ACL 2026 SAN DIEGO JULY 2–7

THE UNIVERSITY OF BRITISH COLUMBIA
Computer Science
Faculty of Science

UBC NLP

VECTOR INSTITUTE

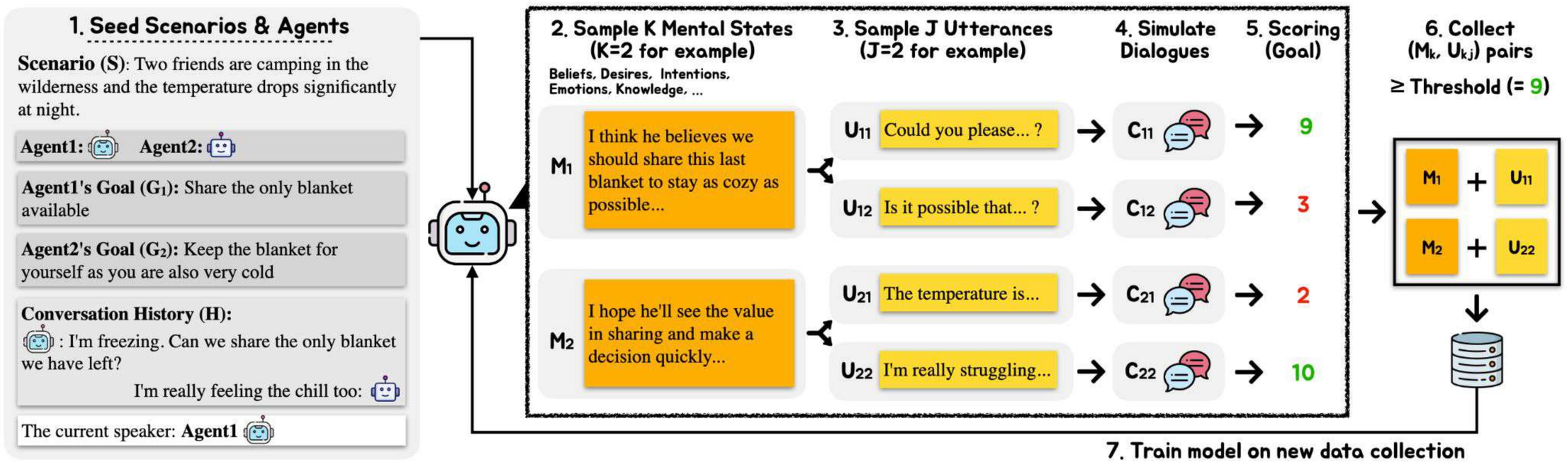
Infusing Theory of Mind into Socially Intelligent LLM Agents

EunJeong Hwang, Yuwei Yin*, Giuseppe Carenini, Peter West, Vered Shwartz*

***Abstract**: Theory of Mind (ToM)-an understanding of the mental states of others-is a key aspect of human social intelligence, yet, chatbots and LLM-based social agents do not typically integrate it. In this work, we demonstrate that LLMs that explicitly use ToM get better at dialogue, achieving goals more effectively. After showing that simply prompting models to generate mental states between dialogue turns already provides significant benefit, we further introduce **ToMAgent (ToMA)**, a ToM-focused dialogue agent. ToMA is trained by pairing ToM with dialogue lookahead to produce mental states that are maximally useful for achieving dialogue goals. Experiments on the Sotopia interactive social evaluation benchmark demonstrate the effectiveness of our method over a range of baselines. Comprehensive analysis shows that ToMA exhibits more strategic, goal-oriented reasoning behaviors, which enable long-horizon adaptation, while maintaining better relationships with their partners. Our results suggest a step forward in integrating ToM for building socially intelligent LLM agents.*

***Research Questions**: Does equipping LLMs with Theory-of-Mind abilities improve their social reasoning? If so, how?*

Method: Theory-of-Mind Agent (ToMA)



Experimental Setups

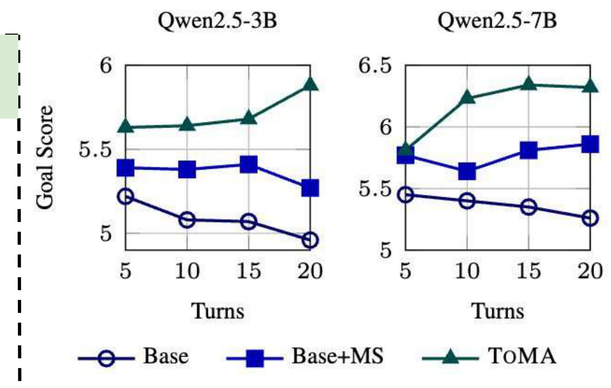
Dataset: Sotopia (All & Hard sets) → Goal-oriented conv between two agents under scenarios.
Evaluation (*LLM-as-a-Judge*: GPT5 & Gemini & DeepSeek & Qwen, with human validation):
(1) ***Goal***: the extent to which the agent achieved their goals (0–10);
(2) ***Rel*** (Relationship): whether the interactions 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 through interactions (0–10).

Baselines:
Base: no fine-tuning, no prompting.
Base+MS: generate MS and then conv.
FT+Uttr: fine-tuning only on conv.
FT+MS: fine-tuning only on MS.
FT+MS+Uttr (ToMA): FT on MS & conv.

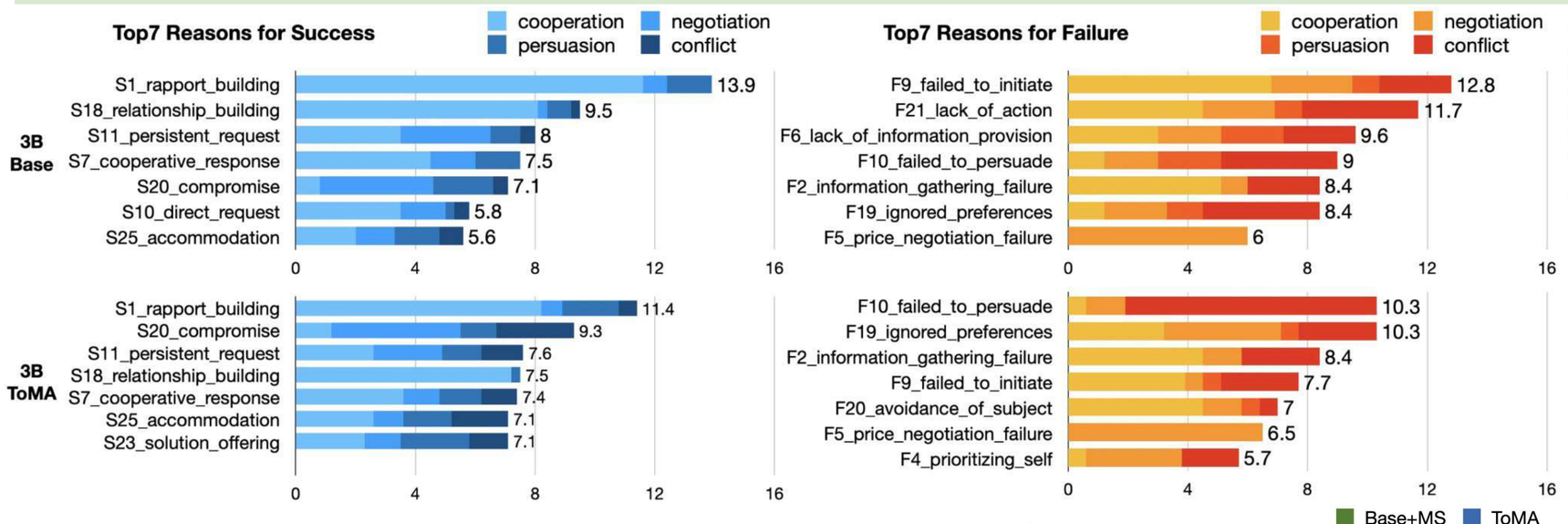
Results: ToM helps with Social Reasoning

Method	Qwen2.5-3B Rel	Know	Goal	Avg.	Qwen2.5-7B Rel	Know	Goal	Avg.	Llama3.1-8B 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+Uttr	1.22	4.10	5.23	3.52	1.36	4.43	5.70	3.83	-0.35	4.91	4.85	3.13
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**

Key Findings:
(Avg.: ToMA > FT+MS > Base+MS > Base)
1. Mental-state conditioning (+MS) boosts social reasoning.
2. Our proposed ToMA model outperforms the baselines.
3. Theory of Mind enables *long-horizon adaptation*.



Analysis: What Strategies Does ToMA Employ?



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.
4. ToMA generates more *1st-order mental states* than the baseline.

Key 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.

Paper

Code