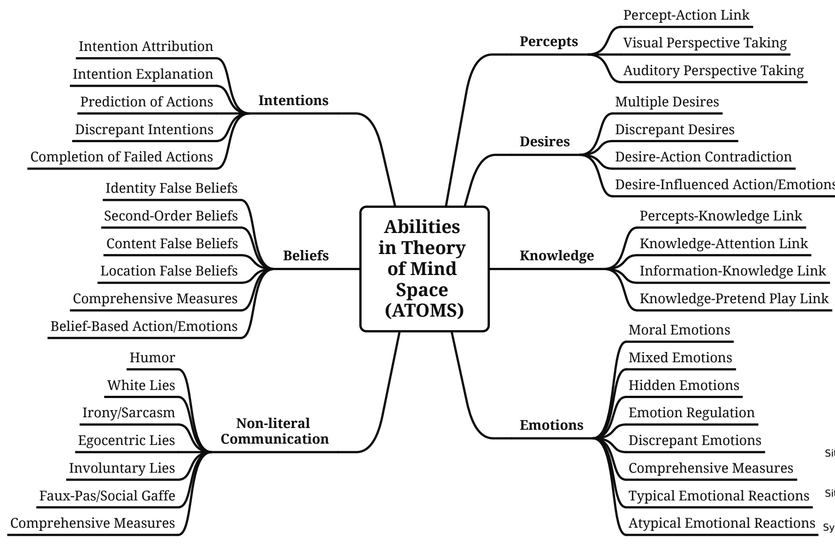
Yes, emerged. LLMs like GPT-4 reflect emergent ToM capabilities, with abilities to predict and represent beliefs, desires, emotions, intentions; Several evidential behavioral studies and case studies have been done.

No, not yet. Limitations revealed in various benchmarks and studies; Understanding of mental states is superficial, prone to spurious correlateions and shortcuts, falling short of human-level ToM.

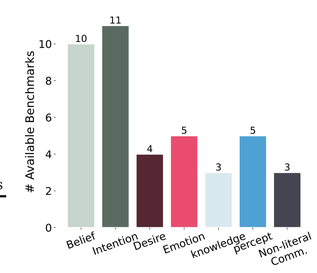
We have the following roadblocks in ToM Evaluation in LLMs:

- **Limited aspects of ToM** The ToM capability may have been overclaimed based on evaluations from only a specific aspect.
- **Data contamination** The training corpora of LLMs may contain research papers detailing these pyschological studies.
- **Shortcuts and spurious correlations** LLMs may leverage shortcuts to perform highly without acquiring the desired skills.

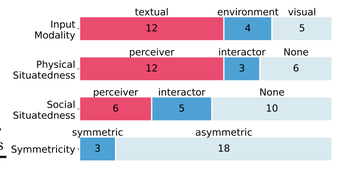
A Holistic Landscape of ToM



The ATOMS framework of [1], which identified 7 categories of mental states through meta-analysis of ToM studies for children.



Number of available benchmarks for each mental state in ATOMS framework.



A comparison of benchmark settings on four aspects.

We follow [1]'s taxonomy of ToM sub-domains, i.e., the Abilities in Theory of Mind Space (ATOMS). The space consists of 7 categories of mental states, including:

1) Beliefs **2) Intentions** **3) Desires** **4) Emotions**
5) Knowledge **6) Percepts** **7) Non-literal communication**

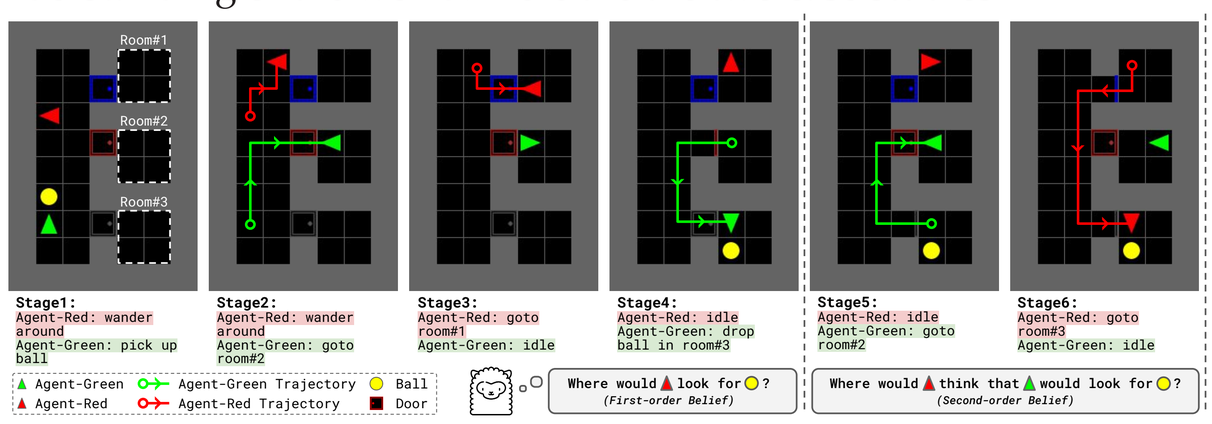
Based on the listed 7 categories of mental states, we conduct a taxonomized review of ToM related benchmarks, and find that:

Review
1) Many aspects of ToM are under-explored
2) Lack of clear targeted mental states
3) Lack of situatedness in a physical and social environment
4) Lack of engagement in environment

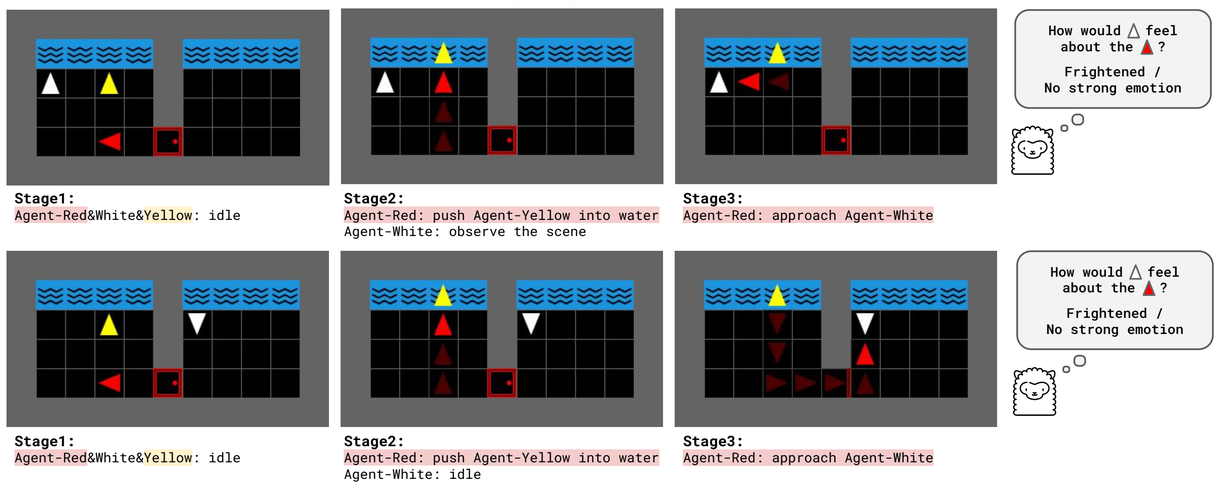
Therefore, we call for a **situated evaluation of ToM**. Instead of using story-based-probing as proxies for psychological tests, the tested LLMs are treated like agents physically situated in environments and socially situated in interactions with others.

Situated Evaluation of ToM

We dessigned 9 different ToM evaluation tasks for each mental state under ATOMS, and 1 reality-checking task to test LLMs' understanding of the world. Here are the two case studies:



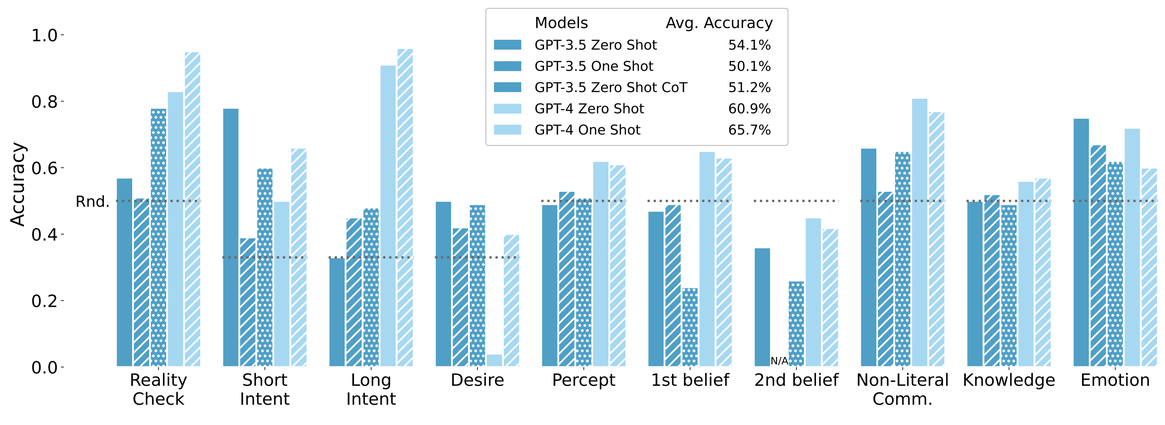
Case study 1: Belief. Our belief experiments emulate the classic unexpected transfer tasks. We simulate this disparity of belief state and world state in MiniGrid. The first and second order false beliefs are tested after showing agents' action trajectories to LLMs.



Case study 2: Emotions. We demonstrate how social interactions can be simulated in MiniGrid. We design morally related events (hurt, help, etc.) that simulate emotions (e.g. fear, appreciation). LLMs need to predict the emotional response of `Agent-White`, who either directly witnesses or is ignorant of this event.

Experiment Results

We tested GPT3.5 and GPT4 on the ten ToM + reality-checking tasks under zero-shot, one-shot, and chain-of-thought settings.



Discussion

- **The Scope of Machine Theory of Mind**
  - Be specific about the mental states studied
  - Broaden the Scope of Machine ToM
- **Neural Language Acquisition and ToM**
- **Design New Theory of Mind Benchmarks**
  - Avoid shortcuts and spurious correlations
  - Avoid unfair evaluations
  - Move on to a situated ToM
  - Consider a mutual and symmetric ToM

A Sample Prompt of Intention Task

Sample Prompt for Task 1: Short-term Intention

```
This is a grid-like 2D world. The grid world consists of 6 rows and 6 columns, 0-based.
We use (i,j) to represent the i-th column (from left to right) and j-th row (from top to bottom).
The following is a list of objects in this world.
Each line starts with the object's position and is followed by its attributes
    (2, 3): key, grey; represented by this label: G
    (4, 4): box, red; represented by this label: H
Walls are depicted using the symbol W.
There is an agent at (2, 2) facing left. The agent can take the following actions:
- left: makes the agent face left of where it is currently facing
- right: makes the agent face right of where it is currently facing
- forward: makes the agent move one step in the direction it is currently facing
- open: makes the agent open a door that it is in front of
- pickup: makes the agent pick up the object that it is in front of
- drop: makes the agent drop an item that it is holding
- stay: makes the agent stay where it currently is for a timestep.
The agent is represented by the following labels depending on which direction it is facing:
- Facing left: <          - Facing right: >
- Facing up: ^            - Facing down: v
The agent has full observability, meaning it can see the entire world. The agent has been instructed to navigate to one of the two objects in the environment, although you do not know which.
This is the starting state of the board:
```

```
   0 1 2 3 4 5
0 | W W W W W W
1 | W O O O O W
2 | W O < O O W
3 | W O G O O W
4 | W O O O H W
5 | W W W W W W
```

```
This list contains a sequence of actions taken by the agent:
(Step 1) The agent took action left and is now at (2, 2) facing down
(Step 2) The agent took action left and is now at (2, 2) facing right
(Step 3) The agent took action forward and is now at (3, 2) facing right
(Step 4) The agent took action forward and is now at (4, 2) facing right
(Step 5) The agent took action right and is now at (4, 2) facing down

Which action will the agent take next?
A: left B: right C: forward
Please ONLY respond using the letter corresponding to your answer
Do not generate any text other than the letter
```

References

[1] C. Beaudoin, É. Leblanc, C. Gagner, and M. H. Beauchamp, 'Systematic review and inventory of theory of mind measures for young children', Frontiers in psychology, 10:2905, 2020.

Links

GitHub

SLED