From Confusion to Clarity: Robust Theory of Mind Reasoning via World-Grounded Interventions

Anonymous ACL submission

Abstract

While large language models (LLMs) excel in mathematical and code reasoning, we observe they struggle with social reasoning tasks. We identify two critical phenomena: *Knowledge Blindness*, where LLMs fail to apply implicit social norms, defaulting to literal interpretations; *Cognitive Dilemmas*, where LLMs become trapped in contradictory reasoning loops when tracking multiple agents' beliefs across time. Both failures stem from inadequate grounding: the former lacks the necessary social knowledge for contextual understanding while the latter conflates objective world states with subjective belief states. To address this, we propose the World-Grounded Theory of Mind (**WG-ToM**) framework, which provides grounded guidance at two stages: injecting elicited social norms before reasoning, and intervening with structured world-state descriptions upon detecting confusion during reasoning. Experiments across five Theory of Mind benchmarks demonstrate that WG-ToM improves accuracy by up to 9.00% while reducing token consumption by up to 49.6%, enabling 32B LLMs to match or exceed state-of-the-art performance.

1 Introduction

Recent advances in large language models (LLMs) have yielded remarkable reasoning capabilities, with systems like OpenAI o3 (OpenAI, 2025b), DeepSeek-R1 (Guo et al., 2025) and Claude Opus 4.7 (Anthropic, 2026) achieving strong performance on mathematical and programming tasks through extended chain-of-thought (CoT) processes (Wei et al., 2022). However, these advances have concentrated in domains with well-defined logical structure. Social reasoning, which requires understanding others' mental states, beliefs and intentions, presents a fundamentally different challenge that remains underexplored (Chen et al., 2025).

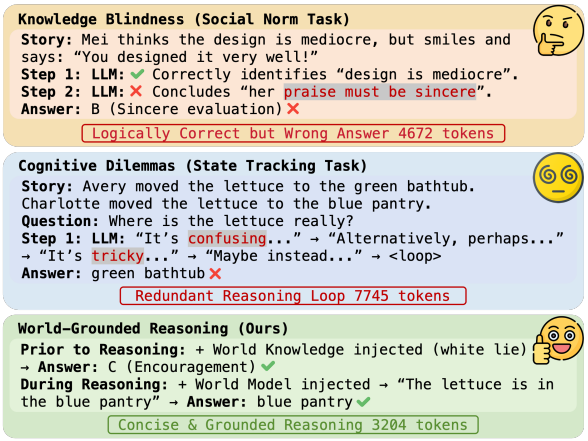
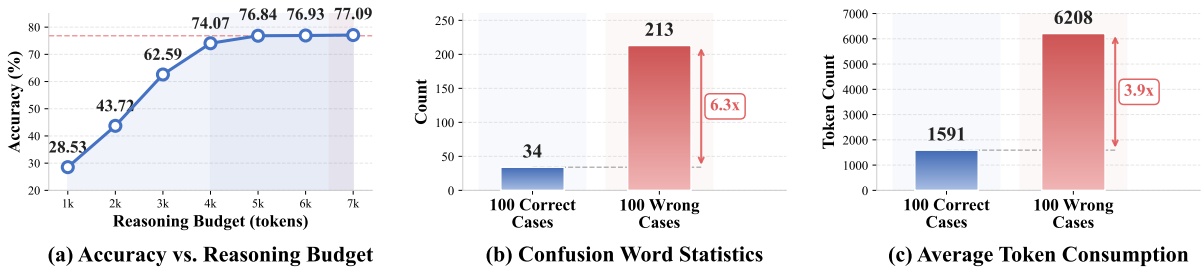


Figure 1: Two failure modes in DeepSeek-R1 social reasoning and our solution. **Top**: Knowledge Blindness. **Middle**: Cognitive Dilemmas. **Bottom**: WG-ToM resolves both failures through World Knowledge injection prior to reasoning and World Model intervention during reasoning.

Theory of Mind (ToM), the capacity to attribute mental states to others, is central to social cognition (Baron-Cohen et al., 1985; Perner et al., 1987). Unlike mathematical reasoning over fixed axioms, social reasoning requires tracking beliefs that diverge from reality and applying unstated cultural norms. This distinction between objective facts and subjective thoughts proves surprisingly difficult for current LLMs to maintain.

We conduct systematic analysis of LLM reasoning trajectories on social tasks and identify two fundamentally distinct failure modes. As shown in Figure 1, *Knowledge Blindness* appears in social interaction scenarios involving implicit norms. When someone praises a mediocre design to preserve feelings, LLMs correctly identify the objective quality but fail to recognize the face-saving norm, concluding that the praise must be sincere. *Cognitive Dilemmas* emerges in dynamic scenarios involving multiple agents and temporal state transitions. When objects move between locations



(a) Accuracy vs. Reasoning Budget (b) Confusion Word Statistics (c) Average Token Consumption

Figure 2: Quantitative analysis of DeepSeek-R1 reasoning trajectories on social reasoning tasks. (a): Accuracy under varying reasoning budgets. (b) and (c): Frequency of confusion words and average token consumption for 100 correct and 100 incorrect cases.

while characters enter and exit scenes, LLMs lose track of objective states. Their reasoning trajectories exhibit frequent confusion markers such as "tricky" and "confused", indicating that LLMs become trapped in unproductive reasoning loops.

These two failure modes share a common cause: inadequate grounding. Knowledge Blindness occurs when LLMs lack the implicit social norms that humans acquire through cultural experience; Cognitive Dilemmas arise when LLMs lose anchor to objective world states and conflate them with subjective belief states. Standard approaches such as CoT or self-reflection (Shinn et al., 2023) cannot resolve either deficit, as they operate within the same ungrounded reasoning space, producing more reasoning without filling the knowledge gap or correcting the fundamental confusion.

Based on the above analysis, we propose the World-Grounded Theory of Mind (**WG-ToM**) framework, which systematically addresses both deficits by providing targeted grounded guidance at different stages of the reasoning process. WG-ToM operates as a unified two-phase pipeline. Prior to reasoning, the framework elicits relevant social norms from advanced LLMs as explicit reasoning premises and conditionally injects them through task-adaptive routing, transforming implicit knowledge into explicit World Knowledge application and thereby eliminating Knowledge Blindness. During reasoning, the framework monitors the reasoning trajectory through a real-time detector based on semantic similarity. Upon detecting confusion signals, it intervenes with a structured World Model to re-anchor LLM in objective facts, breaking Cognitive Dilemmas.

We evaluate WG-ToM across five ToM benchmarks: ToMi (Le et al., 2019), HiToM (Wu et al., 2023), ExploreToM (Sclar et al., 2025), ToMBench (Chen et al., 2024), and EmoBench (Sabour et al., 2024). Results demonstrate consistent improvements, with accuracy gains up to 9.00% on HiToM while reducing token consumption by up to 49.6%. Notably, WG-ToM enables 32B-parameter LLMs to match or exceed the performance of frontier LLMs, offering a cost-effective path to strong social reasoning.

Our contributions are threefold: **(1)** We identify two critical failure modes in LLM social reasoning: *Knowledge Blindness* from missing social norms and *Cognitive Dilemmas* from state-belief conflation. **(2)** We propose WG-ToM, a unified two-phase pipeline integrating task-adaptive World Knowledge Injection and World Model Intervention driven by real-time semantic confusion detection. **(3)** We demonstrate significant improvements in both accuracy and efficiency across five benchmarks, enabling smaller LLMs to achieve performance competitive with state-of-the-art systems.

2 Analysis of Social Reasoning Deficits

To understand the root causes of LLM failures in social reasoning, we systematically collect reasoning LLMs' trajectories from DeepSeek-R1 across ToM datasets and conduct analysis along both qualitative and quantitative dimensions. Qualitatively, we examine linguistic patterns and logical consistency in reasoning chains. Quantitatively, we randomly sample 100 correct and 100 incorrect cases and measure confusion word frequency, token consumption, and accuracy variation under different reasoning budgets. The analysis reveals two fundamentally distinct failure modes. As illustrated in Figure 1, both failures point to a common root cause: the lack of effective grounding mechanisms.

Knowledge Blindness LLMs do not exhibit notable reasoning confusion or loops on tasks involving social norms. Conversely, the reasoning chain is logically coherent. LLM correctly iden-

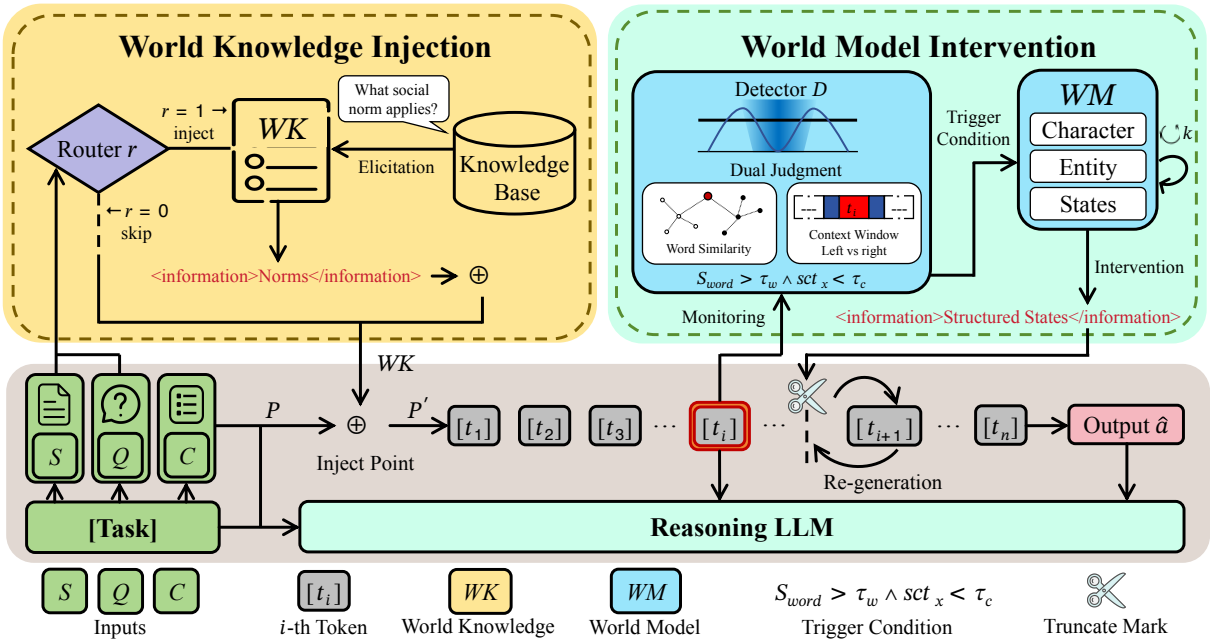


Figure 3: Illustration of the WG-ToM framework. The system provides grounded guidance through two mechanisms: World Knowledge Injection (left), which elicits and injects social norms prior to reasoning via task-adaptive routing, and World Model Intervention (right), which monitors the reasoning trajectory in real time and intervenes with structured state descriptions upon detecting confusion.

tifies the objective assessment but fails to recognize the applicable social norm (e.g., face-saving through white lies), strictly adhering to literal semantics. This indicates that the failure stems from the absence of social common sense rather than insufficient reasoning capacity. To further verify this, we constrain DeepSeek-R1 with varying reasoning budgets on social norm benchmarks. As shown in Figure 2(a), increasing the budget does not improve accuracy, confirming that the failure originates from missing knowledge rather than insufficient reasoning depth.

Cognitive Dilemmas When LLMs encounter reasoning tasks involving multiple agents and object relocations across scenes, their outputs begin to exhibit frequent confusion markers such as "tricky" and "confused". More notably, LLMs become trapped in cycles of self-questioning, producing characteristic patterns like "Alternatively, perhaps..." and "Maybe instead of...", indicating an inability to distinguish between objective world states and subjective belief states. As shown in Figure 2(b) and (c), both metrics are substantially higher in incorrect cases. This demonstrates that DeepSeek-R1 is more prone to cognitive dilemmas when facing complex, error-prone problems, and that such dilemmas incur considerable wasted computation.

3 World-Grounded Theory of Mind Framework

3.1 Task Formulation

To systematically address the challenges in social reasoning, we propose the WG-ToM framework. Given a social reasoning task, the input consists of a story S, a question Q, and candidate answers $\mathcal{C}$. The objective is to predict the correct answer $\hat{a} \in \mathcal{C}$. In standard inference, LLMs generate directly: $\hat{a} = \mathcal{M}(\mathcal{P})$, where $\mathcal{P} = \text{prompt}(S, Q, \mathcal{C})$. WG-ToM constructs a unified two-phase pipeline to augment this process:

$$\hat{a} = \mathcal{M}_{\mathcal{D}}^{k}\Big(\mathcal{P} \oplus r \cdot \text{WK}(S, Q);\ \text{WM}\Big) \quad (1)$$

where $r = \text{Router}(S, Q) \in \{0, 1\}$ is a task-adaptive router that determines whether external World Knowledge is needed, conditionally injecting World Knowledge elicited as explicit reasoning premises to address Knowledge Blindness (Section 3.2); $\mathcal{D}$ is a real-time semantic confusion detector that continuously monitors the generation trajectory during reasoning (Section 3.3); WM is World Model that provides structured state descriptions for intervention upon detector activation, re-anchoring LLM in objective facts to break Cognitive Dilemmas (Section 3.4); and the superscript

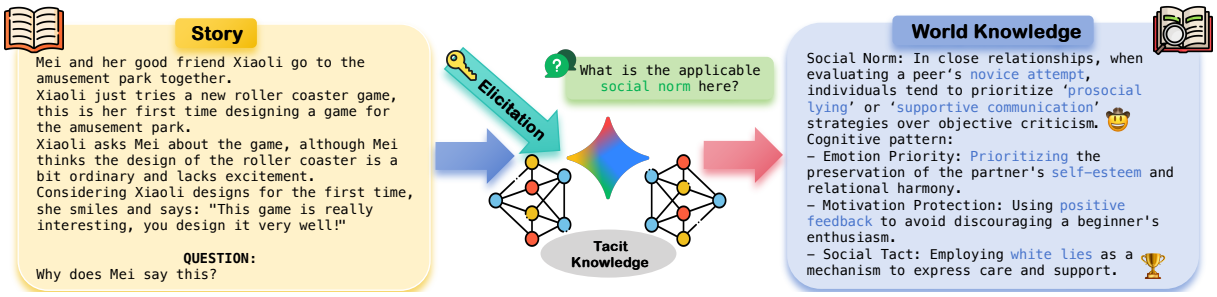


Figure 4: The process of extracting World Knowledge from social reasoning scenarios via Knowledge Elicitation.

k denotes the maximum number of intervention iterations.

3.2 World Knowledge Injection

The analysis in Section 2 reveals that Knowledge Blindness originates from LLMs' lack of implicit social common sense rather than insufficient reasoning capacity. However, such knowledge exists in the form of tacit knowledge within the pretrained parameters of advanced LLMs, requiring specific activation steps to be extracted. We therefore leverage advanced LLMs not merely as reasoning engines but as knowledge bases.

Knowledge Elicitation As illustrated in Figure 4, instead of prompting advanced LLMs for direct task answers, we instruct them to identify the social norms applicable to the given context (e.g., What is the applicable social norm here?). Through this elicitation process, knowledge that exists in tacit form is explicitly activated and structured. We term the resulting output World Knowledge, which transforms the role of advanced LLMs from passive reasoning engines to active knowledge providers.

Task-Adaptive Routing After acquiring World Knowledge, a critical consideration arises: not all task instances benefit equally from external knowledge injection. We therefore introduce a Task-Adaptive Router r. The router presents the current task (story and question) to the reasoning LLM itself, asking whether the task requires external guidance on social norms. When $r = 1$, the corresponding World Knowledge is appended to the original prompt: $\mathcal{P}' = \mathcal{P} \oplus \mathrm{WK}(S, Q)$ in a structured format: <**information**> **{Social Norms from World Knowledge}** <**/information**>. This routing mechanism avoids over-injection on simple tasks while ensuring necessary knowledge supplementation for complex social reasoning instances.

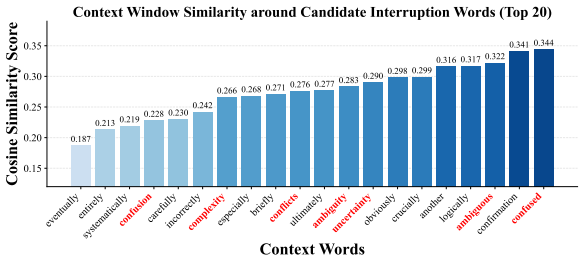


Figure 5: Cosine similarity between preceding and following context windows at positions of candidate interruption words in DeepSeek-R1's reasoning trajectories.

3.3 Semantic Confusion Detection

The foregoing analysis establishes a correlation between confusion words and reasoning failure. To further explore the relationship between confusion words and their surrounding context, we conduct contextual semantic similarity analysis on reasoning trajectories. As shown in Figure 5, after removing some words clearly unsuitable as intervention points (e.g., pronouns like "he" that do not signal reasoning breakdown), we find that confusion words exhibit high confusion degree while maintaining low semantic similarity, indicating that they not only carry confusion semantics themselves but also increase the semantic divergence between preceding and following contexts.

Based on this observation, we propose a semantic confusion detection mechanism $\mathcal{D}$ that jointly evaluates the local semantics of the currently generated word and the contextual coherence at its position. The mechanism employs all-MiniLM-L6-v2 (Wang et al., 2020; Reimers and Gurevych, 2019) as the sentence encoder $f(\cdot)$, projecting text segments into a unified semantic space with cosine similarity as the metric. We select the top 15 most frequently occurring confusion words from the sampled trajectories described in Section 2 to form the set $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$ (As shown in Table 1), encode each word independently and com-

Confusion Words
"ambiguous", "complicating", "confusion", "confusing", "confused", "perplexity", "puzzle", "puzzled", "puzzling", "perplexed", "complication", "troubled", "tricky", "conflicts", "ambiguity"

Table 1: Confusion words used to construct the semantic centroid.

pute the normalized centroid:

$$\mathbf{c} = \text{normalize}\left(\frac{1}{n}\sum_{i=1}^{n} f(v_i)\right) \tag{2}$$

The centroid serves as a semantic anchor for the concept of "confusion" in the embedding space. During the reasoning generation process, for each newly generated word t_i, we compute its semantic proximity to the confusion centroid:

$$s_{\text{word}}(i) = \cos\left(f(t_i),\ \mathbf{c}\right) \tag{3}$$

Simultaneously, to capture changes in contextual coherence around that position, we define a sliding window of size W on each side of position i and compute the semantic similarity between the preceding and following context segments:

$$s_{\text{ctx}}(i) = \cos\left(f(t_{i-W:i}),\ f(t_{i+1:i+1+W})\right) \tag{4}$$

The former measures whether the current word carries confusion semantics; the latter measures whether a semantic break occurs at that position. An intervention is triggered if and only if both conditions are jointly satisfied:

$$\mathcal{D}(i) = \left(s_{\text{word}}(i) > \tau_w\right) \ \wedge\ \left(s_{\text{ctx}}(i) < \tau_c\right) \tag{5}$$

where τ_w and τ_c are thresholds. The conjunction ensures that the detector activates only when a confusion word co-occurs with a genuine contextual disruption, effectively filtering false positives from benign keyword occurrences. During the reasoning generation process, the detector operates with a delayed evaluation scheme: after each new token t_i is produced, the system waits until the subsequent tokens (t_{i+1} through t_{i+W}) have also been generated before applying the dual judgment at position i. Once the trigger condition is met, the ongoing generation is immediately interrupted and the intervention process is initiated. To prevent redundant triggers within the same confusion region, a cooldown window of d tokens is enforced after each activation.

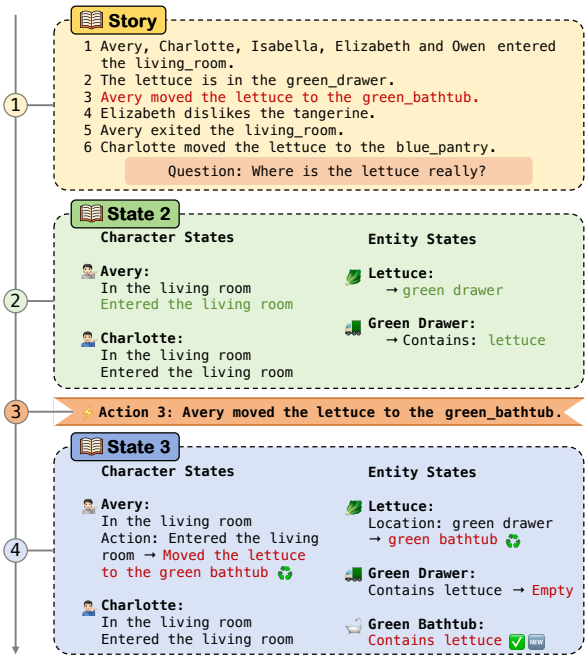


Figure 6: World Model generation process. Character and entity states are dynamically updated as the story progresses. Green denotes initial states and Red highlights updates.

3.4 World Model Intervention

State Representation We define the World Model WM as a structured textual representation that provides an objective factual anchor when reasoning confusion occurs. As illustrated in Figure 6, World Model is constructed as a dynamic state tracker: it parses the narrative into a sequence of discrete behavioral events and maps each action to a structured representation containing Character States and Entity States. As the narrative progresses, the World Model explicitly updates these attributes, maintaining a precise and globally consistent environment description at every time step.

Intervention Procedure When the detector described in Equation (5) triggers at position i, the system executes a three-step procedure. First, the current reasoning trajectory is interrupted at the trigger point, preserving the prefix up to and including the trigger word. Then, the system instructs the LLM to construct a structured World Model based on the current narrative context, producing character states and entity states that reflect the objective facts at the corresponding point in the story. Fi-

Model	ToMi		Hi-ToM		ExploreToM		ToMBench		EmoBench	
	Acc	Token	Acc	Token	Acc	Token	Acc	Token	Acc	Token
Non-reasoning LLMs										
DeepSeek-v3.2	79.67	425	69.33	1394	72.67	417	77.54	440	54.50	829
Llama-3.1-405B-Instruct	77.33	356	68.00	1490	68.68	362	73.95	460	54.50	917
Llama-3.3-70B-Instruct	76.50	345	65.80	1450	64.20	375	74.50	495	56.75	849
Qwen2.5-72B-Instruct	79.67	396	63.68	1548	55.00	402	73.84	505	52.00	1017
Qwen-Max	76.33	327	66.67	1397	75.33	419	75.04	504	56.50	992
GPT-4o	74.00	670	63.00	1402	56.68	512	75.22	415	62.25	852
Reasoning LLMs										
o3	84.33	1430	86.67	3248	75.00	2647	81.80	2142	62.75	2661
DeepSeek-R1	93.33	2846	77.33	7710	78.00	2124	80.13	4571	67.50	4197
GPT-5.5	96.67	717	95.67	945	95.33	863	83.78	1264	70.50	1394
Claude-Opus-4.7	97.00	764	92.67	1512	90.33	942	82.58	1936	69.75	1417
Gemini-3.1-Pro	96.00	547	95.00	1223	95.67	737	81.39	1380	71.50	1448
DS-R1-DQ-7B	55.33	3552	43.33	6080	60.67	4087	54.19	3418	41.50	3771
+WG-ToM (ours)	57.67$_{\uparrow 2.34}$	2548$_{\downarrow 1004}$	47.67$_{\uparrow 4.34}$	5072$_{\downarrow 1008}$	61.67$_{\uparrow 1.00}$	2895$_{\downarrow 1192}$	65.32$_{\uparrow 11.13}$	2337$_{\downarrow 1081}$	49.50$_{\uparrow 8.00}$	2504$_{\downarrow 1267}$
DS-R1-DQ-14B	70.33	3706	59.00	7889	65.33	3097	71.79	4152	54.25	4428
+WG-ToM (ours)	73.00$_{\uparrow 2.67}$	3007$_{\downarrow 699}$	62.33$_{\uparrow 3.33}$	6194$_{\downarrow 1695}$	66.67$_{\uparrow 1.34}$	2239$_{\downarrow 858}$	76.27$_{\uparrow 4.48}$	3194$_{\downarrow 958}$	60.75$_{\uparrow 6.50}$	3520$_{\downarrow 908}$
DS-R1-DQ-32B	79.33	3504	69.33	7541	69.67	3185	77.18	4830	61.25	4139
+WG-ToM (ours)	84.33$_{\uparrow 5.00}$	2709$_{\downarrow 795}$	78.33$_{\uparrow 9.00}$	5183$_{\downarrow 2358}$	73.67$_{\uparrow 4.00}$	2515$_{\downarrow 670}$	83.39$_{\uparrow 6.21}$	2434$_{\downarrow 2396}$	67.25$_{\uparrow 6.00}$	2803$_{\downarrow 1336}$
Qwen3-8B-Reasoning	74.67	2713	74.67	4795	67.67	2359	69.41	3048	55.25	3529
+WG-ToM (ours)	77.00$_{\uparrow 2.33}$	2173$_{\downarrow 540}$	78.33$_{\uparrow 3.66}$	3297$_{\downarrow 1498}$	71.00$_{\uparrow 3.33}$	1729$_{\downarrow 630}$	75.33$_{\uparrow 5.92}$	1950$_{\downarrow 1098}$	60.25$_{\uparrow 5.00}$	2783$_{\downarrow 746}$
Qwen3-14B-Reasoning	76.00	2619	83.33	4269	71.00	3061	74.34	3554	60.50	3828
+WG-ToM (ours)	79.00$_{\uparrow 3.00}$	2061$_{\downarrow 558}$	87.33$_{\uparrow 4.00}$	3287$_{\downarrow 982}$	74.00$_{\uparrow 3.00}$	2421$_{\downarrow 640}$	79.25$_{\uparrow 4.91}$	2268$_{\downarrow 1286}$	64.75$_{\uparrow 4.25}$	2507$_{\downarrow 1321}$
Qwen3-32B-Reasoning	81.33	2942	86.00	4316	74.33	3193	77.54	3729	63.50	2541
+WG-ToM (ours)	85.33$_{\uparrow 4.00}$	2309$_{\downarrow 633}$	90.67$_{\uparrow 4.67}$	3252$_{\downarrow 1064}$	77.67$_{\uparrow 3.34}$	2384$_{\downarrow 809}$	83.59$_{\uparrow 6.05}$	2071$_{\downarrow 1658}$	68.50$_{\uparrow 5.00}$	1454$_{\downarrow 1087}$

Table 2: Performance of LLMs on Social Reasoning Datasets. DS-R1-DQ refers to DeepSeek-R1-Distill. For LLMs with dual entries, the first row shows baseline performance while the second row shows results after improvements.

nally, the constructed state description is injected in a structured format:<**information**> **{State from World Model}** <**/information**>. LLM then re-examines its previous reasoning and conducts secondary reasoning, correcting the erroneous judgments previously caused by conflation of objective states and subjective beliefs.

This intervention process can be applied iteratively for up to k rounds. After each intervention, the detector scans the newly generated output; if confusion signals persist, a subsequent intervention is triggered with an updated state description. The iterative process terminates when either no further confusion is detected or the maximum intervention count k is reached.

4 Experiments

4.1 Experimental Setup

To systematically evaluate ToM, we select ToMi, HiToM, ExploreToM, ToMBench and EmoBench as evaluation benchmarks. We select multiple mainstream LLMs including (i) **Non-Reasoning LLMs:** DeepSeek-v3.2 (DeepSeek-AI, 2025), Llama-3.1-405B-Instruct (Team, 2024), Llama-3.3-70B-Instruct (Team, 2024), Qwen2.5-72B-Instruct (Yang et al., 2024), Qwen-Max (Qwen Team, 2025), and GPT-4o (OpenAI, 2024); (ii) **Reasoning LLMs:** OpenAI o3 (OpenAI, 2025b), DeepSeek-R1 (Guo et al., 2025), GPT-5.5 (OpenAI, 2026), Claude-Opus-4.7 (Anthropic, 2026), Gemini-3.1-Pro (Google DeepMind, 2026), DeepSeek-R1-Distill series (7B/14B/32B) (Guo et al., 2025), and Qwen3-Reasoning series (8B/14B/32B) (Yang et al., 2025). We employ accuracy and total token consumption as metrics. Notably, **the token consumption metric explicitly includes the computational overhead incurred by the generation of World Model or World Knowledge**, ensuring a fair cost-benefit assessment. The sliding window size W is set to 60, the word similarity threshold $\tau_w = 0.65$ and the context coherence threshold $\tau_c = 0.35$. The cooldown window $d = 200$ tokens. The maximum intervention threshold k for the World Model mechanism is set to 3 and the World Knowledge is generated by Gemini-3-Pro (Google DeepMind, 2026). All reported results are averaged over three runs with different random seeds (standard deviations within ±0.5% for accuracy).

Method	ToMi		Hi-ToM		ExploreToM		ToMBench		EmoBench	
	Acc	Token	Acc	Token	Acc	Token	Acc	Token	Acc	Token
DeepSeek-R1-Distill-Qwen-32B										
Baseline	79.33	3504	69.33	7541	69.67	3185	77.18	4830	61.25	4139
+CCoT (Renze and Guven, 2024)	79.67$_{\uparrow 0.34}$	3128$_{\downarrow 376}$	69.67$_{\uparrow 0.34}$	6812$_{\downarrow 729}$	69.00$_{\downarrow 0.67}$	2874$_{\downarrow 311}$	76.75$_{\downarrow 0.43}$	4131$_{\downarrow 699}$	60.75$_{\downarrow 0.50}$	3486$_{\downarrow 653}$
+ReAct (Yao et al., 2023)	81.00$_{\uparrow 1.67}$	2198$_{\downarrow 306}$	70.67$_{\uparrow 1.34}$	7234$_{\downarrow 307}$	69.67$_{-0.00}$	3256$_{\uparrow 71}$	76.84$_{\downarrow 0.34}$	4558$_{\downarrow 272}$	61.75$_{\uparrow 0.50}$	3919$_{\downarrow 220}$
+Reflexion (Shinn et al., 2023)	80.67$_{\uparrow 1.34}$	3076$_{\downarrow 428}$	71.33$_{\uparrow 2.00}$	6687$_{\downarrow 854}$	70.33$_{\uparrow 0.66}$	2934$_{\downarrow 251}$	77.50$_{\uparrow 0.32}$	4219$_{\downarrow 611}$	62.25$_{\uparrow 1.00}$	3661$_{\downarrow 478}$
+s1 (Muennighoff et al., 2025)	80.67$_{\uparrow 1.34}$	4123$_{\uparrow 619}$	70.33$_{\uparrow 1.00}$	7855$_{\uparrow 314}$	69.33$_{\downarrow 0.34}$	3750$_{\uparrow 565}$	76.45$_{\downarrow 0.73}$	5328$_{\uparrow 498}$	61.50$_{\uparrow 0.25}$	4451$_{\uparrow 312}$
+COD (Xu et al., 2025)	80.33$_{\uparrow 1.00}$	3157$_{\downarrow 347}$	70.00$_{\uparrow 0.67}$	6620$_{\downarrow 921}$	69.33$_{\downarrow 0.34}$	3120$_{\downarrow 65}$	77.30$_{\uparrow 0.12}$	5283$_{\uparrow 453}$	62.00$_{\uparrow 0.75}$	4089$_{\downarrow 50}$
+α1 (Zhang et al., 2025b)	80.67$_{\uparrow 1.34}$	2839$_{\downarrow 665}$	71.33$_{\uparrow 2.00}$	5591$_{\downarrow 1950}$	69.67$_{-0.00}$	2734$_{\downarrow 451}$	75.82$_{\downarrow 1.36}$	3306$_{\downarrow 1524}$	61.75$_{\uparrow 0.50}$	3424$_{\downarrow 715}$
+S2F & T2M (Gong et al., 2026)	81.33$_{\uparrow 2.00}$	**2118**$_{\downarrow 1386}$	73.67$_{\uparrow 4.34}$	**4769**$_{\downarrow 2772}$	70.67$_{\uparrow 1.00}$	**2204**$_{\downarrow 981}$	63.10$_{\downarrow 14.08}$	3302$_{\downarrow 1528}$	52.75$_{\downarrow 8.50}$	3309$_{\downarrow 830}$
+PercepToM (Jung et al., 2024)	82.00$_{\uparrow 2.67}$	4795$_{\uparrow 1291}$	73.33$_{\uparrow 4.00}$	7745$_{\uparrow 204}$	71.00$_{\uparrow 1.33}$	4708$_{\uparrow 1523}$	75.52$_{\downarrow 1.66}$	5620$_{\uparrow 790}$	58.75$_{\downarrow 2.50}$	4803$_{\uparrow 664}$
+SymbolicToM (Sclar et al., 2023)	80.67$_{\uparrow 1.34}$	3779$_{\uparrow 275}$	70.33$_{\uparrow 1.00}$	7229$_{\downarrow 312}$	70.33$_{\uparrow 0.66}$	3043$_{\downarrow 142}$	78.45$_{\uparrow 1.27}$	4502$_{\downarrow 328}$	62.50$_{\uparrow 1.25}$	4225$_{\uparrow 86}$
+SimToM (Wilf et al., 2024)	80.66$_{\uparrow 1.33}$	4552$_{\uparrow 1048}$	72.00$_{\uparrow 2.67}$	7804$_{\uparrow 263}$	70.33$_{\uparrow 0.66}$	4370$_{\uparrow 1185}$	77.26$_{\uparrow 0.08}$	5469$_{\uparrow 639}$	61.00$_{\downarrow 0.25}$	4617$_{\uparrow 478}$
+WG-ToM (ours)	**84.33**$_{\uparrow 5.00}$	2709$_{\downarrow 795}$	**78.33**$_{\uparrow 9.00}$	5183$_{\downarrow 2358}$	**73.67**$_{\uparrow 4.00}$	2515$_{\downarrow 670}$	**83.39**$_{\uparrow 6.21}$	**2434**$_{\downarrow 2396}$	**67.25**$_{\uparrow 6.00}$	**2803**$_{\downarrow 1336}$
Qwen3-32B-Reasoning										
Baseline	81.33	2942	86.00	4316	74.33	3193	77.54	3729	63.50	2541
+CCoT (Renze and Guven, 2024)	81.67$_{\uparrow 0.34}$	2618$_{\downarrow 324}$	85.67$_{\downarrow 0.33}$	3624$_{\downarrow 692}$	74.00$_{\downarrow 0.33}$	2871$_{\downarrow 322}$	76.96$_{\downarrow 0.58}$	3168$_{\downarrow 561}$	63.00$_{\downarrow 0.50}$	2094$_{\downarrow 447}$
+ReAct (Yao et al., 2023)	82.00$_{\uparrow 0.67}$	2614$_{\downarrow 328}$	87.00$_{\uparrow 1.00}$	4153$_{\downarrow 163}$	74.67$_{\uparrow 0.34}$	3028$_{\downarrow 165}$	77.93$_{\uparrow 0.39}$	3612$_{\downarrow 117}$	64.50$_{\uparrow 1.00}$	2439$_{\downarrow 102}$
+Reflexion (Shinn et al., 2023)	82.67$_{\uparrow 1.34}$	2845$_{\downarrow 97}$	87.33$_{\uparrow 1.33}$	4209$_{\downarrow 107}$	75.33$_{\uparrow 1.00}$	3110$_{\downarrow 83}$	78.05$_{\uparrow 0.51}$	3588$_{\downarrow 141}$	64.00$_{\uparrow 0.50}$	2504$_{\downarrow 37}$
+s1 (Muennighoff et al., 2025)	81.67$_{\uparrow 0.34}$	3412$_{\uparrow 470}$	86.33$_{\uparrow 0.33}$	4655$_{\uparrow 339}$	74.00$_{\downarrow 0.33}$	3589$_{\uparrow 396}$	77.15$_{\downarrow 0.39}$	4301$_{\uparrow 572}$	63.25$_{\downarrow 0.25}$	2876$_{\uparrow 335}$
+COD (Xu et al., 2025)	81.67$_{\uparrow 0.34}$	2689$_{\downarrow 253}$	86.33$_{\uparrow 0.33}$	3912$_{\downarrow 404}$	74.67$_{\uparrow 0.34}$	2950$_{\downarrow 243}$	77.85$_{\uparrow 0.31}$	3521$_{\downarrow 208}$	63.75$_{\uparrow 0.25}$	2415$_{\downarrow 126}$
+α1 (Zhang et al., 2025b)	82.33$_{\uparrow 1.00}$	2456$_{\downarrow 486}$	87.00$_{\uparrow 1.00}$	3489$_{\downarrow 827}$	75.00$_{\uparrow 0.67}$	2512$_{\downarrow 681}$	77.62$_{\uparrow 0.08}$	2945$_{\downarrow 784}$	63.75$_{\uparrow 0.25}$	2188$_{\downarrow 353}$
+S2F & T2M (Gong et al., 2026)	82.00$_{\uparrow 0.67}$	**1856**$_{\downarrow 1086}$	87.67$_{\uparrow 1.67}$	**3102**$_{\downarrow 1214}$	74.67$_{\uparrow 0.34}$	**2145**$_{\downarrow 1048}$	65.30$_{\downarrow 12.24}$	2645$_{\downarrow 1084}$	55.50$_{\downarrow 8.00}$	2412$_{\downarrow 129}$
+PercepToM (Jung et al., 2024)	82.67$_{\uparrow 1.34}$	3848$_{\uparrow 906}$	87.66$_{\uparrow 1.66}$	5230$_{\uparrow 914}$	75.33$_{\uparrow 1.00}$	4226$_{\uparrow 1033}$	74.48$_{\downarrow 3.06}$	4571$_{\uparrow 842}$	61.25$_{\downarrow 2.25}$	3280$_{\uparrow 739}$
+SymbolicToM (Sclar et al., 2023)	82.00$_{\uparrow 0.67}$	3091$_{\uparrow 149}$	87.33$_{\uparrow 1.33}$	4173$_{\downarrow 143}$	74.67$_{\uparrow 0.34}$	2740$_{\downarrow 453}$	77.18$_{\downarrow 0.36}$	3504$_{\downarrow 225}$	64.00$_{\uparrow 0.50}$	2725$_{\uparrow 184}$
+SimToM (Wilf et al., 2024)	82.00$_{\uparrow 0.67}$	3649$_{\uparrow 707}$	87.33$_{\uparrow 1.33}$	4887$_{\uparrow 571}$	75.33$_{\uparrow 1.00}$	3508$_{\uparrow 315}$	77.18$_{\downarrow 0.36}$	4229$_{\uparrow 500}$	64.00$_{\uparrow 0.50}$	3092$_{\uparrow 551}$
+WG-ToM (ours)	**85.33**$_{\uparrow 4.00}$	2309$_{\downarrow 633}$	**90.67**$_{\uparrow 4.67}$	3252$_{\downarrow 1064}$	**77.67**$_{\uparrow 3.34}$	2384$_{\downarrow 809}$	**83.59**$_{\uparrow 6.05}$	**2071**$_{\downarrow 1658}$	**68.50**$_{\uparrow 5.00}$	**1454**$_{\downarrow 1087}$

Table 3: Comparison with different training-free Test-time Scaling methods on social reasoning tasks.

Due to space constraints, additional evaluations are provided in the Appendix C, including broader model comparisons (C.1), hyperparameter analysis (C.2 - C.4), multi-source knowledge ablation (C.5), and advanced LLM results (C.6).

4.2 Main Results

Reasoning LLMs Significantly Outperform Non-Reasoning LLMs. As shown in Table 2, reasoning LLMs (e.g., GPT-5.5, o3) consistently outperform non-reasoning LLMs (e.g., GPT-4o, Llama-3.3) across all datasets. On ToMi, Claude-Opus-4.7 achieves 97.00% accuracy, substantially surpassing GPT-4o (74.00%). This indicates that complex social inference necessitates a reasoning mechanism based on CoT.

WG-ToM Consistently Improves Both Accuracy and Efficiency Across Social Reasoning Task. WG-ToM yields consistent accuracy gains and token reduction across all five benchmarks. For cognitive dilemmas, WG-ToM breaks reasoning loops through real-time confusion detection and World Model intervention: DS-R1-DQ-32B achieves a 9.00% accuracy increase on Hi-ToM with a reduction of 2358 tokens. For knowledge blindness, WG-ToM compensates for social norm deficits through World Knowledge injection: Qwen3-32B-Reasoning achieves a 6.05% accuracy boost on ToMBench with a reduction of 1658 tokens. The simultaneous gains in accuracy and token reduction confirm that WG-ToM improves performance not through additional computation, but by providing grounded guidance that prevents unproductive reasoning.

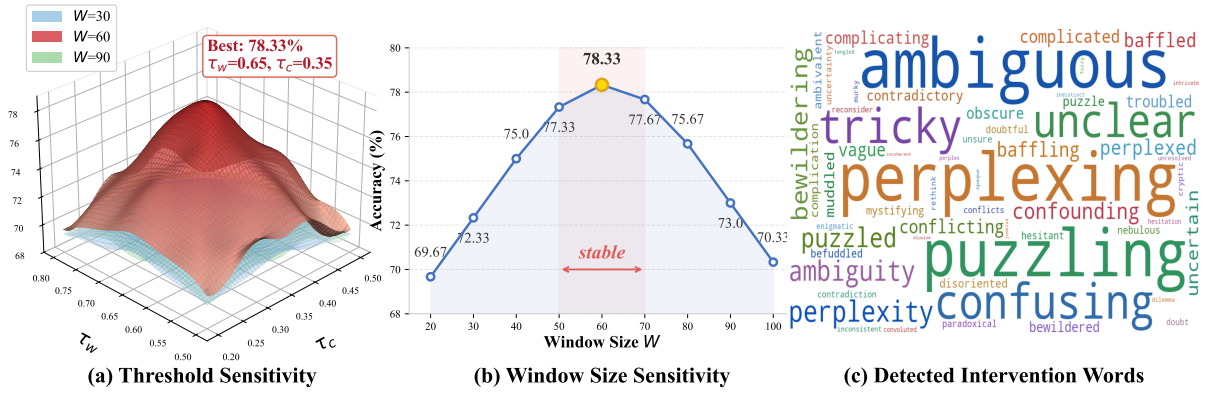
4.3 Results on Test-time Scaling Methods

Table 3 compares WG-ToM with various training-free Test-time Scaling methods, including CCoT, ReAct, Reflexion, s1, COD, α1, S2F & T2M, PercepToM, SymbolicToM and SimToM. The results show that WG-ToM consistently achieves significant improvements in both accuracy and efficiency. Regarding S2F & T2M, although it achieves more substantial token reduction through strict inference interruption and removal of options, it incurs a severe accuracy penalty on option-dependent benchmarks such as ToMBench and EmoBench.

More broadly, methods that operate within the same ungrounded reasoning space (CCoT, Reflexion, s1, COD) yield marginal gains ($\leq$ 2.00% on most benchmarks), confirming the bottleneck lies in missing knowledge and state-belief conflation rather than reasoning depth or format. WG-ToM addresses these deficits directly through targeted grounding, achieving the largest accuracy gains and token reduction simultaneously.

Method	ToMi		Hi-ToM		ExploreToM		ToMBench		EmoBench	
	Acc	Token	Acc	Token	Acc	Token	Acc	Token	Acc	Token
DeepSeek-R1-Distill-Qwen-32B										
Baseline	79.33	3504	69.33	7541	69.67	3185	77.18	4830	61.25	4139
+WK	79.67	3379	70.00	7225	70.33	3018	82.75	2972	66.75	3167
+WM	83.67	2874	77.00	5428	73.00	2667	78.42	4429	62.50	3790
+WG-ToM	**84.33**	**2709**	**78.33**	**5183**	**73.67**	**2515**	**83.39**	**2434**	**67.25**	**2803**
Qwen3-32B-Reasoning										
Baseline	81.33	2942	86.00	4316	74.33	3193	77.54	3729	63.50	2541
+WK	82.00	2757	86.67	4087	74.67	2934	83.13	2781	67.75	1662
+WM	84.67	2490	89.67	3551	77.33	2473	78.15	3508	64.00	2417
+WG-ToM	**85.33**	**2309**	**90.67**	**3252**	**77.67**	**2384**	**83.59**	**2071**	**68.50**	**1454**

Table 4: Ablation study of WG-ToM components. +WK denotes World Knowledge Injection only, +WM denotes World Model Intervention only, and +WG-ToM denotes the full framework.



(a) Threshold Sensitivity (b) Window Size Sensitivity (c) Detected Intervention Words

Figure 7: Ablation study of WG-ToM parameters. (a) Threshold sensitivity over (τ_w, τ_c) under different window sizes W; (b) Accuracy across window sizes W, where the highlighted band marks the stable region; (c) Word cloud of the detected intervention words.

4.4 Ablation Study

The Two Mechanisms Are Complementary and Indispensable. As shown in Table 4, applying World Knowledge alone (+WK) yields notable gains on social norm tasks (e.g., +5.57% on ToMBench for DS-R1-DQ-32B) but provides negligible benefit for state-tracking tasks. Conversely, applying World Model alone (+WM) improves dynamic scenarios (e.g., +7.67% on Hi-ToM) while contributing little to social norm tasks. Combining both mechanisms (+WG-ToM) achieves the best performance across all benchmarks. Neither alone is sufficient to achieve the full gains.

Parameter Sensitivity and Robustness. As shown in Figure 7, the threshold sensitivity in (a) shows that the optimal configuration ($\tau_w = 0.65$, $\tau_c = 0.35$) achieves 78.33%, and that as the thresholds approach extreme values, performance only smoothly falls back to the baseline (69.33%) rather than lower. The window sensitivity in (b) shows that accuracy remains stable within $W \in [50, 70]$ and peaks at $W = 60$, indicating insensitivity to this hyperparameter. Finally, the word cloud of detected intervention words in (c) is highly concentrated on confusion-related terms such as *ambiguous*, *perplexing*, and *confusing*, confirming that the confusion detector captures genuine confusion signals rather than random noise.

5 Conclusions

In this paper, we address the critical limitations of LLMs in social reasoning, specifically Knowledge Blindness and Cognitive Dilemmas that lead to conflation of objective facts and subjective thoughts. We propose WG-ToM framework, which integrates World Knowledge Injection and World Model Intervention to provide grounded guidance. The approach effectively helps LLMs compensate for knowledge deficits and navigate confusion. Extensive results demonstrate WG-ToM significantly improves accuracy while reducing token consumption, offering a simple yet effective solution for deploying LLMs in social contexts.

Limitations

Applicability to Closed-Source LLMs The World Model Intervention mechanism requires real-time access to the LLM's intermediate reasoning tokens in order to monitor confusion signals and inject structured state descriptions at the point of detection. This level of granular control over the generation process is only available in open-source LLMs with accessible decoding interfaces. For closed-source LLMs, the internal reasoning trajectory is typically not exposed to users, making it infeasible to apply real-time monitoring and mid-generation intervention. Future work could explore post-hoc intervention strategies that operate on completed reasoning segments, or leverage future API features that expose intermediate reasoning tokens.

Fixed Knowledge Injection The current implementation of world knowledge injection is relatively static. Future work aims to establish a comprehensive Social Norm Knowledge Base that allows the framework to adaptively retrieve and dynamically match specific norms based on the unique attributes of each task. By shifting from fixed prompts to an automated elicitation and matching process, we can better leverage the intelligence of advanced LLMs to activate a broader range of tacit social knowledge.

Generalizability Across Domains Our evaluation is currently centered on a targeted set of LLMs within theory of mind benchmarks. Future work can extend the framework to a wider range of LLMs and broader social domains, such as moral reasoning, to establish more comprehensive generalizability.

Ethics Statement

Dataset All the LLMs and datasets used in this work are publicly published with permissible licenses. The datasets used in this study are widely recognized open-source for social reasoning. These datasets consist of neutral and controlled social scenarios designed to evaluate pure cognitive reasoning, which inherently contain no personally identifiable information or sensitive personal data. Furthermore, we conducted a manual inspection of a random sample of reasoning chains and found no instances of offensive, biased, or harmful content.

AI Assistance for Writing We used AI assistants to help with grammar checking and formatting in this paper, but all technical content, analysis and conclusions represent the original work of the authors.

References

Anthropic. 2025. Claude 4 system card. Technical report, Anthropic.

Anthropic. 2026. Claude Opus 4.7 System Card. Technical report, Anthropic.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind" ? *Cognition*, 21(1):37–46.

Ruirui Chen, Weifeng Jiang, Chengwei Qin, and Cheston Tan. 2025. Theory of mind in large language models: Assessment and enhancement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 31539–31558. Association for Computational Linguistics.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024. 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, Bangkok, Thailand, August 11-16, 2024*, pages 15959–15983. Association for Computational Linguistics.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *CoRR*, abs/2412.19437.

DeepSeek-AI. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *CoRR*, abs/2512.02556.

Yifu Ding, Wentao Jiang, Shunyu Liu, Yongcheng Jing, Jinyang Guo, Yingjie Wang, Jing Zhang, Zengmao Wang, Ziwei Liu, Bo Du, Xianglong Liu, and Dacheng Tao. 2025. Dynamic parallel tree search for efficient LLM reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 11233–11252. Association for Computational Linguistics.

Nanxu Gong, Haotian Li, Sixun Dong, Jianxun Lian, Yanjie Fu, and Xing Xie. 2026. To think or not to think, that is the question for large reasoning models in theory of mind tasks. *Preprint*, arXiv:2602.10625.

Google DeepMind. 2026. Gemini 3.1 Pro model card. Technical report, Google DeepMind.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, Proceedings of Machine Learning Research, pages 1321–1330. PMLR.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nat.*, 645(8081):633–638.

Jennifer Hu, Felix Sosa, and Tomer D. Ullman. 2025. Re-evaluating theory of mind evaluation in large language models. *CoRR*, abs/2502.21098.

Chani Jung, Dongkwan Kim, Jiho Jin, Jiseon Kim, Yeon Seonwoo, Yejin Choi, Alice Oh, and Hyunwoo Kim. 2024. 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.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, Hong Kong, China. Association for Computational Linguistics.

Cheng Li, Jiexiong Liu, Yixuan Chen, Yanqin Jia, and Zhepeng Li. 2025a. Kunlunbaize: LLM with multi-scale convolution and multi-token prediction under transformerx framework. *CoRR*, abs/2503.04784.

Peiji Li, Kai Lv, Yunfan Shao, Yichuan Ma, Linyang Li, Xiaoqing Zheng, Xipeng Qiu, and Qipeng Guo. 2025b. Fastmcts: A simple sampling strategy for data synthesis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 24405–24422. Association for Computational Linguistics.

Baohao Liao, Yuhui Xu, Hanze Dong, Junnan Li, Christof Monz, Silvio Savarese, Doyen Sahoo, and Caiming Xiong. 2025. Reward-guided speculative decoding for efficient LLM reasoning. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel J. Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 20275–20321. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o system card. *CoRR*, abs/2410.21276.

OpenAI. 2025a. GPT-5 System Card. Technical report, OpenAI.

OpenAI. 2025b. OpenAI o3 and o4-mini System Card. Technical report, OpenAI.

OpenAI. 2025c. OpenAI o3-mini System Card. Technical report, OpenAI.

OpenAI. 2026. GPT-5.5 System Card. Technical report, OpenAI.

Josef Perner, Susan R. Leekam, and Heinz Wimmer. 1987. Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology*, 5(2):125–137.

Hongjin Qian, Zheng Liu, Peitian Zhang, Kelong Mao, Defu Lian, Zhicheng Dou, and Tiejun Huang. 2025. Memorag: Boosting long context processing with global memory-enhanced retrieval augmentation. In *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025-2 May 2025*, pages 2366–2377. ACM.

Qwen Team. 2025. Qwen2.5-Max: Exploring the intelligence of large-scale MoE model. `https://qwen.ai/blog?id=qwen2.5-max`.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Matthew Renze and Erhan Guven. 2024. The benefits of a concise chain of thought on problem-solving in large language models. In *2nd International Conference on Foundation and Large Language Models, FLLM 2024, Dubai, United Arab Emirates, November 26-29, 2024*, pages 476–483. IEEE.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M. Liu, Jinfeng Zhou, Alvionna S. Sunaryo, Tatia M. C. Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages

5986–6004. Association for Computational Linguistics.

Melanie Sclar, Jane Dwivedi-Yu, Maryam Fazel-Zarandi, Yulia Tsvetkov, Yonatan Bisk, Yejin Choi, and Asli Celikyilmaz. 2025. Explore theory of mind: program-guided adversarial data generation for theory of mind reasoning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. 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 (Volume 1: Long Papers)*, pages 13960–13980, Toronto, Canada. Association for Computational Linguistics.

ByteDance Seed. 2025. Seed1.5-thinking: Advancing superb reasoning models with reinforcement learning. *CoRR*, abs/2504.13914.

Claude E. Shannon. 2001. A mathematical theory of communication. *ACM SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55.

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 *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 1510–1519. AAAI Press.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261.

Llama Team. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Xu Wan, Wei Wang, Wenyue Xu, Wotao Yin, Jie Song, and Mingyang Sun. 2025. Adapthink: Adaptive thinking preferences for reasoning language model. *CoRR*, abs/2506.18237.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yiming Wang, Pei Zhang, Siyuan Huang, Baosong Yang, Zhuosheng Zhang, Fei Huang, and Rui Wang. 2025. Sampling-efficient test-time scaling: Self-estimating the best-of-n sampling in early decoding. *CoRR*, abs/2503.01422.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Alex Wilf, Sihyun Shawn 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), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 8292–8308. Association for Computational Linguistics.

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, Singapore, December 6-10, 2023*, Findings of ACL, pages 10691–10706. Association for Computational Linguistics.

Silei Xu, Wenhao Xie, Lingxiao Zhao, and Pengcheng He. 2025. Chain of draft: Thinking faster by writing less. *CoRR*, abs/2502.18600.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jiajie Zhang, Nianyi Lin, Lei Hou, Ling Feng, and Juanzi Li. 2025a. Adaptthink: Reasoning models can learn when to think. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November*

4-9, 2025, pages 3716–3730. Association for Computational Linguistics.

Junyu Zhang, Runpei Dong, Han Wang, Xuying Ning, Haoran Geng, Peihao Li, Xialin He, Yutong Bai, Jitendra Malik, Saurabh Gupta, and Huan Zhang. 2025b. Alphaone: Reasoning models thinking slow and fast at test time. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 11329–11354. Association for Computational Linguistics.

Zhining Zhang, Chuanyang Jin, Mung Yao Jia, and Tianmin Shu. 2025c. Autotom: Automated bayesian inverse planning and model discovery for open-ended theory of mind. *CoRR*, abs/2502.15676.

APPENDIX

A Related Work 13

B Experiment Settings 13
- B.1 Datasets 13
- B.2 Test-time Scaling Methods 14

C Supplementary Experimental Results 15
- C.1 Additional Results 15
- C.2 Evaluation of Maximum Intervention Count 15
- C.3 Evaluation of Intervention Words 15
- C.4 Evaluation of Confidence and Entropy 16
- C.5 Evaluation of Multi-Source World Knowledge 16
- C.6 Evaluation on Advanced LLMs 17

D Qualitative Examples 18
- D.1 Sample Example of World Model Generation 18
- D.2 Sample Example of World Knowledge Generation 18
- D.3 Prompts Design 19
- D.4 Case Study 19

A Related Work

Strategies for Enhancing LLMs' Cognitive Development in Social Domain Recent methods for improving LLM performance on ToM tasks continue to emerge. SimToM and PercepToM improve belief reasoning through perspective-taking and perception decoupling (Wilf et al., 2024; Jung et al., 2024), respectively. AutoToM applies Bayesian inverse planning for open-ended belief inference (Zhang et al., 2025c), MuMA-ToM extends Bayesian frameworks to multimodal multi-agent settings (Shi et al., 2025), and S2F & T2M designs slow-fast reasoning switching and option decoupling mechanisms specifically for ToM tasks (Gong et al., 2026). Recent work has also systematically assessed and enhanced ToM capabilities in LLMs (Chen et al., 2025), and re-evaluated the limitations of ToM evaluation methodologies from a meta-analytic perspective (Hu et al., 2025). However, all these methods operate before or after reasoning rather than intervening during the reasoning process. Our framework uniquely monitors reasoning trajectories in real time and injects factual corrections at the point of confusion.

Test-time scaling Test-time scaling methods improve reasoning capabilities without updating model parameters, with recent work focusing on adaptive control and efficiency optimization. s1 explores budget-controlled scaling (Muennighoff et al., 2025), AlphaOne adaptively adjusts reasoning depth (Zhang et al., 2025b), AdaptThink dynamically switches between deliberative and fast inference modes (Zhang et al., 2025a), FastMCTS leverages Monte Carlo tree search to prioritize high-confidence trajectories (Li et al., 2025b), DPTS improves tree-structured reasoning efficiency through parallelization (Ding et al., 2025), Self-Truncation BoN introduces early stopping to improve sampling efficiency (Wang et al., 2025), and Reward-Guided Speculative Decoding uses process reward models to filter high-quality reasoning results (Liao et al., 2025). These methods either generate more reasoning within the same ungrounded space or apply post-hoc filtering, without injecting structured external knowledge at the point of failure. Our method provides targeted factual intervention precisely when confusion is detected, addressing the root cause rather than the symptom.

Knowledge Augmentation for Social Reasoning Knowledge augmentation methods supplement reasoning context through external knowledge. Self-RAG introduces self-reflection to determine when to retrieve and whether the retrieved content is useful (Asai et al., 2024), and MemoRAG uses a memory model to generate clues that guide retrieval (Qian et al., 2025). However, existing social reasoning methods rely on the LLM's own knowledge without introducing external social norms. In contrast, our World Knowledge injection mechanism explicitly elicits social norms from advanced LLMs and conditionally injects them as structured premises through task-adaptive routing, bridging the LLM's inherent knowledge gap rather than compensating for retrieval insufficiency.

B Experiment Settings

B.1 Datasets

ToMi (Le et al., 2019) ToMi is a controlled benchmark designed to evaluate first-order theory-of-mind reasoning in short textual narratives. It primarily focuses on whether LLM can correctly distinguish an agent's belief from objective reality, especially under false-belief settings. The dataset features concise stories with minimal distractors, making it suitable for isolating core belief attribu-

tion capabilities without requiring complex multi-step reasoning or extensive world knowledge.

HiToM (Wu et al., 2023) HiToM extends traditional ToM evaluation to higher-order belief reasoning, requiring LLMs to infer nested mental states such as "what one agent believes about another agent's belief." Compared to ToMi, HiToM introduces increased structural complexity and deeper recursion depth, posing a greater challenge for LLMs that rely on shallow pattern matching or surface-level heuristics. This benchmark is particularly useful for assessing whether LLMs can consistently maintain and update multiple layers of belief representations.

ExploreToM (Sclar et al., 2025) ExploreToM is a stress-test benchmark for theory-of-mind reasoning constructed via program-guided adversarial data generation. Rather than relying on fixed templates, it systematically explores diverse belief configurations and misleading cues to expose failure modes in LLM reasoning. As a result, ExploreToM emphasizes robustness and generalization, often revealing brittle behaviors that remain hidden on more curated ToM datasets.

ToMBench (Chen et al., 2024) ToMBench provides a comprehensive and structured evaluation of theory-of-mind abilities across multiple dimensions, including belief reasoning, intention understanding, and perspective taking. It includes both English and Chinese data and organizes tasks into fine-grained categories, enabling detailed diagnostic analysis of LLM strengths and weaknesses. ToMBench serves as a unifying benchmark that bridges classic ToM formulations with large-scale evaluation protocols tailored to modern LLMs.

EmoBench (Sabour et al., 2024) EmoBench evaluates emotional intelligence and affective reasoning, which are closely related to theory-of-mind capabilities in social contexts. The dataset assesses LLMs' ability to recognize, interpret, and reason about emotions expressed in text, as well as to infer emotional states underlying social interactions. Unlike belief-centric ToM benchmarks, EmoBench emphasizes affective grounding and socio-emotional understanding, providing a complementary view of social cognition beyond belief attribution alone.

B.2 Test-time Scaling Methods

CCoT (Renze and Guven, 2024) Concise Chain-of-Thought (CCoT) encourages LLMs to generate shorter and more compact reasoning traces by explicitly constraining the verbosity of intermediate reasoning. Rather than expanding multiple reasoning branches, CCoT focuses on producing a single concise chain, aiming to reduce inference cost while largely preserving answer correctness.

ReAct (Yao et al., 2023) ReAct interleaves reasoning steps with action executions, enabling LLMs to alternate between natural language reasoning and tool usage. By grounding reasoning in external actions or observations, ReAct allows LLMs to dynamically acquire missing information and adjust subsequent reasoning based on feedback from the environment.

Reflexion (Shinn et al., 2023) Reflexion introduces an explicit self-reflection mechanism at test time, where LLMs critique their own previous outputs and iteratively refine them. This approach treats self-generated feedback as an additional signal to guide subsequent attempts, improving robustness without requiring parameter updates.

s1 (Muennighoff et al., 2025) S1 (Simple Test-Time Scaling) explores the effect of increasing inference-time computation without modifying model parameters or prompts. By allocating a larger decoding budget, S1 demonstrates that reasoning performance can be improved purely through test-time scaling, highlighting the importance of compute allocation strategies.

COD (Xu et al., 2025) Chain-of-Draft (COD) encourages LLMs to produce extremely concise intermediate reasoning in the form of short drafts rather than detailed CoT explanations. By compressing reasoning traces into minimal decision-relevant tokens, COD aims to reduce inference cost while retaining sufficient information for accurate answer generation.

ALPHAONE (Zhang et al., 2025b) AlphaOne frames test-time reasoning as a controlled search process guided by α moment. It adaptively adjusts the reasoning depth and decoding strategy based on linear anneal, allowing LLMs to balance reasoning quality and computational efficiency.

S2F & T2M (Gong et al., 2026) S2F (Slow-to-Fast) and T2M (Think-to-Match) are training-free test-time reasoning control methods designed for Theory-of-Mind tasks. S2F monitors the reasoning process and adaptively switches from slow, deliberative reasoning to fast inference. T2M decouples reasoning from answer options by first performing option-free reasoning and then matching the resulting conclusion to the provided choices.

PercepToM (Jung et al., 2024) PercepToM treats LLM as an objective observer that first maps raw

environmental inputs into a structured list of "perceptual facts". By explicitly decoupling physical observation from mental inference, it minimizes the "omniscience bias" where LLM inadvertently uses information the protagonist could not have known.

SymbolicToM (Sclar et al., 2023) SymbolicToM converts the internal mental states of multiple agents into a formal, symbolic database. Instead of relying on LLM's probabilistic next-token prediction for complex reasoning, it uses structured prompts to track and update beliefs via logical rules.

SimToM (Wilf et al., 2024) SimToM employs a perspective-taking mechanism. It operates in two stages: first, it filters the global context to retain only the information accessible to a specific character; second, it prompts the LLM to "simulate" that character's response within the restricted environment.

C Supplementary Experimental Results

C.1 Additional Results

To broaden our evaluation, we conducted additional tests on a diverse set of LLMs, encompassing both reasoning and non-reasoning LLMs. As shown in Figure 8, the evaluated LLMs include GPT-5 (OpenAI, 2025a), Claude-Sonnet-4 (Anthropic, 2025), Gemini-2.5-Pro (Team, 2025), Xingchen-Plus-RL (Li et al., 2025a), Xingchen-Plus (Li et al., 2025a), o3-mini (OpenAI, 2025c), DeepSeek-v3 (DeepSeek-AI, 2024), Doubao-1.5-Pro (Seed, 2025), Llama-3.1-70B-Instruct (Team, 2024), Qwen2.5-32B-Instruct (Yang et al., 2024), and Qwen2.5-7B-Instruct (Yang et al., 2024).

The results illustrate that large-scale foundation LLMs, such as GPT-5 and Claude-Sonnet-4, maintain competitive performance across the evaluated benchmarks. Reasoning-specific LLMs, such as Xingchen-Plus-RL and o3-mini, also demonstrate robust capabilities, particularly in tasks requiring logical deduction. However, a consistent performance drop is observed across all LLMs on the EmoBench, suggesting that capturing subtle emotional dynamics remains a challenging task for current LLMs.

C.2 Evaluation of Maximum Intervention Count

To evaluate the effectiveness of World Model Intervention, as shown in Table 5, we conduct research by varying the maximum intervention threshold k (ranging from 1 to 5). Since the ablation study in Section 4.4 reveals that World Model Intervention yields the most substantial gains on state-tracking tasks (ToMi, Hi-ToM, and ExploreToM), we select these three benchmarks for this analysis. This parameter controls the maximum frequency for world model to intervene when confusion is detected.

Complex Tasks Require Frequent Interventions. While performance on the simpler ToMi dataset plateaus at $k = 3$, the challenging Hi-ToM benchmark shows continuous significant improvements up to $k = 5$. This indicates that intricate social scenarios trigger more confusion, necessitating more active interventions to sustain reasoning.

Appropriate interventions simultaneously boost accuracy and reduce costs. By timely intervening reasoning paths, LLMs break free from redundant loops, leading to higher accuracy and significantly lower token consumption (e.g., +10.00% accuracy with about 2580 fewer tokens on Hi-ToM for DS-R1-DQ-32B). Notably, the token savings exhibit diminishing marginal returns as k increases: from $k = 3$ to $k = 5$, Hi-ToM token consumption decreases by only 222 tokens compared to 2358 tokens from Baseline to $k = 3$. In principle, an excessively large k may even lead to a net increase in token consumption, as each additional intervention incurs the overhead of constructing and injecting a World Model state, which can eventually outweigh the tokens saved by truncating redundant reasoning. This motivates our default setting of $k = 3$, which strikes an effective balance between accuracy gains and computational cost.

C.3 Evaluation of Intervention Words

To compare the impact of intervention words selection, we compared several categories of commonly used intervention words described in Wan et al. (2025) with our confusion words: Pause-Validation (referred to as PV) and Branch-Extension (referred to as BE) as comparisons.

- **Pause-Validation**: wait, check, make sure, hold on, verify, let me see, confirm, ensure, evaluate, examine.
- **Branch-Extension**: alternatively, another, instead, however, while, yet, though, rather, otherwise, on the other hand.
- **Confusion Words**: ambiguous, complicating, confusing, confused, perplexity, puzzle, puz-

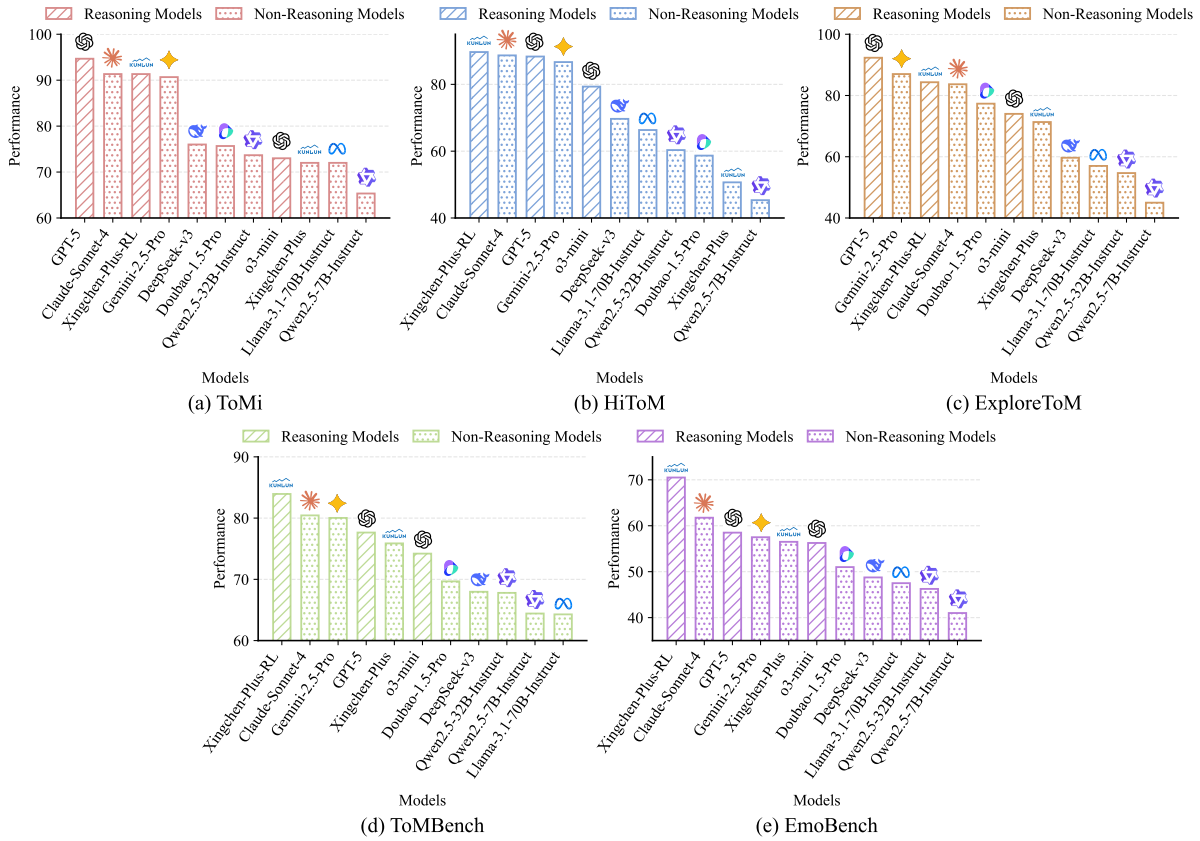


Figure 8: Extended performance comparison on five social reasoning benchmarks. We report the accuracy of reasoning LLMs and non-reasoning LLMs.

zled, puzzling, perplexed, complication, troubled, tricky, conflicts, ambiguity.

As shown in Figure 9 and 10, we evaluate three dimensions: word set selection (PV, BE, CW), detection mechanism (keyword matching vs. embedding-based). First, replacing PV/BE with our confusion words (+CW) under the same keyword-matching mechanism consistently yields higher accuracy, confirming that semantically targeted words are more effective triggers. Second, upgrading from keyword matching to embedding-based detection (+PV Emb, +BE Emb) improves performance even with suboptimal word sets, demonstrating the value of contextual similarity over exact string matching. Finally, our full method (+Ours), which combines the confusion word set with dual embedding-based detection, achieves the best accuracy and lowest token consumption across all benchmarks and both LLMs, validating that both the choice of intervention words and the detection mechanism contribute to the overall effectiveness.

C.4 Evaluation of Confidence and Entropy

To evaluate the efficacy of our method, we compared it against two standard uncertainty-based metrics: Confidence (Guo et al., 2017) and Entropy (Shannon, 2001). As illustrated in Figure 11, these approaches exhibit a suboptimal trade-off between efficiency and performance. Specifically, these methods fail to timely detect and avoid redundant reasoning loops. Furthermore, their accuracy gains are marginal and consistently lag behind our method (+Ours). This demonstrates that generic uncertainty signals lack the specificity required to resolve social cognitive dilemmas, whereas our world model effectively minimizes computational cost while maximizing reasoning robustness.

C.5 Evaluation of Multi-Source World Knowledge

To verify the impact of knowledge quality on social reasoning, we conducted a comparative study using different knowledge generators, ranging from individual advanced LLMs to an aggregated Multi-Source setting and a self-generating open-source model (Qwen3-32B). Since the ablation study in

Model	Threshold	ToMi		Hi-ToM		ExploreToM	
		Acc	Token	Acc	Token	Acc	Token
DeepSeek-R1 Distill Qwen -32B	Baseline	79.33	3504	69.33	7541	69.67	3185
	$k = 1$	$80.33_{\uparrow 1.00}$	$3296_{\downarrow 208}$	$71.33_{\uparrow 2.00}$	$6657_{\downarrow 884}$	$70.33_{\uparrow 0.66}$	$2942_{\downarrow 243}$
	$k = 2$	$82.67_{\uparrow 3.34}$	$2970_{\downarrow 534}$	$74.33_{\uparrow 5.00}$	$5804_{\downarrow 1737}$	$71.67_{\uparrow 2.00}$	$2796_{\downarrow 389}$
	$k = 3$	$84.33_{\uparrow 5.00}$	$2709_{\downarrow 795}$	$78.33_{\uparrow 9.00}$	$5183_{\downarrow 2358}$	$73.67_{\uparrow 4.00}$	$2515_{\downarrow 670}$
	$k = 4$	$84.33_{\uparrow 5.00}$	$2709_{\downarrow 795}$	$79.00_{\uparrow 9.67}$	$5063_{\downarrow 2478}$	$74.33_{\uparrow 4.66}$	$2338_{\downarrow 847}$
	$k = 5$	$84.33_{\uparrow 5.00}$	$2709_{\downarrow 795}$	$79.33_{\uparrow 10.00}$	$4961_{\downarrow 2580}$	$74.33_{\uparrow 4.66}$	$2338_{\downarrow 847}$
Qwen3-32B Reasoning	Baseline	81.33	2942	86.00	4316	74.33	3193
	$k = 1$	$82.33_{\uparrow 1.00}$	$2713_{\downarrow 229}$	$87.33_{\uparrow 1.33}$	$3927_{\downarrow 389}$	$75.33_{\uparrow 1.00}$	$2836_{\downarrow 357}$
	$k = 2$	$84.00_{\uparrow 2.67}$	$2550_{\downarrow 392}$	$88.67_{\uparrow 2.67}$	$3608_{\downarrow 708}$	$76.67_{\uparrow 2.34}$	$2548_{\downarrow 645}$
	$k = 3$	$85.33_{\uparrow 4.00}$	$2309_{\downarrow 633}$	$90.67_{\uparrow 4.67}$	$3252_{\downarrow 1064}$	$77.67_{\uparrow 3.34}$	$2384_{\downarrow 809}$
	$k = 4$	$85.33_{\uparrow 4.00}$	$2309_{\downarrow 633}$	$91.00_{\uparrow 5.00}$	$3082_{\downarrow 1234}$	$78.00_{\uparrow 3.67}$	$2261_{\downarrow 932}$
	$k = 5$	$85.33_{\uparrow 4.00}$	$2309_{\downarrow 633}$	$91.33_{\uparrow 5.33}$	$3015_{\downarrow 1301}$	$78.00_{\uparrow 3.67}$	$2261_{\downarrow 932}$

Table 5: Performance under different maximum intervention count k. The setting $k = 3$ represents the default configuration used in our main results.

Model	**Baseline**	**Gemini-3-pro**	**GPT-5.2**	**Claude-Sonnet-4.5**	**Multi-Source**	**Qwen3-32B**
DeepSeek-R1-Distill-Qwen-7B	54.19	$65.32_{\uparrow +11.13}$	$65.51_{\uparrow +11.32}$	$64.89_{\uparrow +10.70}$	$\mathbf{66.78}_{\uparrow +12.59}$	$59.62_{\uparrow +5.43}$
DeepSeek-R1-Distill-Qwen-14B	71.79	$76.27_{\uparrow +4.48}$	$76.11_{\uparrow +4.32}$	$75.63_{\uparrow +3.84}$	$\mathbf{78.14}_{\uparrow +6.35}$	$74.08_{\uparrow +2.29}$
DeepSeek-R1-Distill-Qwen-32B	77.18	$83.39_{\uparrow +6.21}$	$83.18_{\uparrow +6.00}$	$82.57_{\uparrow +5.39}$	$\mathbf{84.73}_{\uparrow +7.55}$	$79.84_{\uparrow +2.66}$
Qwen3-8B-Reasoning	69.41	$75.33_{\uparrow +5.92}$	$75.06_{\uparrow +5.65}$	$74.71_{\uparrow +5.30}$	$\mathbf{76.28}_{\uparrow +6.87}$	$72.12_{\uparrow +2.71}$
Qwen3-14B-Reasoning	74.34	$79.25_{\uparrow +4.91}$	$79.38_{\uparrow +5.04}$	$78.67_{\uparrow +4.33}$	$\mathbf{80.49}_{\uparrow +6.15}$	$77.21_{\uparrow +2.87}$
Qwen3-32B-Reasoning	77.54	$83.59_{\uparrow +6.05}$	$83.28_{\uparrow +5.74}$	$82.63_{\uparrow +5.09}$	$\mathbf{84.75}_{\uparrow +7.21}$	$80.73_{\uparrow +3.19}$

Table 6: Results on ToMBench. Multi-Source denotes the setting where World Knowledge is jointly aggregated from Gemini-3-pro, GPT-5.2, and Claude-Sonnet-4.5. For each row, the best-performing result is highlighted in **bold**.

Section 4.4 reveals that World Knowledge injection yields the most substantial gains on social norm tasks (ToMBench and EmoBench), we select these two benchmarks for this analysis. The results on ToMBench (Table 6) and EmoBench (Table 7) reveal two fundamental insights:

Comprehensiveness Drives Performance (Multi-Source as Upper Bound): The Multi-Source setting, which aggregates insights from Gemini-3-pro, GPT-5.2, and Claude-Sonnet-4.5, consistently achieves the highest accuracy (e.g., +11.32% on ToMBench). Trajectory analysis indicates that this aggregation provides a more comprehensive set of social norms compared to any single model. By covering a wider spectrum of behavioral motivations and potential outcomes, the Multi-Source knowledge effectively broadens the reasoning LLM's perspective, allowing it to navigate complex social scenarios that require multi-faceted consideration.

Capability Gap in Self-Generated Knowledge (Qwen3-32B): We also tested the feasibility of using an open-source LLM (Qwen3-32B) as a standalone knowledge generator. While it outperforms the baseline, it significantly lags behind the advanced LLMs' settings. A closer examination of the generated trajectories reveals that Qwen3-32B tends to interpret context at a literal level, failing to extract the subtle, implicit social rules necessary for emotional reasoning. This suggests that current open-source models still lack the inherent depth required to serve as robust, independent knowledge bases, highlighting the necessity of leveraging advanced sources to bridge the reasoning gap.

C.6 Evaluation on Advanced LLMs

WG-ToM is designed as a lightweight framework that empowers smaller-scale LLMs to achieve reasoning capabilities comparable to those of advanced LLMs through a cost-effective approach. The original intent of WG-ToM is to strike a balance between performance and cost, significantly reducing expensive token consumption while simultaneously enhancing reasoning accuracy.

As shown in Table 8, to further validate the efficacy of our method, we also tested the performance of WG-ToM on advanced LLMs. However, the dynamic state-update logic of the World Model

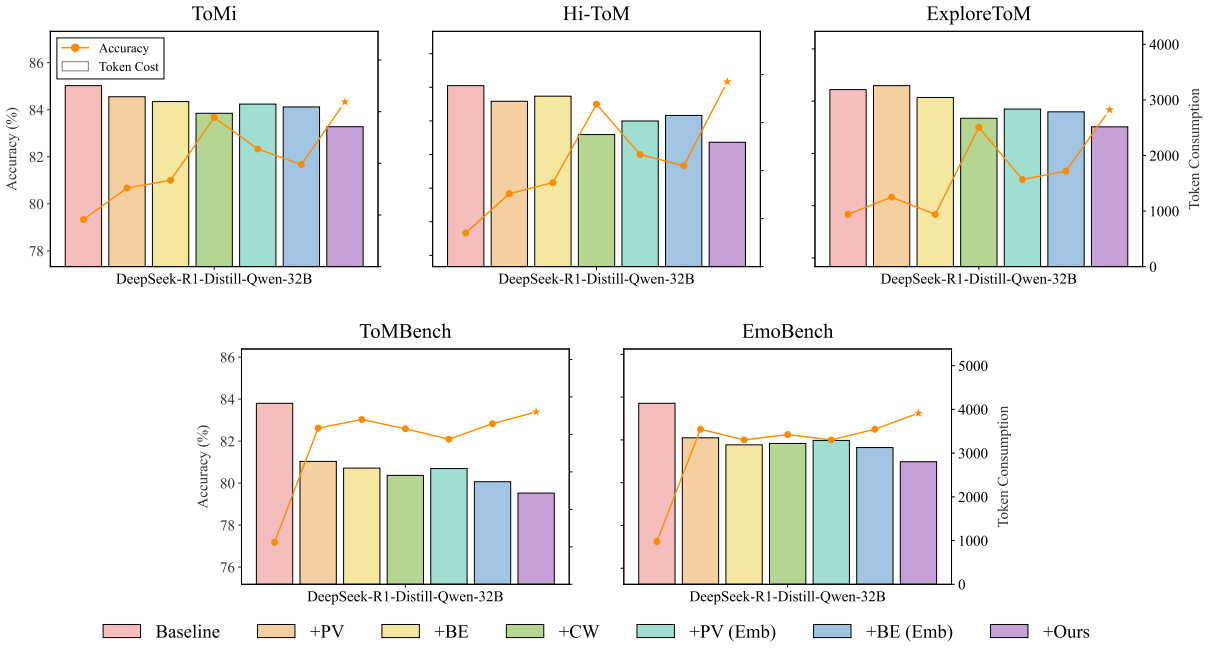


Figure 9: Performance with different intervention words on DeepSeek-R1-Distill-Qwen-32B. PV denotes pause-validation words, BE indicates branch-extension words, CW denotes confusion words (Table 1), and Emb indicates embedding-based detection.

Model	Baseline	Gemini-3-pro	GPT-5.2	Claude-Sonnet-4.5	Multi-Source	Qwen3-32B
DeepSeek-R1-Distill-Qwen-7B	41.50	49.50$_{\uparrow+8.00}$	50.00$_{\uparrow+8.50}$	49.25$_{\uparrow+7.75}$	**51.25**$_{\uparrow+9.75}$	45.75$_{\uparrow+4.25}$
DeepSeek-R1-Distill-Qwen-14B	54.25	60.75$_{\uparrow+6.50}$	60.25$_{\uparrow+6.00}$	59.75$_{\uparrow+5.50}$	**62.00**$_{\uparrow+7.75}$	57.50$_{\uparrow+3.25}$
DeepSeek-R1-Distill-Qwen-32B	61.25	67.25$_{\uparrow+6.00}$	67.00$_{\uparrow+5.75}$	66.25$_{\uparrow+5.00}$	**68.00**$_{\uparrow+6.75}$	64.75$_{\uparrow+3.50}$
Qwen3-8B-Reasoning	55.25	60.25$_{\uparrow+5.00}$	60.50$_{\uparrow+5.25}$	59.50$_{\uparrow+4.25}$	**61.50**$_{\uparrow+6.25}$	57.75$_{\uparrow+2.50}$
Qwen3-14B-Reasoning	60.50	64.75$_{\uparrow+4.25}$	64.50$_{\uparrow+4.00}$	64.00$_{\uparrow+3.50}$	**65.50**$_{\uparrow+5.00}$	63.25$_{\uparrow+2.75}$
Qwen3-32B-Reasoning	63.50	68.50$_{\uparrow+5.00}$	68.25$_{\uparrow+4.75}$	67.50$_{\uparrow+4.00}$	**69.50**$_{\uparrow+6.00}$	66.25$_{\uparrow+2.75}$

Table 7: Results on EmoBench. Multi-Source denotes the setting where World Knowledge is jointly aggregated from Gemini-3-pro, GPT-5.2, and Claude-Sonnet-4.5. For each row, the best-performing result is highlighted in **bold**.

requires granular control over the LLM's internal reasoning process and is currently optimized for open-source architectures. Thus we report only the improvement results for advanced LLMs after incorporating World Knowledge. The experimental data indicate that WG-ToM consistently provides gains even for advanced LLMs. This further demonstrates the generalizability of WG-ToM beyond small-scale LLMs.

D Qualitative Examples

D.1 Sample Example of World Model generation

To better illustrate the effectiveness of adaptive world model-enhanced reasoning mechanism, we present a comprehensive set of examples demonstrating textual world model generation, as illustrated in the Figure 12. The world model maintains dynamic tracking of character and entity states, which are continuously updated in response to different actions and events. This dynamic updating capability enables our method to flexibly adapt to diverse story scenarios, ensuring that during the intervention process, the system can provide accurate information corresponding to the current state of the narrative.

D.2 Sample Example of World Knowledge generation

Figures 13 and 14 present two qualitative case studies illustrating how World Knowledge is generated from the given contexts.

These cases are included to demonstrate the gen-

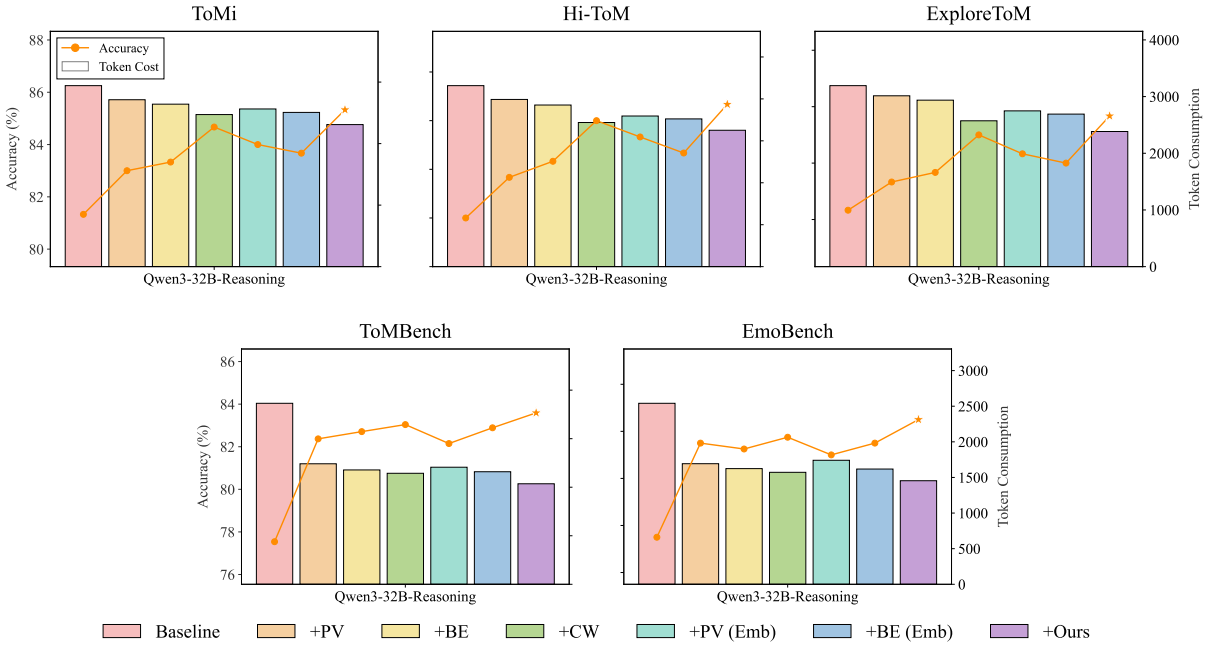


Figure 10: Performance with different intervention words on Qwen3-32B-Reasoning. PV denotes pause-validation words, BE indicates branch-extension words, CW denotes confusion words (Table 1), and Emb indicates embedding-based detection.

Model	ToMBench	EmoBench
GPT-5.5	83.78	70.50
+ Ours	**88.37**	**74.50**
Claude-Opus-4.7	82.58	69.75
+ Ours	**86.45**	**74.25**
Gemini-3.1-Pro	81.39	71.50
+ Ours	**87.71**	**75.25**
DS-R1-DQ-32B	77.18	61.25
+ Ours	83.39	67.25
Qwen3-32B	77.54	63.50
+ Ours	83.59	68.50

Table 8: Evaluation on advanced LLMs (ToMBench and EmoBench).

eration process of world knowledge across different social reasoning scenarios and benchmarks.

D.3 Prompts Design

Figure 15 in our framework are designed exclusively to perform their designated functions: Task-Adaptive Router for binary task classification; World Model for structured state tracking; World Knowledge for social norm elicitation, rather than providing any reasoning guidance. We employed highly structured and constrained instruction formats to ensure the consistency of output across diverse task scenarios, thereby minimizing sensitivity to subtle variations in prompt phrasing, ensuring the role and robustness of prompts.

D.4 Case Study

Figure 16 to 21 present a collection of qualitative case studies illustrating the reasoning behaviors of DeepSeek-R1-Distill-Qwen-32B and Qwen3-32B across multiple social reasoning benchmarks, including ToMi, HiToM, ExploreToM, ToMBench, and EmoBench.

These examples cover a diverse range of reasoning scenarios and task types from each benchmark and are included to complement the quantitative results reported in the main paper.

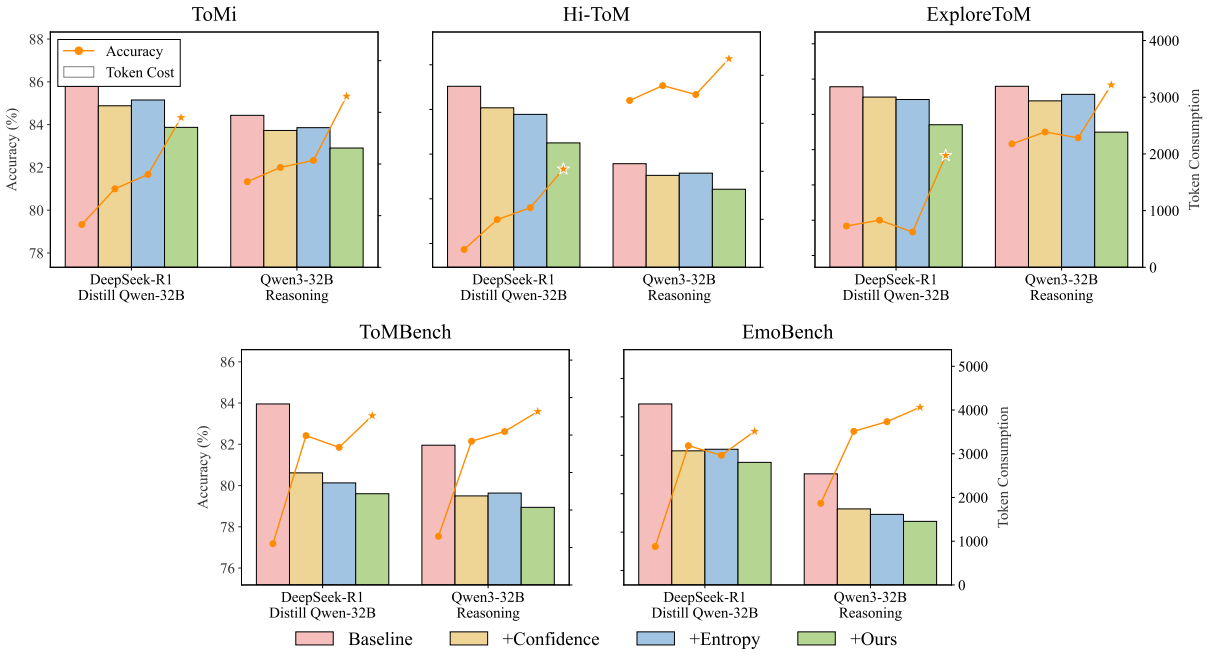


Figure 11: Efficiency and accuracy comparison with Confidence and Entropy-based metrics.

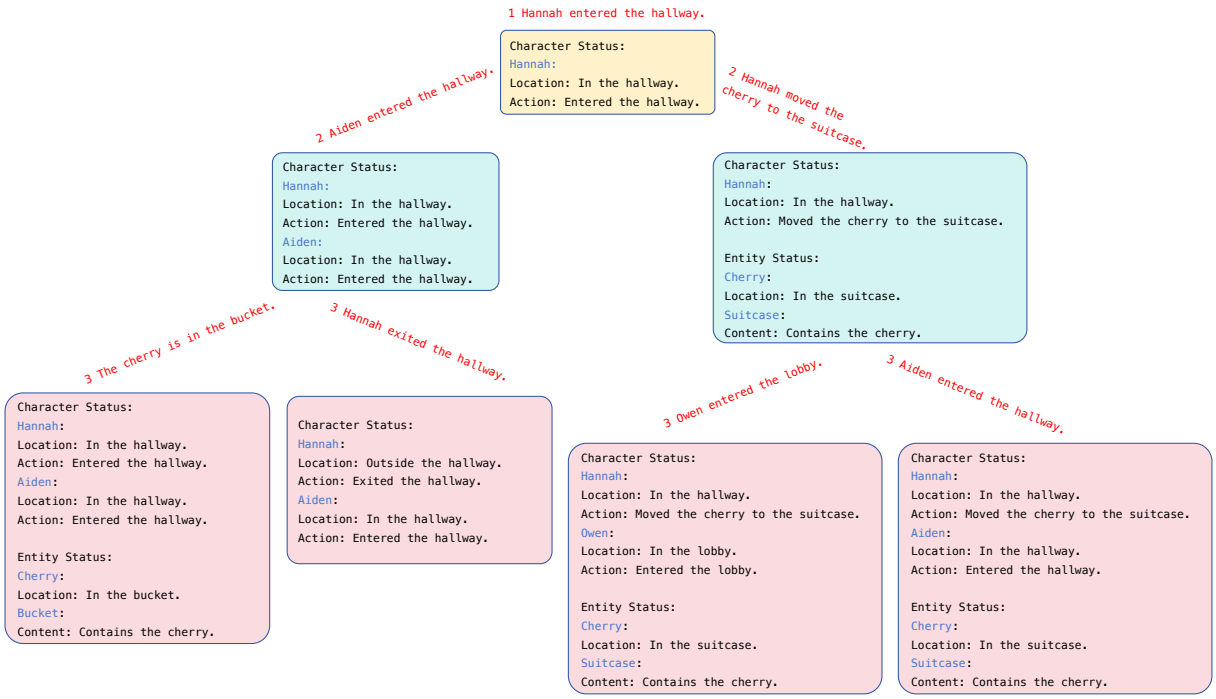


Figure 12: Sample example of World Model Generation

ToMBench

In the school's sports field, Xiao Gang is the sprinter of the school track team, and he always rushes to the finish line quickly in every race. After a race, classmate Xiao Ming says, "Xiao Gang, you definitely have the superpower of teleportation."

QUESTION: Why does Xiao Gang say this?

OPTION-A: He thinks Xiao Gang has the superpower of teleportation.
OPTION-B: He uses a metaphor to express his admiration for Xiao Gang's sprinting speed, congratulating him on winning first place again.
OPTION-C: He thinks Xiao Gang is an outstanding long-distance runner.
OPTION-D: He uses a metaphor to describe how fast Xiao Gang runs in the race.

World Knowledge

Social Norm:
In informal peer interactions, when facing exceptional abilities or extreme performance demonstrated by others, individuals tend to use extreme hyperbole or supernatural metaphors to express strong admiration.

Cognitive pattern:
- Hyperbolic Attribution: Mapping real-world high-level traits to unrealistic abilities to highlight the shocking nature of the trait through cognitive contrast.

- Humorous Compliment: Conveying positive evaluation through illogical jokes to express admiration while lightening the social atmosphere and bridging psychological distance.

- Rationalization of Expectation Violation: Using supernatural concepts to provide a rhetorical explanation for phenomena that exceed conventional expectations.

Figure 13: Sample example of World Knowledge Generation on ToMBench

Emobench

On his way to work, Andre was stuck in traffic when a person started shouting at him for being a bad driver. Once he came out of the car to confront that person, he realized it was his dad.

Question 1: What emotion(s) would Andre ultimately feel in this situation?

Choices for Question 1: (A) Disapproval (B) Amusement (C) Annoyance (D) Nervousness (E) Fear (F) Unbothered

Question 2: Why would Andre feel these emotions in this situation?

Choices for Question 2:
(A) He loves traffic
(B) He enjoys confrontations
(C) The person shouting at him turned out to be his dad
(D) He enjoys being called a bad driver.

World Knowledge

Social norm:
Confrontational behavior from an apparent stranger typically triggers defensiveness or annoyance; however, unexpectedly identifying the antagonist as a close family member radically shifts the interpretation of the event.

Reasoning pattern:
- Hostility De-escalation: Recognizing a close tie neutralizes the perception of genuine threat or malice associated with anonymous conflict.

- Incongruity-Resolution: The sudden mismatch between the hostile context (road rage) and the intimate relationship creates a humorous absurdity (benign violation).

- Context Reframing: The interaction shifts from a public dispute to a personal coincidence, transforming the emotional response from aggression to amusement or surprise.

Figure 14: Sample example of World Knowledge Generation on EmoBench

Prompt for Task-Adaptive Router

Story: [Insert Story Here]
Question: [Insert Question Here]
Instructions: Determine whether answering this question requires understanding implicit social norms, cultural conventions, or interpersonal behavioral expectations beyond the literal content of the story.
Output `1` if the task involves social judgment such as interpreting motivations, emotions, politeness strategies, or culturally conditioned behaviors.
Output `0` if the task only requires tracking factual state changes or physical object locations.
Output only a single number: `1` or `0`.

Prompt for World Model

Story: [Insert Story Here]
Question: [Insert Question Here]
Instructions: Analyze the provided story and update the structured state-tracking information for each step. Output the result strictly in the following format:
State [Number]
Character States:
- [Character Name]:
- Location: [Current Room]
- Action: [Specific action performed]
Entity States:
- [Container Name]:
- Content: [Current contents]
- [Object Name]:
- Location: Located inside the [Container Name]

Prompt for World Knowledge

Story: [Insert Story Here]
Question: [Insert Question Here]
Instructions: Analyze the provided story and question to elicit the tacit social knowledge and cognitive patterns involved. Output the result strictly in the following format:
Social Norm: [Provide the applicable social rule or expectation here]
Cognitive pattern:
- Emotion Priority: [Describe the prioritization of feelings or relationships]
- Motivation Protection: [Describe the protection of enthusiasm or effort]
- Social Tact: [Describe the mechanism used for support]

Figure 15: Prompt designs for the three components of WG-ToM. **Top:** Task-Adaptive Router determines whether World Knowledge injection is needed. **Bottom left:** World Model prompt for structured state tracking. **Bottom right:** World Knowledge prompt for social norm elicitation.

Case Study of DeepSeek-R1-Distill-Qwen-32B on ToMi

Story:
1 Hannah entered the hallway.
2 Aiden entered the hallway.
3 The cherry is in the bucket.
4 The bucket is in the hallway.
5 Aiden exited the hallway.
6 Aiden entered the hallway.
7 Hannah moved the cherry to the suitcase.
8 The suitcase is in the hallway.

Question:
Where does Hannah think that Aiden searches for the cherry?

Response from Init:
The story begins with Hannah and Aiden entering the hallway, where the cherry is located inside a bucket. At this moment, the environment appears stable, and the containment relation between the cherry and the bucket is clear.
Aiden then exits the hallway and later returns. This interruption introduces uncertainty about whether his belief should be treated as continuous or reset upon re-entry. I pause and wait to determine whether this exit meaningfully affects what Hannah assumes about Aiden's knowledge.
After Aiden's return, Hannah moves the cherry from the bucket to the suitcase, which is also in the hallway. However, it is not explicitly stated whether Aiden observes this movement. Alternatively, Hannah might assume that mere co-presence implies awareness, but the narrative does not confirm this assumption.
.....
At this point, reasoning about Hannah's belief regarding Aiden's future search location requires balancing temporal order, observation, and belief attribution. I wait again and consider alternative belief updates; however, the inference path remains unresolved and increasingly feels ambiguous

World Model:
`<information>`
Character Status:
Hannah: Location: Hallway; Action: Moved the cherry to the suitcase.
Aiden: Location: Hallway; Action: Re-entered the hallway after exiting.
Entity Status:
Cherry: Current Location: Suitcase.
Bucket: Location: Hallway; Contains: None (previously contained the cherry).
Suitcase: Location: Hallway; Contains: Cherry.
`</information>`

Response after Intervention:
With this information available, the reasoning process becomes more coherent. The critical point is to reason from Hannah's perspective about Aiden's belief rather than reconstructing Aiden's actual observation history.
.....
Hannah moves the cherry to the suitcase after Aiden has re-entered the hallway. From her point of view, Aiden is present during the relocation, and the containment relation between the cherry and the suitcase is therefore accessible to him.
Given this understanding, Hannah would naturally expect Aiden to search for the cherry in the suitcase rather than in its previous container.

Final Answer:
suitcase

Figure 16: Case Study of DeepSeek-R1-Distill-Qwen-32B on ToMi

Case Study of Qwen3-32B on HiToM

Story:
1 Isabella, Carter, Elizabeth, Hannah and Jacob entered the bedroom.
2 The tomato is in the green_bottle.
3 Isabella moved the tomato to the green_box.
4 Isabella exited the bedroom.
5 Carter made no movements and stayed in the bedroom for 1 minute.
6 Carter exited the bedroom.
7 Elizabeth moved the tomato to the blue_container.
8 Elizabeth exited the bedroom.
9 Hannah made no movements and stayed in the bedroom for 1 minute.
10 Jacob lost his watch.
11 Hannah exited the bedroom.
12 Jacob made no movements and stayed in the bedroom for 1 minute.
13 Jacob exited the bedroom.
14 Isabella, Carter, Elizabeth, Hannah and Jacob entered the waiting_room.
Question:
Where is the tomato really?
Options: [A. green_bottle, B. red_basket, C. green_bucket, D. green_box, E. blue_container, F. red_drawer, G. blue_bucket, H. green_envelope, I. blue_bathtub, J. blue_crate, K. red_envelope, L. red_box, M. green_crate, N. green_basket, O. red_bucket]
Response from Init:
The tomato starts in the green_bottle, and then Isabella moves it to the green_box. At this point, the tomato's location has changed once, while several characters remain in the bedroom.
Isabella exits, Carter stays briefly, and later Carter also exits. After that, Elizabeth moves the tomato again, this time to the blue_container, and then she exits as well. The remaining actions (Hannah staying, Jacob losing a watch, and subsequent exits) appear unrelated to the tomato's physical location.
.....
So the true location should follow the most recent move. I wait and re-check the sequence: the last explicit relocation is performed by Elizabeth to the blue_container; however, the presence and exits of multiple characters makes it easy to accidentally anchor to an earlier container. Alternatively, one might over-weight the initial placement in the green_bottle, but the later moves overwrite it. I wait once more and confirm the final explicit container assignment is confusing
World Model:
<information>
Character Status:
Isabella: Location: Bedroom Exited; Later entered the waiting_room.
Carter: Location: Bedroom Exited; Later entered the waiting_room.
Elizabeth: Location: Bedroom; Action: Moved the tomato to the blue_container; Then exited; Later entered the waiting_room.
Hannah: Location: Bedroom Exited; Later entered the waiting_room.
Entity Status:
Tomato: Current Location: Blue_container.
Green_bottle: Location: Bedroom; Contains: None (previously contained the tomato).
Green_box: Location: Bedroom; Contains: None (previously contained the tomato).
Blue_container: Location: Bedroom; Contains: Tomato.
</information>

Figure 17: Case Study of Qwen3-32B on HiToM

Case Study of Qwen3-32B on HiToM (Continue)

Response after Intervention:
Given the explicit state summary, the answer follows directly from the latest relocation event. The tomato is moved from the green_bottle to the green_box by Isabella, and then later moved again from the green_box to the blue_container by Elizabeth.
.....
No subsequent action in the story indicates another movement of the tomato. The later events (waiting, losing a watch, and characters exiting) do not modify the tomato's location.
Therefore, the tomato is really in the blue_container.
Final Answer:
blue_container

Figure 18: Case Study of Qwen3-32B on HiToM (Continue)

Case Study of DeepSeek-R1-Distill-Qwen-32B on ExploreToM

Story:
1 Kaylee entered the hotel lobby.
2 Kaylee moved the silver letter opener to the wooden desk drawer, which is also located in the hotel lobby.
3 While this action was happening, Liam witnessed this action in secret (and only this action).
4 Kaylee left the hotel lobby.
5 Liam entered the hotel lobby.
6 Kaylee entered the hotel lobby.
7 Liam moved the silver letter opener to the leather briefcase, which is also located in the hotel lobby.
Question:
In which container will Liam search for the silver letter opener?
Response from Init:
Kaylee first places the silver letter opener into the wooden desk drawer in the hotel lobby. During this action, Liam witnesses the placement in secret, and the story explicitly limits his observation to only this placement event.
Kaylee then leaves, and Liam later enters the hotel lobby. At this point, Liam has prior knowledge that the letter opener is in the wooden desk drawer, and there is no indication that he has observed any subsequent movement by Kaylee.
Later, Kaylee re-enters the hotel lobby, but the crucial event is that Liam himself moves the silver letter opener to the leather briefcase.
.....
I wait and verify that no other relocation occurs after Liam moves it. However, the earlier secret witnessing clause can pull attention toward the wooden desk drawer as the "known" location. Alternatively, one might over-focus on what Kaylee knows, but the question asks where Liam will search, and his own action overrides earlier placements. I wait again and confirm Liam's search should align with his latest move, which feels complicating
World Model:
`<information>`
Character Status:
Kaylee: Location: Hotel lobby.
Liam: Witnessed (in secret) the placement into the wooden desk drawer; Later entered the hotel lobby; Action: Moved the silver letter opener to the leather briefcase.
Entity Status:
Silver letter opener: Current Location: Leather briefcase.
Wooden desk drawer: Location: Hotel lobby; Contains: None (previously contained the letter opener).
Leather briefcase: Location: Hotel lobby; Contains: Silver letter opener.
`</information>`
Response after Intervention:
The decisive factor is the most recent physical relocation of the silver letter opener. Liam secretly observes the initial placement into the wooden desk drawer, so he knows it is there at that time.
.....
Later, Liam personally moves the letter opener into the leather briefcase. Because he performs this action, his belief about the object's location should update to the briefcase, and there is no subsequent move that would invalidate this update.
Therefore, when Liam searches for the silver letter opener, he will search in the leather briefcase.
Final Answer:
leather briefcase

Figure 19: Case Study of DeepSeek-R1-Distill-Qwen-32B on ExploreToM

Case Study of Qwen3-32B on ToMBench

Story: Dehua and Sanming are company employees, they are competing for a promotion opportunity. Today is Sanming's birthday, he enjoys a small birthday cake alone. Lingling is the company supervisor, she is having a private conversation with Dehua in the pantry. Sanming sees from a distance that Lingling quietly gives Dehua a document, and with a smile, gently pats Dehua's back, then returns to her office. Dehua quickly checks the document, shows a smile, and carefully puts the document into his file bag.
Question: Why does Lingling give Dehua that document?
Option A: Lingling returns Dehua's personal document left in the conference room.
Option B: Lingling gives Dehua a daily notice unrelated to promotion.
Option C: Lingling and Dehua are planning a surprise for Sanming.
Option D: Lingling gives Dehua extra information.
Map: ["A": "B", "B": "C", "C": "D", "D": "A"]
World Knowledge: Social Norm: In organizational hierarchies involving competition for scarce resources, authority figures may transfer critical competitive resources to specific subordinates through informal, covert channels, thereby disrupting fair competition. Cognitive pattern:
- Establishment of Patronage: Power holders support specific individuals by privately granting privileged resources to form an interest alliance.
- Signaling of Power & Affection: Using intimate non-verbal behaviors (patting back, smiling) to reinforce the trust contract.
- Reciprocal Collusion: The receiver tacitly cooperates in secrecy (hiding the file), forming an information barrier against third parties (competitors).
Response after Intervention: Alright, so I'm trying to figure out why Lingling gave Dehua that document. Let me break this down step by step based on the story and the norms provided.
First, the story sets up a competitive situation between Dehua and Sanming, who are both vying for a promotion.
.....
Now, the knowledge talks about social norms in organizational hierarchies, especially when there's competition for resources like promotions. It mentions that authority figures might provide critical resources informally to specific subordinates to disrupt fair competition. The cognitive patterns listed are: establishing patronage, signaling power and affection, and reciprocal collusion.
Looking at the options:
A. Lingling gives Dehua a daily notice unrelated to promotion. If it's unrelated, why the secrecy? The story mentions it's a document, and Lingling gives it privately, which suggests it's important. A daily notice seems too routine for such behavior.
B. Lingling and Dehua are planning a surprise for Sanming. The story doesn't mention anything about a surprise. Also, Lingling is the supervisor; planning a surprise for a subordinate on his birthday might not fit the context of the competitive promotion scenario.
C. Lingling gives Dehua extra information. This aligns with the norms about transferring critical resources. If the document contains extra info, it could be strategic, giving Dehua an edge in the competition. The secretive manner (private conversation, putting it away) supports this.
D. Lingling returns Dehua's personal document left in the conference room. If it's just returning something, why the smile and the pat? That seems like a more casual interaction, not requiring secrecy.
Considering the knowledge's points: Lingling as an authority figure giving a document privately suggests she's providing critical info (extra information) to Dehua to help him in the competition. The non-verbal cues like the smile and pat indicate a supportive relationship.
Final Answer: [[C]] Lingling gives Dehua extra information.

Figure 20: Case Study of Qwen3-32B on ToMBench

Case Study of DeepSeek-R1-Distill-Qwen-32B on Emobench

Scenario: After weeks of hard work and determination, Andy was finally able to open his art exhibition and display his portraits. While he invited all of the people at work, none of them showed up. As he was packing up to leave, his boss showed up, pointed to his painting, and praised his skills.
Question 1: What emotion(s) would Andy ultimately feel in this situation?
Options 1: [(A) Gratitude & Anger, (B) Embarrassment & Joy, (C) Sadness & Anger, (D) Gratitude & Joy, (E) Embarrassment & Sadness, (F) Embarrassment & Anger]
Question 2: Why would Andy feel these emotions in this situation?
Options 2: [(A) Nobody from work showed up, (B) His boss showed up and praised his work, (C) He didn't like his own drawings, (D) His coworkers disliked his artwork]
World Knowledge: Social norm: When an individual faces group non-attendance after high effort, validation from a high-status figure often triggers a significant emotional rebound. Reasoning pattern:
- Authority weighting: Recognition from a superior or expert often outweighs peer absence in validating competence.
- Contrast & Compensation: Validation received against a backdrop of disappointment acts as a rescue, amplifying feelings of gratitude and joy.
- Competence repair: Specific praise effectively counteracts the negative impact of social exclusion."
Response after Intervention: Okay, so I'm trying to figure out what emotions Andy would feel in this scenario and why. Let me break it down step by step.
First, the scenario: Andy worked really hard and opened his art exhibition. He invited people from work, but none of them showed up. That must be pretty disappointing, right? I mean, he put in a lot of effort, and it's natural to feel sad or maybe even angry that his coworkers didn't support him. But then, as he's packing up, his boss shows up and praises his work. So, on one hand, Andy is feeling let down by his coworkers, which probably makes him feel sad or maybe even angry.
.....
Looking at the emotions, the initial reaction would be sadness because his peers didn't come. But then, the boss's positive feedback could counterbalance that. So maybe he feels a mix of sadness and some positive emotion. Or perhaps the boss's praise makes him feel better, so he's more grateful and joyful.
In terms of the options, option D is Gratitude & Joy. That seems to fit if the boss's praise is enough to make him feel good despite the coworkers' absence. Option E is Embarrassment & Sadness, but I'm not sure if embarrassment is the right emotion here. He might feel sad about the lack of support, but maybe not embarrassed. Embarrassment usually comes from feeling awkward or self-conscious, which isn't clearly indicated here.
Option A is Gratitude & Anger. He might feel grateful to the boss but maybe also angry at the coworkers for not showing up. But the question asks for the ultimate emotions, so I'm not sure if anger would be the dominant one after the boss's praise.
Option B is Embarrassment & Joy. I'm not sure about the embarrassment part. Option C is Sadness & Anger, which could be possible, but again, the boss's praise might mitigate some of that.
The social norm mentioned says that when someone faces group non-attendance but gets validation from a high-status figure, there's a significant emotional rebound. So, the boss's praise would likely make Andy feel more grateful and joyful, even though he was initially sad.
So, for Question 1, the emotions would probably be Gratitude & Joy, which is option D.
For Question 2, the reason is because his boss showed up and praised his work, which is option B.
Final Answer: Question 1: D. Gratitude & Joy Question 2: B. His boss showed up and praised his work

Figure 21: Case Study of DeepSeek-R1-Distill-Qwen-32B on Emobench