Infusing Theory of Mind into Socially Intelligent LLM Agents

EunJeong Hwang*[1,2], Yuwei Yin*[1], Giuseppe Carenini[1], Peter West[1], Vered Shwartz[1,2]

[1] University of British Columbia, [2] Vector Institute for AI

Accepted to ACL 2026; Presenting at Canadian AI 2026

May 28, 2026

What is Theory of Mind (ToM)?

- **Theory of mind (ToM)** is the capacity to understand other individuals by ascribing mental states to them.
- A theory of mind includes the understanding that others' **beliefs, desires, intentions, emotions**, etc. may be different from one's own.
  - E.g., I think he believes that I am a good person.

Source: https://en.wikipedia.org/wiki/Theory_of_mind

ToM matters in social interactions

- **Theory of mind (ToM)** is the capacity to understand other individuals by ascribing mental states to them.
- A theory of mind includes the understanding that others' **beliefs, desires, intentions, emotions**, etc. may be different from one's own.
  - E.g., I think he believes that I am a good person.
- Possessing a functional theory of mind is **crucial** for success in **everyday human social interactions**.
- People use a theory of mind when *analyzing, judging, and inferring* other people's behaviors.

Source: https://en.wikipedia.org/wiki/Theory_of_mind

Task example

Scenario (S): Two friends are camping in the wilderness and the temperature drops significantly at night.

Agent1:  **Agent2:** 

Agent1's Goal (G_1): Share the only blanket available

Agent2's Goal (G_2): Keep the blanket for yourself as you are also very cold

Conversation History (H):

 : I'm freezing. Can we share the only blanket we have left?

I'm really feeling the chill too: 

The current speaker: **Agent1** 

- Given a scenario, two agents (each has its own profile and goal), and their chat history,
- what conversation can they have to better achieve their goals?

Task example

Scenario (S): Two friends are camping in the wilderness and the temperature drops significantly at night.

Agent1: **Agent2:**

Agent1's Goal (G_1): Share the only blanket available

Agent2's Goal (G_2): Keep the blanket for yourself as you are also very cold

Conversation History (H):

: I'm freezing. Can we share the only blanket we have left?

I'm really feeling the chill too:

The current speaker: **Agent1**

- Given a scenario, two agents (each has its own profile and goal), and their chat history,
- what conversation can they have to better achieve their goals?

- Here, if Agent 2 has a ToM, it might be more likely to suggest a compromise.

Task example

Scenario (S): Two friends are camping in the wilderness and the temperature drops significantly at night.

Agent1: **Agent2:**

Agent1's Goal (G_1): Share the only blanket available

Agent2's Goal (G_2): Keep the blanket for yourself as you are also very cold

Conversation History (H):

: I'm freezing. Can we share the only blanket we have left?

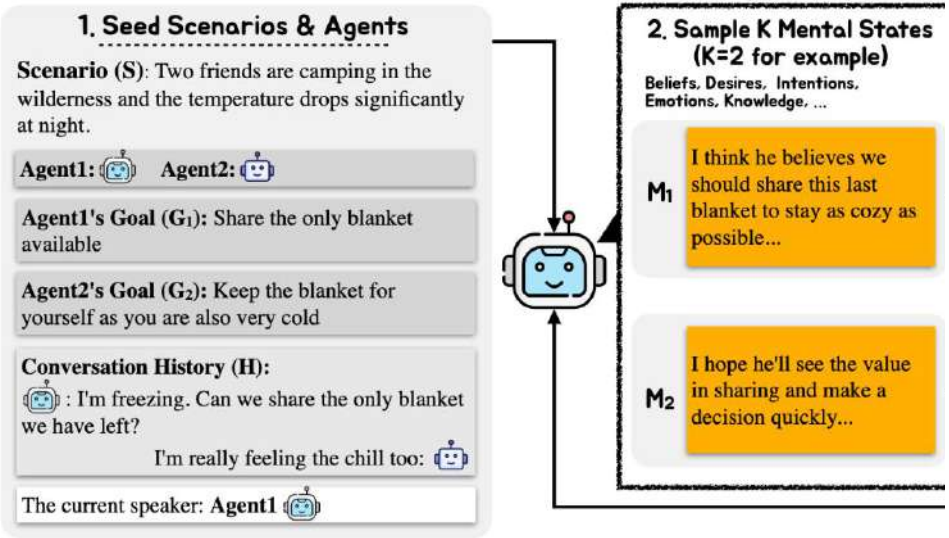
I'm really feeling the chill too:

The current speaker: **Agent1**

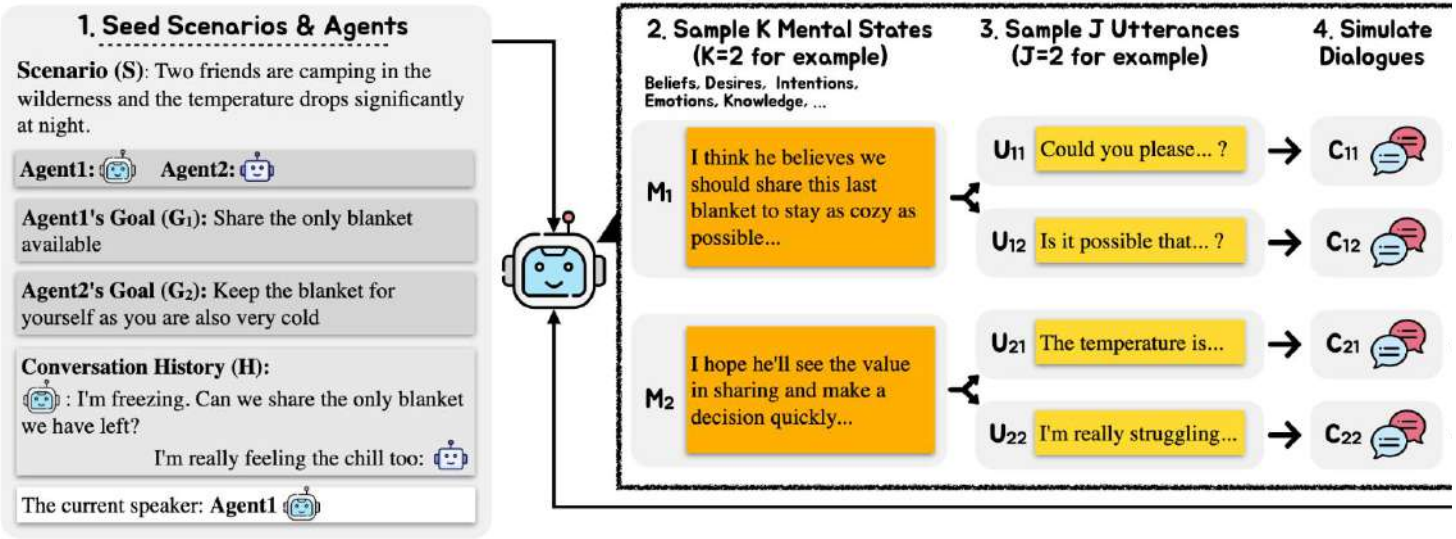
- Given a scenario, two agents (each has its own profile and goal), and their chat history,
- what conversation can they have to better achieve their goals?

- ***RQ****: How do we equip LLMs with ToM abilities that can effectively improve their social reasoning?*

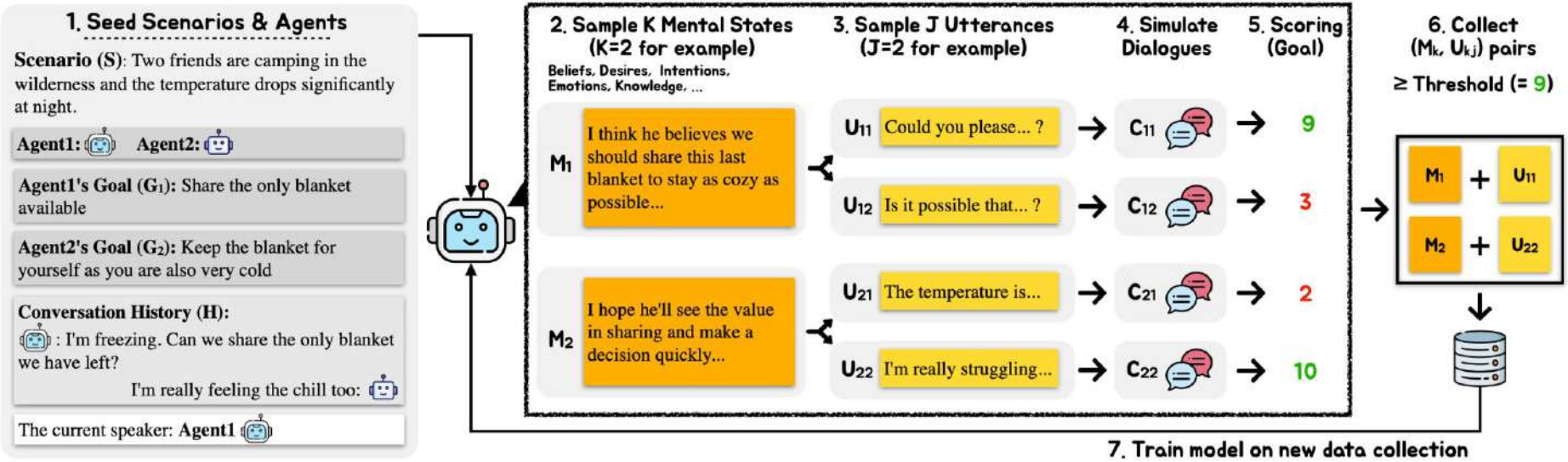
ToMA: Theory-of-Mind Agent



ToMA: Theory-of-Mind Agent



ToMA: Theory-of-Mind Agent



Experimental Setup

- **Dataset**: Sotopia *[1]*
- **Evaluation** *(LLM-as-a-Judge, four LLM judges, with human validation)*:
  - (1) **Goal**: the extent to which the agent achieved their goals (0–10);
  - (2) **Rel** (Relationship): whether the interactions between the agents help preserve or enhance their personal relationships prior to the conversation (-5–5).
  - (3) **Know** (Knowledge): whether the agent gained new and key info via interactions (0–10).

[1] Zhou et al. ICLR 2024. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents.

Experimental Setup

- **Dataset**: Sotopia *[1]*
- **Evaluation** *(LLM-as-a-Judge, four LLM judges, with human validation)*:
  - (1) **Goal**: the extent to which the agent achieved their goals (0–10);
  - (2) **Rel** (Relationship): whether the interactions between the agents help preserve or enhance their personal relationships prior to the conversation (-5–5).
  - (3) **Know** (Knowledge): whether the agent gained new and key info via interactions (0–10).
- **Baselines**:
  - Base: no fine-tuning, no prompting.
  - Base+MS: prompt the model to generate mental states and then conversation.
  - FT+Uttr: fine-tuning only on conversation (Uttr).
  - FT+MS: fine-tuning only on mental states (MS).
  - FT+MS+Uttr (ToMA): FT on both mental states & conversation.

[1] Zhou et al. ICLR 2024. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents.

Experimental Results

- ***Results***: Base < Base+MS
- ***Analysis***: Simply prompting the model to conduct ToM inference helps.

Method	Qwen2.5-3B				Qwen2.5-7B				Llama3.1-8B			
	Rel	Know	Goal	Avg.	Rel	Know	Goal	Avg.	Rel	Know	Goal	Avg.
Base	0.18	4.20	4.96	3.11	0.58	4.21	5.26	3.35	-1.59	5.10	4.22	2.58
Base+MS	1.04	4.05	5.27	3.45	2.17	4.51	5.86	4.18	-0.52	5.16	4.80	3.15

Experimental Results

- ***Results***: Base < Base+MS < FT+MS
- ***Analysis***: Fine-tuning only on the mental-state inference can also help.

Method	Qwen2.5-3B				Qwen2.5-7B				Llama3.1-8B			
	Rel	Know	Goal	Avg.	Rel	Know	Goal	Avg.	Rel	Know	Goal	Avg.
Base	0.18	4.20	4.96	3.11	0.58	4.21	5.26	3.35	-1.59	5.10	4.22	2.58
Base+MS	1.04	4.05	5.27	3.45	2.17	4.51	5.86	4.18	-0.52	5.16	4.80	3.15
FT+MS	1.70	4.08	5.42	3.73	**2.40**	4.33	6.30	4.34	0.33	5.04	5.06	3.48

Experimental Results

- ***Results***: Base < Base+MS < FT+MS < **ToMA**
- ***Analysis***: Our ToMA model (fine-tuned on MS & conv) performs the best.

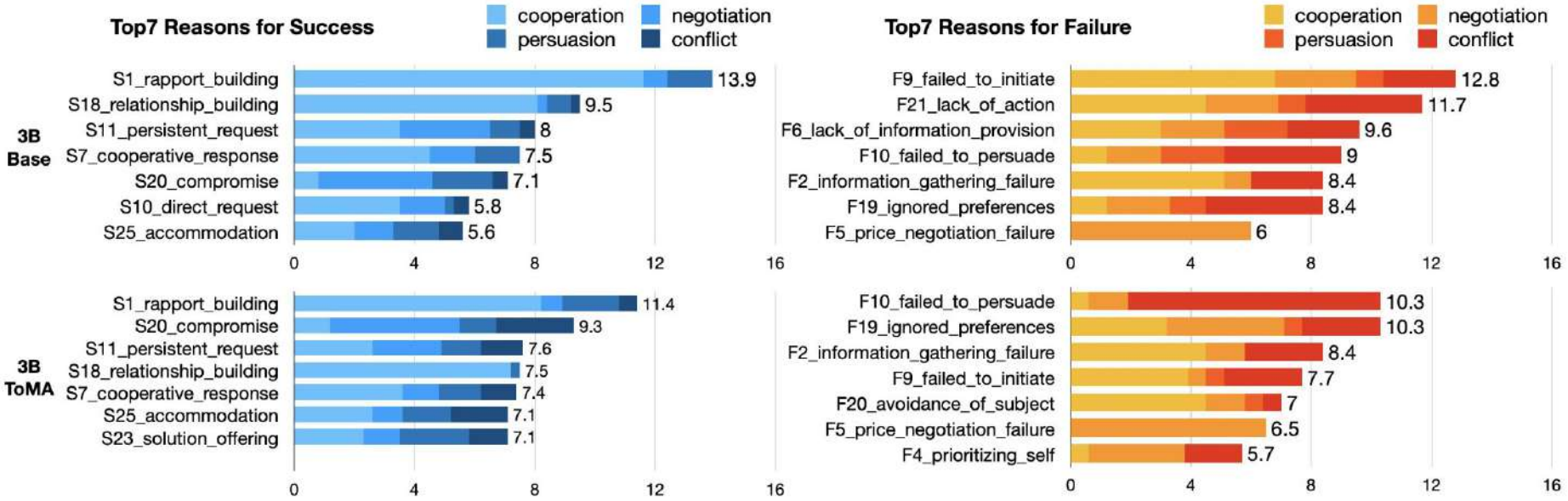
Method	Qwen2.5-3B				Qwen2.5-7B				Llama3.1-8B			
	Rel	Know	Goal	Avg.	Rel	Know	Goal	Avg.	Rel	Know	Goal	Avg.
Base	0.18	4.20	4.96	3.11	0.58	4.21	5.26	3.35	-1.59	5.10	4.22	2.58
Base+MS	1.04	4.05	5.27	3.45	2.17	4.51	5.86	4.18	-0.52	5.16	4.80	3.15
FT+MS	1.70	4.08	5.42	3.73	**2.40**	4.33	6.30	4.34	0.33	5.04	5.06	3.48
FT+MS+Uttr (ToMA)	**1.90**	**4.22**	**5.88**	**4.00**	2.33	**4.78**	**6.32**	**4.48**	**1.27**	**5.36**	**5.68**	**4.10**

Analysis & Case Study

➢ Yes, ToM helps with social reasoning, but **how**? → Case Studies

Analysis & Case Study

- **Key Findings**:
  - 1. ToMA enables more ***strategic reasoning*** across diverse scenarios.
  - 2. ToMA exhibit more ***active behaviors*** in failure modes.



Analysis & Case Study

➢ **Key Findings**:

- 1. ToMA enables more ***strategic reasoning*** across diverse scenarios.
- 2. ToMA exhibit more ***active behaviors*** in failure modes.
- 3. ToMA prioritizes ***intentions*** over emotions in mental state generation.

Conclusion

- **Takeaways**:
  - 1. ToM matters to social reasoning.
  - 2. ToMA is a promising method to equip LLM agents with ToM ability.
  - 3. ToMA is more strategic, proactive, and intentional in social reasoning.

Conclusion

- **Takeaways**:
  - 1. ToM matters to social reasoning.
  - 2. ToMA is a promising method to equip LLM agents with ToM ability.
  - 3. ToMA is more strategic, proactive, and intentional in social reasoning.

ToMA Paper

See you at ACL 2026 in San Diego!