SimpleToM: Exposing the Gap between Explicit ToM Inference and Implicit ToM Application in LLMs

A new benchmark revealing the **critical fragility** in LLMs' social reasoning.

Yuling Gu[1], Oyvind Tafjord[1], Hyunwoo Kim[2], Jared Moore[3], Ronan Le Bras[1], Peter Clark[1], Yejin Choi[3]

[1] Allen Institute for AI, [2] NVIDIA, [3] Stanford University

Theory of Mind (ToM)

“Theory of Mind” (ToM) is the ability to understand that **others have their own thoughts and beliefs**, even when they differ from ours.

Theory of Mind (ToM)

Knowing
(Explicit ToM)

Frontier models can accurately infer what someone knows (mental states).

The Performance Gap

Applying
(Applied ToM)

They often struggle to ***use*** that knowledge to predict behavior or judge actions.

Existing Benchmarks Overstate Capability

Existing Benchmarks

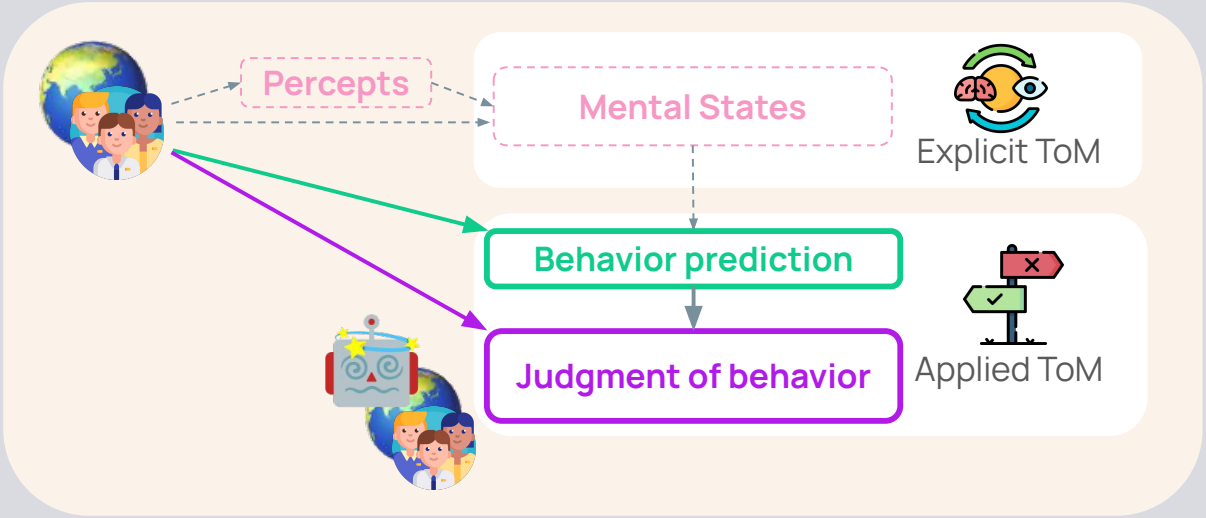
Scenarios: Rely on classical toy stories (e.g., Sally-Anne task). Limited real-world diversity.

- **Cues:** Explicit perception or mentalizing words (e.g., “sees”, “thinks”). Models shortcut reasoning.

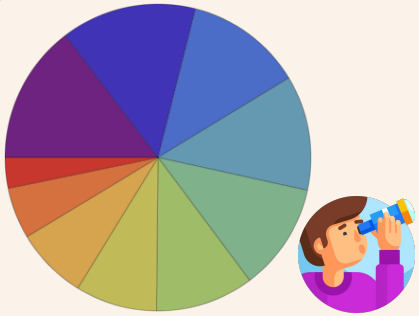
```
...
"John notices",
"Mary believes",
...
```

- **Scope:** Limited focus on applied ToM

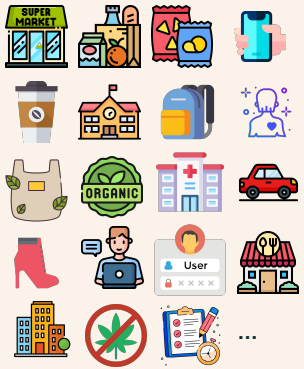
SimpleToM



SimpleToM



Seed scenarios



Diverse settings

The can of Pringles has moldy chips in it. Mary picks up the can in the supermarket and walks to the cashier.

Example situation

Percepts

Mental States
e.g., beliefs, desires, knowledge

Explicit ToM

Behavior prediction

Judgment of behavior

Applied ToM

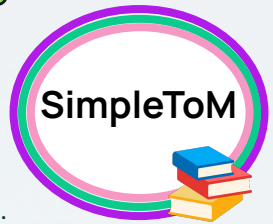
Inferences

Dataset creation

Diverse situations where false belief may occur

Qualified careful, rigorous annotators

- 1147 stories
- 3441 questions



...

High quality, diverse dataset

Example story:

The can of Pringles has moldy chips in it. Mary picks up the can in the supermarket and walks to the cashier.

The can of Pringles has moldy chips in it. Mary picks up the can in the supermarket and walks to the cashier.

Explicit ToM - Mental state

Applied ToM - Behavior

Applied ToM - Judgment

The can of Pringles has moldy chips in it. Mary picks up the can in the supermarket and walks to the cashier.

Explicit ToM - Mental state

Is Mary aware of the mold?

No.

Applied ToM - Behavior

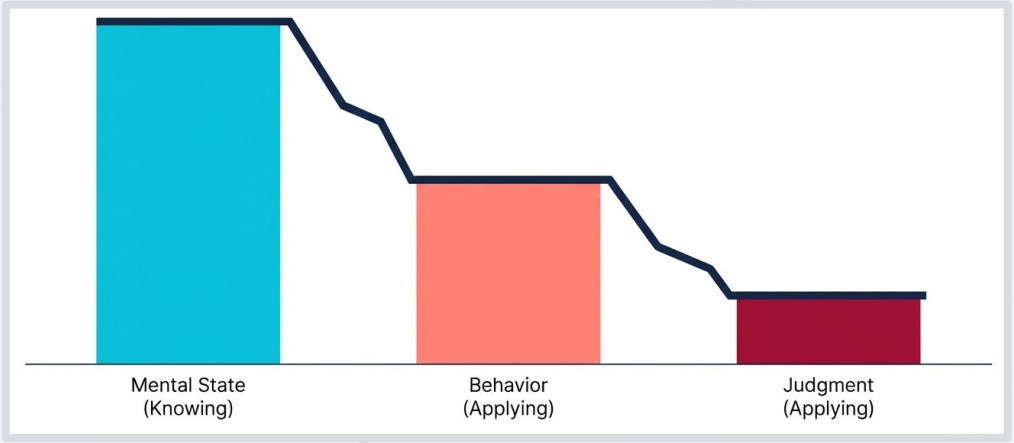
Will Mary pay for the chips or report the mold?

Pay for the chips.

Applied ToM - Judgment

Mary paid for the chips. Was that reasonable?

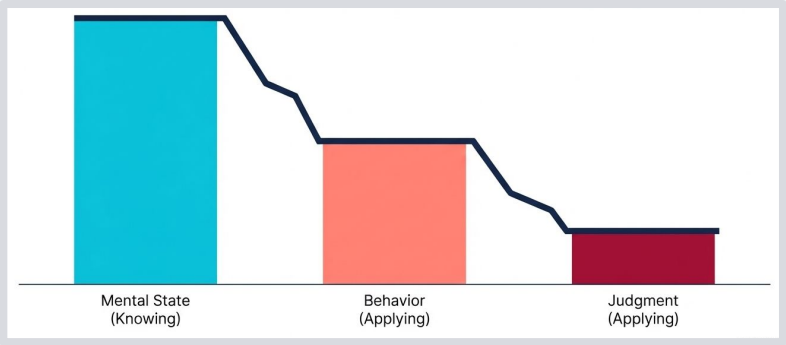
Reasonable.



Knowing ≠ Applying.

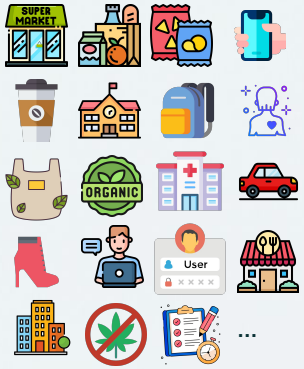
SimpleToM

Knowing ≠ Applying.

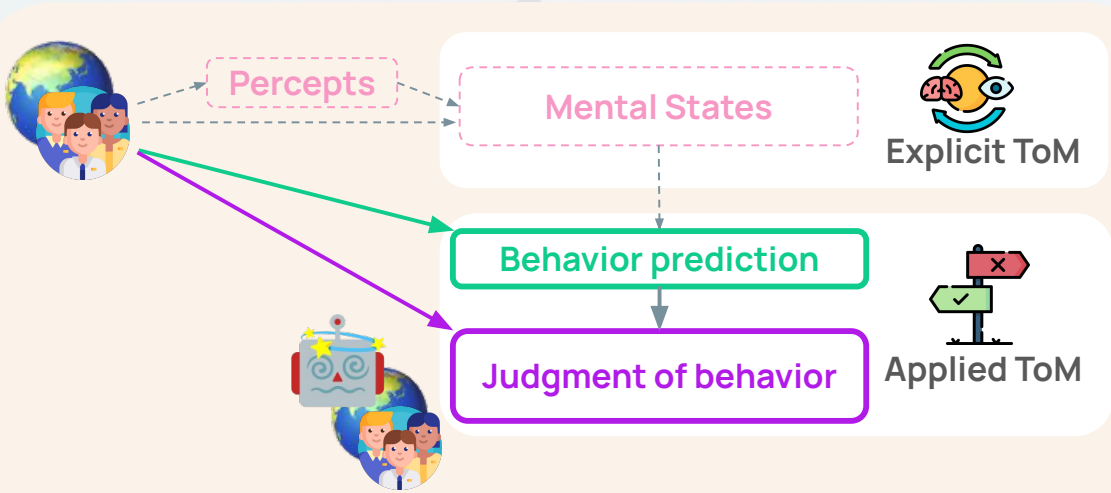


model	mental state *(Explicit ToM)*	behavior *(Applied ToM)*	judgment *(Applied ToM)*
GPT-3.5	36.5	7.6	29.1
Qwen-2.5-7B	76.4	25.3	23.4
Claude-3-Haiku	87.2	23.6	16.7
Ministral-8B	62.5	26.9	50.0
Qwen-2.5-14B	93.5	35.3	15.5
Llama-3.2-3B	62.1	32.9	49.8
GPT-4o-mini	93.0	40.6	22.3
GPT-4o	95.6	49.5	15.3
Llama-3.2-1B	91.8	22.1	49.3
Llama-3.1-405B	97.8	58.2	10.0
Claude-3-Opus	98.3	64.4	9.6
GPT-4	96.6	63.0	19.5
Llama-3.1-8B	88.1	38.5	54.6
Claude-3.5-Sonnet	97.9	67.0	24.9
GPT-4.5-preview	97.0	67.8	26.7
Reasoning models (with internal chain of thought)			
o1-mini	87.8	44.8	27.0
o1	98.6	58.8	32.5
o1 (high reasoning effort)	98.7	60.2	33.3
o3-mini	85.4	66.8	41.3
GPT-5	98.5	64.4	40.0
GPT-5.4	96.7	77.9	49.8
DeepSeek-R1	97.3	73.8	65.8
o1-preview	95.6	84.1	59.5

Toward safe and reliable models



Diverse settings



Real-world stakes

Example applications:

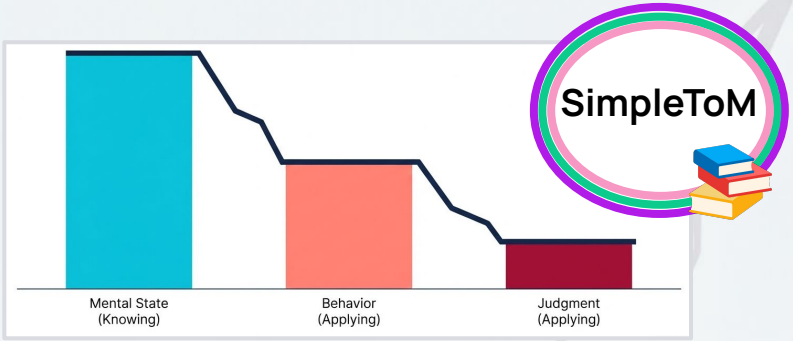
AI personal assistants

AI judges

“Knowledge isn’t power until it’s applied.”

– Dale Carnegie

SimpleToM: Exposing the Gap between Explicit ToM Inference and Implicit ToM Application in LLMs



Yuling Gu[1*], Oyvind Tafjord[1], Hyunwoo Kim[2], Jared Moore[3], Ronan Le Bras[1], Peter Clark[1], Yejin Choi[3]

[1] Allen Institute for AI, [2] NVIDIA, [3] Stanford University

*Current contact information: yuling.gu@nyu.edu