MINDWORLD: ITERATIVE CO-OPTIMIZATION OF MENTAL WORLD MODELS AND POLICIES FOR SOCIAL INTELLIGENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Language agents increasingly operate in interactive environments, yet their understanding of users and counterparts is often not maintained as an explicit, evolving component of the decision process, limiting their ability to adapt as interactions evolve. We introduce **MindWorld**, a framework that explicitly couples a Mental World Model with a Policy Model. The Mental World Model dynamically tracks beliefs, desires, and intentions from the evolving interaction history, while the Policy Model conditions on these mental states together with the task context to generate actions. To improve long-horizon credit assignment, MindWorld combines trajectory-level outcome feedback with dedicated step-level rewards and converts them into step-wise relative advantages for reinforcement learning. We further introduce an iterative co-optimization procedure. The two models are trained in alternating phases using a trajectory-level outcome reward and step-level rewards from fixed scorers. Experiments on SOTOPIA, SOTOPIA-Hard, ALFWorld, and WebShop demonstrate consistent gains across social, household, and web-based interaction. With Qwen3-8B under self-play, MindWorld obtains GOAL/OVERALL scores of 9.27/4.03 on SOTOPIA and 8.16/3.94 on SOTOPIA-Hard. On ALFWorld and WebShop, the reported results exceed those of the evaluated general RL baselines, and integration with task-specific methods yields additional gains. These results show that explicitly modeling evolving mental states, together with fine-grained credit assignment and iterative policy optimization, provides an effective and general approach for improving interactive language agents.

1 INTRODUCTION

With the rapid development of large language models (LLMs), language agents have shown increasing competence in reasoning, planning, and goal-directed interaction. Recent benchmarks such as SOTOPIA (Zhou et al., 2024a) and AgentSense (Mou et al., 2025) have further shifted the study of social intelligence from static question answering to interactive settings, where agents must pursue private goals, interpret evolving context, and adapt their behavior over multiple turns. Existing work has improved such agents through interactive learning (Wang et al., 2024), explicit strategy planning (Deng et al., 2024; Liu et al., 2025a), and reinforcement learning (Yu et al., 2025). Yet effective social interaction depends on information that is not directly observable from dialogue alone. An agent must continually infer what it believes about the current situation, what it wants to achieve, and what it intends to do next; when another participant is involved, analogous mental states of the counterpart also shape subsequent decisions. Although recent methods improve strategic behavior, these mental states are typically left implicit in the policy or represented only indirectly through planning traces. This motivates the explicit modeling of an agent's evolving mental world as a separate component of decision making.

Recent work has started to make mental states more explicit in language agents. Theory-of-Mind methods (Jafari et al., 2025) infer beliefs, desires, and intentions from interaction history, while social-agent methods improve behavior through explicit strategy planning or adaptive reasoning (Wang et al., 2026a). These directions are closely related, but they are usually studied separately. Mental-state modeling is often used only as an auxiliary reasoning component, while policy optimization focuses on action quality and task reward. As a result, the mental state itself is rarely

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

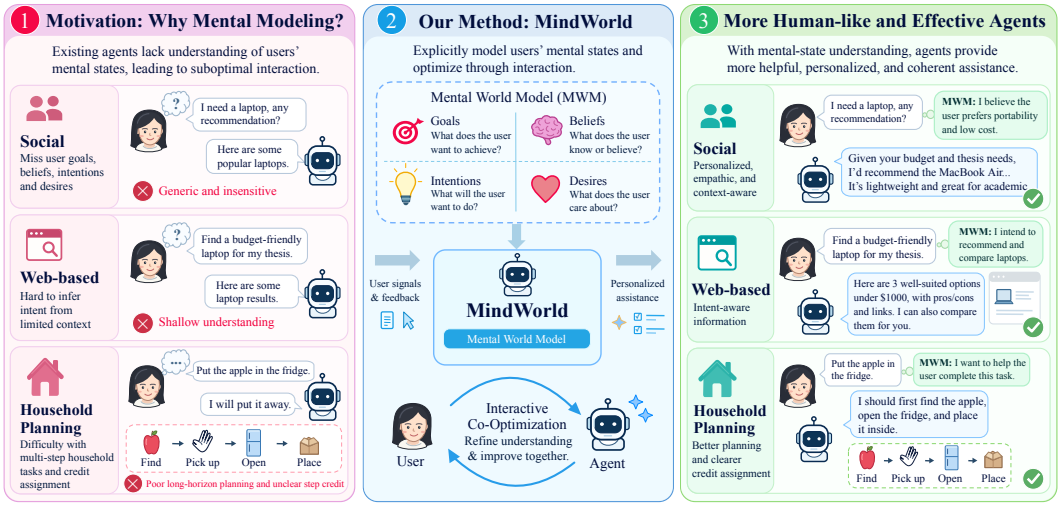


Figure 1: Overview of **MindWorld**: explicit mental-state modeling and iterative co-optimization enable more adaptive and effective interactive agents.

treated as a learnable model that evolves together with the policy. This leaves an important gap: the agent may reason about what to do, but it does not explicitly maintain and refine a model of why it should do so. Moreover, without interaction-level optimization, errors in mental-state estimation cannot be corrected according to their downstream influence on decisions.

To bridge this gap, we propose **MindWorld**, which pairs two independently parameterized models for mental-state estimation and action generation. The Mental World Model (Fei & Zhao, 2026) infers beliefs, desires, and intentions (BDI) (Rao & Georgeff, 1995) from the task context and interaction history: what the agent believes about the situation, what it seeks to achieve, and what it intends to do. In social interactions, it also estimates the partner's BDI. The Policy Model uses these mental states alongside the original context and history to generate the next action. New observations update the history, allowing the Mental World Model to revise its estimates before the next decision. Figure 1 illustrates this interaction loop.

Training alternates between optimizing the Mental World Model with the Policy Model fixed and optimizing the Policy Model with the updated Mental World Model fixed. Trajectory-level rewards measure overall task achievement but provide limited guidance on individual outputs, motivating step-level feedback (Liu et al., 2025b). Two offline-trained, frozen scorers provide step-level feedback: one measures the usefulness of the predicted mental state, and the other measures whether the action contributes to task completion. Format feedback additionally checks the required output structure. We combine outcome advantages computed across rollouts of the same task with step and format advantages computed within each rollout, providing both models with global supervision and local relative credit.

We evaluate MindWorld on four interactive benchmarks, including SOTOPIA and SOTOPIA-Hard (Zhou et al., 2024a) for social intelligence, and ALFWorld (Shridhar et al., 2021) and WebShop (Yao et al., 2022) for task-oriented decision making. Under the Self-Play setting with Qwen3-8B, MindWorld achieves 9.27/4.03 on SOTOPIA and 8.16/3.94 on SOTOPIA-Hard in terms of GOAL/OVERALL, improving over the corresponding backbone by 1.23/0.42 and 1.45/0.75 points, respectively. Beyond social interaction, MindWorld consistently improves general RL baselines on ALFWorld and WebShop, and further enhances specialized agent methods when combined with their existing training frameworks.

Our contributions are threefold. **(1)** We introduce MindWorld, a framework that explicitly models agents' evolving mental states through a dedicated Mental World Model and integrates them into policy decision making. **(2)** We develop an iterative training framework that combines trajectory-level outcome rewards with step-level feedback and format rewards, enabling more effective credit assignment and jointly improving mental-state modeling and action generation during interaction. **(3)** We conduct extensive experiments on social, embodied, and web-based interaction benchmarks.

MindWorld consistently improves strong baselines across SOTOPIA, ALFWorld, and WebShop, demonstrating the effectiveness of explicit mental-state modeling for interactive language agents.

2 Related Work

Social Intelligence and Mental-State Modeling. Recent work has moved social intelligence toward explicit modeling of cognition, strategy, and interaction dynamics. ToMBench (Chen et al., 2024b) and COKE (Wu et al., 2024) study Theory-of-Mind abilities and structured mental-state reasoning, while SOTOPIA-Ω (Zhang et al., 2025), SDPO (Kong et al., 2025), ARIA (Yang et al., 2025b), and SAVOIR (Feng et al., 2026) improve social agents through strategy injection, preference optimization, intention-aware rewards, and utterance-level attribution. These methods expose increasingly rich social signals, but rarely maintain mental states as a dedicated model that evolves together with action generation.

World Models for Language Agents. Language agent world models represent aspects of an environment that support planning, including web transitions, causal dependencies, and embodied interactions. Recent methods model web transitions (Chae et al., 2025), causal dynamics (Gkountouras et al., 2025), and embodied environments (Yoo et al., 2025), while WebEvolver (Fang et al., 2025), CoEx (Kim & Hwang, 2025), and SWAP (Xiong et al., 2025) further couple world modeling with agent learning or reasoning. These approaches mainly model external environment dynamics, whereas MindWorld models latent cognitive states that influence decision making.

Reinforcement Learning and Credit Assignment. Long-horizon agent learning has motivated reinforcement-learning methods that go beyond a single terminal reward. ArCHer (Zhou et al., 2024b) and REFUEL (Gao et al., 2025) address multi-turn optimization, while SVPO (Chen et al., 2024a), Tree-PLV (He et al., 2024), Rewarding Progress (Setlur et al., 2025), and Free Process Rewards (Yuan et al., 2025) provide finer supervision through step-level values or process rewards. MindWorld follows this direction by combining global outcome supervision with local step rewards, and applies both signals to the Mental World Model and the Policy Model.

3 Method

MindWorld couples a Mental World Model with a Policy Model. The former maintains an explicit mental state that evolves with interaction, while the latter uses this state to guide action generation. The two models are optimized alternately with trajectory-level and step-level reinforcement signals. We next formalize the interaction process and present the main components of MindWorld.

3.1 Task Formulation

We consider an interactive decision-making task in which an agent acts over multiple steps to achieve a task-specific goal. Each task is specified by static context x, such as the scenario, profile, and goal, together with an interaction history H_t containing the observations and actions available before the decision at step t. We denote the observable interaction state as $S_t = (x, H_t)$, with H_1 containing the initial observation. MindWorld introduces an explicit mental state between the interaction context and action generation.

At each step, the Mental World Model infers a mental state M_t, and the Policy Model generates an action P_t conditioned on both the original context and this state:

$$M_t \sim p_\phi(\cdot \mid x, H_t), \qquad P_t \sim \pi_\theta(\cdot \mid x, H_t, M_t), \tag{1}$$

where p_ϕ and π_θ are the conditional generation distributions of the Mental World Model and Policy Model, parameterized by ϕ and θ, respectively. After executing P_t, the environment or interaction partner returns an observation O_{t+1}. The history is updated as

$$H_{t+1} = H_t \oplus (P_t, O_{t+1}), \tag{2}$$

where $\oplus$ denotes appending the latest interaction. Conditioned on x and H_1, this process produces a trajectory $\tau = (M_1, P_1, O_2, \ldots, M_T, P_T, O_{T+1})$, where T is the number of interaction steps. The mental state is reconstructed from the updated history at each step rather than fixed at the beginning of the episode.

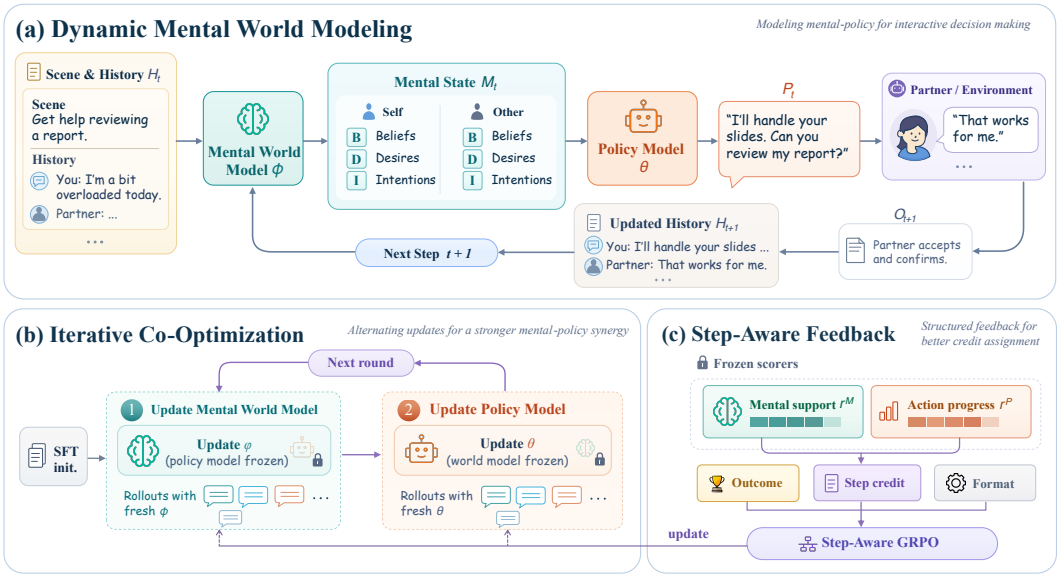


Figure 2: Framework of **MindWorld**. (a) The Mental World Model receives scene and history; the Policy Model receives the same context and history together with the BDI conditioning text. (b) The two generation models are updated in alternating phases. (c) Benchmark outcomes, fixed step scorers, and deterministic format checks supply distinct reward components.

3.2 Dynamic Mental World Modeling

The Mental World Model maintains an explicit representation of the agent's mental state throughout interaction. Drawing on the belief-desire-intention view of agent reasoning, we denote the mental state at step t as

$$M_t = (B_t, D_t, I_t), \tag{3}$$

where B_t, D_t, and I_t represent belief, desire, and intention, respectively. Belief captures the agent's current understanding of the situation, desire reflects the outcome it seeks, and intention describes its intended course of action. For social scenarios, the representation can additionally contain counterpart-related mental states.

The mental state is revised as the interaction unfolds. After action P_t and observation O_{t+1}, the updated history is used to infer $M_{t+1} \sim p_\phi(\cdot \mid x, H_{t+1})$, yielding the interaction loop

$$M_t \to P_t \to O_{t+1} \to M_{t+1}. \tag{4}$$

The Policy Model uses the latest mental state together with the original task context and interaction history. Mental-state estimation and action generation therefore remain coupled as new information becomes available.

Figure 2 provides an overview of the MindWorld framework, including dynamic mental world modeling, iterative co-optimization, and step-aware credit assignment feedback.

3.3 Step-Aware Credit Assignment

Outcome rewards evaluate whether an entire trajectory succeeds, but they provide limited information about the contribution of individual steps. In long-horizon interactions, a final success or failure cannot directly indicate whether a predicted mental state was useful or whether a specific action improved the situation. This problem is particularly important for MindWorld because the Mental World Model and the Policy Model are optimized for different intermediate behaviors: one should maintain useful mental states, while the other should select actions that advance the task.

Therefore, MindWorld introduces step-aware credit assignment by combining trajectory-level outcome supervision with step-level feedback. The outcome reward provides global task-level guid-

ance, while step rewards evaluate intermediate decisions from the perspective of each model. This allows the two models to receive more informative learning signals without relying on dense human annotations.

Step-Level Reward Modeling.

MindWorld uses two offline-trained scorers to provide complementary step-level feedback. For the Mental World Model, we construct *with-BDI* and *without-BDI* training pairs using GPT-4o, inspired by counterfactual reward analysis (Reber et al., 2025). The scorer s_M evaluates whether the predicted mental state improves the subsequent decision:

$$r_{i,t}^{M} = s_M(x, H_{i,t}, M_{i,t}) \in [0, 1], \tag{5}$$

where $H_{i,t}$ and $M_{i,t}$ denote the interaction history and the predicted mental state at step t. A higher value indicates that the mental state provides more useful information for the next decision.

For the Policy Model, the scorer s_P evaluates the progress caused by the current action based on the transition $(S_{i,t}, P_{i,t}, S_{i,t+1})$. It predicts one of five progress levels, $\mathcal{V} = \{-1, -0.5, 0, 0.5, 1\}$, where positive values indicate task progress, zero indicates no significant change, and negative values indicate regression. The predicted class value is directly used as the policy step reward $r_{i,t}^{P}$.

Both scorers use `all-mpnet-base-v2` with lightweight task-specific heads (Song et al., 2020; Reimers & Gurevych, 2019). They are trained separately before reinforcement learning and remain frozen during optimization.

Advantage Estimation and Credit Assignment.

For each task x, we sample G rollouts and compute a trajectory-level outcome advantage:

$$\widehat{A}_{i}^{\text{out}} = \frac{R_{i}^{\text{out}} - \mu_{x}^{\text{out}}}{\sigma_{x}^{\text{out}} + \epsilon_{\text{norm}}}, \tag{6}$$

where R_i^{out} is the final task reward. This global signal is shared across all steps of the same rollout.

To distinguish the contribution of individual steps, we further compute within-trajectory relative advantages. For model $z \in \{M, P\}$ and feedback type $d \in \{\text{step}, \text{fmt}\}$:

$$\widehat{A}_{i,t}^{d,z} = \frac{r_{i,t}^{d,z} - \mu_{i}^{d,z}}{\sigma_{i}^{d,z} + \epsilon_{\text{norm}}}. \tag{7}$$

Here, $r_{i,t}^{\text{step},M} = r_{i,t}^{M}$ and $r_{i,t}^{\text{step},P} = r_{i,t}^{P}$. The statistics are computed over the steps within the same rollout, meaning that each decision is compared with other decisions in the same trajectory rather than with the same position across different rollouts. Therefore, these local terms provide relative credit signals instead of estimating an absolute action value. A negative local advantage only indicates below-average contribution within the same trajectory and does not imply a negative raw reward.

The normalized components are combined as:

$$\widehat{A}_{i,t}^{z} = \lambda_{\text{out}}^{z}\widehat{A}_{i}^{\text{out}} + \lambda_{\text{step}}^{z}\widehat{A}_{i,t}^{\text{step},z} + \lambda_{\text{fmt}}^{z}\widehat{A}_{i,t}^{\text{fmt},z}. \tag{8}$$

The outcome advantage provides trajectory-level optimization direction, while the local terms identify which mental-state predictions and actions contribute more within the interaction.

Step-Aware Policy Optimization.

We optimize the Mental World Model and Policy Model separately using a clipped GRPO objective (Shao et al., 2024):

$$\mathcal{L}_z = -\mathbb{E}_{i,t,j}\left[\min\left(\rho_{i,t,j}^{z}\widehat{A}_{i,t}^{z}, \text{clip}\left(\rho_{i,t,j}^{z}, 1-\epsilon_{\text{clip}}, 1+\epsilon_{\text{clip}}\right)\widehat{A}_{i,t}^{z}\right)\right]. \tag{9}$$

The likelihood ratio is defined as:

$$\rho_{i,t,j}^{z} = \frac{q_z(y_{i,t,j}^{z} \mid c_{i,t}^{z}, y_{i,t,<j}^{z})}{q_{z,\text{old}}(y_{i,t,j}^{z} \mid c_{i,t}^{z}, y_{i,t,<j}^{z})}. \tag{10}$$

For the Mental World Model, $q_M = p_\phi$, while for the Policy Model, $q_P = \pi_\theta$. All tokens generated within the same step share the corresponding step advantage. During each update, rollout samples and advantages are fixed, and only the parameters of the model being optimized are updated.

Algorithm 1 Iterative Co-Optimization of MindWorld

Require: SFT-initialized models $\phi^{(0)}, \theta^{(0)}$; task set $\mathcal{D}_0$; fixed scorers s_M, s_P; rounds K; rollouts per task G; threshold δ

for $k = 0, \ldots, K-1$ **do**
  Freeze $\theta^{(k)}$; $\phi^{(k+1)} \leftarrow \mathcal{U}_M(\phi^{(k)}; \theta^{(k)}, \mathcal{D}_k)$
  Freeze $\phi^{(k+1)}$; $\theta^{(k+1)} \leftarrow \mathcal{U}_P(\theta^{(k)}; \phi^{(k+1)}, \mathcal{D}_k)$
  Generate G fresh rollouts per task using $(\phi^{(k+1)}, \theta^{(k+1)})$
  Form $\mathcal{D}_{k+1}$ by retaining tasks with outcome-reward variance $\geq \delta$
  if $\mathcal{D}_{k+1} = \emptyset$ **then**
    return $\phi^{(k+1)}, \theta^{(k+1)}$
  end if
end for
return $\phi^{(K)}, \theta^{(K)}$

3.4 Iterative Co-Optimization

MindWorld optimizes the Mental World Model and Policy Model alternately after supervised initialization. Each round first updates the Mental World Model with the Policy Model fixed, then updates the Policy Model with the revised Mental World Model fixed.

Let $\phi^{(k)}$ and $\theta^{(k)}$ denote the model parameters at round k. The updates are

$$\phi^{(k+1)} = \mathcal{U}_M\left(\phi^{(k)}; \theta^{(k)}, \mathcal{D}_k\right), \qquad \theta^{(k+1)} = \mathcal{U}_P\left(\theta^{(k)}; \phi^{(k+1)}, \mathcal{D}_k\right), \tag{11}$$

where $\mathcal{D}_k$ is the retained task set. Each update operator $\mathcal{U}_z$ collects fresh rollouts with the active model and its frozen counterpart, computes the rewards and advantages in Section 3.3, and applies Eq. 9. Rollouts are refreshed throughout each training phase. The complete training procedure is summarized in Algorithm 1.

Updated mental states change the information available to the Policy Model, while updated actions produce new histories for subsequent mental-state learning. To maintain variation in outcome-based supervision, tasks with rollout outcome variance below a threshold δ are removed before the next round. This filtering avoids spending optimization budget on tasks where all sampled trajectories provide nearly identical signals. The scorers remain fixed throughout.

4 Experiments

We evaluate MindWorld across three types of interactive tasks, including social interaction, embodied decision making, and web navigation. We compare against representative agent methods and further analyze the effectiveness of the proposed components and iterative optimization.

4.1 Experimental Setup

Benchmarks. We evaluate MindWorld on four interactive benchmarks: SOTOPIA and SOTOPIA-Hard (Zhou et al., 2024a), ALFWorld (Shridhar et al., 2021), and WebShop (Yao et al., 2022). SOTOPIA evaluates goal-oriented social interaction, while SOTOPIA-Hard focuses on more challenging scenarios requiring stronger strategic reasoning. ALFWorld evaluates embodied task planning through language-based interaction, and WebShop evaluates multi-step decision making in a web environment. Training data are constructed from the official benchmark scenarios, expert trajectories, and released interaction data, with detailed preprocessing and splits provided in Appendix A.

Baselines. We implement MindWorld on Llama3.1-8B-Instruct (Llama Team, 2024) and Qwen3-8B (Yang et al., 2025a), and compare against four groups of baselines: (1) **Frontier LLMs**, including GPT-5.6 Sol (OpenAI, 2026), Claude Fable 5.1 (Anthropic, 2026), Gemini 3.7 Flash (Google DeepMind, 2026), DeepSeek V4 Pro (DeepSeek-AI, 2026), and Grok 4.6 (xAI, 2026); (2) **Social-Agent Methods** on SOTOPIA and SOTOPIA-Hard, including DSI (Zhang et al., 2025), EPO (Liu et al., 2025a), Sotopia-RL (Yu et al., 2025), AMPO (Wang et al., 2026a), SAVOIR (Feng et al., 2026), and iStar (Liu et al., 2025b); (3) **General RL Methods** on ALFWorld and WebShop, including

Table 1: Main results on SOTOPIA and SOTOPIA-Hard under Self-Play and GPT-4o-as-Partner. We report GOAL and OVERALL. Bold denotes the best result within each backbone group.

Models	Self-Play				GPT-4o-as-Partner			
	SOTOPIA		SOTOPIA-Hard		SOTOPIA		SOTOPIA-Hard	
	GOAL↑	OVERALL↑	GOAL↑	OVERALL↑	GOAL↑	OVERALL↑	GOAL↑	OVERALL↑
Frontier LLMs								
GPT-5.6 Sol	8.78	3.83	7.66	3.54	8.61	3.79	7.44	3.49
Claude Fable 5.1	8.86	3.84	7.58	3.51	8.69	3.80	7.49	3.53
Gemini 3.7 Flash	8.39	3.72	7.02	3.33	8.28	3.69	6.94	3.30
DeepSeek V4 Pro	8.54	3.77	7.25	3.42	8.40	3.73	7.13	3.38
Grok 4.6	8.46	3.71	7.11	3.36	8.35	3.70	7.03	3.34
Llama Backbone								
Llama3.1-8B-Instruct	6.99	3.23	5.11	2.36	7.68	3.62	6.21	3.05
w/ DSI	8.08	3.64	6.93	3.31	8.08	3.67	6.84	3.41
w/ EPO	7.94	3.78	6.79	3.27	8.37	3.83	6.91	3.53
w/ Sotopia-RL	8.12	3.68	7.58	3.69	8.28	3.87	6.77	3.31
w/ AMPO	9.08	3.95	8.06	3.68	8.75	**3.98**	7.68	3.74
w/ SAVOIR	8.41	3.84	7.89	**3.94**	8.40	3.93	7.16	3.52
w/ iStar	9.20	3.91	8.06	3.70	8.45	3.89	7.16	3.50
w/ MindWorld (Ours)	**9.23**	**3.98**	**8.12**	3.89	**8.81**	3.94	**7.79**	**3.81**
Qwen Backbone								
Qwen3-8B	8.04	3.61	6.71	3.19	7.18	3.22	6.20	3.05
w/ DSI	8.48	3.78	7.42	3.56	8.26	3.75	7.02	3.47
w/ EPO	8.31	3.66	7.01	3.31	8.44	3.89	7.03	3.57
w/ Sotopia-RL	8.35	3.76	7.82	3.78	8.46	3.92	7.04	3.38
w/ AMPO	9.13	3.96	8.11	3.70	8.73	3.93	7.74	3.73
w/ SAVOIR	8.74	3.91	8.10	3.91	8.58	3.96	7.38	3.58
w/ iStar	8.72	3.88	7.56	3.60	8.57	3.91	7.12	3.52
w/ MindWorld (Ours)	**9.27**	**4.03**	**8.16**	**3.94**	**8.83**	**3.98**	**7.80**	**3.83**

PPO (Schulman et al., 2017), GRPO (Shao et al., 2024), RLOO (Ahmadian et al., 2024), and REINFORCE++ (Hu, 2025); and (4) **Task-Specific Agent Methods**, including GiGPO (Feng et al., 2025), StraTA (Xue et al., 2026), BEACON (Wang et al., 2026b), and SCPO (Xu et al., 2026). For the task-specific methods, we additionally report their MindWorld-augmented variants to evaluate the complementarity of MindWorld with specialized agent-learning approaches.

Evaluation. For SOTOPIA and SOTOPIA-Hard, we follow SOTOPIA-EVAL (Zhou et al., 2024a) and report GOAL (0–10) and OVERALL, where OVERALL is the average over seven social dimensions. We consider both Self-Play, where the evaluated agent interacts with itself, and GPT-4o-as-Partner, where it interacts with GPT-4o (OpenAI, 2024), with GPT-4o serving as the evaluator. For ALFWorld, we report task Success Rate (%). For WebShop, we report both Success Rate (%) and the average environment Score. Appendix A describes implementation details, optimization settings, and computational resources and Appendix C defines the evaluation metrics.

4.2 Main Results

MindWorld consistently improves social interaction performance. As shown in Table 1, MindWorld achieves competitive performance and obtains the best results on most metrics across both backbone groups. With Qwen3-8B, it reaches 9.27 GOAL and 4.03 OVERALL on SOTOPIA under self-play, and 8.16 GOAL and 3.94 OVERALL on SOTOPIA-Hard. The OVERALL comparison is more mixed for Llama3.1-8B-Instruct. MindWorld scores 3.89 on SOTOPIA-Hard under self-play, below SAVOIR's 3.94, and 3.94 on SOTOPIA with GPT-4o as the partner, below AMPO's 3.98. Across the two backbone groups, MindWorld has the highest reported value in 14 of the 16 metric comparisons, indicating that explicit mental-state modeling remains effective across different interaction settings and model backbones.

MindWorld generalizes to diverse interactive decision-making tasks. On ALFWorld and WebShop, standalone MindWorld improves on the general RL methods reported in Table 2. For Qwen3-8B, its ALFWorld success rates are 88.9% on seen tasks and 87.4% on unseen tasks, compared with 82.2% and 81.0% for GRPO. Its WebShop success rate is 84.3%, compared with 81.0% for GRPO. These correspond to gains of 6.7, 6.4, and 3.3 percentage points, respectively.

Table 2: Performance on ALFWorld and WebShop. For general RL methods, we report their standalone performance. For task-specific agent methods, we report the performance after augmentation with MindWorld, with the absolute improvement over the original method shown in parentheses.

Method	Llama3.1-8B-Instruct				Qwen3-8B			
	ALFWorld		WebShop		ALFWorld		WebShop	
	Seen↑	Unseen↑	Success↑	Score↑	Seen↑	Unseen↑	Success↑	Score↑
General RL Methods								
Base Model	20.3	18.7	15.4	34.1	23.1	21.6	18.2	38.0
+ PPO	78.4	76.8	78.2	86.3	79.6	78.1	79.0	87.0
+ RLOO	79.5	78.6	77.6	87.4	80.3	79.4	78.5	88.1
+ GRPO	81.1	79.9	80.1	89.2	82.2	81.0	81.0	89.8
+ REINFORCE++	82.3	81.2	80.8	89.7	83.5	82.1	81.6	90.2
+ MindWorld (Ours)	**87.6**	**86.2**	**83.5**	**91.1**	**88.9**	**87.4**	**84.3**	**91.8**
Task-Specific Agent Methods (+MindWorld)								
+ GiGPO	92.5 (+1.5)	92.4 (+2.0)	77.8 (+2.2)	87.3 (+1.3)	93.4 (+1.6)	93.0 (+2.1)	78.2 (+3.4)	86.6 (+1.3)
+ StraTA	94.9 (+1.1)	94.1 (+1.4)	85.6 (+1.2)	91.8 (+0.6)	95.8 (+1.2)	95.0 (+1.6)	87.6 (+2.5)	92.4 (+0.7)
+ BEACON	95.8 (+0.9)	95.2 (+1.3)	82.6 (+1.8)	89.7 (+0.7)	96.3 (+0.9)	96.0 (+1.2)	83.9 (+2.6)	89.8 (+1.0)
+ SCPO	96.3 (+0.7)	95.4 (+1.4)	82.2 (+2.4)	90.4 (+0.9)	96.8 (+0.7)	95.8 (+1.3)	80.9 (+1.7)	90.2 (+1.0)

Table 3: Ablation of MindWorld under the Self-Play setting. We report GOAL and OVERALL on SOTOPIA and SOTOPIA-Hard. Bold denotes the full MindWorld results.

Variant	Llama3.1-8B-Instruct (Self-Play)				Qwen3-8B (Self-Play)			
	SOTOPIA		SOTOPIA-Hard		SOTOPIA		SOTOPIA-Hard	
	GOAL↑	OVERALL↑	GOAL↑	OVERALL↑	GOAL↑	OVERALL↑	GOAL↑	OVERALL↑
w/ SFT	6.76	3.25	5.14	2.24	7.73	3.49	6.18	3.01
w/o Mental World Model	8.31	3.70	6.98	3.39	8.60	3.81	7.37	3.57
w/o Step Reward	8.84	3.82	7.49	3.57	8.98	3.91	7.72	3.67
w/o Outcome Reward	8.67	3.76	7.63	3.64	8.79	3.88	7.58	3.71
Shuffled BDI	8.18	3.65	6.87	3.34	8.46	3.78	7.24	3.52
MindWorld (Ours)	**9.23**	**3.98**	**8.12**	**3.89**	**9.27**	**4.03**	**8.16**	**3.94**

The task-specific rows report each method combined with MindWorld, with gains relative to the corresponding original method. The largest success-rate increase in these rows is 3.4 percentage points for Qwen3-8B with GiGPO on WebShop. These results concern the reported combinations; they should not be read as standalone MindWorld outperforming every specialized method.

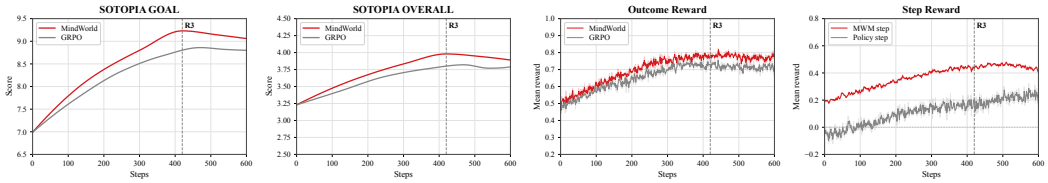
4.3 Ablation and Iterative Analysis

Ablation Study. As shown in Table 3, removing the Mental World Model, step-level reward, or outcome reward reduces performance on both backbones, with larger drops generally observed on SOTOPIA-Hard. The reward ablations suggest that trajectory-level supervision and local feedback contribute complementary learning signals. Shuffling the predicted BDI states performs worse than removing the Mental World Model, suggesting that contextually mismatched mental states can interfere with action selection. The full MindWorld achieves the best results across all reported metrics, supporting the combined use of explicit mental-state modeling and step-aware credit assignment.

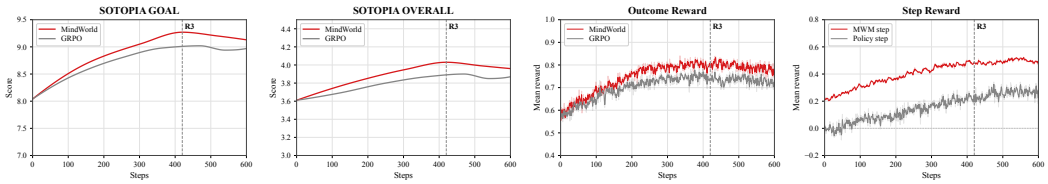
Iterative Analysis. Table 4 shows that the gains extend beyond the first round, which includes SFT initialization and initial co-optimization. From Round 1 to Round 3, SOTOPIA GOAL increases from 8.17 to 9.23 on Llama and from 8.72 to 9.27 on Qwen, with further improvements also observed on SOTOPIA-Hard. Both backbones peak at Round 3, whose checkpoints are used for the final results in Table 1. The slight decline in Rounds 4 and 5 indicates that additional optimization does not necessarily improve evaluation performance, supporting checkpoint selection based on held-out validation.

Table 4: Iterative analysis of MindWorld under the Self-Play setting. We report GOAL and OVERALL on SOTOPIA and SOTOPIA-Hard across five training rounds. Bold denotes the best checkpoint, which is used for the final evaluation.

Iteration	Llama3.1-8B-Instruct (Self-Play)				Qwen3-8B (Self-Play)			
	SOTOPIA		SOTOPIA-Hard		SOTOPIA		SOTOPIA-Hard	
	GOAL↑	OVERALL↑	GOAL↑	OVERALL↑	GOAL↑	OVERALL↑	GOAL↑	OVERALL↑
Baseline	6.99	3.23	5.11	2.36	8.04	3.61	6.71	3.19
Round 1	8.17	3.60	6.79	3.08	8.72	3.81	7.49	3.56
Round 2	8.82	3.84	7.57	3.55	9.06	3.95	7.91	3.81
Round 3 (Ours)	**9.23**	**3.98**	**8.12**	**3.89**	**9.27**	**4.03**	**8.16**	**3.94**
Round 4	9.14	3.94	8.04	3.82	9.20	3.99	8.09	3.88
Round 5	9.06	3.89	7.95	3.76	9.13	3.96	8.01	3.84



(a) Results on Llama3.1-8B-Instruct



(b) Results on Qwen3-8B

Figure 3: Training dynamics of MindWorld on SOTOPIA. GOAL and OVERALL are evaluated on the datasets, with the dashed line marking Round 3, the best validation checkpoint. Outcome Reward is the trajectory-level SOTOPIA goal reward normalized to $[0, 1]$.

4.4 Analysis

Figure 3 illustrates the training dynamics of MindWorld on SOTOPIA. GOAL and OVERALL improve rapidly in the early stages, remain consistently above GRPO, and reach their best validation performance around Round 3, after which they decline slightly, consistent with Table 4. The normalized outcome reward follows a similar overall trend, while the step-level signals reveal how local behavior improves during training: the Mental World Model reward remains positive and generally increases, indicating more useful mental-state predictions, whereas the Policy Model reward gradually shifts from near-zero or slightly negative values to positive values, reflecting actions that increasingly contribute to task progress.

5 Conclusion

We presented **MindWorld**, a framework for improving social and interactive reasoning by explicitly coupling a Mental World Model with a Policy Model. MindWorld dynamically tracks beliefs, desires, and intentions throughout interaction, uses step-aware credit assignment to combine step-level signals with trajectory-level outcomes, and iteratively co-optimizes mental-state modeling and action generation. Experiments on SOTOPIA, ALFWorld, and WebShop show that this design substantially improves social interaction performance, outperforms general RL baselines, and remains competitive with task-specific agent methods across household and web-based tasks. These results highlight the value of explicitly modeling evolving mental states and jointly optimizing them with policy learning for more adaptive and effective language agents.

AI USE STATEMENT

AI tools assisted with language editing and the preparation of revision suggestions and code examples. The authors are responsible for checking the manuscript, implementation, references, and empirical claims. GPT-4o supervision used within the research method is described separately in Appendices B and E.

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 publicly available benchmarks for social interaction, embodied tasks, and web-based tasks. 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.

Resources The study uses open-weight models, hosted model APIs, and benchmark datasets. These are distinct forms of access and are governed by their respective licenses or service terms. API availability is not an open-weight release. The model and dataset references identify the resources used.

REPRODUCIBILITY STATEMENT

We describe the model configurations, training procedure, reward design, evaluation protocol, and key hyperparameters in the main paper and appendix. All experiments are conducted on publicly available benchmarks and base models under consistent evaluation settings. We report the random seeds and computational setup used in our experiments, and will release the implementation, training prompts, scorer configurations, and reproduction scripts upon publication. Implementation details, optimization settings, and computational resources are provided in Appendices A to E.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 12248–12267. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.662. URL https://doi.org/10.18653/v1/2024.acl-long.662.

Anthropic. Claude fable 5.1 and claude mythos 5.1. https://www.anthropic.com/claude-fable-and-mythos-5-1, September 2026.

Hyungjoo Chae, Namyoung Kim, Kai Tzu-iunn Ong, Minju Gwak, Gwanwoo Song, Jihoon Kim, Sunghwan Kim, Dongha Lee, and Jinyoung Yeo. Web agents with world models: Learning and leveraging environment dynamics in web navigation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=moWiYJuSGF.

Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. Step-level value preference optimization for mathematical reasoning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, volume EMNLP 2024 of *Findings of ACL*, pp. 7889–7903. Association for Computational Linguistics, 2024a. doi: 10.18653/V1/2024.FINDINGS-EMNLP.463. URL https://doi.org/10.18653/v1/2024.findings-emnlp.463.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. ToMBench: Benchmarking theory of mind in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, pp. 15959–15983. Association for Computational Linguistics, 2024b. doi: 10.18653/V1/2024.ACL-LONG.847. URL https://doi.org/10.18653/v1/2024.acl-long.847.

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=mZn2Xyh9Ec.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026. URL https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro.

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

Tianqing Fang, Hongming Zhang, Zhisong Zhang, Kaixin Ma, Wenhao Yu, Haitao Mi, and Dong Yu. WebEvolver: Enhancing web agent self-improvement with co-evolving world model. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 8959–8975. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.454. URL https://doi.org/10.18653/v1/2025.emnlp-main.454.

Hao Fei and Yiran Zhao. Mental world modeling. *CoRR*, abs/2607.27201, 2026. doi: 10.48550/ARXIV.2607.27201. URL https://doi.org/10.48550/arXiv.2607.27201.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/420c9f777c0b4f78d515e53cf74d58b2-Abstract-Conference.html.

Xiachong Feng, Yi Jiang, Xiaocheng Feng, Deyi Yin, Libo Qin, Yangfan Ye, Lei Huang, Weitao Ma, Yuxuan Gu, Chonghan Qin, Bing Qin, and Lingpeng Kong. SAVOIR: learning social savoir-faire via shapley-based reward attribution. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 14276–14290. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.FINDINGS-ACL.699. URL https://doi.org/10.18653/v1/2026.findings-acl.699.

Zhaolin Gao, Wenhao Zhan, Jonathan Daniel Chang, Gokul Swamy, Kianté Brantley, Jason D. Lee, and Wen Sun. Regressing the relative future: Efficient policy optimization for multi-turn RLHF. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=cVyELMpMRS.

John Gkountouras, Matthias Lindemann, Phillip Lippe, Efstratios Gavves, and Ivan Titov. Language agents meet causality - bridging llms and causal world models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=y9A2TpaGsE.

Google DeepMind. Gemini 3.7 flash: Model card. https://deepmind.google/models/model-cards/gemini-3-7-flash/, August 2026.

Mingqian He, Yongliang Shen, Wenqi Zhang, Zeqi Tan, and Weiming Lu. Advancing process verification for large language models via tree-based preference learning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in*

Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024, pp. 2086–2099. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.125. URL https://doi.org/10.18653/v1/2024.emnlp-main.125.

Jian Hu. REINFORCE++: A simple and efficient approach for aligning large language models. *CoRR*, abs/2501.03262, 2025. doi: 10.48550/ARXIV.2501.03262. URL https://doi.org/10.48550/arXiv.2501.03262.

Mehdi Jafari, Yuncheng Hua, Hao Xue, and Flora D. Salim. Beyond words: Integrating theory of mind into conversational agents for human-like belief, desire, and intention alignment. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pp. 5489–5508. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-ACL.287. URL https://doi.org/10.18653/v1/2025.findings-acl.287.

Minsoo Kim and Seung-won Hwang. Coex - co-evolving world-model and exploration. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 21629–21651. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-EMNLP.1179. URL https://doi.org/10.18653/v1/2025.findings-emnlp.1179.

Aobo Kong, Wentao Ma, Shiwan Zhao, Yongbin Li, Yuchuan Wu, Ke Wang, Xiaoqian Liu, Qicheng Li, Yong Qin, and Fei Huang. SDPO: segment-level direct preference optimization for social agents. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *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*, pp. 12409–12423. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.607. URL https://doi.org/10.18653/v1/2025.acl-long.607.

Xiaoqian Liu, Ke Wang, Yongbin Li, Yuchuan Wu, Wentao Ma, Aobo Kong, Fei Huang, Jianbin Jiao, and Junge Zhang. EPO: explicit policy optimization for strategic reasoning in llms via reinforcement learning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *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*, pp. 15371–15396. Association for Computational Linguistics, 2025a. doi: 10.18653/V1/2025.ACL-LONG.747. URL https://doi.org/10.18653/v1/2025.acl-long.747.

Xiaoqian Liu, Ke Wang, Yuchuan Wu, Fei Huang, Yongbin Li, Junge Zhang, and Jianbin Jiao. Agentic reinforcement learning with implicit step rewards. *CoRR*, abs/2509.19199, 2025b. doi: 10.48550/ARXIV.2509.19199. URL https://doi.org/10.48550/arXiv.2509.19199.

Llama Team. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Xinyi Mou, Jingcong Liang, Jiayu Lin, Xinnong Zhang, Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen, Haoyu Kuang, Xuanjing Huang, and Zhongyu Wei. Agentsense: Benchmarking social intelligence of language agents through interactive scenarios. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pp. 4975–5001. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.NAACL-LONG.257. URL https://doi.org/10.18653/v1/2025.naacl-long.257.

OpenAI. Gpt-4o system card, August 2024. URL https://openai.com/index/gpt-4o-system-card/.

OpenAI. Gpt-5.6: Frontier intelligence that scales with your ambition, 2026. URL https://openai.com/index/gpt-5-6/.

Anand S. Rao and Michael P. Georgeff. BDI agents: From theory to practice. In Victor R. Lesser and Les Gasser (eds.), *Proceedings of the First International Conference on Multiagent Systems, June 12-14, 1995, San Francisco, California, USA*, pp. 312–319. The MIT Press, 1995.

David Reber, Sean M. Richardson, Todd Nief, Cristina Garbacea, and Victor Veitch. RATE: causal explainability of reward models with imperfect counterfactuals. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/reber25a.html.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese bert-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *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*, pp. 3980–3990. Association for Computational Linguistics, 2019. doi: 10.18653/V1/D19-1410. URL https://doi.org/10.18653/v1/D19-1410.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for LLM reasoning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=A6Y7AqlzLW.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/ARXIV.2402.03300. URL https://doi.org/10.48550/arXiv.2402.03300.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew J. Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=0IOX0YcCdTn.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c3a690be93aa602ee2dc0ccab5b7b67e-Abstract.html.

Minzheng Wang, Yongbin Li, Haobo Wang, Xinghua Zhang, Nan Xu, Bingli Wu, Fei Huang, Haiyang Yu, and Wenji Mao. Adaptive social learning via mode policy optimization for language agents. In *Proceedings of ICLR*, 2026a. URL https://openreview.net/forum?id=GG7YQnsdhp.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. Sotopia-π: Interactive learning of socially intelligent language agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 12912–12940. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.698. URL https://doi.org/10.18653/v1/2024.acl-long.698.

Zixuan Wang, Yuchen Yan, Hongxing Li, Teng Pan, Dingming Li, Ruiqing Zhang, Weiming Lu, Jun Xiao, Yueting Zhuang, and Yongliang Shen. Milestone-guided policy learning for long-horizon language agents. *CoRR*, abs/2605.06078, 2026b. doi: 10.48550/ARXIV.2605.06078. URL https://doi.org/10.48550/arXiv.2605.06078.

Jincenzi Wu, Zhuang Chen, Jiawen Deng, Sahand Sabour, Helen Meng, and Minlie Huang. COKE: A cognitive knowledge graph for machine theory of mind. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 15984–16007. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.848. URL https://doi.org/10.18653/v1/2024.acl-long.848.

xAI. Introducing grok 4.6. https://x.ai/news/grok-4-6, August 2026.

Siheng Xiong, Ali Payani, Yuan Yang, and Faramarz Fekri. Deliberate reasoning in language models as structure-aware planning with an accurate world model. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31900–31931, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1540. URL https://aclanthology.org/2025.acl-long.1540/.

Peng Xu, Sijia Chen, Junzhuo Li, and Xuming Hu. Semantic consistency policy optimization for reinforcement learning of LLM agents. *CoRR*, abs/2606.25852, 2026. doi: 10.48550/ARXIV.2606.25852. URL https://doi.org/10.48550/arXiv.2606.25852.

Xiangyuan Xue, Yifan Zhou, Zidong Wang, Shengji Tang, Philip Torr, Wanli Ouyang, Lei Bai, and Zhenfei Yin. StraTA: Incentivizing agentic reinforcement learning with strategic trajectory abstraction. *CoRR*, abs/2605.06642, 2026. doi: 10.48550/ARXIV.2605.06642. URL https://doi.org/10.48550/arXiv.2605.06642.

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, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Ruihan Yang, Yikai Zhang, Aili Chen, Xintao Wang, Jiangjie Chen, Siyu Yuan, Deqing Yang, and Yanghua Xiao. ARIA: training language agents with intention-driven reward aggregation. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025b. URL http://papers.nips.cc/paper_files/paper/2025/hash/9e13188461115e17a4e719d5c00cb35d-Abstract-Conference.html.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *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*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/82ad13ec01f9fe44c01cb91814fd7b8c-Abstract-Conference.html.

Minjong Yoo, Jinwoo Jang, Sihyung Yoon, and Honguk Woo. World model implanting for test-time adaptation of embodied agents. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second*

International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/yoo25a.html.

Haofei Yu, Zhengyang Qi, Yining Zhao, Kolby Nottingham, Keyang Xuan, Bodhisattwa Prasad Majumder, Hao Zhu, Paul Pu Liang, and Jiaxuan You. Sotopia-rl: Reward design for social intelligence. *CoRR*, abs/2508.03905, 2025. doi: 10.48550/ARXIV.2508.03905. URL https://doi.org/10.48550/arXiv.2508.03905.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. Free process rewards without process labels. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/yuan25c.html.

Wenyuan Zhang, Tianyun Liu, Mengxiao Song, Xiaodong Li, and Tingwen Liu. SOTOPIA-Ω: dynamic strategy injection learning and social instruction following evaluation for social agents. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *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*, pp. 24669–24697. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.1203. URL https://doi.org/10.18653/v1/2025.acl-long.1203.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. PyTorch FSDP: experiences on scaling fully sharded data parallel. *Proc. VLDB Endow.*, 16(12):3848–3860, 2023. doi: 10.14778/3611540.3611569. URL https://www.vldb.org/pvldb/vol16/p3848-huang.pdf.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024a. URL https://openreview.net/forum?id=mM7VurbA4r.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. ArCHer: Training language model agents via hierarchical multi-turn RL. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62178–62209. PMLR / OpenReview.net, 2024b. URL https://proceedings.mlr.press/v235/zhou24t.html.

A Training and Implementation Details

A.1 General Setup

MindWorld contains two independently parameterized models: a Mental World Model and a Policy Model. Our main experiments use Qwen3-8B and Llama3.1-8B-Instruct, trained on $4\times$ NVIDIA RTX PRO 6000 96GB GPUs.

Both SFT and RL use full-parameter optimization. During each RL phase, only the active model is updated while the other model and both step scorers remain frozen. We use BF16, TF32, FlashAttention-2 (Dao, 2024), gradient checkpointing, and FSDP full sharding (Zhao et al., 2023).

For SFT, we use AdamW (Loshchilov & Hutter, 2019) with learning rate 1×10^{-5}, weight decay 0.01, cosine decay, warmup ratio 0.03, gradient clipping 1.0, and effective batch size 8. Prompt tokens are masked and only target tokens contribute to the language-modeling loss.

For RL, we use AdamW with learning rate 5×10^{-7} on SOTOPIA and 1×10^{-6} on ALFWorld and WebShop, with cosine decay, warmup ratio 0.03, and gradient clipping 1.0. For each task, we sample $G = 4$ trajectories. We set $\epsilon_{\text{norm}} = 10^{-6}, \epsilon_{\text{clip}} = 0.2$. All benchmarks use five alternating co-optimization rounds. After each round, tasks with normalized outcome-reward variance below $\delta = 2 \times 10^{-3}$ are removed. Round 3 is selected according to held-out validation performance and used for the main results. Step-scorer training is described in Appendix B.

A.2 SOTOPIA AND SOTOPIA-HARD

We construct SOTOPIA training data from the official scenarios, profiles, private goals, and interaction trajectories. For each generation model, we use approximately 1,250 SFT examples and 250 validation examples. The initial RL pool contains about 2,200 complete interaction episodes.

The Mental World Model predicts Self/Counterpart beliefs, desires, and intentions, while the Policy Model generates the next utterance conditioned on the latest mental state. The maximum context length is 6144 tokens, with at most 512 and 256 generated tokens for the Mental World Model and Policy Model, respectively. SFT is performed for three epochs. RL generation uses temperature 0.8 and top-$p = 0.95$.

The trajectory reward is the normalized SOTOPIA GOAL score,

$$R_i^{\text{out}} = \frac{\text{GOAL}_i}{10}. \tag{12}$$

The normalized advantages are combined using

$$\lambda_{\text{out}} = 1.0, \qquad \lambda_{\text{step}} = 0.65, \qquad \lambda_{\text{fmt}} = 0.20. \tag{13}$$

The five rounds contain approximately $160/145/115/100/80$ global updates, with Round 3 ending at update 420. We evaluate both Self-Play and GPT-4o-as-Partner and report GOAL and OVERALL following SOTOPIA-EVAL.

A.3 ALFWORLD

We use the official text-only ALFWorld environment. For each generation model, approximately 1,800 SFT examples and 200 task-disjoint validation examples are constructed from expert trajectories. The RL pool contains about 1,500 complete household tasks.

The Mental World Model predicts the acting agent's BDI state and the Policy Model predicts one executable action. The maximum context length is 4,096 tokens, with 256/128 maximum generated tokens for the Mental World Model and Policy Model. SFT is performed for three epochs; RL uses temperature 0.8 and top-$p = 0.95$.

The trajectory-level reward is binary task success,

$$R_i^{\text{out}} = \mathbb{1}[\text{task completed}], \tag{14}$$

and the advantage weights are

$$\lambda_{\text{out}} = 1.0, \qquad \lambda_{\text{step}} = 0.55, \qquad \lambda_{\text{fmt}} = 0.20. \tag{15}$$

Five co-optimization rounds are used and Round 3 is selected for final evaluation. We report Success Rate (%) on the official seen and unseen splits.

A.4 WEBSHOP

We use the official text-based WebShop environment and expert interaction trajectories. For each generation model, approximately 1,700 SFT examples and 200 trajectory-disjoint validation examples are used. The RL pool contains about 1,500 goal-conditioned shopping sessions.

The Mental World Model tracks user requirements and search intent, while the Policy Model generates `search[...]` or `click[...]` actions. The context length is 4,096 tokens, with maximum generation lengths of 256/128 tokens. SFT is performed for three epochs and RL uses temperature 0.8 and top-$p = 0.95$.

Table 5: Versions and deployment methods of the LLMs used in our experiments.

Model	Version	Deployment method
Frontier LLMs		
GPT-5.6 Sol	`gpt-5.6-sol`	API
Claude Fable 5.1	`claude-fable-5-1`	API
Gemini 3.7 Flash	`gemini-3.7-flash`	API
DeepSeek V4 Pro	`deepseek-v4-pro`	API
Grok 4.6	`grok-4.6`	API
GPT-4o	`gpt-4o-2024-08-06`	API
Open-source LLMs		
Qwen3-8B	`Qwen3-8B`	vLLM
Qwen3-4B	`Qwen3-4B`	vLLM
Qwen3-1.7B	`Qwen3-1.7B`	vLLM
Llama3.1-8B-Instruct	`LLaMA-3.1-8B-Instruct`	vLLM
Llama3.2-3B-Instruct	`LLaMA-3.2-3B-Instruct`	vLLM

The official WebShop environment reward is used as the trajectory reward:

$$R_i^{\text{out}} = R_i^{\text{WebShop}} \in [0, 1]. \tag{16}$$

We use

$$\lambda_{\text{out}} = 1.0, \qquad \lambda_{\text{step}} = 0.50, \qquad \lambda_{\text{fmt}} = 0.20. \tag{17}$$

Five co-optimization rounds are performed and Round 3 is used for final reporting. We report Success Rate (%) and the official environment Score.

A.5 Smaller Backbones

We additionally evaluate Qwen3-4B, Qwen3-1.7B, and Llama3.2-3B-Instruct on $4\times$ RTX A6000 48GB GPUs. Both SFT and RL remain full-parameter training. The data splits, reward definitions, advantage coefficients, group size, filtering criterion, and five-round training procedure are kept unchanged. We use learning rate 2×10^{-5} for SFT and 1×10^{-6}/2×10^{-6} for RL on SOTOPIA and ALFWorld/WebShop, respectively.

A.6 Model Implementation

Table 5 lists the detailed model versions and deployment methods used in our experiments.

B Step Scorer Details

MindWorld uses two fixed step scorers: a Mental World Model scorer s_M for mental-state usefulness and a Policy Model scorer s_P for action progress. Both scorers are trained offline with GPT-4o supervision and remain frozen throughout RL.

B.1 Mental World Model Scorer

Let q^{with} and q^{without} denote usefulness scores under matched with-BDI and without-BDI decision conditions. Both conditions use the same scale. The normalized nonnegative target is

$$y^M = \text{clip}\left(\frac{q^{\text{with}} - q^{\text{without}}}{10}, 0, 1\right), \tag{18}$$

which measures the marginal benefit of explicitly conditioning on the mental state.

The scorer uses `all-mpnet-base-v2` with attention-mask mean pooling and a $768 \to 256 \to 1$ regression head. A sigmoid activation produces

$$r_{i,t}^M = s_M(x, H_{i,t}, M_{i,t}) \in [0, 1]. \tag{19}$$

We construct approximately 3.2K training pairs for SOTOPIA and 4K pairs for each of ALFWorld and WebShop, with roughly 10% reserved for validation using episode/task-disjoint splits.

B.2 Policy Progress Scorer

For each transition (S_t, P_t, S_{t+1}), GPT-4o evaluates the progress caused by the action and assigns one of five labels:

$$\mathcal{V} = \{-1, -0.5, 0, 0.5, 1\}. \tag{20}$$

These correspond to significant regression, slight regression, no change, effective progress, and substantial progress.

The scorer uses the same MPNet encoder with a $768 \to 256 \to 5$ classification head and cross-entropy loss. At inference time, the predicted class value is used directly as

$$r^P_{i,t} \in \mathcal{V}. \tag{21}$$

We use approximately 4K labeled transitions for each benchmark, with task- or episode-level separation between training and validation sets. GPT-4o uses deterministic decoding for all annotations.

B.3 Training and Use

Both scorers are trained with AdamW using learning rate 2×10^{-5}, weight decay 0.01, warmup ratio 0.06, gradient clipping 1.0, batch size 8, BF16 precision, and five epochs. SOTOPIA uses a maximum sequence length of 512 tokens, while ALFWorld and WebShop use 256 tokens. The Mental World Model scorer is selected by validation regression performance, while the Policy scorer is selected by macro-F1.

During RL, both scorers remain fixed and independently produce

$$r^M_{i,t} = s_M(x, H_{i,t}, M_{i,t}), \qquad r^P_{i,t} = s_P(x, S_{i,t}, P_{i,t}, S_{i,t+1}). \tag{22}$$

These raw rewards are converted into within-trajectory relative advantages before being combined with the trajectory-level outcome advantage.

C Experiment Details

C.1 Benchmarks

SOTOPIA and SOTOPIA-Hard. SOTOPIA (Zhou et al., 2024a) is an open-ended social-interaction benchmark in which two role-playing agents interact under a shared scenario while pursuing role-specific private goals. Each task specifies the social context, character profiles, and private goals, and agents interact through natural-language utterances and actions over multiple turns. The benchmark covers diverse social situations, including negotiation, collaboration, persuasion, conflict resolution, relationship management, and resource exchange. An interaction terminates when an agent leaves the conversation or when the maximum number of turns is reached.

We evaluate on both the full SOTOPIA benchmark and SOTOPIA-Hard, a challenging subset identified by the original benchmark in which strong language models exhibit substantially lower social-task performance. SOTOPIA serves as the primary benchmark for evaluating the social reasoning and mental-state modeling capabilities of MindWorld.

ALFWorld. ALFWorld (Shridhar et al., 2021) evaluates embodied household decision making through a text-based interface. An agent receives a natural-language household instruction and interacts with an environment through textual observations and discrete actions. Tasks include object manipulation, placement, cleaning, heating, cooling, and multi-object planning. We report task Success Rate (%) separately on the standard seen and unseen evaluation splits, following the official benchmark protocol.

WebShop. WebShop (Yao et al., 2022) evaluates language agents on multi-step web-shopping tasks. Given a natural-language user request containing product preferences and constraints, an agent must search for products, inspect candidate items, select appropriate options, and complete the shopping task through textual web interactions. We report both task Success Rate (%) and the official WebShop environment Score. The former measures complete task success, while the latter provides partial credit according to how well the selected product satisfies the target constraints.

C.2 SOTOPIA Evaluation Protocol

We follow the official SOTOPIA-EVAL framework (Zhou et al., 2024a) for evaluating social interaction. SOTOPIA-EVAL performs episode-level evaluation after the complete interaction and scores each agent independently along seven social dimensions. In all experiments, we use GPT-4o (OpenAI, 2024) as the evaluator with temperature 0 to provide deterministic judgments. The same evaluator model, evaluation prompt, and decoding configuration are used for all compared methods.

For each evaluation, GPT-4o receives the scenario information, role-specific profile and goal, together with the complete interaction trajectory. It produces a short justification and an integer score for each SOTOPIA-EVAL dimension. We use the same evaluation protocol under both interaction settings:

- **Self-Play**: two copies of the evaluated agent interact with each other under their respective private roles and goals;
- **GPT-4o-as-Partner**: the evaluated agent interacts with GPT-4o as the counterpart, while GPT-4o is also used through a separate evaluation call to score the completed interaction.

All reported SOTOPIA results are averaged over the evaluated episodes and agent roles.

C.3 SOTOPIA-EVAL Dimensions

SOTOPIA-EVAL evaluates each interaction using seven dimensions. We follow the score ranges and definitions from the original SOTOPIA benchmark:

- **Goal Completion (GOAL),** $[0, 10]$**.** Measures the extent to which the agent achieves its private social goal, where a higher score indicates greater goal completion.
- **Believability (BEL),** $[0, 10]$**.** Measures whether the agent behaves naturally and consistently with its assigned character profile and social role.
- **Knowledge (KNO),** $[0, 10]$**.** Measures how effectively the agent acquires relevant new information or knowledge during the interaction.
- **Relationship (REL),** $[-5, 5]$**.** Measures how the interaction changes the relationship between the participants; positive values indicate improvement and negative values indicate deterioration.
- **Financial and Material Benefits (FIN),** $[-5, 5]$**.** Measures gains or losses in short- or long-term financial and material benefits resulting from the interaction.
- **Social Rules (SOC),** $[-10, 0]$**.** Measures violations of social norms, moral expectations, or legal rules. A score of 0 indicates no identified violation, while more negative values indicate more severe violations.
- **Secret (SEC),** $[-10, 0]$**.** Measures whether the agent successfully preserves private information, secrets, or hidden intentions. A score of 0 indicates that no relevant secret is revealed, while negative values reflect increasingly serious information leakage.

Following SOTOPIA-EVAL, we compute the aggregate social-intelligence score as the arithmetic mean of all seven dimensions:

$$\text{OVERALL} = \frac{\text{GOAL} + \text{BEL} + \text{KNO} + \text{REL} + \text{FIN} + \text{SOC} + \text{SEC}}{7}. \tag{23}$$

Given the official ranges of the seven dimensions, the theoretical range of OVERALL is

$$\text{OVERALL} \in \left[-\frac{30}{7}, \frac{40}{7}\right] \approx [-4.29, 5.71]. \tag{24}$$

In the main experiments, we primarily report **GOAL** and **OVERALL**. GOAL directly reflects whether the agent successfully completes its private social objective, while OVERALL summarizes broader social quality across goal achievement, realism, information acquisition, relationship effects, material consequences, social-rule compliance, and privacy preservation. The remaining six individual dimensions are retained during evaluation and can be reported separately in the supplementary results.

D Baseline Methods and Comparison Scope

We provide brief descriptions of the learning-based baselines used in our experiments. We group them according to their primary design motivation: social-agent learning, general reinforcement learning, and task-specific long-horizon agent optimization.

D.1 Methods Compared on SOTOPIA

DSI. DSI (Zhang et al., 2025) is introduced in SOTOPIA-Ω and improves social agents through *Dynamic Strategy Injection*. It dynamically incorporates diverse social strategies into expert agents to construct high-quality social-interaction demonstrations, which are subsequently used to train smaller language agents.

EPO. EPO (Liu et al., 2025a) introduces an explicit strategy model that generates open-ended strategies for social interaction. The strategy model is optimized through multi-turn reinforcement learning with process rewards and iterative self-play, and its generated strategy is used to guide the downstream acting agent.

SOTOPIA-RL. SOTOPIA-RL (Yu et al., 2025) focuses on reward design for social-agent reinforcement learning. It decomposes coarse episode-level social feedback into utterance-level credit and combines multiple social reward dimensions, providing denser supervision for individual dialogue actions.

AMPO. AMPO (Wang et al., 2026a) is the reinforcement-learning component of the Adaptive Social Learning framework. It defines multiple reasoning modes ranging from intuitive responses to deeper deliberation and combines mode-level and sample-level advantages to learn context-dependent reasoning-mode selection.

SAVOIR. SAVOIR (Feng et al., 2026) addresses utterance-level credit assignment in multi-turn social interaction using Shapley-value-based reward attribution. It combines prospective expected utility with Shapley values to estimate how individual utterances contribute to future social outcomes.

iStar. iStar (Liu et al., 2025b) introduces implicit step rewards for agentic reinforcement learning. It alternately optimizes an implicit process reward model and the policy, learns step-level rewards through a trajectory-based preference objective, and combines the resulting step-level advantage with the episode-level advantage for policy optimization.

D.2 General RL Baselines

PPO. PPO (Schulman et al., 2017) is a widely used policy-gradient algorithm that stabilizes policy optimization through a clipped surrogate objective. The clipping mechanism limits excessively large policy updates while allowing multiple optimization steps over sampled interaction data.

GRPO. GRPO (Shao et al., 2024) replaces an explicit critic with group-relative reward normalization. For multiple responses sampled for the same input, GRPO standardizes their rewards within the group and uses the resulting relative advantages for clipped policy optimization.

RLOO. RLOO (Ahmadian et al., 2024) is a REINFORCE-style optimization method that uses a leave-one-out reward baseline. For each sampled response, the average reward of the remaining responses provides the baseline, reducing variance without requiring a learned value model.

REINFORCE++. REINFORCE++ (Hu, 2025) extends the classical REINFORCE algorithm with optimization techniques motivated by PPO while remaining critic-free. Its design aims to improve optimization stability and computational efficiency without introducing a separate value network.

D.3 Task-Oriented Agent Baselines

GiGPO. GiGPO (Feng et al., 2025) extends group-relative optimization to multi-step language agents through a two-level credit-assignment mechanism. It computes episode-level relative advantages from complete trajectories and constructs step-level groups from repeated anchor states, enabling finer action-level credit without a learned critic.

StraTA. StraTA (Xue et al., 2026) introduces Strategic Trajectory Abstraction for long-horizon agentic RL. It first generates a compact trajectory-level strategy from the initial task state and then conditions subsequent actions on this strategy, jointly optimizing high-level strategy generation and low-level action execution.

BEACON. BEACON (Wang et al., 2026b) improves long-horizon credit assignment by decomposing trajectories according to task milestones. It applies reward shaping within milestone-bounded segments and uses multi-scale advantage estimation so that distant terminal failures do not incorrectly penalize useful intermediate progress.

SCPO. SCPO (Xu et al., 2026) addresses semantic inconsistency in step-level credit assignment. It compares steps from failed trajectories with semantically related steps from successful trajectories in the same rollout group and recovers positive credit for intermediate progress that would otherwise be discarded by trajectory-level rewards.

E Prompts and Output Formats

We use a shared set of prompt templates across SFT and RL. SFT uses the corresponding reference output as the training target, whereas RL samples the model output online. The Mental World Model and Policy Model use fixed output formats to facilitate parsing and format-based feedback.

E.1 Mental World Model Prompt

Prompt template for the Mental World Model

You are the Mental World Model of an interactive agent.
Your task is to produce a concise BDI representation for the next decision.
Task / Scenario:
{task_context}
Agent Goal:
{agent_goal}
Completed Interaction History:
{history_before_current_observation}
Current Observation:
{current_observation}
Reason about the current interaction state and infer the beliefs, desires, and intentions that are relevant to the next decision.
For social interaction, also infer the corresponding mental states of the opponent when they are supported by the dialogue. Do not introduce information that is not grounded in the current context.
Return exactly one complete <think> section and one <answer> section. The think section is a brief evidence summary with relevant turn IDs. Do not put tags inside fields or text outside the two sections.
<think>
Brief evidence summary.
</think>

<answer>
I believe ...
I want ...
I intend to ...
The opponent believes ...
The opponent wants ...
The opponent intends to ...
</answer>
For non-social environments, omit the three opponent-related fields.

Figure 4: Prompt template and output format used by the Mental World Model during SFT and RL.

E.2 Policy Model Prompt

Prompt template for the Policy Model

You are an autonomous agent operating in an interactive environment.
Task / Scenario:
{task_context}
Agent Goal:
{agent_goal}
Completed Interaction History:
{history_before_current_observation}
Current Observation:
{current_observation}
Current Mental State:
{mwm_answer_payload}
Available Actions:
{available_actions}
Based on the task context, interaction history, and current mental state, select the action that best advances the task.
Return exactly one complete <think> section and one <answer> section. Do not place tags inside the answer or text outside the two sections.
<think>
Briefly reason about the next decision.
</think>
<answer>
{next_action}
</answer>
Only the answer payload is executed.
SOTOPIA: one utterance under the documented task interface.
ALFWorld: one executable environment action.
WebShop: one valid search[...] or click[...] action.

Figure 5: Prompt template and output format used by the Policy Model during SFT and RL.

E.3 GPT-4o Supervision for the Mental-State Scorer

For the Mental World Model scorer, we construct supervision from existing training-step trajectories. For each step, we form matched with-BDI and without-BDI decision conditions while keeping the task context and trajectory state fixed. GPT-4o evaluates both conditions on a continuous 0–10 usefulness scale.

GPT-4o annotation prompt for the Mental World Model scorer

You are evaluating whether an explicit mental state improves an agent's decision at a given interaction step.

Task / Scenario:
{task_context}
Agent Goal:
{agent_goal}
Completed Interaction History:
{history_before_current_observation}
Current Observation:
{current_observation}
Predicted Mental State:
{mental_state}
Decision WITH Mental State:
{action_with}
Decision WITHOUT Mental State:
{action_without}
Observed Subsequent State:
{state_after}
Evaluate how useful each decision is for advancing the task while remaining consistent with the current context.
Assign each decision a score from 0 to 10, where:
0 means completely unhelpful or clearly harmful, and
10 means highly useful and strongly supportive of the next decision.
Focus on the marginal benefit provided by the explicit mental state, rather than linguistic style.
Return only:
{"with": score_with, "without": score_without}

E.4 GPT-4o Supervision for the Policy Progress Scorer

The Policy Model scorer is trained directly from existing step-level state transitions. GPT-4o evaluates the progress caused by the current action rather than the absolute quality of the resulting state.

GPT-4o annotation prompt for the Policy Model scorer

You are evaluating the effect of one agent action on task progress.
Task / Scenario:
{task_context}
Agent Goal:
{agent_goal}
Completed Interaction History:
{history_before_current_observation}
Current Observation:
{current_observation}
Current Mental State:
{mwm_answer_payload}
Action:
{action}
State After Action:
{state_after}
Evaluate only the progress caused by this action.
Assign exactly one of the following labels:
-1: significant regression;
-0.5: slight regression;
0: no meaningful change;
0.5: effective progress;
1: substantial progress.
Do not judge the absolute quality of the state. Judge only how much the current action changes progress toward the specified task goal.
Return only:
{"progress": label}

E.5 Output Format Summary

The required generation formats are summarized in Table 6.

Table 6: Output formats used by the two MindWorld components.

Component	Environment	Required output
Mental World Model	SOTOPIA	`<think>` + Self/Opponent BDI in `<answer>`
Mental World Model	ALFWorld/WebShop	`<think>` + Self BDI in `<answer>`
Policy Model	SOTOPIA	`<think>` + utterance in `<answer>`
Policy Model	ALFWorld	`<think>` + executable action in `<answer>`
Policy Model	WebShop	`<think>` + web action in `<answer>`

F Additional Experimental Results

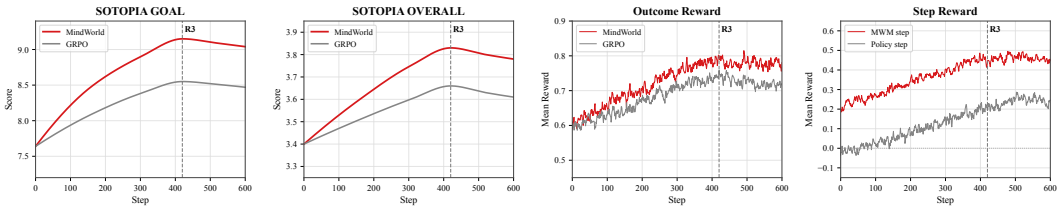
To further evaluate the effectiveness and generalizability of MindWorld across different model scales, we conduct additional experiments with three smaller backbone models: Qwen3-4B, Qwen3-1.7B, and Llama3.2-3B-Instruct. We evaluate these models on both social interaction and task-oriented agent benchmarks. Specifically, SOTOPIA and SOTOPIA-Hard are used to evaluate social interaction capabilities under both Self-Play and GPT-4o-as-Partner settings, while ALFWorld and WebShop are used to evaluate performance in embodied and web-based agent tasks. The social interaction results for Qwen3-4B, Qwen3-1.7B, and Llama3.2-3B-Instruct are reported in Tables 7, 9, and 11, respectively, while their task-oriented agent results on ALFWorld and WebShop are presented in Tables 8, 10, and 12, respectively. We additionally report the corresponding training dynamics in Figures 6, 7, and 8 to illustrate the optimization process of MindWorld across different backbone scales.

Table 7: Results on SOTOPIA and SOTOPIA-Hard with Qwen3-4B. Bold indicates column maxima.

Method	Self-Play				GPT-4o-as-Partner			
	SOTOPIA		SOTOPIA-Hard		SOTOPIA		SOTOPIA-Hard	
	GOAL↑	OVERALL↑	GOAL↑	OVERALL↑	GOAL↑	OVERALL↑	GOAL↑	OVERALL↑
Qwen3-4B	7.64	3.40	6.25	2.97	6.93	3.11	5.89	2.82
w/ DSI	8.24	3.64	7.15	3.37	8.07	3.63	6.82	3.30
w/ EPO	8.10	3.61	6.84	3.20	8.22	3.70	6.86	3.41
w/ SOTOPIA-RL	8.16	3.68	7.57	3.54	8.27	3.71	6.88	3.23
w/ AMPO	8.99	3.80	7.86	3.61	8.61	3.82	7.46	3.61
w/ SAVOIR	8.49	3.78	7.81	**3.80**	8.43	**3.83**	7.10	3.46
w/ iStar	8.61	3.72	7.49	3.49	8.46	3.79	6.98	3.41
w/ GRPO	8.55	3.66	7.22	3.27	8.17	3.62	6.90	3.24
w/ MindWorld (Ours)	**9.15**	**3.83**	**8.02**	3.76	**8.70**	3.81	**7.64**	**3.69**

Table 8: Results on ALFWorld and WebShop with Qwen3-4B. Task-specific rows report the corresponding method combined with MindWorld, with absolute gains over the original method shown in parentheses. All metrics use a 0–100 scale. Bold highlights standalone MindWorld.

Method	ALFWorld		WebShop	
	Seen↑	Unseen↑	Success↑	Score↑
General RL Methods				
Base Model	20.0	18.1	14.2	34.4
+ PPO	76.5	75.2	75.6	84.9
+ RLOO	77.4	74.9	74.8	85.3
+ GRPO	79.8	78.6	78.3	88.0
+ REINFORCE++	80.6	79.1	77.9	88.4
+ MindWorld (Ours)	**86.7**	**85.1**	**82.2**	**90.5**
Task-Specific Agent Methods (+MindWorld)				
+ GiGPO	91.4 (+1.8)	90.7 (+2.0)	75.9 (+2.7)	85.5 (+1.3)
+ StraTA	93.8 (+1.1)	93.1 (+1.5)	85.5 (+1.7)	91.4 (+0.8)
+ BEACON	94.4 (+1.0)	93.8 (+1.7)	80.8 (+1.8)	88.7 (+0.9)
+ SCPO	94.9 (+0.8)	94.3 (+1.5)	79.6 (+2.0)	88.6 (+1.1)



Results on Qwen3-4B

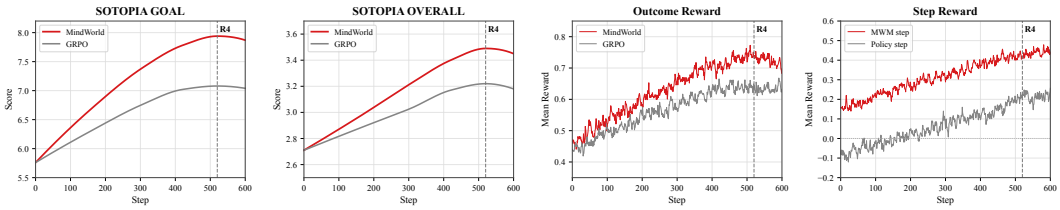
Figure 6: Training dynamics of Qwen3-4B with MindWorld under Self-Play. The curves report the evolution of the outcome score, MWM usefulness, and policy progress across training rounds. The reported checkpoint corresponds to the MindWorld GOAL and OVERALL results in Table 7. Faint traces denote raw batch means.

Table 9: Results on SOTOPIA and SOTOPIA-Hard with Qwen3-1.7B. Bold indicates column maxima.

Method	Self-Play				GPT-4o-as-Partner			
	SOTOPIA		SOTOPIA-Hard		SOTOPIA		SOTOPIA-Hard	
	GOAL↑	OVERALL↑	GOAL↑	OVERALL↑	GOAL↑	OVERALL↑	GOAL↑	OVERALL↑
Qwen3-1.7B	5.76	2.71	3.91	1.86	5.96	2.88	4.80	2.33
w/ DSI	6.63	3.02	4.96	2.37	6.42	3.08	5.18	2.54
w/ EPO	6.42	2.93	4.68	2.21	6.79	3.23	5.42	2.65
w/ SOTOPIA-RL	6.85	3.15	5.72	2.88	6.91	3.24	5.54	2.71
w/ AMPO	7.66	3.40	6.37	3.12	7.45	3.42	6.11	2.98
w/ SAVOIR	7.25	3.39	6.29	3.23	7.19	**3.45**	5.98	3.06
w/ iStar	7.43	3.29	6.42	3.09	7.28	3.35	6.02	2.95
w/ GRPO	7.08	3.22	5.84	2.82	6.85	3.22	5.66	2.72
w/ MindWorld (Ours)	**7.94**	**3.49**	**6.69**	**3.28**	**7.52**	3.41	**6.24**	**3.11**

Table 10: Results on ALFWorld and WebShop with Qwen3-1.7B. Task-specific rows report the corresponding method combined with MindWorld, with absolute gains over the original method shown in parentheses. All metrics use a 0–100 scale. Bold highlights standalone MindWorld.

Method	ALFWorld		WebShop	
	Seen↑	Unseen↑	Success↑	Score↑
General RL Methods				
Base Model	10.6	9.1	6.1	18.2
+ PPO	59.6	55.8	55.2	71.9
+ RLOO	62.9	59.7	53.8	72.4
+ GRPO	66.0	62.5	61.5	77.2
+ REINFORCE++	65.7	63.2	62.7	76.9
+ MindWorld (Ours)	**75.4**	**71.3**	**68.6**	**82.5**
Task-Specific Agent Methods (+MindWorld)				
+ GiGPO	79.6 (+2.3)	76.0 (+2.8)	61.7 (+2.9)	78.1 (+1.8)
+ StraTA	83.0 (+1.6)	80.2 (+2.4)	75.5 (+2.1)	86.2 (+1.2)
+ BEACON	85.8 (+1.4)	82.1 (+2.6)	67.5 (+2.5)	82.0 (+1.3)
+ SCPO	86.5 (+1.2)	83.0 (+2.1)	68.2 (+1.9)	82.8 (+1.1)



Results on Qwen3-1.7B

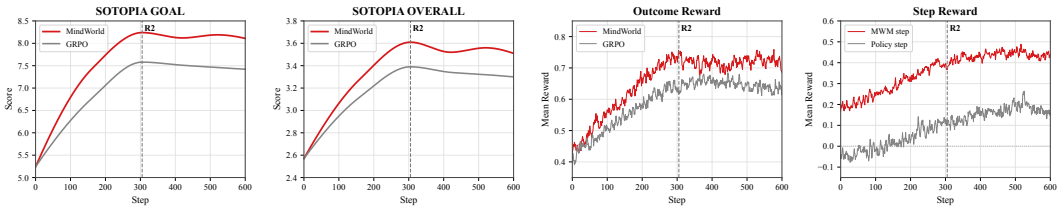
Figure 7: Training dynamics of Qwen3-1.7B with MindWorld under Self-Play. The curves report the evolution of the outcome score, MWM usefulness, and policy progress across training rounds. The reported checkpoint corresponds to the MindWorld GOAL and OVERALL results in Table 9. Faint traces denote raw batch means.

Table 11: Results on SOTOPIA and SOTOPIA-Hard with Llama3.2-3B-Instruct. Bold indicates column maxima.

Method	Self-Play				GPT-4o-as-Partner			
	SOTOPIA		SOTOPIA-Hard		SOTOPIA		SOTOPIA-Hard	
	GOAL↑	OVERALL↑	GOAL↑	OVERALL↑	GOAL↑	OVERALL↑	GOAL↑	OVERALL↑
Llama3.2-3B-Instruct	5.24	2.57	3.70	1.69	6.04	2.99	4.91	2.42
w/ DSI	6.82	3.13	5.41	2.65	6.91	3.19	5.63	2.79
w/ EPO	6.96	3.31	5.60	2.62	7.26	3.38	5.94	2.91
w/ SOTOPIA-RL	7.24	3.28	6.10	3.08	7.33	3.39	5.78	2.82
w/ AMPO	8.03	3.51	6.59	3.14	7.69	3.51	6.46	3.18
w/ SAVOIR	7.72	3.48	6.47	**3.34**	7.44	**3.53**	6.25	3.15
w/ iStar	8.10	3.43	6.64	3.19	7.51	3.42	6.19	3.09
w/ GRPO	7.58	3.39	5.97	2.90	7.12	3.32	5.87	2.94
w/ MindWorld (Ours)	**8.24**	**3.61**	**6.83**	3.27	**7.81**	3.50	**6.59**	**3.29**

Table 12: Results on ALFWorld and WebShop with Llama3.2-3B-Instruct. Task-specific rows report the corresponding method combined with MindWorld, with absolute gains over the original method shown in parentheses. All metrics use a 0–100 scale. Bold highlights standalone MindWorld.

Method	ALFWorld		WebShop	
	Seen↑	Unseen↑	Success↑	Score↑
General RL Methods				
Base Model	13.8	12.0	8.9	29.1
+ PPO	66.9	64.2	63.3	77.9
+ RLOO	68.5	66.1	61.9	78.3
+ GRPO	72.0	68.9	66.3	81.2
+ REINFORCE++	71.6	69.6	67.0	81.9
+ MindWorld (Ours)	**80.2**	**76.8**	**73.7**	**85.6**
Task-Specific Agent Methods (+MindWorld)				
+ GiGPO	84.0 (+2.0)	81.9 (+2.6)	67.3 (+2.4)	81.5 (+1.5)
+ StraTA	87.4 (+1.4)	85.3 (+1.8)	77.5 (+1.5)	87.6 (+0.9)
+ BEACON	89.7 (+1.0)	87.2 (+2.1)	72.7 (+1.9)	84.2 (+1.1)
+ SCPO	90.4 (+0.9)	87.8 (+1.7)	71.0 (+2.2)	84.3 (+0.9)



Results on Llama-3.2-3B-Instruct

Figure 8: Training dynamics of Llama3.2-3B-Instruct with MindWorld under Self-Play. The curves report the evolution of the outcome score, MWM usefulness, and policy progress across training rounds. The reported checkpoint corresponds to the MindWorld GOAL and OVERALL results in Table 11. Faint traces denote raw batch means.

G CASE STUDY

We present three extended interaction traces to illustrate how the mental-state representation is revised as new observations arrive and how the Policy Model conditions on the resulting mental state. For every case, both components are instantiated with Qwen3-8B. The benchmark task specification is retained, while each round displays the current observation, the structured mental state, and the selected action. The SOTOPIA case is shown in Figure 9, the ALFWorld case in Figure 10, and the WebShop case in Figure 11.

G.1 SOTOPIA

Case Study of Qwen3-8B MindWorld on SOTOPIA

Benchmark source. The scenario is based on the camping example presented in the original SOTOPIA paper (Zhou et al., 2024a).

Shared scenario. William and Mia are friends camping in the wilderness. The temperature drops sharply at night, and they have only one blanket available.

Relationship. Friends.

Agent 1 profile: William Brown.

- Occupation: chef;
- Pronouns: he/him;
- Personality: open to experience, conscientious, and extraverted;
- Decision style: strategic;
- Public information: enjoys exploring food and trying new recipes.

William's private goal. Retain exclusive use of the only blanket.

Agent 2 profile: Mia Davis.

- Occupation: high-school principal;
- Pronouns: she/her;
- Personality: extraverted and somewhat neurotic;
- Decision style: decisive;
- Public information: owns two cats;
- Private information: played in a rebellious punk-rock band when she was younger.

Mia's private goal. Persuade William to share the blanket.

Evaluated agent. Mia is controlled by the Qwen3-8B Mental World Model and Qwen3-8B Policy Model; William acts as the interaction partner.

Round 1
Observation / Partner: The temperature has dropped further. William pulls the blanket tightly around his sleeping bag and says, "I only brought one, and I really need it tonight."
Mental World Model (Qwen3-8B):
<think> William's refusal may be driven by both fear of losing warmth and discomfort with physical closeness, so I should identify which constraint matters most before proposing a compromise. </think>
<answer> I believe William views the blanket as his only reliable protection from the cold; I want to obtain shared warmth without dismissing his need for comfort; I intend to ask which aspect of sharing concerns him; The opponent believes sharing may leave him colder or require unwanted closeness; The opponent wants warmth and personal space; The opponent intends to reject an unrestricted request to share. </answer>
Policy Model (Qwen3-8B):
<think> A direct demand is likely to strengthen his resistance, so I should first acknowledge his concern and clarify it. </think>
<answer> I understand why you do not want to lose warmth; are you more worried about having less blanket coverage, or about our sleeping bags being too close together? </answer>

Round 2
Observation / Partner: William replies, "Both, honestly. If we split it, neither of us may be warm, and I do not want to sleep shoulder-to-shoulder."
Mental World Model (Qwen3-8B):

<think> William has stated two concrete constraints, which makes a distance-preserving arrangement more promising than asking him to divide the blanket equally. </think>
<answer> I believe William needs full upper-body coverage and physical separation; I want both of us to benefit from the blanket while respecting those constraints; I intend to propose an arrangement across separated sleeping bags; The opponent believes conventional sharing means reduced coverage and excessive proximity; The opponent wants control over his side of the blanket; The opponent intends to consider only a clearly specified alternative. </answer>
Policy Model (Qwen3-8B):
<think> I should propose a concrete spatial arrangement rather than repeat the general request. </think>
<answer> We could keep our sleeping bags about two feet apart and lay the blanket across only the upper halves, so you keep most of it over your shoulders without us lying close together. </answer>

Round 3
Observation / Partner: William looks at the gap and says, “Cold air would still come through the middle, and the blanket might slide off.”
Mental World Model (Qwen3-8B):
<think> His objection is practical rather than purely oppositional, so solving the draft and stability problems may move the negotiation forward. </think>
<answer> I believe William is testing whether the proposal can work in practice; I want to remove the remaining sources of heat loss; I intend to use available jackets and backpacks to seal and anchor the shared cover; The opponent believes the central gap and movement will make sharing ineffective; The opponent wants evidence that the setup will remain stable; The opponent intends to reject it unless the draft and slipping problems are addressed. </answer>
Policy Model (Qwen3-8B):
<think> A revised physical plan should directly answer both objections. </think>
<answer> We can roll our spare jackets along the inner edges to block the draft and place our backpacks on the outer corners so the blanket cannot slide away. </answer>

Round 4
Observation / Partner: William says, “That sounds possible, but I do not want to commit to it for the whole night if it makes things worse.”
Mental World Model (Qwen3-8B):
<think> William is willing to experiment but wants a reversible decision, so a short trial lowers the perceived risk. </think>
<answer> I believe William’s resistance has shifted from rejection to risk management; I want him to test the arrangement without feeling locked in; I intend to propose a brief trial with an explicit option to stop; The opponent believes the arrangement may work but remains uncertain; The opponent wants an easy way to reverse the decision; The opponent intends to cooperate only if the trial preserves his control. </answer>
Policy Model (Qwen3-8B):
<think> A time-limited trial creates a low-cost path to agreement. </think>
<answer> Let’s test it for five minutes; if either of us feels colder or uncomfortable, we will immediately return the blanket to you. </answer>

Round 5
Observation / Partner: William nods and says, “Five minutes is fine. Show me where you want the sleeping bags.”
Mental World Model (Qwen3-8B):
<think> He has accepted the trial, so I should give simple cooperative instructions and avoid taking over his side. </think>
<answer> I believe William is conditionally cooperating and expects the agreed boundaries to be honored; I want to set up the test efficiently; I intend to keep the bags separated and let William position his own side; The opponent believes the test is temporary and reversible; The opponent wants control over his personal space; The opponent intends to participate while monitoring comfort. </answer>
Policy Model (Qwen3-8B):
<think> The next response should turn the verbal agreement into a joint physical action. </think>
<answer> Let’s place the bags parallel with this gap between them; you can hold your side in place while I roll the jackets along the center. </answer>

Round 6
Observation / Partner: They arrange the sleeping bags and drape the blanket across them. William says, “The draft is better, but my right shoulder is not fully covered.”

Mental World Model (Qwen3-8B):
<think> The trial is partly successful, and a small asymmetric adjustment may satisfy William without removing Mia's access. </think>
<answer> I believe the shared configuration is working but distributes the blanket unevenly around William's shoulder; I want to preserve the arrangement while correcting that local problem; I intend to rotate the blanket diagonally toward William's upper body; The opponent believes the idea is viable but needs more coverage on his side; The opponent wants his shoulders fully protected; The opponent intends to continue if the adjustment does not reduce his warmth. </answer>
Policy Model (Qwen3-8B):
<think> A small adjustment is preferable to renegotiating the entire setup. </think>
<answer> I'll shift the top corner toward your shoulder and keep only the lower edge over my side, so your upper body remains fully covered. </answer>

Round 7
Observation / Partner: After the adjustment, William says, "My shoulders are warm now, but the bottom of my sleeping bag still feels cold."
Mental World Model (Qwen3-8B):
<think> The remaining problem concerns his feet rather than blanket sharing, so other insulation can solve it without reclaiming the shared cover. </think>
<answer> I believe William now accepts the shared blanket but still has heat loss near his feet; I want to protect the agreement by solving that problem with another resource; I intend to use spare clothing as foot insulation; The opponent believes the blanket arrangement is mostly effective; The opponent wants warmer feet; The opponent intends to keep sharing if that final comfort issue is addressed. </answer>
Policy Model (Qwen3-8B):
<think> I should use the available clothes instead of moving the blanket away from its successful position. </think>
<answer> Let's put your spare sweater and my rain jacket around the foot of your sleeping bag; that should add insulation without changing the blanket coverage. </answer>

Round 8
Observation / Partner: William tries the additional insulation and says, "That is much better, although the wind is still hitting us from the left."
Mental World Model (Qwen3-8B):
<think> Both agents are warmer, and rotating the setup toward the tent wall can reduce wind exposure without changing their agreed distance. </think>
<answer> I believe the shared arrangement is now acceptable and the remaining cold comes from wind direction; I want to improve warmth without reopening the interpersonal negotiation; I intend to rotate both bags toward the windbreak together; The opponent believes sharing is no longer the main problem; The opponent wants protection from the side wind; The opponent intends to preserve the current arrangement if the wind can be blocked. </answer>
Policy Model (Qwen3-8B):
<think> The policy should frame the next move as a cooperative environmental adjustment. </think>
<answer> We can rotate both sleeping bags slightly toward the tent wall so it blocks the left-side wind while keeping the same distance and blanket layout. </answer>

Round 9
Observation / Partner: They rotate the setup. William says, "This is warmer than I expected, and the space between us is still comfortable."
Mental World Model (Qwen3-8B):
<think> William explicitly confirms that both of his original constraints have been satisfied, so I should reinforce the agreement without asking for more. </think>
<answer> I believe William now sees that sharing can preserve both warmth and personal space; I want to stabilize the cooperative outcome; I intend to confirm the arrangement and respect his boundary; The opponent believes the compromise works better than anticipated; The opponent wants the arrangement to remain stable; The opponent intends to continue sharing under the agreed conditions. </answer>
Policy Model (Qwen3-8B):
<think> A concise confirmation will preserve trust and avoid appearing opportunistic. </think>
<answer> Great, let's keep it exactly like this, and if either of us gets uncomfortable we can adjust it without changing the space between us. </answer>

Round 10
Observation / Partner: Several minutes later, William says, "I am comfortable with this; we can share it this way for the rest of the night."

Mental World Model (Qwen3-8B):
<think> Mia's explicit social goal has been achieved, and the interaction can close with appreciation rather than further negotiation. </think>
<answer> I believe William has voluntarily accepted sustained blanket sharing; I want to end the negotiation on a positive and reciprocal note; I intend to thank him and maintain the arrangement; The opponent believes the shared solution protects his warmth and personal space; The opponent wants a calm end to the discussion; The opponent intends to keep sharing the blanket. </answer>
Policy Model (Qwen3-8B):
<think> The final response should acknowledge cooperation and avoid adding new demands. </think>
<answer> Thank you for working it out with me; this keeps both of us warmer, and I appreciate that we found a setup that respects your space. </answer>

Final outcome. William agrees to share the blanket for the rest of the night. The arrangement preserves physical distance, keeps both agents warm, and avoids worsening their friendship. Mia achieves her private goal through constraint-aware negotiation rather than repeated pressure.

Figure 9: Case study of Qwen3-8B MindWorld on the SOTOPIA camping scenario.

G.2 ALFWorld

Case Study of Qwen3-8B MindWorld on ALFWorld

Benchmark source. The task follows the official ALFWorld text-environment example (Shridhar et al., 2021).

Task instruction. Put a hot potato in the sink basin.

Initial environment. The agent begins in a kitchen containing a sink basin, a microwave, a refrigerator, a garbage can, a coffee machine, a toaster, two countertops, several cabinets, and four stove burners. The target potato is not visible from the initial position.

Models. Both the Mental World Model and Policy Model are Qwen3-8B. Only the acting agent's belief, desire, and intention are modeled, because ALFWorld has no social counterpart.

Round 1
Observation: The agent is in the center of the kitchen; the current view lists the major receptacles but does not reveal the potato.
Mental World Model (Qwen3-8B):
<think> The task requires locating a potato, heating it, and finally placing it in the sink basin, so the first step is systematic object search. </think>
<answer> I believe the potato is hidden on or inside one of the kitchen receptacles; I want to find a potato and preserve the required heat-then-place order; I intend to inspect the visible countertops before opening more costly containers. </answer>
Policy Model (Qwen3-8B):
<think> Countertops are plausible food locations and can be checked without opening a receptacle. </think>
<answer> go to countertop 1 </answer>

Round 2
Observation: At countertop 1, the environment shows bread, an egg, utensils, a plate, a pan, and other kitchen objects, but no potato.
Mental World Model (Qwen3-8B):
<think> The first candidate location is exhausted, and countertop 2 is the next low-cost food location. </think>
<answer> I believe countertop 1 does not contain the target object; I want to continue the search without repeating actions; I intend to inspect countertop 2 next. </answer>
Policy Model (Qwen3-8B):
<think> Moving directly to the second countertop is more efficient than opening unrelated cabinets. </think>
<answer> go to countertop 2 </answer>

Round 3
Observation: Countertop 2 contains several objects, including potato 1 and potato 2.
Mental World Model (Qwen3-8B):

<think> The target object has been found, and either potato can satisfy the task if it is heated and placed correctly. </think>
<answer> I believe potato 1 is currently accessible on countertop 2; I want to acquire one valid potato; I intend to take potato 1 and then move toward the microwave. </answer>
Policy Model (Qwen3-8B):
<think> Taking the first accessible potato avoids unnecessary comparison between equivalent objects. </think>
<answer> take potato 1 from countertop 2 </answer>

Round 4
Observation: The environment confirms that potato 1 has been picked up.
Mental World Model (Qwen3-8B):
<think> Before leaving the location, verifying the held object reduces the risk of planning from an incorrect inventory state. </think>
<answer> I believe potato 1 should now be in my inventory; I want to confirm that the pickup succeeded; I intend to inspect the inventory before beginning the heating subtask. </answer>
Policy Model (Qwen3-8B):
<think> A single inventory check provides an explicit state confirmation. </think>
<answer> inventory </answer>

Round 5
Observation: The inventory reports that the agent is carrying potato 1.
Mental World Model (Qwen3-8B):
<think> The possession condition is satisfied, so the next required subgoal is heating the potato with the microwave. </think>
<answer> I believe I am holding the correct potato; I want to make it hot before placement; I intend to navigate to microwave 1. </answer>
Policy Model (Qwen3-8B):
<think> The microwave is the environment's designated heating appliance. </think>
<answer> go to microwave 1 </answer>

Round 6
Observation: The agent arrives at microwave 1, which is currently closed.
Mental World Model (Qwen3-8B):
<think> Opening the appliance creates a clear, valid precondition for the heating action and avoids relying on an implicit state transition. </think>
<answer> I believe microwave 1 is available but closed; I want to prepare it for heating potato 1; I intend to open the microwave before issuing the heat action. </answer>
Policy Model (Qwen3-8B):
<think> Opening the microwave is the safest next executable action. </think>
<answer> open microwave 1 </answer>

Round 7
Observation: The environment confirms that microwave 1 is open.
Mental World Model (Qwen3-8B):
<think> The potato is held and the heating appliance is ready, so the heating subgoal can now be completed. </think>
<answer> I believe the preconditions for heating are satisfied; I want potato 1 to acquire the hot state; I intend to heat it using microwave 1. </answer>
Policy Model (Qwen3-8B):
<think> The next action should directly change the potato's temperature rather than move to the sink prematurely. </think>
<answer> heat potato 1 with microwave 1 </answer>

Round 8
Observation: The environment reports that potato 1 has been heated.
Mental World Model (Qwen3-8B):
<think> The temperature requirement is complete, leaving only transport and placement in the sink basin. </think>
<answer> I believe potato 1 is now hot and still held; I want to satisfy the destination condition; I intend to travel to sinkbasin 1 without interacting with unrelated objects. </answer>
Policy Model (Qwen3-8B):
<think> The shortest useful continuation is direct navigation to the target receptacle. </think>
<answer> go to sinkbasin 1 </answer>

Round 9

Observation: At sinkbasin 1, the environment shows a knife, an egg, and a dish sponge; the basin itself is available as a receptacle.
Mental World Model (Qwen3-8B):
<think> The destination is reached, but an inventory confirmation before the terminal action makes the final state transition explicit. </think>
<answer> I believe sinkbasin 1 is the correct destination and potato 1 should still be in my inventory; I want to avoid a failed placement caused by losing the object; I intend to verify the inventory once more. </answer>
Policy Model (Qwen3-8B):
<think> A final inventory check is inexpensive and confirms the remaining precondition. </think>
<answer> inventory </answer>

Round 10
Observation: The inventory confirms that the agent is carrying the hot potato 1 while standing at sinkbasin 1.
Mental World Model (Qwen3-8B):
<think> All task predicates are now satisfied except the final containment relation. </think>
<answer> I believe potato 1 is hot, held, and adjacent to sinkbasin 1; I want to complete the task; I intend to place the potato in the basin now. </answer>
Policy Model (Qwen3-8B):
<think> The terminal action should establish the required object–receptacle relation. </think>
<answer> move potato 1 to sinkbasin 1 </answer>

Environment result. The environment returns won=True. The final state satisfies both requirements: potato 1 is hot and is located in sinkbasin 1.

Observed role of the Mental World Model. The mental state preserves the ordering constraint

$$\text{locate} \rightarrow \text{pick up} \rightarrow \text{heat} \rightarrow \text{place},$$

preventing the Policy Model from placing the potato before completing the heating subgoal.

Figure 10: Case study of Qwen3-8B MindWorld on an ALFWorld household task.

G.3 WEBSHOP

Case Study of Qwen3-8B MindWorld on WebShop

Benchmark source. The instruction is the compositional shopping example shown on the official WebShop project page (Yao et al., 2022).

User instruction. "I'm looking for a quick-release replacement fitness strap band; it should match my chic teal Fitbit, and price lower than $40.00."

Hard constraints.

- product type: replacement fitness strap band;
- mechanism: quick release;
- compatibility: Fitbit;
- color: chic teal;
- price: strictly below $40.00.

Models. The Mental World Model and Policy Model are both Qwen3-8B. The page observations below follow the WebShop search, result, item, and item-detail structure.

Round 1
Observation: The agent is on the initial WebShop search page with an empty query field.
Mental World Model (Qwen3-8B):
<think> The request contains five simultaneous constraints, so the search query should retain the rare attributes "quick release," "Fitbit," and "teal." </think>
<answer> I believe the user needs a replacement Fitbit strap with a quick-release mechanism, chic-teal color, and price below $40; I want to preserve every hard constraint throughout navigation; I intend to begin with a specific attribute-rich query. </answer>
Policy Model (Qwen3-8B):
<think> A generic query such as "fitness band" would return too many irrelevant products. </think>

<answer> search[quick release replacement band teal fitbit] </answer>

Round 2
Observation: The result page lists several candidates, including a generic teal silicone watch band, a quick-release nylon band without explicit Fitbit compatibility, and a Fitbit replacement band whose mechanism is not visible in the title.
Mental World Model (Qwen3-8B):
<think> No title alone proves all constraints, so the first inexpensive teal candidate should be inspected but not purchased immediately. </think>
<answer> I believe the generic teal band satisfies color and price but has uncertain compatibility and release mechanism; I want to test it as a candidate without committing; I intend to open its product page and inspect the details. </answer>
Policy Model (Qwen3-8B):
<think> Opening a plausible candidate is more informative than choosing based only on a result-page title. </think>
<answer> click[Generic Teal Silicone Fitness Band] </answer>

Round 3
Observation: The item page shows a price of $9.99 and a teal color, but describes the product only as a universal 20-mm watch band.
Mental World Model (Qwen3-8B):
<think> The low price and color match are insufficient because Fitbit compatibility and quick-release support remain unverified. </think>
<answer> I believe this product satisfies the budget and color constraints but may fail compatibility and mechanism requirements; I want explicit evidence before rejecting it; I intend to open the description. </answer>
Policy Model (Qwen3-8B):
<think> The description is the appropriate source for technical compatibility information. </think>
<answer> click[Description] </answer>

Round 4
Observation: The description states that the band uses a standard spring-bar connector for generic watches and does not list Fitbit compatibility.
Mental World Model (Qwen3-8B):
<think> Fitbit compatibility remains unverified and may not provide the required quick-release mechanism. </think>
<answer> I believe the current product is not a reliable Fitbit replacement despite its teal color and low price; I want to avoid a partial-match purchase; I intend to return to search and reformulate the query more specifically. </answer>
Policy Model (Qwen3-8B):
<think> Purchasing this item would leave two hard constraints unverified. </think>
<answer> click[Back to Search] </answer>

Round 5
Observation: The agent returns to the search page with the original instruction still visible.
Mental World Model (Qwen3-8B):
<think> The failed candidate shows that “universal” results dominate the initial query, so the refined query should prioritize Fitbit and the exact color phrase. </think>
<answer> I believe the previous query under-specified device compatibility; I want results that explicitly mention both Fitbit and chic teal; I intend to use a narrower reordered query. </answer>
Policy Model (Qwen3-8B):
<think> Placing the device name and exact color earlier should improve result precision. </think>
<answer> search[fitbit chic teal quick release replacement strap] </answer>

Round 6
Observation: The refined result page contains: a chic-teal Fitbit replacement band with quick-release pins priced at $18.99; an aqua Fitbit band without quick-release wording; and a black quick-release Fitbit band.
Mental World Model (Qwen3-8B):
<think> The first candidate is the only result whose visible attributes jointly satisfy product type, mechanism, compatibility, color, and budget. </think>
<answer> I believe the first refined result is the strongest candidate because all five constraints appear in its title or summary; I want to verify the attributes on the item page; I intend to open that candidate. </answer>

Policy Model (Qwen3-8B):
<think> The other candidates either have the wrong color or lack explicit confirmation of a quick-release mechanism. </think>
<answer> click[Quick-Release Replacement Band for Fitbit] </answer>

Round 7
Observation: The item page shows a price of $18.99, a Chic Teal color option, other color options, a description link, an overview link, and a Buy Now button.
Mental World Model (Qwen3-8B):
<think> The price is valid and the required color exists, but compatibility and mechanism should be checked in the detailed description before selecting options. </think>
<answer> I believe the candidate meets the budget and offers chic teal, while the detailed Fitbit compatibility still requires confirmation; I want to verify all semantic constraints before purchase; I intend to read the description. </answer>
Policy Model (Qwen3-8B):
<think> Buying now would be premature because a visible option does not by itself prove product compatibility. </think>
<answer> click[Description] </answer>

Round 8
Observation: The description identifies the item as a replacement band for the listed Fitbit series and states that it uses quick-release pins.
Mental World Model (Qwen3-8B):
<think> The description confirms the two previously uncertain hard constraints, leaving only selection of the requested color. </think>
<answer> I believe the product is a compatible Fitbit replacement with the required quick-release mechanism and a price below $40; I want to complete the remaining color selection; I intend to return to the item page and choose chic teal. </answer>
Policy Model (Qwen3-8B):
<think> The next action should preserve the verified product and navigate back to its selectable options. </think>
<answer> click[< Prev] </answer>

Round 9
Observation: The item page is restored; Chic Teal is available but not yet selected.
Mental World Model (Qwen3-8B):
<think> All non-option constraints have been verified, so the required color can now be chosen safely. </think>
<answer> I believe this is the correct product and only the color option remains unresolved; I want the selected configuration to match the user's Fitbit; I intend to choose chic teal. </answer>
Policy Model (Qwen3-8B):
<think> Selecting the requested option before purchase is necessary for full constraint satisfaction. </think>
<answer> click[Chic Teal] </answer>

Round 10
Observation: The page now marks Chic Teal as selected and continues to show the $18.99 price.
Mental World Model (Qwen3-8B):
<think> The selected option and price are correct, but the overview can provide one final check that no conflicting product property is hidden. </think>
<answer> I believe all visible hard constraints are currently satisfied; I want one final consistency check before committing; I intend to inspect the overview while preserving the selected option. </answer>
Policy Model (Qwen3-8B):
<think> A final verification step reduces the chance of buying a superficially matching but functionally different accessory. </think>
<answer> click[Overview] </answer>

Round 11
Observation: The overview confirms that the item is a replacement fitness strap band, supports Fitbit devices, uses quick-release pins, and remains priced below $40.
Mental World Model (Qwen3-8B):
<think> Every hard constraint is now supported by the page content, and the chosen color remains valid. </think>
<answer> I believe the selected product satisfies type, mechanism, compatibility, color, and price constraints; I want to complete the shopping task without changing the configuration; I intend to return to the item page and purchase it. </answer>

Policy Model (Qwen3-8B):
<think> The item page must be restored before the terminal purchase action. </think>
<answer> click[< Prev] </answer>

Round 12
Observation: The verified product page shows Chic Teal as the selected color, a price of $18.99, and the Buy Now button.
Mental World Model (Qwen3-8B):
<think> The product has passed every required check, so further exploration would add cost without improving the decision. </think>
<answer> I believe the current configuration fully matches the user's request; I want to complete the purchase; I intend to select Buy Now. </answer>
Policy Model (Qwen3-8B):
<think> The terminal action is justified because every hard constraint has been explicitly verified. </think>
<answer> click[Buy Now] </answer>

Final outcome. The episode terminates successfully with a product that is a quick-release replacement Fitbit band, offers the chic-teal option, and costs less than $40.

Observed role of the Mental World Model. The mental representation retains all five constraints across search, inspection, rejection, reformulation, option selection, and purchase. This prevents the Policy Model from buying the first low-cost teal band when compatibility and mechanism remain unsupported.

Figure 11: Case study of Qwen3-8B MindWorld on a WebShop shopping task.