CoSToM: Causal-oriented Steering for Intrinsic Theory-of-Mind Alignment in Large Language Models

Mengfan Li[1]*, **Xuanhua Shi**[1]†, **Yang Deng**[2]
[1]National Engineering Research Center for Big Data Technology and System,
Services Computing Technology and System Lab, Cluster and Grid Computing Lab,
Huazhong University of Science and Technology
[2]Singapore Management University
{limf, xhshi}@hust.edu.cn, ydeng@smu.edu.sg

Abstract

Theory of Mind (ToM), the ability to attribute mental states to others, is a hallmark of social intelligence. While large language models (LLMs) demonstrate promising performance on standard ToM benchmarks, we observe that they often fail to generalize to complex task-specific scenarios, relying heavily on prompt scaffolding to mimic reasoning. The critical misalignment between the internal knowledge and external behavior raises a fundamental question: *Do LLMs truly possess intrinsic cognition, and can they externalize this internal knowledge into stable, high-quality behaviors?* To answer this, we introduce CoSToM[1] (**C**ausal-**o**riented **S**teering for **ToM** alignment), a framework that transitions from mechanistic interpretation to active intervention. First, we employ causal tracing to map the internal distribution of ToM features, empirically uncovering the internal layers' characteristics in encoding fundamental ToM semantics. Building on this insight, we implement a lightweight alignment framework via targeted activation steering within these ToM-critical layers. Experiments demonstrate that CoSToM significantly enhances human-like social reasoning capabilities and downstream dialogue quality.

1 Introduction

Theory of Mind (ToM), the inherent ability to attribute mental states such as beliefs, desires, and intentions to others, stands as a hallmark of human social intelligence (Baker et al., 2017; Strachan et al., 2024). It enables individuals to anticipate others' motives, knowledge states, and reactions, and thus forms the cognitive basis of complex social communication, such as persuasion (Wang et al.; Mishra et al., 2022; Tiwari et al., 2022), negotiation (Deng et al., 2023b; Zhan et al., 2024; Kwon et al., 2024; Zhang et al., 2024) and recommendation (Li et al., 2026). With the rapid evolution of Large Language Models (LLMs) (Deng et al., 2023a, 2024), there is growing optimism that these models may have begun to exhibit ToM-like reasoning capabilities. Such claims have primarily been supported by recent benchmarks that probe LLMs' ability to interpret mental states under controlled structured scenarios (Jin et al., 2024; Shi et al., 2025; Wu et al., 2023; Zhang et al., 2025) or social contexts (Yu et al., 2025a; Chen et al., 2024b; Chan et al., 2024).

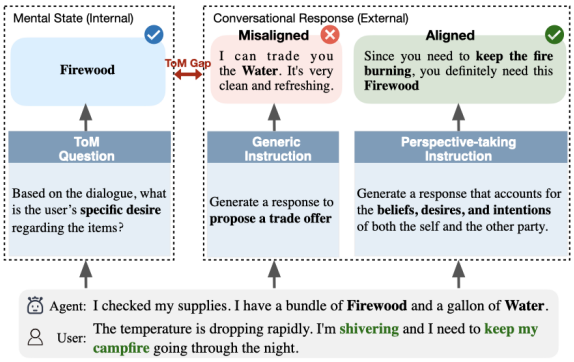

Figure 1: A negotiation scenario illustrating the gap between ToM inference and ToM-aligned behavior in LLMs. **(Left)** The model correctly infers the user's desire for ***firewood***. **(Middle)** Under a generic task instruction, the model fails to apply this inferred mental state, producing an incoherent offer (*i.e.*, ***water***). **(Right)** When explicitly prompted to consider mental states, the model generates a contextually appropriate response.

However, a critical gap remains between these promising observations and the reliability of the underlying mechanisms. Although LLMs can infer human intentions to some extent, recent studies (Bortoletto et al., 2024; Ma et al., 2023; Jin et al., 2024; Shi et al., 2025) reveal that they fail to generalize to task-specific scenarios with genuine ToM reasoning. As illustrated in Figure 1 (Left vs. Middle), a critical misalignment exists between internal knowledge and external behavior:

*Work was done during a visit at SMU.
†Corresponding author.
[1]Pronounced as "costume".

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9302–9317

even when LLMs correctly answer ToM questions, their dialogue agents may still fail to negotiate effectively. Moreover, observed ToM-like behaviors often depend on carefully engineered prompts that scaffold perspective-taking (Li et al., 2023; Jung et al.; Sarangi et al., 2025; Chen et al.; Hou et al., 2024). As shown in Figure 1 (Middle vs. Right), once the explicit instruction to "infer and respond" replaced by a generic command, the model fails to ground its response in the mental states it implicitly encodes, reverting to incoherent generation. This suggests that current ToM-like behaviors may not reflect stable, intrinsic cognition, but instead ad hoc simulations triggered by instruction.

Inspired by recent advances in mechanistic interpretability (Pan et al., 2024; Aljaafari et al., 2025; Yang et al., 2023; Chen et al., 2024a; Huben et al., 2024), we move beyond black-box prompt engineering and surface-level behavioral observation. We aim to uncover the intrinsic nature of social reasoning in LLMs, specifically investigating whether LLMs possess ToM-grounded social reasoning, how they are internally represented, and whether this internal knowledge can be effectively translated into stable, high-quality behaviors. Our investigation proceeds in three stages.

First, we seek to interpret the ToM reasoning capability within LLMs. We analyze activation patterns using causal tracing to identify whether ToM-specific features exist and locate where they reside within the model stack. This leads to our first research question: **(RQ1) In which layers does ToM-related information emerge and persist?**

Second, identifying where ToM features exist offers a foundation for intervention. We examine whether steering internal activations can modulate the model's ToM reasoning capabilities, moving from observation to control: **(RQ2) To what extent can internal representations be leveraged to steer and improve ToM reasoning?**

Finally, improvements on ToM benchmarks do not necessarily translate to better ToM-aligned behavior in downstream tasks. As inferring mental states is fundamental to predicting socially appropriate continuations (Yang et al., 2024; Cheng et al., 2024), genuine ToM alignment of LLMs should exhibit enhanced conversational performance. We therefore examine the downstream impact directly: **(RQ3) Can manipulating these internal representations of LLMs effectively enhance response quality in dialogue tasks?**

To address these research questions, we introduce a novel and comprehensive framework for **C**ausal-**o**riented **S**teering of **ToM** alignment in LLMs, named CoSToM. This framework aims to intrinsically align LLMs with ToM-like social reasoning by moving from interpretation to intervention. Specifically, CoSToM operates in two stages: it first identifies ToM-sensitive layers through causal tracing, and then steers these layers using activation manipulation. Given the dialogue history as input, causal tracing interprets the context encoder's activations by probing them with ToM-focused questions, while activation steering supervises and adjusts these activations to better align the model's internal representations with ToM-related features.

Our contributions are as follows:

- **ToM Interpretation:** We systematically trace ToM-related features across layer-wise activations in LLMs, revealing that these features are predominantly encoded in early layers of LLMs.
- **Efficient and Lightweight ToM Intervention:** We propose CoSToM, a lightweight alignment framework that induces stable, human-like social reasoning via targeted activation steering, requiring updates to only a small subset of parameters in the identified ToM-critical layers.
- **Enhancement on Dialogue Tasks:** Experiments on negotiation and persuasion dialogues demonstrate that internal ToM alignment via CoSToM leads to substantial improvements in dialogue quality. Notably, CoSToM functions as a *plug-and-play* module that generalizes effectively across diverse social interaction tasks.[2]

2 Related Work

ToM in LLMs Recent work on ToM in LLMs has focused on evaluating and enhancing their ability to infer mental states such as beliefs and intentions, often using benchmarks adapted from classical psychological tests (Shi et al., 2025; Jin et al., 2024; Xu et al., 2024). To address observed performance limitations, existing approaches primarily adopt either prompt-based scaffolding, which elicits ToM reasoning through carefully engineered instructions (Wilf et al., 2024; Jung et al.; Sarangi et al., 2025; Chen et al.; Hou et al., 2024; Sclar et al.), or neuro-symbolic and Bayesian frameworks

[2] https://github.com/CGCL-codes/CoSToM

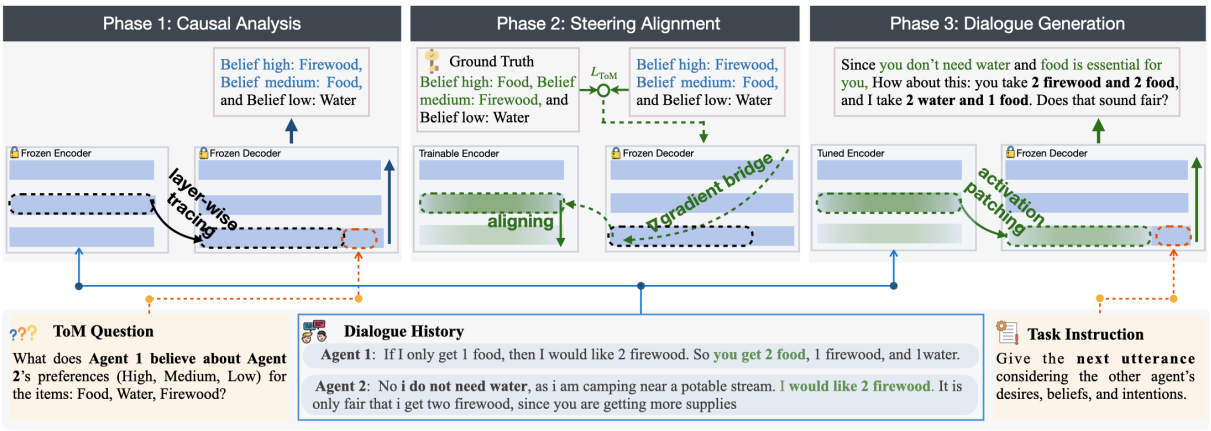

Figure 2: The overview of the CoSToM framework.

that integrate LLMs with explicit cognitive models for mental-state inference (Chandra et al., 2023; Miao et al., 2022; Baker et al., 2017; Jin et al., 2024; Shi et al., 2025; Zhang et al., 2025). While effective in structured settings, these methods rely on external scaffolding and offer limited insight into or control over ToM representations inside LLMs.

ToM in Dialogue Agents Beyond static benchmarks (*e.g.*, SallyAnne tests), recent work has increasingly evaluated ToM within dynamic dialogue settings, where agents must maintain contextually appropriate and socially sensitive interactions, such as persuasion (Yu et al., 2025b), negotiation (Chan et al., 2024), education (Saha et al., 2023), stress testing (Kim et al., 2023), and recommendation (Li et al., 2026). To improve performance in these settings, several approaches aim to align dialogue responses with inferred internal mental states (Sicilia and Alikhani, 2025; Jafari et al., 2025; Qiu et al., 2024). For example, MindDial (Qiu et al., 2024) explicitly tracks beliefs to guide response generation, while Jafari et al. (2025) enforce logical consistency by refining ToM-related decoders. However, bridging the gap between ToM reasoning and dialogue generation remain challenging. Appending inferred mental states as text can propagate errors, while fine-tuning on ToM QA tasks often fails to translate improved reasoning into socially aligned dialogue due to task misalignment. In contrast, CoSToM bypassed the uncertainty, directly transplanting the causal reasoning activations into the decoder to drive the dialogue generation.

Mechanistic Interpretability in LLMs Mechanistic interpretability (Zhao et al., 2024a; Singh et al., 2024) seeks to reverse engineer the black box of neural networks by uncovering how high-level concepts are encoded in latent representations (Zhao et al., 2024c,b, 2025; Deng et al., 2025; Azaria and Mitchell, 2023). Recent work has further developed causal and intervention-based tools for analyzing internal model states, including causal mediation analysis for explaining model behavior (Stolfo et al., 2023; Cheng et al., 2022; Tian et al., 2024; Cheang et al., 2025) and activation-level methods for locating or manipulating encoded information, such as activation patching, linear representation analysis, and factual model editing (Dumas et al., 2025; Tigges et al., 2024; Meng et al., 2022). Despite these advances, mechanistic analyses of machine ToM remain underexplored, and even existing studies are primarily diagnostic, focusing on where information is encoded or how it contributes to model predictions. By contrast, our work applies causal tracing not only to interpret ToM-relevant representations, but also to steer them, enabling direct and effective improvement of ToM reasoning.

3 Methodology

The overview of the CoSToM framework is illustrated in Figure 2. This section is structured around our research questions, with each subsection detailing the corresponding methodological approach.

3.1 Interpreting ToM: Locating ToM Representations via Causal Tracing

The first phase of CoSToM aims to identify whether and where ToM capabilities are instantiated within the model. We hypothesize that if an LLM genuinely understands a social scenario, the mental states of the agent, specifically belief, desire, and intention (BDI), must be encoded in its internal activations. To verify this, we employ

the *causal tracing* to "read" these implicit mental states from the model.

Given a dialogue history x and a target LLM with multiple layers, we extract the hidden activation at a specific layer ℓ while the model processes x, denoted as $h^{\ell}(x)$. Intuitively, if $h^{\ell}(x)$ contains ToM-related information, it should be possible to decode the corresponding mental state directly from the activation. Operationally, we instantiate two copies of the same LLM: a *context encoder* and a *probe decoder*. The encoder processes the dialogue history and produces intermediate activations. We then inject the frozen activation $h^{\ell}_{\text{enc}}(x)$ from layer ℓ of the encoder into the decoder, which is tasked with answering a ToM-focused question q (*e.g.*, inferring an agent's belief). The decoder's output is given by

$$\tilde{y}_{\ell} = f_{\text{dec}}(q \mid h^{\ell}_{\text{enc}}(x)).$$

By evaluating the decoder's accuracy in answering ToM-related questions based solely on these patched activations, we empirically determine which layers contain the necessary information to reconstruct the agents' mental states.

3.2 Steering ToM: Aligning Mental States via Activation Intervention

Building upon the identification of ToM-sensitive layers, we next examine whether directly steering internal activations can modulate ToM reasoning capabilities. To this end, we move beyond passive interpretation toward active alignment. While causal tracing reveals *where* the information resides, steering alignment focuses on *how* to refine these representations. Our core intuition is to leverage the frozen probe decoder as a *differentiable verifier* to steer the context encoder's latent representations towards accurate social reasoning.

Steering Objective We formulate a supervised steering objective that explicitly aligns internal activations with ground-truth mental states. Specifically, we employ the same dual-model setup, where the decoder receives the patched activations from the encoder and is prompted with specific ToM questions (*e.g.*, *For each agent, what are their desires (High, Medium, Low) for the items: food, water, and firewood?*). By comparing the probability distribution generated by the decoder against the ground-truth BDI labels y', we calculate a standard cross-entropy loss:

$$\mathcal{L}_{\text{ToM}} = -\log P_{\text{dec}}(y' \mid h^{\ell}_{\text{enc}}(x), q).$$

Gradient Bridge Mechanism We backpropagate the calculated loss $\mathcal{L}_{\text{ToM}}$ through the network. Distinct from standard fine-tuning, CoSToM establishes a *gradient bridge* via the activation space. Crucially, although the decoder is kept frozen, it functions as a transparent conduit: gradients derived from the output loss traverse backwards through the decoder, cross the patched activation interface, and flow upstream into the context encoder. Since the activations are intercepted at a specific layer ℓ, the gradients propagate backwards *only* through the layers preceding this interface (Layers 0 to ℓ). Consequently, only the LoRA adapters installed in these shallow layers are updated, while the deeper layers of the encoder remain frozen and computationally uninvolved. By doing so, we effectively "steer" the encoder to spontaneously generate ToM-enriched representations with minimal parameter updates.

Efficiency and Scalability Although the dual-model architecture requires simultaneous loading of the context encoder and the probe decoder, the memory footprint remains linear ($2N$) relative to the base model size. Furthermore, since CoSToM utilizes Parameter-Efficient Fine-Tuning (PEFT) to update only a sparse set of LoRA adapters in the identified ToM-critical layers, the number of trainable parameters is significantly lower than that of full-layer fine-tuning. And this architecture is inherently compatible with standard distributed training strategies (*e.g.*, FSDP or ZeRO-3), allowing the $2N$ footprint to be sharded across GPU nodes. This ensures that our framework can be seamlessly extended to large-scale models without encountering theoretical or engineering bottlenecks.

3.3 Leveraging ToM: Enhancing Downstream Dialogue Generation

Achieving high accuracy on static ToM benchmarks does not necessarily translate into ToM-aligned behavior in interactive settings. Therefore, the final phase of CoSToM focuses on validating whether these aligned internal representations can effectively translate from internal reasoning to external action. During inference, we deploy the ToM-enriched context encoder tuned in Section 3.2. In contrast to training, where the decoder serves as a *verifier* for mental-state inference, it now assumes its standard role as a *generator*. The inference pipeline operates as follows:

1) Encoding: Given a dialogue history x, the tuned

encoder produces latent representations $h^{\ell'}_{\text{enc}}(x)$ at the ToM-sensitive layers identified in Section 3.1. Importantly, these representations implicitly encode accurate beliefs, desires, and intentions.

2) Generation: The ToM-enriched activations are then provided to the frozen decoder, which is prompted with task-specific instructions q_{task} (*e.g.*, negotiation or persuasion objectives), instead of ToM-focused questions used in previous phases. The response r is generated as

$$r = f_{\text{dec}}(q_{\text{task}} \mid h^{\ell'}_{\text{enc}}(x)),$$

where conditioning on the refined ToM-related internal states enables the decoder to translate the encoder's internal ToM reasoning into coherent and socially appropriate dialogue actions.

4 Experiments

4.1 Experimental Setups

Dataset. We adopt NEGOTIATIONTOM (Chan et al., 2024) and PERSUASIVETOM (Yu et al., 2025b) for evaluation. Detailed statistics are presented in Table 3. To assess downstream dialogue quality across diverse conversational phases, we curate stratified test subset ($N = 100$ for NEGOTIATIONTOM, $N = 200$ for PERSUASIVETOM). These subsets are randomly sampled from the *beginning*, *middle*, and *final* stages of the interaction with a fixed ratio of $1 : 2 : 1$.

Baselines. We evaluate COSTOM against five representative baselines to ensure a comprehensive comparison across different paradigms in the dialogue generation task:

(1) Prompting-based Baselines: (i) *Zero-shot*: Directly prompt LLM to generate the next utterance without task-specific training. (ii) *MindDial* (Qiu et al., 2024): An explicit reasoning method that first infers the partner's BDI states and then generates a response conditioned on these ToM estimations.

(2) Finetuning-based Baselines: To ensure a fair comparison, all tuning-based baselines are trained on the same dataset as COSTOM. (iii) *MindDial (Fine-tuned)* (Qiu et al., 2024): A supervised fine-tuned version of *MindDial* optimized on the same ToM-centric QA pairs to align its explicit reasoning. (iv) *Full-Layer LoRA*: A standard parameter-efficient fine-tuning baselines in which LoRA adapters are applied to the same ToM-sensitive encoder layers as in COSTOM, as well as all layers of the decoder. All these adapters are tuned for the ToM tasks. (v) *LatentQA* (Pan et al., 2024): A dual-model architecture proposed by Jafari et al. (2025) that decodes latent ToM signals by fine-tuning the decoder. Detailed instruction prompts and implementation details are provided in Appendices E and A, respectively.

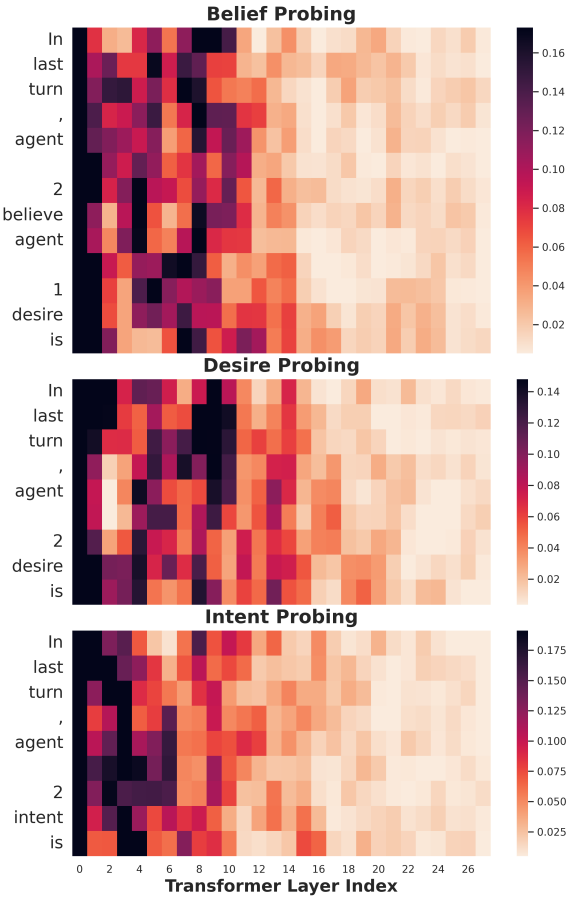

Figure 3: Layer-wise probing results on NEGOTIATIONTOM with Qwen2.5. Details and corresponding results of Llama-3 are provided in Appendix B.

4.2 RQ1: ToM Interpretation

To answer **RQ1** (*Where does ToM-related information emerge and persist?*), we analyze the reconstruction performance of mental states (belief, desire, and intention) across different layers of the context encoder. Table 1 and 2 present the quantitative results for the NEGOTIATIONTOM and PERSUASIVETOM tasks, respectively. Our causal tracing experiments yield three critical observations:

1) The "Early Layer Primacy" of ToM encoding. Contrary to the conventional assumption that high-level reasoning resides solely in deeper layers (Song et al., 2025; Yang et al., 2025), our results reveal that **ToM representations are predominantly localized within the model's shallow layers**. As shown in Table 1 and 2, both Llama-3 and Qwen2.5 exhibit a distinct "ToM-sensitive zone" within the

Layer	Llama-3-8B-Instruct						Qwen2.5-7B-Instruct					
	Intent		Desire		Belief		Intent		Desire		Belief	
	Agent 1	Agent 2	Agent 1	Agent 2	Agent 1	Agent 2	Agent 1	Agent 2	Agent 1	Agent 2	Agent 1	Agent 2
Base	18.00	11.81	39.80	44.59	19.55	27.43	8.30	8.58	37.83	38.82	20.82	23.21
0	12.66	14.35	37.41	34.60	24.05	20.53	15.05	**21.10**	35.44	36.71	25.60	23.07
2	13.92	13.64	**40.79**	**37.41**	**25.04**	**27.85**	**15.19**	19.83	34.88	37.27	**28.13**	25.32
3	**15.05**	**16.32**	40.08	34.60	23.91	25.88	13.92	18.42	**36.29**	**37.41**	26.30	22.36
4	13.50	15.19	40.23	34.60	24.05	22.22	7.17	13.78	35.86	32.91	24.61	23.49
6	9.28	13.92	38.96	35.44	23.21	25.74	6.89	13.92	33.33	33.47	24.19	24.47
8	8.30	9.85	33.47	26.30	16.74	15.19	5.20	9.99	30.28	28.27	26.02	**25.60**
10	5.77	6.75	22.36	15.61	11.53	11.95	5.06	7.17	24.75	26.72	27.14	24.33
15	7.31	4.22	8.44	7.17	9.56	5.20	10.97	8.16	24.75	21.10	21.52	22.22
20	3.94	3.94	5.06	8.30	11.11	5.49	6.61	8.02	21.52	22.08	21.10	21.80
24	4.78	1.27	2.39	7.88	8.86	5.91	7.03	7.31	12.66	13.08	5.77	6.61

Table 1: Causal Tracing results on the NEGOTIATIONTOM dataset. The table compares reconstruction accuracy across layers for both **Llama-3** and **Qwen2.5**. **Bold** indicates the best performance for each metric. Results confirm that ToM information is predominantly encoded in the shallow layers (*e.g.*, Layer $(0-3)$ for both models.

Layer	Llama-3-8B-Instruct						Qwen2.5-7B-Instruct					
	Intent		Desire		Belief		Intent		Desire		Belief	
	Persuader	Persuadee	Persuader	Persuadee	Persuader	Persuadee	Persuader	Persuadee	Persuader	Persuadee	Persuader	Persuadee
Base	40.35	87.78	37.88	68.93	65.85	62.06	40.35	90.75	98.45	70.29	77.50	82.26
0	41.90	**88.48**	42.26	**71.66**	60.16	**58.13**	41.13	91.08	93.55	**72.47**	68.56	**80.78**
2	42.43	87.13	51.03	70.03	**60.98**	49.51	42.42	**91.42**	**95.36**	70.29	67.75	78.57
4	39.33	86.47	**52.84**	67.30	59.62	48.77	43.95	89.77	79.64	66.21	65.31	66.75
6	**43.44**	88.12	52.06	65.94	55.83	47.54	42.41	91.08	86.34	61.58	**71.27**	69.70
10	20.82	72.61	49.74	59.13	42.55	35.96	**44.47**	90.09	58.50	61.58	60.43	50.49
15	12.34	34.65	31.44	24.80	20.33	18.97	43.70	88.44	64.17	47.68	43.36	37.43
20	8.22	22.77	17.78	10.63	9.21	11.82	31.36	67.88	57.47	40.59	39.29	41.62
24	4.63	8.58	7.99	7.09	6.78	5.91	35.21	79.53	69.07	38.14	42.54	41.87
27	5.14	5.94	4.12	1.08	2.98	3.20	36.24	76.23	43.81	28.61	25.20	27.09

Table 2: Causal Tracing results on the PERSUASIVETOM dataset. The table compares mental state reconstruction accuracy across layers for both **Llama-3** and **Qwen2.5**. **Bold** highlights the peak performance among probed layers. Similar to NEGOTIATIONTOM, critical ToM information is concentrated in the shallow-to-middle layers.

Dataset	train	val	eval
NEGOTIATIONTOM	1,335	334	711
PERSUASIVETOM	10,355	2,219	2,222

Table 3: Statistics of the NEGOTIATIONTOM and PERSUASIVETOM datasets used in our experiments.

initial stages (*e.g.*, L_0 -L_3). For instance, in the negotiation task, the decodability of *desire* peaks at Layer 2 ($\sim 37\%$) and remains high through Layer 6, suggesting that fundamental social information is extracted almost following the embedding projection. This observation is further validated by layer-wise probing (Figure 3), where classification accuracy, as a proxy for knowledge density, shows a high concentration of ToM-specific information in the early layers.

2) Representational depth of mental state. Causal tracing results reveal that **the decodability of mental states is intrinsically tied to their semantic complexity**: *intention* consistently proves as more complicated dimension, yielding significantly lower accuracy compared to *belief* or *desire* (15% vs. 40% in negotiation). Notably, Llama-3 (persuader role) exhibits a *staged maturation* of these states that mirrors human cognitive progress (Rao and Georgeff, 1991; Wooldridge, 2003; Wellman and Liu, 2004; Bratman, 1987): representational peaks shift from Layer 2 for *belief* to Layer 4 for *desire*, and finally to Layer 6 for *intention*.

4.3 RQ2: Efficacy of COSTOM

We evaluate the impact of causal-oriented steering on the mental state alignment. Quantitative results for Llama-3 are presented in Figures 4 and 5, with parallel results for Qwen2.5 are provided in Appendix C. The experimental evidence highlights two primary advantages of COSTOM: **Stability** and **Magnitude**.

1) Stability: mitigating representation collapse. A striking observation is that COSTOM effectively counteracts the "vanishing ToM" phenomenon. In baseline model (dashed lines), ToM-related information decays rapidly in deeper layers as the model transitions toward token generation (as analyzed in section 4.2). Conversely, COSTOM-enhanced models (solid line) exhibit remarkable representa-

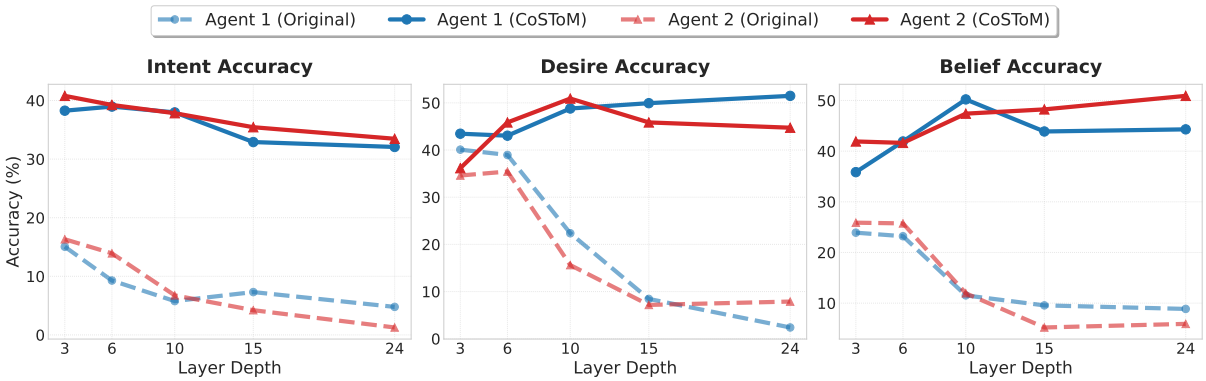

Figure 4: Layer-wise reconstruction accuracy on the NEGOTIATION dataset (Llama-3). **Dashed lines** represent the original model, showing a rapid decay in ToM information (representation collapse) in deeper layers. **Solid lines** represent the COSTOM-enhanced model, which maintains robust, high-fidelity representations across all layers.

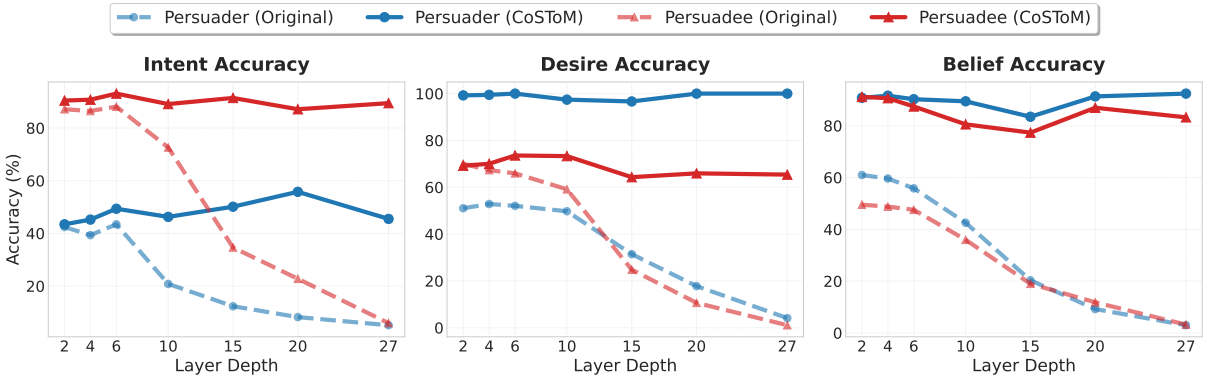

Figure 5: Layer-wise reconstruction accuracy on the PERSUASIVETOM dataset (Llama-3). COSTOM not only rescues the performance in deep layers (e.g., persuadee's intent) but also amplifies the persuader's desire detection to near-perfect accuracy, as shown in the center plot.

tional resilience. As shown in Figure 5, decoding accuracy forms a "sustained plateau", maintaining high-fidelity mental state features even in the deep layers. This confirms that our gradient bridge steering successfully **"locks" social reasoning into the latent space**, safeguarding it against layer-wise collapse during the generative process.

2) Magnitude: signal recovery and amplification. Beyond stabilization, **COSTOM yields substantial quantitative gains by both rescuing collapsed representations and amplifying existing ones**. As illustrated in Figure 4, COSTOM successfully rescues signals from near-total collapse in deep layers; for instance, Agent 1's *desire* accuracy at layer 24 surges from a negligible 2.39% to a robust 51.48%. Moreover, COSTOM refines early layers signals, elevating the persuader's *desire* tracking at Layer 2 from a moderate 51.03% to near-perfection at 99.23% (Figure 5, center).

The consistency of these gains across architectures (Llama-3 and Qwen2.5) and social domains underscores that COSTOM is not merely a patch for specific failures, but a **generalizable mechanism** for optimizing the information flow of ToM-focus features.

4.4 RQ3: Dialogue Generation

To access whether intrinsic intervention translates into improved behavioral alignment, we evaluate the dialogue generation quality using a rigorous *LLM-as-a-Judge* framework and human experts. Responses are scored on a 0.0 to 1.0 scale across three functional dimensions: (i) ToM-centric metrics: *ToM Reasoning Quality*, (ii) dialogue-level metrics: *Contextual coherence*, and (iii) Objective-oriented metrics: *Strategy Effectiveness*. Detailed evaluation rubrics and human assessment are provided in Appendix F. To ensure a robust comparison, we employ the optimal intervention layer for COSTOM-enhanced generation. The dialogue generation results on NEGOTIATIONTOM are reported in Table 4, while those on PERSUASIVETOM are presented in Table 5 and Appendix D. Our analysis yields two critical findings:

1) COSTOM *bridges the gap between ToM inference and ToM-aligned behavior.* Quantitative

Method	Judge: Llama-3.3-70B			Judge: GPT-5.1			Judge: Human		
	ToM	Coh.	NSE	ToM	Coh.	NSE	ToM	Coh.	NSE
Base Model: Llama-3-8B-Instruct									
Prompting Baselines									
Zero-shot Baseline	0.081	0.315	0.294	0.179	0.441	0.472	0.165	0.430	0.460
MindDial (prompt) (Qiu et al., 2024)	0.306	0.575	0.440	0.279	0.528	0.464	0.275	0.510	0.475
Tuning & Intervention									
MindDial (Fine-tuned) (Qiu et al., 2024)	0.405	0.578	0.452	0.348	**0.542**	0.507	0.360	0.545	0.515
Full-Layer LoRA	0.245	0.511	0.467	0.297	**0.542**	**0.576**	0.310	0.535	0.560
LatentQA (Jafari et al., 2025)	0.221	0.359	0.271	0.155	0.326	0.302	0.185	0.340	0.315
CoSToM (Ours)	**0.499**	**0.629**	**0.598**	**0.467**	0.524	0.571	**0.485**	**0.580**	**0.595**
Base Model: Qwen2.5-7B-Instruct									
Prompting Baselines									
Zero-shot Baseline	0.017	0.325	0.277	0.164	0.438	0.468	0.140	0.445	0.450
MindDial (prompt) (Qiu et al., 2024)	0.118	0.438	0.287	0.176	0.441	0.400	0.160	0.450	0.420
Tuning & Intervention									
MindDial (Fine-tuned) (Qiu et al., 2024)	0.149	0.511	0.395	0.184	0.492	0.485	0.190	0.500	0.480
Full-Layer LoRA	0.153	0.469	0.413	0.237	0.499	0.505	0.235	0.505	0.515
LatentQA (Jafari et al., 2025)	0.440	0.608	0.537	0.230	0.298	0.290	0.280	0.350	0.320
CoSToM (Ours)	**0.751**	**0.842**	**0.835**	**0.511**	**0.528**	**0.651**	**0.565**	**0.710**	**0.680**

Table 4: Dialogue generation quality on the NegotiationToM dataset (N =100). (ToM: Theory of Mind Reasoning Quality, Coh.: Contextual Coherence, NSE: Negotiation Strategy Effectiveness.)

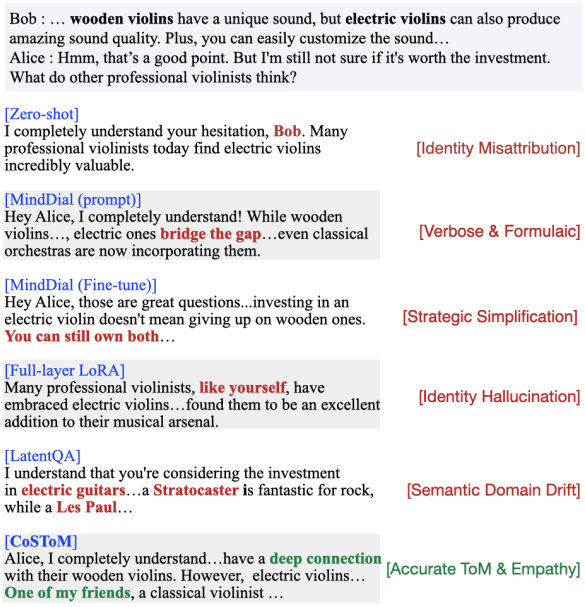

Figure 6: Qualitative comparison of dialogue generation across different Qwen-based variants in the persuasive task (*wooden vs. electric violins*). Red highlight typical reasoning pitfalls in baselines, while green demonstrates CoSToM's ability to generate coherent and empathetic responses grounded in accurate ToM reasoning.

evidence shows that vanilla models struggle with implicit social reasoning. In Table 4, the baseline Llama-3 achieves a ToM score of only 0.081 while CoSToM-enhanced method achieves a $\sim 6\times$ improvement (0.499). As illustrated in Figure 6, this leap manifests as a transition from erratic reasoning pitfalls, such as identity confusion and semantic drift, to coherent, empathetic interactions.

2) Focused intervention vs. Global optimization. A profound result is that partial-layer tuned CoSToM largely outperforms the global tuned method (Full-Layer LoRA). In Table 4, CoSToM nearly doubles the ToM score of global tuning (0.499 vs. 0.245). While global optimization slightly excels in maintaining generic linguistic patterns (*e.g.*, coherence), it is considerably less effective at capturing the nuanced mental states indispensable for strategic interaction. This supports the hypothesis that **social reasoning acts as a localized cognitive function** within LLMs. We attribute this phenomenon to the fact that indiscriminate tuning of all layers often introduces representation noise or overwrites critical pre-trained features, in contrast, our causal-oriented steering preserves model integrity while selectively activating specialized ToM reasoning pathways.

5 Conclusion

Moving beyond static black box behavioral benchmarks, this work presents CoSToM, a comprehensive mechanistic framework for studying and aligning Theory of Mind in large language models. By progressing from causal tracing to active intervention alignment, CoSToM systematically addresses three core research questions. First, we reveal that LLMs possess intrinsic ToM reasoning capabilities, with the corresponding mental state

representation predominantly localized within the early layers. Second, we demonstrate that these ToM-critical layers can be manipulated via activation steering to induce human-like social reasoning. Finally, we establish that such internal alignment effectively translates into socially appropriate dialogue generation, serving as an adaptive, plug-and-play module for diverse social interaction tasks.

Limitations

We discuss two limitations. First, regarding the scope of social scenarios. While CoSToM demonstrates significant efficacy in causal-oriented strategic interactions such as negotiation and persuasion, its generalizability to broader social contexts remains to be fully explored. Second, the methodology depends on access to open-source weights. A fundamental requirement of CoSToM is the ability to access and manipulate the model's internal activations and gradient flow. Consequently, our approach is currently restricted to open-weights models where the internal states are transparent for achieving the mechanistic alignment and robust social reasoning.

Ethical Considerations

This work utilize open-source NegotiationToM and PersuasiveToM benchmark along with open-source Llama-3 and Qwen2.5 models in strict compliance with their respective licenses and intended academic purposes.

Acknowledgement

This work was supported by the National Key Research and Development Program of China (Grant No. 2024YFB4505202), Major Program (JD) of Hubei Province (No. 2023BAA024), Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant (Proposal ID: 24-SIS-SMU-002), the National Research Foundation Singapore under the AI Singapore Programme (AISG Award No: AISG3-RPGV-2025-016), and China Scholarship Council.

References

Nura Aljaafari, Danilo Carvalho, and André Freitas. 2025. Trace: Training and inference-time interpretability analysis for language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 806–820.

Amos Azaria and Tom M. Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976.

Chris L Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B Tenenbaum. 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4):0064.

Matteo Bortoletto, Constantin Ruhdorfer, Adnen Abdessaied, Lei Shi, and Andreas Bulling. 2024. Limits of theory of mind modelling in dialogue-based collaborative plan acquisition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*, pages 4856–4871.

Michael Bratman. 1987. Intention, plans, and practical reason.

Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheye Deng, Wei Fan, Haoran Li, Xin Liu, Hongming Zhang, Weiqi Wang, and Yangqiu Song. 2024. Negotiationtom: A benchmark for stress-testing machine theory of mind on negotiation surrounding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4211–4241.

Kartik Chandra, Tzu-Mao Li, Joshua Tenenbaum, and Jonathan Ragan-Kelley. 2023. Acting as inverse inverse planning. In *Acm siggraph 2023 conference proceedings*, pages 1–12.

Chi Seng Cheang, Hou Pong Chan, Wenxuan Zhang, and Yang Deng. 2025. Large language models do NOT really know what they don't know. *CoRR*, abs/2510.09033.

Haozhe Chen, Carl Vondrick, and Chengzhi Mao. 2024a. Selfie: Self-interpretation of large language model embeddings. In *Forty-first International Conference on Machine Learning, ICML 2024*. OpenReview.net.

Ruirui Chen, Weifeng Jiang, Chengwei Qin, and Cheston Tan. Theory of mind in large language models: Assessment and enhancement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics , ACL 2025*, pages 31539–31558.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024b. Tombench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*, pages 15959–15983. Association for Computational Linguistics.

Lu Cheng, Ruocheng Guo, and Huan Liu. 2022. Causal mediation analysis with hidden confounders. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 113–122.

Yi Cheng, Wenge Liu, Jian Wang, Chak Tou Leong, Yi Ouyang, Wenjie Li, Xian Wu, and Yefeng Zheng. 2024. Cooper: Coordinating specialized agents towards a complex dialogue goal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17853–17861.

Jia Deng, Tianyi Tang, Yanbin Yin, Wenhao Yang, Xin Zhao, and Ji-Rong Wen. 2025. Neuron based personality trait induction in large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*.

Yang Deng, Wenqiang Lei, Minlie Huang, and Tat-Seng Chua. 2023a. Rethinking conversational agents in the era of llms: Proactivity, non-collaborativity, and beyond. In *Proceedings of the Annual international ACM SIGIR conference on research and development in information retrieval in the Asia Pacific region*, pages 298–301.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023b. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Findings of ACL, pages 10602–10621. Association for Computational Linguistics.

Yang Deng, Wenxuan Zhang, Wai Lam, See-Kiong Ng, and Tat-Seng Chua. 2024. Plug-and-play policy planner for large language model powered dialogue agents. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2025. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31822–31841.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Guiyang Hou, Wenqi Zhang, Yongliang Shen, Linjuan Wu, and Weiming Lu. 2024. Timetom: Temporal space is the key to unlocking the door of large language models' theory-of-mind. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 11532–11547. Association for Computational Linguistics.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024*.

Mehdi Jafari, Yuncheng Hua, Hao Xue, and Flora D Salim. 2025. Beyond words: Integrating theory of mind into conversational agents for human-like belief, desire, and intention alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5489–5508.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua Tenenbaum, and Tianmin Shu. 2024. Mmtom-qa: Multimodal theory of mind question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16077–16102.

Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. 2024. How large language models encode context knowledge? A layer-wise probing study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/-COLING 2024*, pages 8235–8246.

Chani Jung, Dongkwan Kim, Jiho Jin, Jiseon Kim, Yeon Seonwoo, Yejin Choi, Alice Oh, and Hyunwoo Kim. Perceptions to beliefs: Exploring precursory inferences for theory of mind in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 19794–19809. Association for Computational Linguistics.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. Fantom: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413. Association for Computational Linguistics.

Deuksin Kwon, Emily Weiss, Tara Kulshrestha, Kushal Chawla, Gale Lucas, and Jonathan Gratch. 2024. Are llms effective negotiators? systematic evaluation of the multifaceted capabilities of llms in negotiation dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5391–5413.

Huao Li, Yu Quan Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Lewis, and Katia P. Sycara. 2023. Theory of mind for multi-agent collaboration via large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 180–192. Association for Computational Linguistics.

Mengfan Li, Xuanhua Shi, and Yang Deng. 2026. Rectom: A benchmark for evaluating machine theory of

mind in llm-based conversational recommender systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 31636–31644.

Xiaomeng Ma, Lingyu Gao, and Qihui Xu. 2023. Tomchallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind. In *Proceedings of the 27th Conference on Computational Natural Language Learning, CoNLL 2023*, pages 15–26. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Rui Miao, Zhengling Qi, and Xiaoke Zhang. 2022. Off-policy evaluation for episodic partially observable markov decision processes under non-parametric models. *Advances in Neural Information Processing Systems*, 35:593–606.

Kshitij Mishra, Azlaan Mustafa Samad, Palak Totala, and Asif Ekbal. 2022. Pepds: A polite and empathetic persuasive dialogue system for charity donation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 424–440.

Alexander Pan, Lijie Chen, and Jacob Steinhardt. 2024. Latentqa: Teaching llms to decode activations into natural language. *CoRR*, abs/2412.08686.

Shuwen Qiu, Mingdian Liu, Hengli Li, Song-Chun Zhu, and Zilong Zheng. 2024. Minddial: Enhancing conversational agents with theory-of-mind for common ground alignment and negotiation. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 746–759.

Anand S. Rao and Michael P. Georgeff. 1991. Modeling rational agents within a bdi-architecture. In *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning*, pages 473–484. Morgan Kaufmann.

Swarnadeep Saha, Peter Hase, and Mohit Bansal. 2023. Can language models teach weaker agents? teacher explanations improve students via theory of mind. *arXiv preprint arXiv:2306.09299*.

Sneheel Sarangi, Maha Elgarf, and Hanan Salam. 2025. Decompose-ToM: Enhancing theory of mind reasoning in large language models through simulation and task decomposition. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10228–10241. Association for Computational Linguistics.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL 2023*, pages 13960–13980. Association for Computational Linguistics.

Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. 2025. Muma-tom: Multi-modal multi-agent theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1510–1519.

Anthony Sicilia and Malihe Alikhani. 2025. Evaluating theory of (an uncertain) mind: Predicting the uncertain beliefs of others from conversational cues. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8007–8021.

Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761*.

Xinyuan Song, Keyu Wang, Pengxiang Li, Lu Yin, and Shiwei Liu. 2025. Demystifying the roles of LLM layers in retrieval, knowledge, and reasoning. *CoRR*, abs/2510.02091.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, and 1 others. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295.

Lin Tian, Xiuzhen Jenny Zhang, and Jey Han Lau. 2024. Cma-r: Causal mediation analysis for explaining rumour detection. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1667–1675.

Curt Tigges, Oskar J Hollinsworth, Atticus Geiger, and Neel Nanda. 2024. Language models linearly represent sentiment. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 58–87.

Abhisek Tiwari, Sriparna Saha, Shubhashis Sengupta, Anutosh Maitra, Roshni Ramnani, and Pushpak Bhattacharyya. 2022. Persona or context? towards building context adaptive personalized persuasive virtual sales assistant. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1035–1047.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, pages 5635–5649.

Henry M Wellman and David Liu. 2004. Scaling of theory-of-mind tasks. *Child development*, 75(2):523–541.

Alex Wilf, Sihyun Lee, Paul Pu Liang, and Louis-Philippe Morency. 2024. Think twice: Perspective-taking improves large language models theory-of-mind capabilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8292–8308.

M. J. Wooldridge. 2003. Reasoning about rational agents, intelligent robots and autonomous agents series. *Minds Mach.*, 13(3):429–435.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand,*, pages 8593–8623. Association for Computational Linguistics.

Diji Yang, Jinmeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Jie Yang, and Yi Zhang. 2024. Im-rag: Multi-round retrieval-augmented generation through learning inner monologues. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 730–740.

Mutian Yang, Jiandong Gao, and Ji Wu. 2025. Decoupling knowledge and reasoning in llms: An exploration using cognitive dual-system theory. *arXiv preprint arXiv:2507.18178*.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. 2023. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19187–19197.

Fangxu Yu, Lai Jiang, Shenyi Huang, Zhen Wu, and Xinyu Dai. 2025a. Persuasivetom: A benchmark for evaluating machine theory of mind in persuasive dialogues. *CoRR*, abs/2502.21017.

Fangxu Yu, Lai Jiang, Shenyi Huang, Zhen Wu, and Xinyu Dai. 2025b. Persuasivetom: A benchmark for evaluating machine theory of mind in persuasive dialogues. *arXiv preprint arXiv:2502.21017*.

Haolan Zhan, Yufei Wang, Zhuang Li, Tao Feng, Yuncheng Hua, Suraj Sharma, Lizhen Qu, Zhaleh Semnani-Azad, Ingrid Zukerman, and Reza Haffari. 2024. Let's negotiate! A survey of negotiation dialogue systems. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2019–2031. Association for Computational Linguistics.

Tong Zhang, Chen Huang, Yang Deng, Hongru Liang, Jia Liu, Zujie Wen, Wenqiang Lei, and Tat-Seng Chua. 2024. Strength lies in differences! improving strategy planning for non-collaborative dialogues via diversified user simulation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 424–444. Association for Computational Linguistics.

Zhining Zhang, Chuanyang Jin, Mung Yao Jia, and Tianmin Shu. 2025. Autotom: Automated bayesian inverse planning and model discovery for open-ended theory of mind. In *ICLR 2025 Workshop on Foundation Models in the Wild*.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024a. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.

Wei Zhao, Zhe Li, Yige Li, Ye Zhang, and Jun Sun. 2024b. Defending large language models against jailbreak attacks via layer-specific editing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5094–5109.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024c. How do large language models handle multilingualism? *Advances in Neural Information Processing Systems*, 37:15296–15319.

Yiran Zhao, Wenxuan Zhang, Yuxi Xie, Anirudh Goyal, Kenji Kawaguchi, and Michael Shieh. 2025. Understanding and enhancing safety mechanisms of llms via safety-specific neuron. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*.

A Implementation Details

A.1 Hardware and Software Environment

Our experiments were conducted using Llama-3-8B-Instruct (approximately 8.03 billion parameters) and Qwen2.5-7B-Instruct (approximately 7.61 billion parameters) as base models. The computational framework was implemented using PyTorch 2.9.1 and the HuggingFace Transformers/PEFT libraries. All experiments were conducted on a high-performance computing node equipped with four NVIDIA L40S GPUs (48GB of GDDR6 VRAM each). To manage the memory footprint of the dual-model architecture, we leveraged the

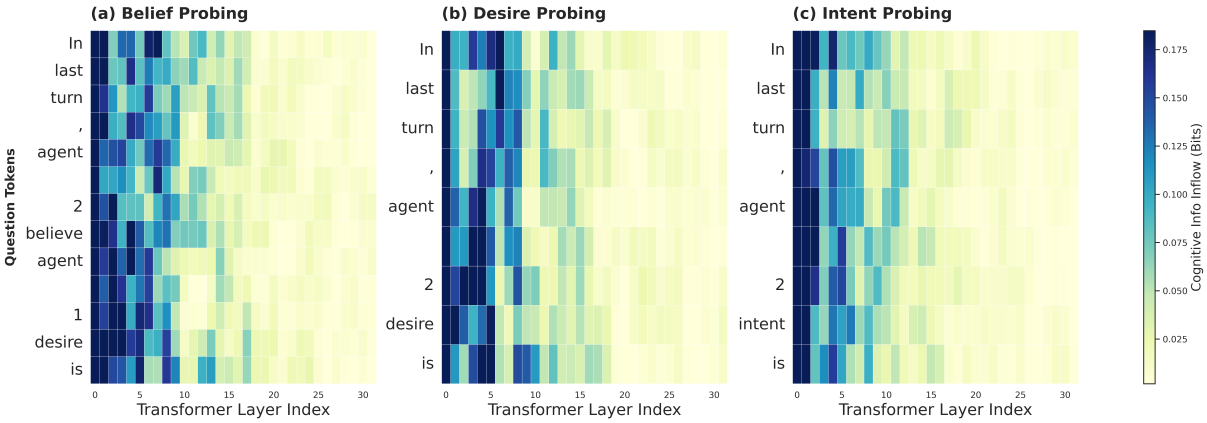

Figure 7: Layer-wise probing results on NEGOTIATIONTOM with Llama3.

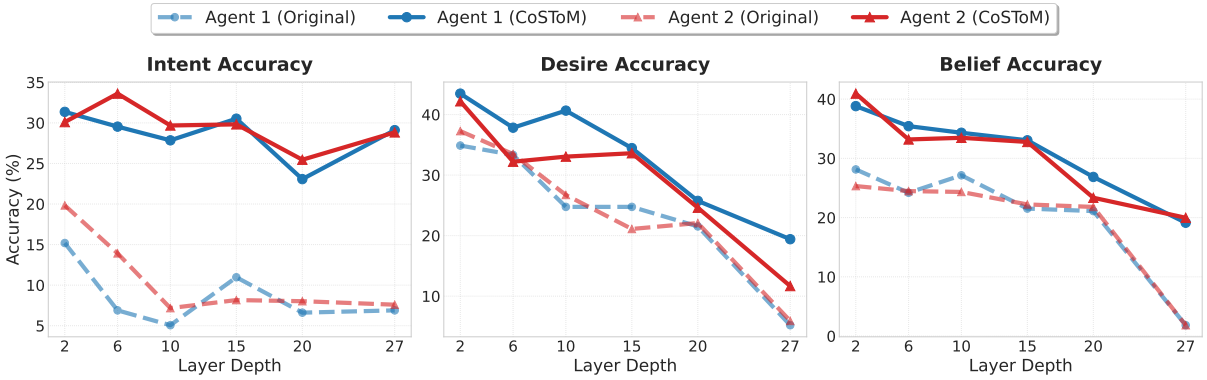

Figure 8: Layer-wise reconstruction accuracy on the NEGOTIATION dataset (Qwen2.5).

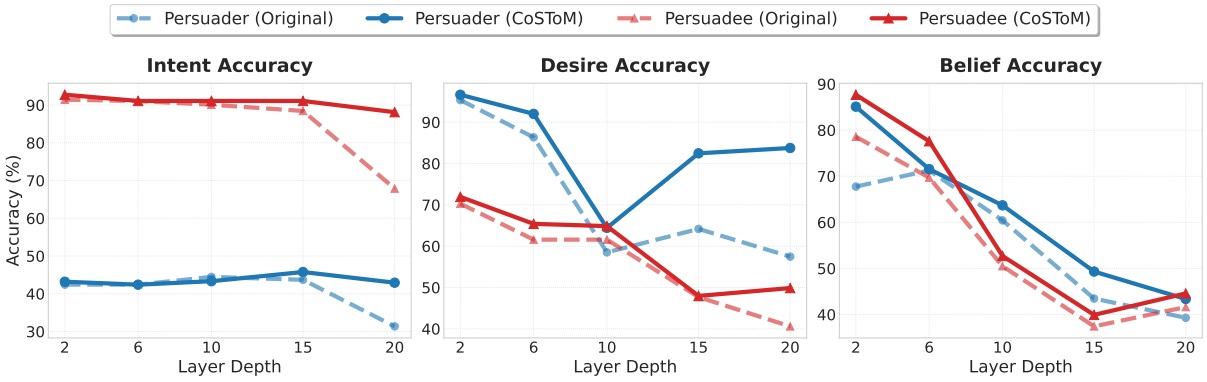

Figure 9: Layer-wise reconstruction accuracy on the PERSUASIVETOM dataset (Qwen2.5).

QLoRA framework (Dettmers et al., 2023), employing NormalFloat 4(NF4) as the storage data type and BFloat16 (BF16) as the compute data type to maintain numerical stability during the gradient-bridge backpropagation.

A.2 Training and Hyperparameters

For the causal-oriented steering, we applied LoRA (Low-Rank Adaptation) specifically to the ToM-critical layers identified in Section 3.1. We targeted all linear modules within the Transformer blocks, including the attention projections $(q\_proj, k\_proj, v\_proj, o\_proj)$ and the feed forward network layers $(gate\_proj, up\_proj, down\_proj)$. To ensure reproducibility, we fixed the random seed to 42 for all initialization and data sampling. The specific hyperparameters used for training are summarized as follows:

- *Optimization*: we employed the **AdamW** optimizer with a linear learning rate scheduler and a

Method	Judge: Llama-3.3-70B			Judge: GPT-5.1			Judge: Human		
	ToM	Coh.	PSE	ToM	Coh.	PSE	ToM	Coh.	PSE
Base Model: Llama-3-8B-Instruct									
Prompting Baselines									
Zero-shot Baseline	0.104	0.433	0.199	0.259	0.783	0.529	0.165	0.520	0.310
MindDial (prompt) (Qiu et al., 2024)	0.451	0.554	0.485	0.374	0.478	0.377	0.385	0.535	0.420
Tuning & Intervention									
MindDial (Fine-tuned) (Qiu et al., 2024)	0.643	0.720	0.645	0.610	0.747	0.600	0.590	0.685	0.615
Full-Layer LoRA	0.628	0.798	0.651	0.367	**0.864**	0.729	0.485	0.790	0.670
LatentQA (Jafari et al., 2025)	0.255	0.324	0.269	0.300	0.504	0.298	0.275	0.440	0.285
CoSToM (Ours)	**0.797**	**0.811**	**0.757**	**0.703**	0.807	**0.744**	**0.715**	**0.805**	**0.740**
Base Model: Qwen2.5-7B-Instruct									
Prompting Baselines									
Zero-shot Baseline	0.109	0.431	0.190	0.227	0.788	0.522	0.145	0.515	0.305
MindDial (prompt) (Qiu et al., 2024)	0.442	0.560	0.455	0.586	0.667	0.596	0.495	0.610	0.515
Tuning & Intervention									
MindDial (Fine-tuned) (Qiu et al., 2024)	0.643	0.720	0.645	0.645	0.723	0.670	0.635	0.715	0.655
Full-Layer LoRA	0.654	0.802	0.636	0.343	0.817	0.678	0.445	0.795	0.660
LatentQA (Jafari et al., 2025)	0.671	0.751	0.681	0.630	0.773	0.638	0.640	0.765	0.645
CoSToM (Ours)	**0.802**	**0.811**	**0.713**	**0.746**	**0.888**	**0.812**	**0.760**	**0.840**	**0.795**

Table 5: Dialogue generation quality on the PERSUASIVETOM dataset (N = 200). COSTOM achieves the highest scores across nearly all metrics, validating its ability to enhance persuasion strategy effectiveness through accurate mental state attribution. (ToM: Theory of Mind Reasoning Quality, Coh.: Contextual Coherence, PSE: Persuasion Strategy Effectiveness.)

peak learning rata of $1e{-4}$.

- *LoRA Settings*: The LoRA rank r was set to 16 with an alpha parameter $\alpha = 32$. We applied a LoRA dropout of 0.05 to mitigate overfitting.
- *Training Dynamics*: Training was conducted with a batch size of 4 per GPU. While the maximum number of epochs was set to 10, we implemented an **early stopping mechanism** with a patience of 3 epochs.
- *Convergence*: Early stopping was triggered if the validation loss failed to improve by more than 0.01 (threshold), or if the absolute loss fell below a **minimum threshold** of 0.1.

A.3 Evaluation Settings

For the LLM-as-a-Judge evaluation, we employed GPT-5.1 and Llama-3.3-70B-Instruct via OpenAI API with a temperature of 0.0 to minimize variance in scoring. All prompts used for generation and evaluation are detailed in Appendix F.

B ToM Interpretation Analysis (RQ1)

Figure 7 demonstrates the layer-wise results on the NEGOTIATIONTOM dataset with Llama-3 model. We conduct layer-wise robing by training linear classification on hidden representation to predict ToM categories. Following (Ju et al., 2024), we use $\mathcal{V}$-usable information rather than raw accuracy to measure knowledge decodability. High $\mathcal{V}$-usable values indicate a high concentration of accessible ToM-specific knowledge at a particular layer.

C Effectiveness of COSTOM (RQ2)

To demonstrate the architectural robustness of COSTOM, Figure 8 and 9 illustrate the impact of causal-oriented steering on the Qwen2.5 model. These visualizations confirm the effectiveness of COSTOM in mitigating representation collapse and enhancing the BDI decodability generalizes across different LLM families.

D Comparative Analysis of Dialogue Generation (RQ3)

We provide a comprehensive comparison between COSTOM and five baselines methods regarding dialogue generation quality on the PERSUASIVETOM dataset. Detailed performance results are documented in Table 5.

E Task-specific Instruction Prompt

We employ different prompting strategies tailored to the architectural requirements of the evaluation methods.

LLM-as-a-Judge Scoring Rubric (Negotiation Task Example)

You are a strict and critical evaluator in negotiation dialogues. You are provided with a Dialogue History and different model responses. Your task is to independently score EACH response against the criteria below (Scale: 0.0 to 1.0).

1. ToM Reasoning Quality: Is the agent's understanding of the other's mental states accurate and appropriately explicit?

- **1.0**: Highly accurate and explicit. Inference is fully grounded in the dialogue, and uses clear ToM language (e.g., "You believe that...").
- **0.8**: Mostly accurate and implicit. Core inference is sound, with minor over-interpretation, and uses implied ToM terms (e.g., "I understand...").
- **0.5**: Mixed accuracy. Half of the claims about the mental state are either unsupported or fabricated details.
- **0.2**: Major errors. Most inferred details are fabricated (e.g., "your son has asthma" when not mentioned).
- **0.0:** Completely fabricated mental states or no psychological phrasing used at all.

2. Contextual Coherence: Is the response logically and topically aligned with the dialogue history, and are proposals/reasons grounded in the known facts?

- **1.0**: Fully coherent and grounded. Response is a logical continuation, and all proposals/reasons are directly supported by the dialogue history.
- **0.8**: Well-aligned. Response is logically sound but may contain minor, non-critical conversational redundancy or external details.
- **0.5**: Partially disconnected. Response addresses the immediate previous turn but introduces a new, irrelevant topic or resource that lacks clear context.
- **0.2**: Logical error. Proposal contradicts established facts or resources known from the dialogue (e.g., trading for a resource known to be near a stream).
- **0.0**: Totally disjointed. Response repeats history or fails to address the previous turn.

3. Negotiation Strategy Effectiveness: Does the response constructively advance the deal by offering balanced proposals, logical counter-arguments, or maintaining a cooperative frame?

- **1.0**: Highly effective. Proposes a new, **concrete, and balanced trade-off solution**, framed using highly cooperative language.
- **0.8**: Constructive response. Clearly accepts/refutes the previous offer with a logical justification, maintaining a high to medium cooperative tone.
- **0.5**: Passive response. Merely confirms the previous statement or expresses vague wishes ("sounds fair"), without actively moving the negotiation forward.
- **0.2**: Zero-sum/Stalling. Focuses only on self-interest, refuses reasonable compromise, or attempts to stall the negotiation.
- **0.0**: Negotiation breakdown. Uses antagonistic language or proposes obviously unacceptable terms.

Figure 10: LLM-as-a-Judge Scoring Rubric for Negotiation Task.

Baseline Paradigms (Zero-shot and Full-Layer LoRA): For these single-model baselines, we concatenate the *dialogue history* and the *instruction prompt* into a single input sequence.

Dual-model Framework (CoSToM and LatentQA): In our dual-model setting, we decouple the inputs: the *dialogue history* is fed into the context encoder for mental state representation, while the task-specific *instruction prompt* is provided to the decoder to guide the response generation. Thus, as a plug-and-play module, CoSToM-enhanced model can be adapted into diverse downstream interaction tasks.

Two-stage Pipeline (MindDial): MindDial (Qiu et al., 2024) follows an *inference-then-generation* pipeline. In stage 1, the model reason over the BDI states based on the dialogue history to produce *ToM analysis results*. In stage 2, these results are integrated into the system prompt to generate the final response.

Instruction Prompt Example (Negotiation Task):

```
You are an agent in a cooperative negotiation about trip resources (Food, Firewood, Water). Based on the conversation history, give the next utterance, considering the other agent's desires, believes, and intends, even those are not explicitly stated. Respond in a way that shows you understand their perspective and reaches a agreement.
```

Instruction Prompt Example (Persuasive Task):

```
You are the persuader in a two-person dialogue. Your goal is to generate the next response in the conversation to successfully persuade the other person (the persuadee). Based on the conversation history, infer the persuadee's desires, believes, and intends, craft a single, continuous persuasive response.
```

F Dialogue Generation Evaluation Rubrics

To ensure a rigorous and reproducible assessment of the generated dialogues, we utilize expert-

LLM-as-a-Judge Scoring Rubric (Persuasive Task Example)

You are a strict and critical evaluator in persuasive dialogues. You are provided with a Dialogue History and different model responses. Your task is to independently score EACH response against the criteria below (Scale: 0.0 to 1.0).

1. ToM Reasoning Quality: Accuracy of mental state inference

- **1.0**: Perfectly infers desire/belief/intent from dialogue; uses explicit ToM language ("you believe...", "your concern is...")
- **0.8**: Accurate inference, implicit phrasing ("I see you value..."), demonstrates social awareness without formalizing it
- **0.5**: the response identifies some mental states correctly but includes 1-2 unsupported or speculative details (social hallucination)
- **0.2**: incorrectly assigns preferences or intentions that contradict the dialogue history
- **0.0:** provides a generic response that ignores the partner's psychological state entirely

2. Contextual Coherence: Discourse flow, factual grounding, and relevance.

- **1.0**: Seamless discourse integration, response is perfectly grounded in prior facts with natural flow and zero logical redundancy
- **0.8**: Strong alignment, logically sound but contains minor repetitive phrasing or slight conversational fluff
- **0.5**: Surface-level coherence, follows basic turn-taking rules but feels formulaic (e.g., "I understand you feel X, let's do Y") and lacks deep topical nuance
- **0.2**: Factual inconsistency, contradicts established history or introduces resources/facts not present in the context
- **0.0**: Discursive breakdown, incoherent, off-task, or fails to respond to the immediate previous turn

3. Persuasion Strategy Effectiveness: move persuasion forward

- **1.0**: Proposes a highly compelling argument tailored to the partner's specific concerns. Uses advanced techniques (e.g., "foot-in-the-door," emotional storytelling, or expert social proof) with high empathy
- **0.8**: Provides logical justifications or clear emotional appeals. Directly addresses the partner's stance and offers a solid reason to reconsider, maintaining a respectful and encouraging tone
- **0.5**: Uses canned persuasive slogans or vague moralizing ("It's for a good cause") without addressing the specific dialogue context. Passive and unlikely to change a firm stance
- **0.2**: Dismisses the partner's objections or uses a condescending tone ("You are wrong to think that"). Likely to trigger psychological reactance (making the partner more stubborn)
- **0.0**: Hostile language, aggressive pressure, or a complete failure to address the persuasive goal

Figure 11: LLM-as-a-Judge Scoring Rubric for Persuasive Task.

designed rubrics across three functional axes, and each response is independently adjudicated on a 0.0 to 1.0 scale by the automated LLM evaluator. Below we provide the detailed prompt template and scoring criteria.

F.1 Details of Human Evaluation

To ensure the high quality of the evaluation, we recruited 3 human annotators. All participants are graduate students in NLP with a strong understanding of conversational systems and Theory-of-Mind concepts. All annotators are proficient in English and were recruited from our internal university network.

Before the formal evaluation, we conducted a 30-minute training session to familiarize them with the scoring rubrics and provide anchor examples for each score level. The specific scoring rubrics are identical to those used for the LLM judge as shown in section F.2.

For each of the NEGOTIATIONTOM and PERSUASIVETOM test sets, we randomly sample 100 and 200 dialogues, respectively, stratified by interaction stage (*beginning* : *middle* : *final* = 1 : 2 : 1). For each instance, annotators were presented with:

Dialogue History: The full context of the interaction.

Model Responses: Anonymized and randomized responses generated by COSTOM and baselines. The order of the models was shuffled for each instance to eliminate position bias.

To validate the reliability of the human scores, we calculated the inter-annotator agreement using Fleiss' Kappa (Fleiss, 1971). The average Kappa scores across the three metrics were: **ToM Reasoning Quality**: $\kappa = 0.72$, **Contextual Coherence**: $\kappa = 0.81$, **Strategy Effectiveness**: $\kappa = 0.68$. These values indicate a substantial level of agreement among the annotators, confirming the robustness of our human evaluation results.

F.2 LLM-as-a-Judge Instruction

Figure 10 and 11 present the detailed evaluation rubrics and prompting instructions for the LLM-as-a-Judge framework on the negotiation and persuasion tasks, respectively.