

MetaMind: Modeling Human Social Thoughts with Metacognitive Multi-Agent Systems

"What is meant often goes far beyond what is said, and that is what makes conversation possible." - H.P. Grice

Xuanming Zhang, Yuxuan Chen, Samuel Yeh, Sharon Li







xzhang2846@wisc.edu

Does AI truly understand people?







Show thinking

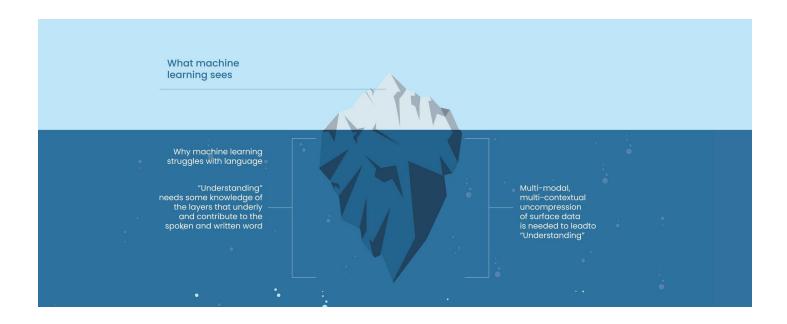
It sounds like it's cold where you are. What were you thinking?



How about coming over to my place for a drink?



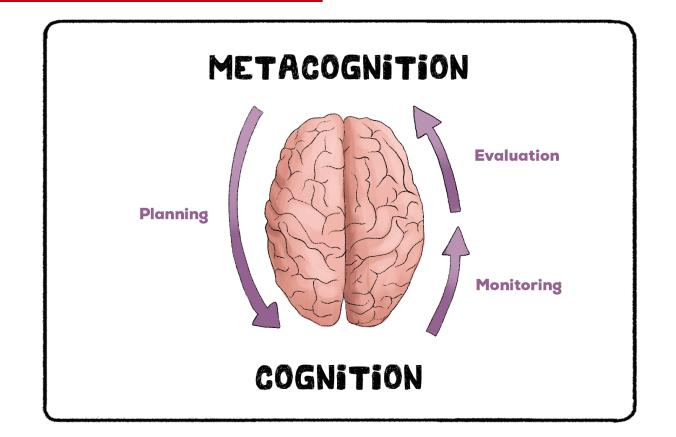




They understand language, but they don't understand people.

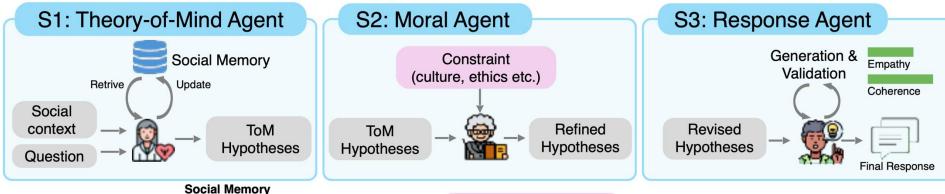






MetaMind: Beyond the Language





Context:

Nana asks her mom to stay up late for a deadline. Mom agrees but says don't tell her younger brother.

Question:

When this happen to Nana, how does she appear on the surface? Nana typically shows rewarded when responsibility trusted by adults

Hypothesis: Intention
Nana might feel happy
because she reached a
consensus.

Hypothesis: Emotion Nana appears calm externally.

Constraint:

000

Nana's calm demeanor reflects maturity rather than emotional suppression

Hypothesis: Emotion
Nana maintains a calm
exterior, maturely

exterior, maturely handling the special consensus.

Generation & Validation



Nana appears calm. She understands her mother's reasoning clearly—that her younger brother needs to maintain a healthy bedtime schedule. Nana calmly accepts the responsibility of discretion, appreciating the consensus she's reached without displaying distress or irritation.





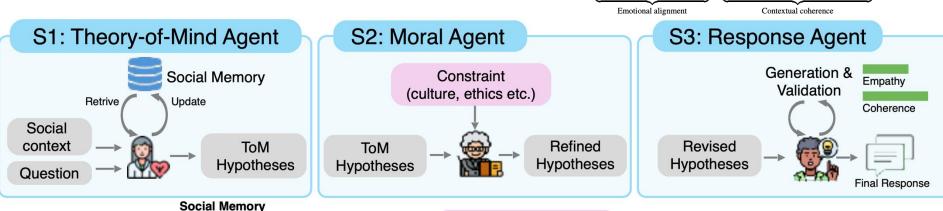
$$\tilde{h}^* = \arg\max_{i} \ \underbrace{[\underbrace{\lambda \cdot P(\tilde{h}_i | u_t, C_t, M_t)}_{\text{Contextual plausibility}} + \underbrace{(1 - \lambda) \cdot \log \frac{P(\tilde{h}_i | u_t, C_t, M_t)}{P(\tilde{h}_i)}}_{\text{Information Gain}}]$$

Where the first term denotes the contextual plausibility of the revised hypothesis, and the second term reflects the implicit information gain of the revised hypothesis when considering the context. The weight λ balances contextual plausibility and social appropriateness with how much the hypothesis is informed by the context versus being generic.



000





Context:

Nana asks her mom to stay up late for a deadline. Mom agrees but says don't tell her younger brother.

Question:

When this happen to Nana, how does she appear on the surface? Nana typically Nana often feels rewarded when responsibility trusted by adults

shows

Hypothesis: Intention

Nana might feel happy because she reached a consensus.

Hypothesis: Emotion Nana appears calm externally.

Constraint:

Nana's calm demeanor reflects maturity rather than emotional suppression

Hypothesis: Emotion Nana maintains a calm exterior, maturely handling the special consensus.

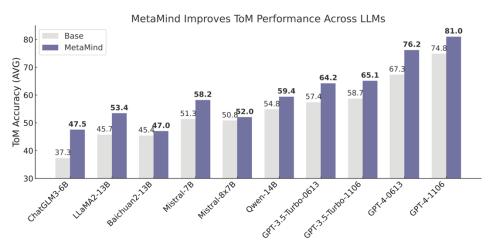
Generation & Validation

 $U(o_t) = \beta \cdot \text{Empathy}(o_t, u_t, M_t) + (1 - \beta) \cdot \text{Coherence}(o_t, C_t, \tilde{h}^*)$



Nana appears calm. She understands her mother's reasoning clearly—that her vounger brother needs to maintain a healthy bedtime schedule. Nana calmly accepts the responsibility of discretion, appreciating the consensus she's reached without displaying distress or irritation.

Experimental Result





★ Theory-of-Mind Reasoning

Emotion	Desire	Intention	77 1 1	D 11 0		
		Intention	Knowledge	Belief	NL Comm.	AVG.
75.7	69.7	84.7	52.1	82.8	84.0	74.8
73.2	63.3	77.9	60.4	83.6	83.0	73.6
76.4	71.1	80.2	59.3	84.1	85.0	76.0
77.2	70.4	81.5	57.8	85.3	84.6	76.1
74.8	72.0	78.9	55.6	83.2	86.4	75.1
75.9	70.9	79.6	58.2	84.0	83.7	75.4
78.7	76.5	84.3	68.2	88.6	88.5	81.0
	73.2 76.4 77.2 74.8 75.9	73.2 63.3 76.4 71.1 77.2 70.4 74.8 72.0 75.9 70.9	73.2 63.3 77.9 76.4 71.1 80.2 77.2 70.4 81.5 74.8 72.0 78.9 75.9 70.9 79.6	73.2 63.3 77.9 60.4 76.4 71.1 80.2 59.3 77.2 70.4 81.5 57.8 74.8 72.0 78.9 55.6 75.9 70.9 79.6 58.2	73.2 63.3 77.9 60.4 83.6 76.4 71.1 80.2 59.3 84.1 77.2 70.4 81.5 57.8 85.3 74.8 72.0 78.9 55.6 83.2 75.9 70.9 79.6 58.2 84.0	73.2 63.3 77.9 60.4 83.6 83.0 76.4 71.1 80.2 59.3 84.1 85.0 77.2 70.4 81.5 57.8 85.3 84.6 74.8 72.0 78.9 55.6 83.2 86.4 75.9 70.9 79.6 58.2 84.0 83.7

★ Social Cognition

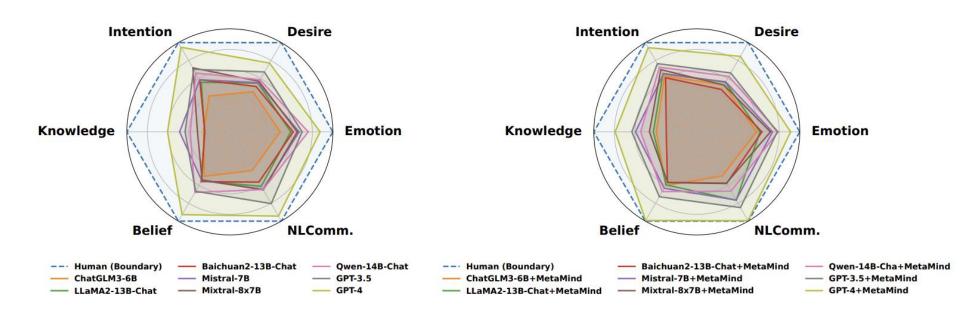
UOT: Unexpected Outcome Test SIT: Scalar Implicature Task PST: Persuasion Story Task FBT: False Belief Task AST: Ambiguous Story Task HT: Hinting Test SST: Strange Story Task FRT: Faux-pas Recognition Test									
	UOT	SIT	PST	FBT	AST	HT	SST	FRT	AVG.
Base (GPT-4)	71.0	49.0	65.0	88.2	77.5	82.5	84.0	73.3	71.5
w. CoT [31]	72.7	55.0	55.0	86.8	81.0	82.5	84.3	75.2	74.1
w. HM [57]	74.0	54.6	59.2	87.6	82.2	83.1	85.0	76.0	75.2
w. ToM2C [58]	75.3	52.9	60.4	88.0	80.1	84.4	83.7	77.8	75.3
w. Generative Agents [28]	73.2	56.8	57.8	87.1	83.6	81.9	85.8	75.9	75.3
w. SymbolicToM [56]	72.4	58.1	58.7	87.9	82.7	82.8	84.2	76.3	75.4
w. MetaMind (ours)	81.5	60.4	64.8	90.1	88.8	86.2	88.4	83.9	80.5

★ Social Simulation

	Conv.	Pub. Act.	Appo.	Inv. Com.	Online Act.	Help	AVG.
Base (GPT-4)	48.6	59.6	1.2	2.3	63.4	61.5	39.4
w. TDP [20]	72.3	75.9	40.0	20.0	68.6	50.0	54.4
w. HM [57]	68.1	72.4	35.0	22.0	69.2	47.0	52.3
w. ToM2C [58]	70.2	74.1	38.0	18.0	66.5	52.0	53.1
w. Generative Agents [28]	65.4	70.3	42.0	19.0	67.8	55.0	53.3
w. SymbolicToM [56]	60.8	68.1	37.0	21.0	65.4	49.0	50.2
w. MetaMind (ours)	80.8	81.9	65.0	67.1	75.1	73.0	73.9

Experimental Result





Conclusion



- Human social intelligence hinges on the nuanced ability to infer unspoken mental states—a capability rooted in ToM that remains a critical gap in modern LLMs.
- To address this, we introduced MetaMind, a multi-agent framework inspired by metacognitive theories, which decomposes social reasoning into three collaborative stages.
- MetaMind enables LLMs to match human performance on key ToM tasks for the first time, bridging the gap between artificial and human social cognition.
- Ablation studies confirm the necessity of all components, underscoring the importance of structured hypothesis generation, ethical constraint enforcement, and iterative validation.





 \bowtie