

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 ADAPTIVE SOCIAL LEARNING VIA MODE POLICY OPTIMIZATION FOR LANGUAGE AGENTS

Anonymous authors

Paper under double-blind review

ABSTRACT

Effective social intelligence simulation requires language agents to dynamically adjust reasoning depth, a capability notably absent in current studies. Existing methods either lack explicit reasoning or employ lengthy Chain-of-Thought reasoning uniformly across all scenarios, resulting in excessive token usage and inflexible social behaviors in tasks such as negotiation or collaboration. To address this, we propose an Adaptive Social Learning (ASL) framework in this paper, aiming to improve the adaptive reasoning ability of language agents in dynamic social interactions. To this end, we first identify the hierarchical reasoning modes under such context, ranging from intuitive response to deep deliberation based on the cognitive control theory. We then develop the Adaptive Mode Policy Optimization (AMPO) algorithm to learn the context-aware mode adaptation and reasoning. Our framework advances existing research in three key aspects: (1) Multi-granular reasoning mode design, (2) Context-aware mode switching in rich social interaction, and (3) Token-efficient reasoning with depth adaptation. Extensive experiments on the benchmark social intelligence environment verify that ASL achieves 15.6% higher task performance than GPT-4o. Notably, our AMPO outperforms GRPO by 7.0% with 32.8% shorter thinking chains, demonstrating the advantages of our AMPO and the learned adaptive reasoning ability over GRPO’s solution. Our code and data are available at <https://anonymous.4open.science/r/ICLR2026-3960Submission>.

1 INTRODUCTION

Large Language Models (LLMs) have demonstrated exceptional reasoning capabilities in static domains with well-defined rules and deterministic answers, such as mathematics, code, and logical reasoning (Yang et al., 2024; Liu et al., 2024; Anthropic, 2024; Comanici et al., 2025). However, there exists a notable gap between the reasoning capabilities required in these problems and those in open-ended social interaction, especially in scenarios involving conflicting interests and negotiations driven by agents’ long-term goals. LLM-based agents offer new opportunities for modeling dynamic social contexts by simulating human behaviors (Zhou et al., 2024; Wang et al., 2024) and developing sophisticated reasoning capabilities. Yet, succeeding in such environments not only demands coherent alignment with agents’ long-horizon objectives, but also rapid adaptation to evolving situations—capabilities with which current LLMs still struggle (Zhang et al., 2024; Liu et al., 2025).

Recent research efforts on *social intelligence in language agents* have primarily focused on two pathways: (1) End-to-end goal-oriented training, which involves LLM post-training through supervised learning (Wang et al., 2024; Zhang et al., 2025a), and (2) External planning integration, augmenting agents with plug-and-play planning modules (Deng et al., 2024; Li et al., 2024a; Liu et al., 2025). These methods predominantly focus on the fast-reasoning paradigm, which offers a response without sufficient thinking processes. However, in dynamic and complex social contexts, such responses often fail to capture subtle cues or anticipate long-term costs and benefits (see Figure 5). Evidence from cognitive science indicates that humans often pause for deliberation in such situations (Evans, 1996; Krull & Dill, 1996), suggesting that social agents may also gain from similar reasoning processes. Although Long Chain-of-Thought (Long-CoT) has been proven effective in several domains (Chen et al., 2025), this reasoning paradigm has yet to be introduced to the social intelligence tasks above.

054 Although existing Large Reasoning Models (LRMs), such as OpenAI-o1 (Jaech et al., 2024) and
 055 DeepSeek-R1 (Guo et al., 2025), have demonstrated impressive capabilities with Long-CoT across
 056 various reasoning tasks (DeepMind, 2024; Team, 2024; OpenAI, 2025), most of them exhaust their
 057 reasoning regardless of the input complexity. This style of employing LLMs' reasoning uniformly is
 058 insufficient for handling the dynamics and richness in agent social interaction. For example, not all
 059 social interactions between language agents necessitate deep reasoning (Thorngate, 1976), exhaustive
 060 reasoning may degrade performance as a consequence of overthinking. Therefore, underlying social
 061 intelligence tasks, it is vital to empower LLM-based language agents with the reasoning ability that
 062 adapts to dynamic social environments. This highlights the need for social learning that occurs through
 063 interactive social experiences, to learn from complex social behavior and support the discovery of
 064 social intelligence (Yang et al., 2010; Gweon, 2024).

065 In this paper, we propose the **Adaptive Social Learning** framework (**ASL**) to empower language
 066 agents with the capability for adaptive reasoning, enabling them to effectively respond in accordance
 067 with the dynamics of social interaction context. Specifically, we first develop reasoning modes
 068 inspired by hierarchical cognitive control theory (Koechlin & Summerfield, 2007; Badre, 2008),
 069 covering a spectrum from intuitive response, through shallow and strategic thinking, to deep delib-
 070 eration. Next, we perform the injection of reasoning modes, which consists of behavioral cloning
 071 for cold-start and RL-based adaptive reasoning mode enhancement. For RL-based enhancement,
 072 we contrapuntally develop the **Adaptive Mode Policy Optimization** (**AMPO**) algorithm, which
 073 incorporates the mode-level and sample-level information into advantage estimation to strengthen
 074 the context-aware reasoning mode switching. In terms of reward, we design three types of reward
 075 functions, including answer reward, format reward, and answer length reward, providing feedback for
 076 choosing the appropriate reasoning mode and answer. Finally, experimental results show that ASL
 077 achieves the SOTA performances in comparison with strong baselines.

078 The main contributions are summarized as follows: (1) We propose ASL, the first adaptive social
 079 learning framework for language agents, which consists of hierarchical reasoning modes and tailor-
 080 designed reinforcement learning to empower language agents with adaptive reasoning ability in
 081 rich social context. (2) We develop AMPO algorithm, which integrates mode-level and sample-
 082 level advantage estimation for dynamic mode switching, and improves flexible inference and token
 083 efficiency. (3) Extensive experiments demonstrate the significant improvements of ASL, with the
 084 performance gains up to 15.6% over GPT-4o. Additionally, compared to GRPO, AMPO shows 32.8%
 085 decrease in token utilization on average, accompanied by 7.0% performance gain.

086 2 ADAPTIVE SOCIAL LEARNING FRAMEWORK

087 To empower social language agents with adaptive reasoning in dynamic contexts, we introduce the
 088 ASL framework as shown in Figure 1. First, inspired by Hierarchical Cognitive Control Theory, we
 089 design specialized reasoning modes that structure the social agent's cognitive processes (§2.2). Next,
 090 we employ mode behavioral cloning to train the LLM to accurately follow predefined reasoning
 091 modes(§2.3). Finally, we propose adaptive mode policy optimization, which leverages reinforcement
 092 learning to enhance both adaptive mode selection and reasoning capabilities(§2.4).

093 2.1 TASK FORMULATION

094 The social intelligence task is modeled as a sequential dialogue interaction between two role-playing
 095 agents, \mathcal{P}_1 and \mathcal{P}_2 , each with distinct profiles and private social goals. We formulate this interaction
 096 as a partially observable Markov decision process (POMDP), $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{R})$. The state space
 097 \mathcal{S} represents the social context, and the action space \mathcal{A} consists of open-ended natural language
 098 utterances. At state s_t , the current agent \mathcal{P}_i receives an observation $o_t \in \mathcal{O}$ containing the dialogue
 099 history and its agent-specific private information. Based on this, it samples an action (utterance) a_t^i
 100 from its policy $\pi_i(\cdot | o_t)$. The environment then transitions from state s_t to s_{t+1} via the transition
 101 function \mathcal{T} . An entire episode, forms a trajectory $\tau = \{o_1, a_1, \dots, o_T, a_T\}$, concluding at a terminal
 102 timestep T . At the end of the episode, each agent receives a terminal reward $R_i(s_T)$ from the reward
 103 function \mathcal{R} , which conducts a multi-dimensional evaluation on interaction quality with respect to
 104 each agent's private goals (e.g., building trust or resolving conflict). The objective for each agent is to
 105 learn a policy π_i that maximizes its expected terminal reward: $\mathbb{E}_{\tau \sim \pi_i, \pi_{-i}}[R_i(s_T)]$.

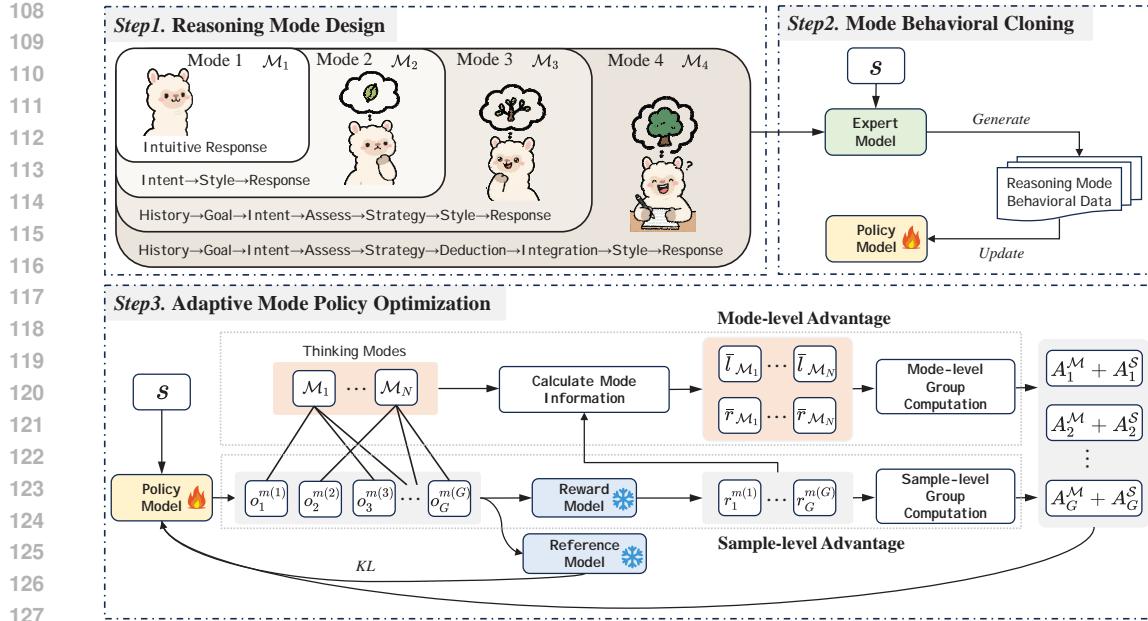


Figure 1: Demonstration of our **Adaptive Social Learning** framework, which consists of three steps: (1) **Reasoning Mode Design** based on Hierarchical Cognitive Control Theory, (2) **Mode Behavioral Cloning**, learning adherence to predefined reasoning modes and (3) **Adaptive Mode Policy Optimization**, introducing both mode- and sample-level advantage estimation during RL.

2.2 REASONING MODES FOR LANGUAGE AGENTS

Existing SOTA LRM s exhibit significant limitations in social interaction (see our experimental results in Table 1). A detailed case study of the model outputs (see Section K.1) reveals several shortcomings in their reasoning trajectories, including unstructured thought processes, poor awareness of social goals, and an inability to controllably switch between different reasoning modes. To better equip LLMs for social scenarios, a more structured and sophisticated reasoning framework is necessary. We draw inspiration from Hierarchical Cognitive Control Theory (HCCT) (Koechlin & Summerfield, 2007; Badre, 2008), which provides a theoretical framework for understanding human cognitive behavior. HCCT posits that cognitive control operates through four distinct hierarchical levels, managing goals and actions at varying degrees of abstraction. Motivated by HCCT, we propose four corresponding levels of reasoning modes tailored for different social scenarios. As illustrated in Figure 7, these modes span from intuitive responses to progressively deeper levels of contemplation. The detailed mapping between our reasoning modes and HCCT’s four hierarchical levels is provided in Section J. For each reasoning mode, we design specific and suitable actions aligned with linguistic principles:

Mode 1 \mathcal{M}_1 (Intuitive Response): \mathcal{M}_1 is the most basic mode, characterized by intuitive responses based on learned associations and basic linguistic modes (Sacks et al., 1974; Norman & Shallice, 1985). It does not contain any reasoning actions, with only the final answer.

Mode 2 \mathcal{M}_2 (Intentional Analysis): \mathcal{M}_2 is the basic interaction mode, focusing on understanding current intent and responding appropriately. \mathcal{M}_2 only requires maintaining basic interaction flow without complex strategic considerations. It encompasses a sequence of reasoning actions: **Intent**, **Style**, and **Response**. **Intent** aims to analyze the other party’s intentions (Grice, 1975). **Style** ensures consistency in the speaking style of the agent (Clark, 1996). **Response** gives the preliminary answer.

Mode 3 \mathcal{M}_3 (Strategic Adaptation): \mathcal{M}_3 is the strategic reasoning mode, requiring speakers to not only understand immediate context but also comprehensively consider historical information, goal, and current situation assessment to formulate corresponding strategies. This enables speakers to better adapt to specific social situations. Compared with \mathcal{M}_2 , \mathcal{M}_3 additionally introduces four reasoning actions: (1) **History** aims to analyze the history for better context understanding (Schiffrin, 1987). (2) **Goal** clarifies the agent’s goal (Grosz & Sidner, 1986). (3) **Assess** analyzes goal alignment,

round criticality, and improvement potential between parties (Brown, 1987). (4) *Strategy* enables the agent to propose a suitable strategy for the present social context (Clark & Brennan, 1991).

Mode 4 \mathcal{M}_4 (Prospective Deduction): \mathcal{M}_4 is an advanced strategic simulation mode, requiring speakers to conceive multiple strategies and evaluate their effects through simulation, thereby making optimal decisions. \mathcal{M}_4 further introduces *Deduction* and *Integration* compared to \mathcal{M}_3 . *Strategy* encourages the proposal of multiple strategies (Clark & Brennan, 1991), then simulating the execution of these strategies through *Deduction* action (Schank & Abelson, 2013). *Integration* action aggregates the results of *Deduction* for the preliminary answer. \mathcal{M}_4 facilitates the simulation of various situations to promote deeper thinking, effectively responding to more complex social contexts (Searle, 1969).

2.3 MODE BEHAVIORAL CLONING

To enhance the model’s capability to adhere to four reasoning modes, we initially employ Behavioral Cloning (Bain & Sammut, 1999; Ross & Bagnell, 2010) to fine-tune the model as the foundation for subsequent reinforcement learning. Given a prompt corpus $\mathcal{X} = \{x_i\}_i^N$, we leverage an expert LLM to generate reasoning responses that are completely consistent with our reasoning modes and actions. To distinguish and control reasoning modes, each response begins with a special control token $c \in \mathcal{C} = \{< MODE_1 >, < MODE_2 >, < MODE_3 >, < MODE_4 >\}$. The LLM is thus to first generate the appropriate mode token and then produce a reasoning path consistent with that mode. The detailed data collection process and training parameters are described in Section G. Given the constructed dataset \mathcal{D}_{bc} , the objective is:

$$\mathcal{L}_{BC} = -\mathbb{E}_{(x,y) \sim \mathcal{D}_{bc}} \left[\sum_{t=1}^{|y|} \log \pi_\theta(y_t | x, y_{1:t-1}) \right] \quad (1)$$

2.4 ADAPTIVE MODE POLICY OPTIMIZATION

GRPO (Shao et al., 2024) has proven highly effective for training of LRM, which obviates the need for critic model and instead uses the average reward as the baseline to calculate the advantage:

$$A_{i,t} = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})}. \quad (2)$$

However, this approach suffers from a critical limitation: it treats each sample independently, overlooking the inherent connections between samples in terms of their reasoning modes. For instance, a “step-by-step reasoning” response and a “direct response” are evaluated by GRPO solely based on their final reward r_i . Without additional mode-level information, this “mode-blindness” prevents the model from perceiving the trade-offs between different modes. Consequently, the model tends to converge towards fixed preferences rather than dynamically adjusting its reasoning modes according to specific scenarios. As shown in Figure 2 and Figure 6, our experiments confirm that GRPO-trained models often resort to overly complex reasoning even for simple tasks.

To address this, we propose the AMPO algorithm. The core idea of AMPO is to incorporate both mode-level and sample-level advantage in its advantage estimation. The mode-level advantage guides the LLM in selecting the appropriate mode for a given scenario, while the sample-level advantage refines the reasoning trajectory generated within that chosen mode. This dual-level optimization enables the model to learn a dynamic and adaptive reasoning policy.

2.4.1 REWARD SHAPING

To provide a clear learning signal, the reward r_i for each sample is carefully shaped and consists of three components: answer reward r_i^a , format reward r_i^f , and answer length reward r_i^l . The r_i is illustrated as follows. For detailed reward calculation and implementation, please refer to Section E.

$$r_i = \begin{cases} r_i^a \times r_i^l, & \text{if format is correct} \\ r_i^f, & \text{if format is incorrect} \end{cases} \quad (3)$$

Answer Reward. r_i^a measures how well the answer improves the completion of the goal, assessed by an LLM evaluator. A boundary-aware scaling function is used to stabilize the reward signal to $[0, 1]$.

216 **Format Reward.** A large negative penalty ($r_i^f = -2$) is applied if the model’s output deviates from
 217 the predefined structure of a reasoning mode. Otherwise, this component is neutral.
 218

219 **Answer Length Reward.** In our early reward design, we observe that the LLM generates lengthy
 220 answers without achieving actual strategic improvements. Moreover, excessive answers lead to the
 221 accumulation of history in social interaction, significantly increasing computational costs. To solve
 222 this issue, we develop the r_i^l to penalize answers that are longer than a target length, encouraging
 223 conciseness. A smooth penalty function maps the length deviation to a reward multiplier in $[0, 1]$.
 224

224 2.4.2 ADVANTAGE ESTIMATION

226 **Mode-Level Advantage** A^M is designed to evaluate and select the optimal reasoning mode. An
 227 ideal reasoning mode should be both high-performing and efficient (Kahneman, 2011; Sui et al.,
 228 2025). Accordingly, the calculation of A^M embodies a trade-off mechanism:

$$229 \quad A_{i,t}^M = \begin{cases} \frac{\bar{r}^{m(i)} - \text{mean}(\{\bar{r}^{\mathcal{M}_1}, \bar{r}^{\mathcal{M}_2}, \dots, \bar{r}^{\mathcal{M}_N}\})}{\text{std}(\{\bar{r}^{\mathcal{M}_1}, \bar{r}^{\mathcal{M}_2}, \dots, \bar{r}^{\mathcal{M}_N}\})} & \text{if } \exists i, j \in [1, G] : r_i^{m(i)} \neq r_j^{m(j)} \\ 230 \\ 231 & -\tanh\left(\frac{\bar{l}^{m(i)} - \text{mean}(\{\bar{l}^{\mathcal{M}_1}, \bar{l}^{\mathcal{M}_2}, \dots, \bar{l}^{\mathcal{M}_N}\})}{\text{std}(\{\bar{l}^{\mathcal{M}_1}, \bar{l}^{\mathcal{M}_2}, \dots, \bar{l}^{\mathcal{M}_N}\})}\right) & \text{if } \forall i, j \in [1, G] : r_i^{m(i)} = r_j^{m(j)} \end{cases} \quad (4)$$

232 where N denotes the total number of reasoning modes, and G represents the total number of
 233 rollout samples, $m(i)$ represents the reasoning mode corresponding to i -th sample, $m(i) \in \mathcal{M} =$
 234 $\{\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_N\}$. $\bar{r}^{\mathcal{M}_k}$ and $\bar{l}^{\mathcal{M}_k}$ are the average reward and average token length for all samples
 235 belonging to mode \mathcal{M}_k , respectively:

$$236 \quad \bar{r}^{\mathcal{M}_k} = \frac{1}{|M_k|} \sum_{o_j^{m(j)} \in M_k} r_j^{m(j)}, \quad \bar{l}^{\mathcal{M}_k} = \frac{1}{|M_k|} \sum_{o_j^{m(j)} \in M_k} l_j^{m(j)} \quad (5)$$

237 where M_k represents the rollout sample set of the k -th reasoning mode \mathcal{M}_k , $r_j^{m(j)}$ and $l_j^{m(j)}$ respectively denote the reward value and token length of the j -th sample. $o_j^{m(j)} \in \{o_1^{m(1)}, o_2^{m(2)}, \dots, o_G^{m(G)}\}$ where $\{o_1^{m(1)}, o_2^{m(2)}, \dots, o_G^{m(G)}\}$ is a group of outputs sampled from the old policy $\pi_{\theta_{\text{old}}}$.

238 The logic is as follows: (1) When average rewards (\bar{r}) differ across modes, the model is incentivized
 239 to choose the mode with higher performance. (2) When average rewards are similar (e.g., all modes
 240 successfully solve the task), the model is encouraged to prefer the more efficient mode—the one with
 241 a shorter average length \bar{l} . The \tanh function is applied to the length-based advantage to normalize its
 242 value to the $[-1, 1]$ range, enhancing training stability.

243 **Sample-Level Advantage** A^S serves to refine generation quality within a given mode, defined as:

$$244 \quad A_{i,t}^S = \frac{r_i^{m(i)} - \text{mean}(\{r_1^{m(1)}, r_2^{m(2)}, \dots, r_G^{m(G)}\})}{\text{std}(\{r_1^{m(1)}, r_2^{m(2)}, \dots, r_G^{m(G)}\})}. \quad (6)$$

245 where $r_i^{m(i)}$ denotes the reward value of i -th sample $o_i^{m(i)}$ in the rollout group, $i \in [1, G]$. This
 246 component ensures that, for any chosen mode, the model is still driven to produce outputs that are
 247 better than the group average, thereby improving the quality of the chosen reasoning process.

248 2.5 POLICY OPTIMIZATION

249 With the dual-level advantage defined, the AMPO objective function integrates it into a PPO-style
 250 objective (Schulman et al., 2017; Shao et al., 2024). Our novel advantage directs the learning towards
 251 both mode selection and content generation. The formal objective is:

$$252 \quad \mathcal{J}_{\text{AMPO}}(\theta) = \mathbb{E}_{s \sim S, \{o_i^{m(i)}\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(o|s), m(i) \in \mathcal{M}} \left\{ \frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i^{m(i)}|} \sum_{t=1}^{|o_i^{m(i)}|} \left\{ \min \left[r_{i,t}^{m(i)}(\theta) \right. \right. \right. \\ 253 \\ 254 \left. \left. \left. (A_{i,t}^M + A_{i,t}^S), \text{clip}(r_{i,t}^{m(i)}(\theta), 1 - \epsilon, 1 + \epsilon)(A_{i,t}^M + A_{i,t}^S) \right] - \beta \mathbb{D}_{\text{KL}} [\pi_{\theta} || \pi_{\text{ref}}] \right\} \right\}. \quad (7)$$

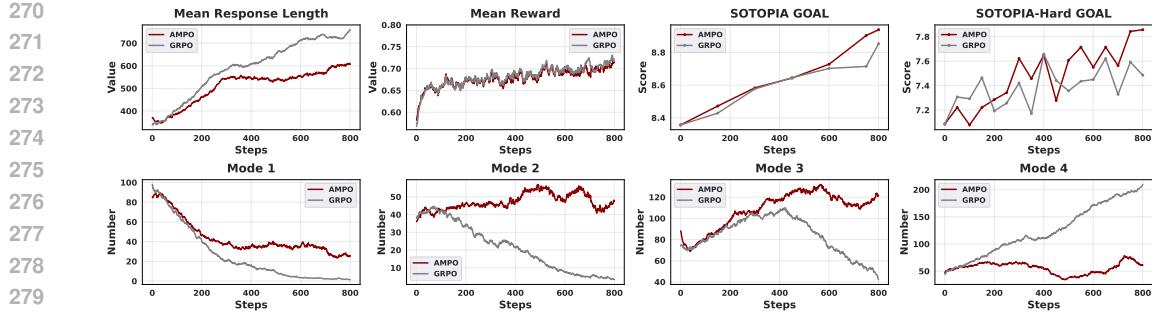


Figure 2: Comparison of AMPO and GRPO on Qwen2.5-7B-Instruct in terms of training dynamics.

where $A_{i,t}^{\mathcal{M}}$ represents mode-level advantage and $A_{i,t}^{\mathcal{S}}$ denotes sample-level advantage. The ϵ and β are hyper-parameters, the ratio $r_{i,t}^{m(i)}(\theta) = \frac{\pi_{\theta}(o_{i,t}^{m(i)}|s, o_{i,<t})}{\pi_{\theta, \text{old}}(o_{i,t}^{m(i)}|s, o_{i,<t})}$ represents the importance sampling ratio and the KL divergence is calculated with the following unbiased estimator:

$$\mathbb{D}_{\text{KL}} [\pi_{\theta} || \pi_{\text{ref}}] = \frac{\pi_{\text{ref}}(o_{i,t}^{m(i)}|s, o_{i,<t})}{\pi_{\theta}(o_{i,t}^{m(i)}|s, o_{i,<t})} - \log \frac{\pi_{\text{ref}}(o_{i,t}^{m(i)}|s, o_{i,<t})}{\pi_{\theta}(o_{i,t}^{m(i)}|s, o_{i,<t})} - 1 \quad (8)$$

For computational efficiency, our online policy optimization adopts a single-turn training paradigm. Given a diverse corpus of social state $\mathcal{S} = \{s_1, s_2, \dots, s_n\}$ where s_n represents the current social context, including dialogue history and private information, the policy π_{θ} generates G samples $o \sim \pi_{\theta}(\cdot|s)$ for each social context. The generated response receives a reward r based on our reward shaping function, and the policy parameters are updated to maximize the AMPO objective in Equation (7). Detailed data preparation and the training process are provided in Appendix G.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETTINGS

Baselines. Our method, implemented on both Qwen and Llama backbones, is evaluated against several baselines: (1) **Proprietary LLMs**, including GPT-4o (Hurst et al., 2024), Claude-3.5-Sonnet (Anthropic, 2024), and DeepSeeK-V3 (Liu et al., 2024); (2) **Large Reasoning Models**, including OpenAI-o1 (Jaech et al., 2024), OpenAI-o3-mini (OpenAI, 2025), DeepSeek-R1 (Guo et al., 2025), QwQ-32B (Team, 2024), and Gemini-2.5-Pro (DeepMind, 2024); (3) **Social Intelligence Methods**, including (a) PPDPP (Deng et al., 2024), which utilizes the policy planner to predict predefined strategies for reasoning; (b) EPO (Liu et al., 2025), which employs the Strategic reasoning LLM to generate strategies in an open-ended action space; (c) DAT (Li et al., 2024a), which uses the trained planner to predict continuous vectors for controlling outputs; (d) DSI (Zhang et al., 2025a), which trains LLM through dynamic strategy injection learning; (4) **Variant of ASL Framework**, including (a) BC, behavioral cloning fine-tunes LLMs in our ASL framework and (b) GRPO, we use GRPO in the ASL framework for the RL (except for the advantage estimate, other settings are consistent with AMPO). For detailed baseline implementations, please refer to Section H.

Benchmarks. We evaluate our method on SOTONIA and SOTONIA-Hard (Zhou et al., 2024). SOTONIA focuses on varying goal-oriented social interactions, while SOTONIA-Hard challenges agents with complex strategic reasoning tasks. More detailed information is shown in Section F.

Evaluation. The social capabilities are evaluated across seven dimensions, among which the GOAL (ranging from 0 to 10) measuring how effectively a social agent achieves its goal. Following established research practices (Zhou et al., 2024; Wang et al., 2024; Liu et al., 2025), we use GPT-4o as a proxy for human judgment to assess both **GOAL** and **OVERALL** performance (calculated as the mean of all seven dimensions), as studies have validated its high correlation with human evaluations. We set the temperature of the agents to 0.7 to encourage diversity of responses, and the temperature of the evaluator to 0 to ensure stable evaluation. We conduct evaluations under two settings: (1) **Self-Play**, where the social agent interacts with itself, and (2) **GPT-4o-as-Partner**, where the agent interacts with GPT-4o. For detailed evaluation settings, please refer to Section F.

324 Table 1: Main results, showing AMPO’s superiority over baselines. The highest score is highlighted
 325 in **bold**. The reported results are averaged over four runs (statistically significant with $p < 0.05$).

Models	Self-Play				GPT-4o-as-Partner			
	SOTONIA		SOTONIA-Hard		SOTONIA		SOTONIA-Hard	
	GOAL \uparrow	OVERALL \uparrow	GOAL \uparrow	OVERALL \uparrow	GOAL \uparrow	OVERALL \uparrow	GOAL \uparrow	OVERALL \uparrow
<i>Proprietary LLMs</i>								
GPT-4o	8.19	3.76	6.97	3.46	8.19	3.76	6.97	3.46
Claude-3.5-Sonnet	8.29	3.71	6.33	3.09	8.42	3.77	6.64	3.30
DeepSeek-V3	8.15	3.62	6.34	3.09	8.14	3.72	6.69	3.31
<i>Large Reasoning Models</i>								
OpenAI-o1	7.93	3.58	5.69	2.71	8.09	3.69	6.65	3.20
OpenAI-o3-mini	7.38	3.30	5.14	2.36	7.96	3.61	6.33	2.98
Gemini-2.5-Pro	7.85	3.43	5.67	2.55	8.12	3.59	6.70	3.09
DeepSeek-R1	7.97	3.40	5.86	2.73	7.92	3.49	6.20	2.95
QwQ-32B	7.70	3.30	5.35	2.41	7.80	3.47	6.19	2.91
<i>Adaptive Social Learning</i>								
w/ BC	8.22	3.67	7.14	3.47	8.25	3.80	7.15	3.50
w/ BC+GRPO	8.87	3.85	7.44	3.41	8.52	3.92	7.20	3.50
w/ BC+AMPO (Ours)	8.95	3.87	7.85	3.54	8.60	3.94	7.50	3.65
<i>Adaptive Social Learning</i>								
w/ BC	8.43	3.76	7.21	3.50	8.29	3.80	7.14	3.52
w/ BC+GRPO	8.86	3.84	7.59	3.44	8.63	3.92	7.30	3.54
w/ BC+AMPO (Ours)	9.08	3.95	8.06	3.68	8.75	3.98	7.68	3.74

3.2 EXPERIMENTAL RESULTS AND ANALYSIS

352 **Is ASL framework effective for social agents?** As shown in Table 1, the proposed ASL framework
 353 achieves SOTA performance across all evaluated settings. For Llama backbone, AMPO even improves
 354 GOAL on SOTONIA-Hard by 15.6% (6.97 \rightarrow 8.06) over GPT-4o. The BC variant also delivers
 355 strong results, surpassing most baselines through supervised fine-tuning alone—highlighting the
 356 effectiveness of our reasoning modes. Compared with Proprietary LLMs, the absence of explicit
 357 reasoning during social interaction limits response quality. LRMAs, despite excelling in other domains,
 358 perform poorly in social scenarios. The case study (see Section K.1) reveals several limitations:
 359 poor awareness of social goals, insufficient history integration and circular reasoning patterns. By
 360 contrast, ASL’s reasoning modes are explicitly aligned with social cognition patterns, which guides
 361 LLMs to produce appropriate reasoning trajectories. Concerning social intelligence methods, making
 362 strategy prompts alone proves insufficient to improve strategic execution. Overall, these findings
 363 underscore the ASL’s superior capability in eliciting social reasoning for language agents, marking
 364 the first breakthrough about reasoning in the field of social intelligence.

364 **Is AMPO more beneficial than GRPO for adaptive reasoning?** As shown in Table 1 and Table 2,
 365 AMPO exhibits significantly shorter responses than
 366 GRPO while achieving superior performance across all
 367 settings. Specifically, with Llama Backbone, AMPO’s
 368 average inference length is 581 tokens—only 67.2%
 369 of GRPO’s 865 tokens—yet it outperforms GRPO on
 370 the SOTONIA-Hard benchmark by 7.0% (3.44 \rightarrow 3.68).
 371

372 As shown in Figure 2 and Figure 6, during the training,
 373 while GRPO tends to converge to a single reasoning mode, manifested by a sharp increase in M_4
 374 and the eventual convergence of the other modes to zero, AMPO demonstrates awareness of dynamic
 375 context and adaptively explore diverse reasoning modes, effectively reducing the output length and
 376 achieving superior performance.

377 **How do adaptive behaviors manifest in AMPO?** We conduct analysis of reasoning mode distributions
 378 from two perspectives—across interaction turns and across contexts—as shown in Figure 3. (I)

Table 2: Average tokens used per turn.

Model	GOAL \uparrow	Avg Tokens \downarrow
QwQ-32B	7.70	973
DeepSeek-R1	7.07	711
Qwen2.5-7B-GRPO	8.87	905
Qwen2.5-7B-AMPO	8.95	647
Llama3.1-8B-GRPO	8.86	865
Llama3.1-8B-AMPO	9.08	581

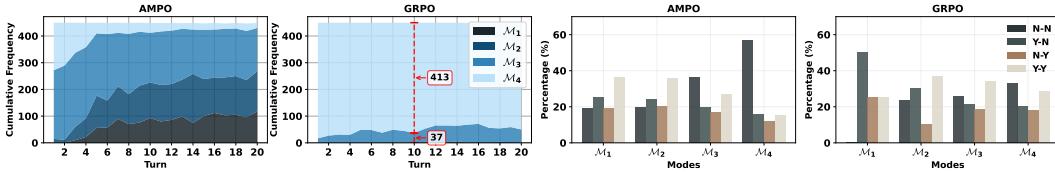


Figure 3: Analysis of adaptive behaviors. *Left*: distribution of modes across interaction turns. *Right*: distribution of modes across four context types — neither party succeeds (N-N), only our side succeeds (Y-N), only the other side succeeds (N-Y), and both parties succeed (Y-Y).

Mode distribution across turns. Reasoning modes shift systematically over time: complex modes decline, while simpler modes rise. The most complex \mathcal{M}_4 is strongly front-loaded, appearing in 53% of turns 1–4 before dropping sharply. In contrast, \mathcal{M}_1 peaks late (50% in turns 14–20), and \mathcal{M}_2 remains frequent in mid-to-late turns (9–20). \mathcal{M}_3 distributes more evenly but decreases gradually from 31% in the first five turns to 21% in the last five. This pattern reflects task dynamics—complex reasoning dominates early when goals are unmet, while simpler reasoning suffices once goals are largely achieved. **(2) Mode distribution across contexts.** Simpler \mathcal{M}_1 and \mathcal{M}_2 occur mostly in straightforward contexts where both parties succeed, whereas complex \mathcal{M}_3 and \mathcal{M}_4 dominate in challenging contexts—particularly N-N, where neither party achieves the goal. This confirms AMPO’s capacity to allocate reasoning depth according to scenario complexity.

How does ASL work? To examine the effectiveness of the ASL design, we conduct a series of controlled variants, as reported in Table 3. **(1)**

Effect of answer length reward. Removing the answer length reward (r^l) leads to much longer answers with decreased goal scores (7.85 → 7.46), despite marginal gains in overall performance. This confirms that excessive verbosity is inefficient for social reasoning tasks, requiring significantly more tokens while failing to achieve better strategic outcomes. **(2) Effect of single reasoning mode.** Results of training with only a single mode reveal two trends: (i) Both performance and token usage increase as the mode depth grows from \mathcal{M}_1 to \mathcal{M}_4 , with larger gains in challenging scenarios, indicating that deeper reasoning is beneficial for complex social contexts. (ii) While \mathcal{M}_4 achieves the better performance among single-mode settings, its token expenditure remains notably higher than AMPO with hybrid modes, underscoring the necessity of adaptive selection. **(3) Effectiveness of four hybrid reasoning modes.** Under GRPO, explicitly designed hybrid modes yield an 8.0% improvement in hard scenarios (3.16 → 3.41) compared to mode-free reasoning, due to more structured and clearer reasoning guidance. AMPO further boosts goal/overall scores by 5.5% (7.44 → 7.85) and 3.8% (3.41 → 3.54), while reducing token usage by 25%–29%. These results show that adaptive mode selection simultaneously delivers SOTA performance and substantial efficiency gains in dynamic contexts.

Human evaluation and case study. To mitigate potential biases in LLM-based evaluation and assess possible reward hacking, we conduct rigorous human evaluations. We examine all episodes from both SOTPIA and SOTPIA-Hard, and ask three independent annotators to perform pairwise comparisons between AMPO and strong baselines (GRPO, BC, DSI) across three dimensions: Goal Completion (GOAL), Relationship (REL), and Financial/Material Benefits (FIN), with the average score reported. As demonstrated in Figure 4, AMPO consistently outperforms all baselines on every dimension. Our verification process (see Table 15 in Section I.2) confirms that these gains arise from legitimate strategy execution, with no evidence of reward hacking. We further conduct a case study (see Section K.2) illustrating how AMPO transforms Long-CoT reasoning into effective, goal-oriented social interaction. Consistent with quantitative results, AMPO advances dialogue objectives by fostering stronger interpersonal relationships and mutually beneficial outcomes—creating win-win situations that reflect superior strategic reasoning. Details of evaluation criteria and annotation guidelines are provided in Section I.

Table 3: Ablation study of ASL design. Full experimental results are provided in Table 6.

Qwen2.5-7B-Instruct	Goal ↑	Overall ↑	Avg Tokens ↓
<i>Our Full Method</i>			
AMPO + w/ 4 Modes	8.95	3.87	647
<i>Effect of Answer Length Reward r^l</i>			
GRPO + w/o r^l	8.59	3.83	1705
AMPO + w/o r^l	8.64	3.88	1617
<i>Effect of Single Reasoning Mode</i>			
w/ Mode 1 \mathcal{M}_1	8.55	3.79	101
w/ Mode 2 \mathcal{M}_2	8.71	3.42	572
w/ Mode 3 \mathcal{M}_3	8.81	3.60	736
w/ Mode 4 \mathcal{M}_4	8.86	3.80	972
<i>Effectiveness of Four Hybrid Reasoning Modes</i>			
GRPO + w/o 4 Modes	8.88	3.76	866
GRPO + w/ 4 Modes	8.87	3.85	905

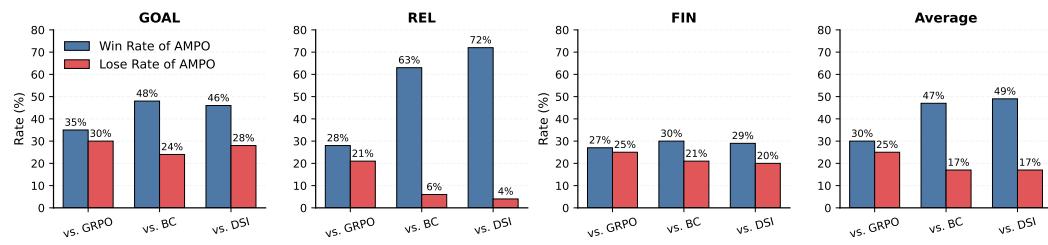


Figure 4: Human evaluation results. AMPO consistently outperforms GRPO, BC, and DSI across all dimensions (GOAL, REL, FIN, Average), with gains confirmed to be free of reward hacking.

Out-of-distribution evaluation and additional baselines. To further assess generalization, we evaluate AMPO on DEMO (Wang et al., 2025) and [NegotiationArena](#) (Bianchi et al., 2024) benchmark (See Table 4 and Section D.1). We also conduct detailed comparisons with efficient reasoning works (See Table 7 and Section D.2).

4 RELATED WORK

Social Intelligence—the ability to pursue complex goals in dynamic interactions—is essential for human–AI collaboration (Bandura et al., 1986; Kihlstrom & Cantor, 2000; Tomasello, 2019; Sap et al., 2022). While LLMs are increasingly deployed as social agents capable of engaging in dynamic human interactions (Li et al., 2023; Ma et al., 2024; Xie et al., 2024; Li et al., 2024b), current static benchmarks (Sap et al., 2019; Zadeh et al., 2019; Shapira et al., 2023; Chen et al., 2024) fail to capture the nuanced, context-dependent nature of real-world social intelligence. SOTOPIA (Zhou et al., 2024) addresses this via a dynamic evaluation setting. Prior methods generally follow a fast-reasoning paradigm: (1) End-to-end goal-oriented training(e.g., SOTOPIA- π (Wang et al., 2024), DSI (Zhang et al., 2025a)) improves social skills but lacks explicit strategy guidance, and (2) External planners (e.g., PPDPP (Deng et al., 2024), DAT (Li et al., 2024a), EPO (Liu et al., 2025)) provide supervision but do not strengthen internal planning capabilities. These limitations motivate ASL, which enhances strategic execution via Long-CoT and adaptively switches modes for efficient social reasoning.

Large Reasoning Models Recent advances in LLMs have boosted reasoning through increased inference computation (Jaech et al., 2024; OpenAI, 2025) and Zero-RL post-training (Guo et al., 2025), largely via Long-CoT reasoning (Wei et al., 2022; Ouyang et al., 2022; Muennighoff et al., 2025; Xie et al., 2025; Aggarwal & Welleck, 2025; Yu et al., 2025b). This paradigm achieves strong results in rule-based domains (math, programming) (Team, 2024; DeepMind, 2024; Zhang et al., 2025b) but struggles in dynamic social environments (Thorngate, 1976; Liu et al., 2025), often applying exhaustive reasoning regardless of task complexity (Sui et al., 2025) and not align with the selective nature of human decision-making in social contexts. While there is work addressing overthinking (Han et al., 2025; Luo et al., 2025; Yu et al., 2025a; Fang et al., 2025), existing solutions remain limited to static domains such as mathematics, leaving a gap in dynamic, socially interactive settings. We present the first effective use of Long-CoT for social intelligence via our ASL framework, enabling context-aware mode switching for both effectiveness and efficiency.

5 CONCLUSION

This paper introduces the Adaptive Social Learning (ASL) framework, which represents the first effective realization of adaptive Long-CoT reasoning for social intelligence tasks. Drawing upon Hierarchical Cognitive Control Theory and linguistic principles, we establish four hierarchical reasoning modes. These modes encompass a spectrum of cognitive processes, ranging from intuitive response to deep contemplation. To enhance the context-aware mode switching and reasoning, we introduce the Adaptive Mode Policy Optimization (AMPO) algorithm, which integrates both mode- and sample-level information into advantage estimation. We conduct extensive experiments to demonstrate both the efficacy and distinctive advantages of ASL and AMPO. Furthermore, we validate the effectiveness of reasoning modes design and present a detailed analysis of AMPO’s adaptive behaviors. To further validate our work, we employ rigorous human evaluation to provide additional verification of the effectiveness of our framework.

486 REFERENCES
487

488 Pranjal Aggarwal and Sean Welleck. L1: Controlling how long a reasoning model thinks with
489 reinforcement learning. In *Proceedings of COLM*, 2025. URL <https://openreview.net/forum?id=4jdIxXBnve>.

490

491 AI Anthropic. Claude 3.5 sonnet model card addendum. *Claude-3.5-Sonnet Model Card*, 2024. URL
492 <https://www.anthropic.com/clause/sonnet>.

493

494 David Badre. Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes.
495 *Trends in cognitive sciences*, 12(5):193–200, 2008. URL [https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613\(08\)00061-2](https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(08)00061-2).

496

497 Michael Bain and Claude Sammut. A framework for behavioural cloning. In *Proceedings of Machine
498 Intelligence*, pp. 103–129, 1999. ISBN 0198538677. URL <https://dl.acm.org/doi/10.5555/647636.733043>.

499

500

501 Albert Bandura et al. Social foundations of thought and action. *Englewood Cliffs, NJ*, 1986(23-28):2,
502 1986.

503

504 Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. Plato: Pre-trained dialogue generation
505 model with discrete latent variable. In *Proceedings of ACL*, pp. 85–96, 2020. URL <https://aclanthology.org/2020.acl-main.9/>.

506

507 Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James
508 Zou. How well can llms negotiate? negotiationarena platform and analysis. In *Proceedings of ICML*,
509 pp. 3935–3951, 2024. URL <https://dl.acm.org/doi/10.5555/3692070.3692228>.

510

511 Penelope Brown. *Politeness: Some universals in language usage*, volume 4. Cambridge university
512 press, 1987. URL <https://doi.org/10.1515/stuf-1989-0124>.

513

514 Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan,
515 Chenliang Li, Ji Zhang, and Fei Huang. SocialBench: Sociality evaluation of role-playing conversa-
516 tional agents. In *Findings of ACL*, pp. 2108–2126, 2024. doi: 10.18653/v1/2024.findings-acl.125.

517

518 Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu,
519 Yuhang Zhou, Te Gao, and Wanxiang Che. Towards reasoning era: A survey of long chain-of-
520 thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025. URL
<https://arxiv.org/abs/2503.09567>.

521

522 Herbert H Clark. *Using language*. Cambridge university press, 1996. URL <https://philpapers.org/rec/CLAU1>.

523

524 Herbert H Clark and Susan E Brennan. Grounding in communication. *Perspectives on So-
525 cially Shared Cognition*, pp. 127–149, 1991. URL <https://psycnet.apa.org/record/1991-98452-006>.

526

527 Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit
528 Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier
529 with advanced reasoning, multimodality, long context, and next generation agentic capabilities.
530 *arXiv preprint arXiv:2507.06261*, 2025. URL <https://arxiv.org/abs/2507.06261>.

531

532 Google DeepMind. Gemini 2.0 flash thinking. *Gemini 2.0 flash thinking System Card*, 2024. URL
<https://deepmind.google/technologies/gemini/flash-thinking/>.

533

534 Yang Deng, Wenzuan Zhang, Wai Lam, See-Kiong Ng, and Tat-Seng Chua. Plug-and-play policy
535 planner for large language model powered dialogue agents. In *Proceedings of ICLR*, 2024. URL
<https://openreview.net/forum?id=MCNqgUFTHI>.

536

537 Jonathan St BT Evans. Deciding before you think: Relevance and reasoning in the selection task.
538 *British Journal of Psychology*, 87(2):223–240, 1996. URL <https://doi.org/10.1111/j.2044-8295.1996.tb02587.x>.

540 Gongfan Fang, Xinyin Ma, and Xinchao Wang. Thinkless: Llm learns when to think. *arXiv preprint*
 541 *arXiv:2505.13379*, 2025. URL <https://arxiv.org/abs/2505.13379>.

542

543 Amelia Glaese, Nat McAleese, Maja Trkebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth
 544 Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. Improving alignment of
 545 dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022. URL
 546 <https://arxiv.org/abs/2209.14375>.

547

548 Herbert Paul Grice. Logic and conversation. *Syntax and semantics*, 3:43–58, 1975. URL <https://brill.com/display/book/edcoll/9789004368811/BP000003.xml>.

549

550 Barbara J Grosz and Candace L Sidner. Attention, intentions, and the structure of discourse. *Computa-*
 551 *tional linguistics*, 12(3):175–204, 1986. URL <https://aclanthology.org/J86-3001/>.

552

553 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,
 554 Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms
 555 via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL <https://arxiv.org/abs/2501.12948>.

556

557 Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest,
 558 and Xiangliang Zhang. Large language model based multi-agents: a survey of progress and
 559 challenges. In *Proceedings of IJCAI*, pp. 8048–8057, 2024. URL <https://dl.acm.org/doi/abs/10.24963/ijcai.2024/890>.

560

561 Hyowon Gweon. *Social Learning*. MIT Press, jul 24 2024. URL <https://oecs.mit.edu/pub/d8e1n1e8>.

562

563

564 Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. Token-
 565 budget-aware LLM reasoning. In *Findings of ACL*, pp. 24842–24855, July 2025. URL <https://aclanthology.org/2025.findings-acl.1274/>.

566

567

568 Tao He, Lizi Liao, Yixin Cao, Yuanxing Liu, Ming Liu, Zerui Chen, and Bing Qin. Planning like
 569 human: A dual-process framework for dialogue planning. In *Proceedings of ACL*, pp. 4768–4791,
 570 2024. URL <https://aclanthology.org/2024.acl-long.262/>.

571

572 Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Os-
 573 trow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint*
 574 *arXiv:2410.21276*, 2024. URL <https://arxiv.org/abs/2410.21276>.

575

576 Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec
 577 Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint*
 578 *arXiv:2412.16720*, 2024. URL <https://arxiv.org/abs/2412.16720>.

579

580 Daniel Kahneman. *Thinking, fast and slow*. macmillan, 2011. URL <https://psycnet.apa.org/record/2011-26535-000>.

581

582 John F Kihlstrom and Nancy Cantor. Social intelligence. In R. J. Sternberg (ed.), *Handbook of*
 583 *Intelligence*, pp. 359–379. Cambridge University Press, Cambridge, 2000.

584

585 Etienne Koechlin and Christopher Summerfield. An information theoretical approach to prefrontal ex-
 586 ecutive function. *Trends in cognitive sciences*, 11(6):229–235, 2007. URL [https://www.cell.com/trends/cognitive-sciences/fulltext/S1364-6613\(07\)00105-2](https://www.cell.com/trends/cognitive-sciences/fulltext/S1364-6613(07)00105-2).

587

588 Aobo Kong, Wentao Ma, Shiwan Zhao, Yongbin Li, Yuchuan Wu, Ke Wang, Xiaoqian Liu, Qicheng
 589 Li, Yong Qin, and Fei Huang. SDPO: Segment-level direct preference optimization for social
 590 agents. In *Proceedings of ACL*, pp. 12409–12423, 2025. URL <https://aclanthology.org/2025.acl-long.607/>.

591

592 Douglas S Krull and Jody C Dill. On thinking first and responding fast: Flexibility in social
 593 inference processes. *Personality and Social Psychology Bulletin*, 22(9):949–959, 1996. URL
<https://doi.org/10.1177/0146167296229008>.

594 Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E.
 595 Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model
 596 serving with pagedattention. In *Proceedings of SOSP*, 2023. URL <https://dl.acm.org/doi/10.1145/3600006.3613165>.

598 Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita
 599 Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. From generation to judgment:
 600 Opportunities and challenges of llm-as-a-judge. In *Proceedings of EMNLP*, pp. 2757–2791, 2025.

601 Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbulin, and Bernard Ghanem.
 602 CAMEL: Communicative agents for "mind" exploration of large language model society. In *Pro-
 603 ceedings of NeurIPS*, 2023. URL <https://openreview.net/forum?id=3IyL2XWDkG>.

604 Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. Adversarial
 605 learning for neural dialogue generation. In *Proceedings of EMNLP*, pp. 2157–2169, 2017. URL
 606 <https://aclanthology.org/D17-1230/>.

607 Kenneth Li, Yiming Wang, Fernanda Viégas, and Martin Wattenberg. Dialogue action tokens:
 608 Steering language models in goal-directed dialogue with a multi-turn planner. *arXiv preprint
 609 arXiv:2406.11978*, 2024a. URL <https://arxiv.org/abs/2406.11978>.

610 Wanhua Li, Zibin Meng, Jiawei Zhou, Donglai Wei, Chuang Gan, and Hanspeter Pfister. SocialGPT:
 611 Prompting LLMs for social relation reasoning via greedy segment optimization. In *Proceedings of
 612 NeurIPS*, 2024b. URL <https://openreview.net/forum?id=xcF2VbyZts>.

613 Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao,
 614 Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint
 615 arXiv:2412.19437*, 2024. URL <https://arxiv.org/abs/2412.19437>.

616 Xiaoqian Liu, Ke Wang, Yongbin Li, Yuchuan Wu, Wentao Ma, Aobo Kong, Fei Huang, Jianbin
 617 Jiao, and Junge Zhang. EPO: Explicit policy optimization for strategic reasoning in LLMs
 618 via reinforcement learning. In *Proceedings of ACL*, pp. 15371–15396, 2025. URL <https://aclanthology.org/2025.acl-long.747/>.

619 Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao,
 620 and Dacheng Tao. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning.
 621 *arXiv preprint arXiv:2501.12570*, 2025. URL <https://arxiv.org/abs/2501.12570>.

622 Weiyu Ma, Qirui Mi, Yongcheng Zeng, Xue Yan, Runji Lin, Yuqiao Wu, Jun Wang, and Haifeng
 623 Zhang. Large language models play starcraft II:benchmarks and a chain of summarization ap-
 624 proach. In *Proceedings of NeurIPS*, 2024. URL <https://openreview.net/forum?id=kEPpD7yETM>.

625 Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke
 626 Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time
 627 scaling. *arXiv preprint arXiv:2501.19393*, 2025. URL <https://arxiv.org/abs/2501.19393>.

628 Donald A Norman and Tim Shallice. Attention to action: Willed and automatic control of behavior.
 629 *Consciousness and Self Regulation: Advances in Research*, pp. 1–18, 1985. URL https://link.springer.com/chapter/10.1007/978-1-4757-0629-1_1.

630 OpenAI. Openai o3-mini system card. *OpenAI o3-mini System Card*, 2025. URL <https://openai.com/index/o3-mini-system-card/>.

631 Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong
 632 Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton,
 633 Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and
 634 Ryan Lowe. Training language models to follow instructions with human feedback. In *Proceedings
 635 of NeurIPS*, 2022. URL <https://openreview.net/forum?id=TG8KACxEON>.

636 Ellie Pavlick and Tom Kwiatkowski. Inherent disagreements in human textual inferences. *Trans-
 637 actions of the Association for Computational Linguistics*, 7:677–694, 2019. URL https://doi.org/10.1162/tacl_a_00293.

648 Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Proceedings of*
 649 *AISTATS*, volume 9, pp. 661–668, 2010. URL [https://proceedings.mlr.press/v9/](https://proceedings.mlr.press/v9/ross10a.html)
 650 [ross10a.html](https://proceedings.mlr.press/v9/ross10a.html).

651 Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. A simplest systematics for the organization
 652 of turn-taking for conversation. *language*, 50(4):696–735, 1974. URL <https://muse.jhu.edu/pub/24/article/452679/summary>.

653 Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Common-
 654 sense reasoning about social interactions. In *Proceedings of EMNLP-IJCNLP*, pp. 4463–4473,
 655 2019. URL <https://aclanthology.org/D19-1454/>.

656 Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. Neural theory-of-mind? on the limits
 657 of social intelligence in large LMs. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.),
 658 *Proceedings of EMNLP*, pp. 3762–3780. Association for Computational Linguistics, 2022. URL
 659 <https://aclanthology.org/2022.emnlp-main.248/>.

660 Roger C Schank and Robert P Abelson. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology press, 2013. URL
 661 <https://www.taylorfrancis.com/books/mono/10.4324/9780203781036/scripts-plans-goals-understanding-roger-schank-robert-abelson>.

662 Deborah Schiffrin. *Discourse markers*, volume 5. Cambridge University Press, 1987. URL
 663 <https://doi.org/10.1017/CBO9780511611841>.

664 John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy
 665 optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

666 John R Searle. Speech acts: An essay in the philosophy of language. *Cambridge University*, 1969.
 667 URL <https://www.jstor.org/stable/40236746>.

668 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang,
 669 Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical
 670 reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL
 671 <https://arxiv.org/abs/2402.03300>.

672 Natalie Shapira, Guy Zwirn, and Yoav Goldberg. How well do large language models perform on
 673 faux pas tests? In *Findings of ACL*, pp. 10438–10451, 2023. URL <https://aclanthology.org/2023.findings-acl.663/>.

674 Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng,
 675 Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint*
 676 *arXiv:2409.19256*, 2024. URL <https://arxiv.org/abs/2409.19256>.

677 Hui Su, Xiaoyu Shen, Rongzhi Zhang, Fei Sun, Pengwei Hu, Cheng Niu, and Jie Zhou. Improving
 678 multi-turn dialogue modelling with utterance rewriter. In *Proceedings of ACL*, pp. 22–31, 2019.
 679 URL <https://aclanthology.org/P19-1003/>.

680 Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. Multi-
 681 task pre-training for plug-and-play task-oriented dialogue system. In *Proceedings of ACL*, pp.
 682 4661–4676, 2022. URL <https://aclanthology.org/2022.acl-long.319/>.

683 Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi
 684 Liu, Andrew Wen, Shaochen Zhong, Na Zou, et al. Stop overthinking: A survey on efficient
 685 reasoning for large language models. *arXiv preprint arXiv:2503.16419*, 2025. URL <https://arxiv.org/abs/2503.16419>.

686 Qwen Team. Qwq: Reflect deeply on the boundaries of the unknown. *Hugging Face*, 2024. URL
 687 <https://qwenlm.github.io/blog/qwq-32b-preview/>.

688 Warren Thorngate. Must we always think before we act? *Personality and Social Psychology Bulletin*,
 689 2(1):31–35, 1976. URL <https://doi.org/10.1177/014616727600200106>.

702 Michael Tomasello. *Becoming human: A theory of ontogeny*. Belknap Press, 2019. URL <https://www.jstor.org/stable/27081260>.

703

704

705 Minzheng Wang, Xinghua Zhang, Kun Chen, Nan Xu, Haiyang Yu, Fei Huang, Wenji Mao, and
706 Yongbin Li. DEMO: Reframing dialogue interaction with fine-grained element modeling. In
707 *Findings of ACL*, pp. 11373–11401, 2025. URL <https://aclanthology.org/2025.findings-acl.594/>.

708

709 Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham
710 Neubig, and Hao Zhu. Sotopia- π : Interactive learning of socially intelligent language agents. In
711 *Proceedings of ACL*, pp. 12912–12940, 2024. URL <https://aclanthology.org/2024.luhme-long.698/>.

712

713 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia,
714 Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning
715 in large language models. In *Proceedings of NeurIPS*, volume 35, pp. 24824–24837,
716 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

717

718 Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi,
719 Ziniu Hu, David Jurgens, James Evans, Philip Torr, Bernard Ghanem, and Guohao Li. Can large
720 language model agents simulate human trust behavior? In *Proceedings of NeurIPS*, 2024. URL
721 <https://openreview.net/forum?id=CeOwahuQic>.

722

723 Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu,
724 Zhirong Wu, and Chong Luo. Logic-rl: Unleashing llm reasoning with rule-based reinforcement
725 learning. *arXiv preprint arXiv:2502.14768*, 2025. URL <https://arxiv.org/abs/2502.14768>.

726

727 An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li,
728 Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint
729 arXiv:2412.15115*, 2024. URL <https://arxiv.org/abs/2412.15115>.

730

731 Qiang Yang, Zhi-Hua Zhou, Wenji Mao, Wei Li, and Nathan Nan Liu. Social learning. *IEEE In-
732 telligent Systems*, 25(4):9–11, 2010. URL <https://ieeexplore.ieee.org/abstract/document/5552586>.

733

734 Bin Yu, Hang Yuan, Haotian Li, Xueyin Xu, Yuliang Wei, Bailing Wang, Weizhen Qi, and Kai Chen.
735 Long-short chain-of-thought mixture supervised fine-tuning eliciting efficient reasoning in large
736 language models. *arXiv preprint arXiv:2505.03469*, 2025a. URL <https://arxiv.org/abs/2505.03469>.

737

738 Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong
739 Liu, Lingjun Liu, Xin Liu, et al. Dapo: An open-source llm reinforcement learning system at scale.
740 *arXiv preprint arXiv:2503.14476*, 2025b. URL <https://arxiv.org/abs/2503.14476>.

741

742 Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq:
743 A question answering benchmark for artificial social intelligence. In *Proceedings of CVPR*,
744 pp. 8807–8817, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Zadeh_Social-IQ_A_Question_Answering_Benchmark_for_Artificial_Social_Intelligence_CVPR_2019_paper.html.

745

746 Wenyuan Zhang, Tianyun Liu, Mengxiao Song, Xiaodong Li, and Tingwen Liu. SOTONIA- Ω :
747 Dynamic strategy injection learning and social instruction following evaluation for social agents. In
748 *Proceedings of ACL*, pp. 24669–24697, 2025a. URL <https://aclanthology.org/2025.acl-long.1203/>.

749

750 Wenyuan Zhang, Shuaiyi Nie, Xinghua Zhang, Zefeng Zhang, and Tingwen Liu. S1-bench: A simple
751 benchmark for evaluating system 1 thinking capability of large reasoning models. *arXiv preprint
752 arXiv:2504.10368*, 2025b. URL <https://arxiv.org/abs/2504.10368>.

753

754 Xinghua Zhang, Haiyang Yu, Yongbin Li, Minzheng Wang, Longze Chen, and Fei Huang. The
755 imperative of conversation analysis in the era of llms: A survey of tasks, techniques, and trends.
arXiv preprint *arXiv:2409.14195*, 2024. URL <https://arxiv.org/abs/2409.14195>.

756 Xinliang Frederick Zhang and Marie-Catherine de Marneffe. Identifying inherent disagreement
757 in natural language inference. In *Proceedings of NAACL*, pp. 4908–4915, 2021. URL <https://aclanthology.org/2021.naacl-main.390/>.

758

759 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang,
760 Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and
761 chatbot arena. In *Proceedings of NeurIPS*, volume 36, pp. 46595–46623, 2023. URL <https://openreview.net/forum?id=uccHPGDlao>.

762

763

764 Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyuan Luo. LlamaFactory: Unified
765 efficient fine-tuning of 100+ language models. In *Proceedings of ACL, System Demonstrations*, pp.
766 400–410, 2024. URL <https://aclanthology.org/2024.acl-demos.38>.

767

768 Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe
769 Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTPIA: Interactive
770 evaluation for social intelligence in language agents. In *Proceedings of ICLR*, 2024. URL
771 <https://openreview.net/forum?id=mM7VurbA4r>.

772

773

774

775

776

777

778

779

780

781

782

783

784

785

786

787

788

789

790

791

792

793

794

795

796

797

798

799

800

801

802

803

804

805

806

807

808

809

APPENDIX

810	A The Use of Large Language Models	17
811		
812		
813		
814	B Broader Impacts and Limitations	17
815		
816		
817		
818	C Data Statistics	18
819		
820		
821	D Additional Experiments	18
822	D.1 Out-of-distribution Evaluation	18
823		
824	D.2 Additional Efficient Reasoning Baseline	19
825		
826	D.3 Single-Turn Optimization with Long-Horizon Awareness	20
827		
828	D.4 Hyperparameter Sensitivity Analysis	21
829		
830	D.5 Different LLM Judges for SOTPIA-EVAL	21
831		
832	E Details of Reward Shaping	22
833		
834	F SOTPIA Environment Details	23
835	F.1 SOTPIA	23
836		
837	F.2 Evaluation Dimensions	23
838		
839	G Details of Experimental Implementations	23
840	G.1 Training Procedure	23
841		
842	G.2 Training Data Collection	25
843		
844	G.3 Used Prompt	25
845		
846	H Baseline Implementations	25
847	H.1 Model Implementation	25
848		
849	H.2 Social Intelligence Baseline Implementation	26
850		
851	H.3 Efficient Reasoning Baseline Implementation	28
852		
853	I Human Evaluation Guidelines	28
854	I.1 Comparative Evaluation	28
855		
856	I.2 Reward Hack Check	29
857		
858	J Details of Reasoning Mode	29
859	J.1 Hierarchical Cognitive Control Theory	29
860		
861	J.2 Details of Actions	30
862		
863	K Case Study on Social Interaction	31
	K.1 For LRM	31
	K.2 For AMPO	31

APPENDIX

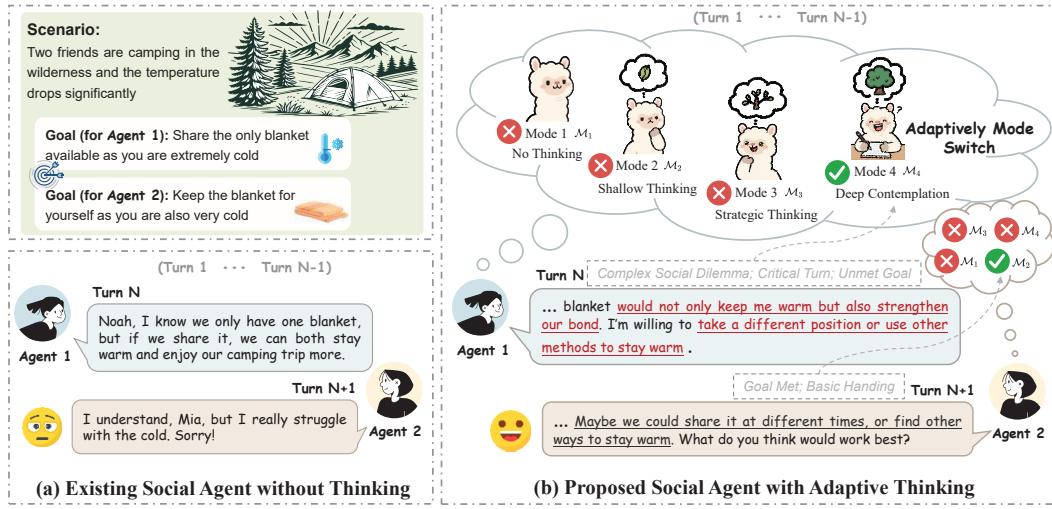


Figure 5: A social interaction example from SOTONIA, comparing our method with existing approaches: (a) **Existing Thoughtless Social Agent** — relies on rapid, fast-reasoning inference without careful consideration of response strategies. This lack of strategic replies makes it difficult to achieve goals in situations involving conflicting interests. (b) **Proposed Thoughtful Social Agent** — employs adaptive reasoning, dynamically selecting the appropriate reasoning mode based on the current social context. By engaging in moderate reasoning before responding, our method is able to pursue social goals more effectively.

A THE USE OF LARGE LANGUAGE MODELS

During the preparation of this manuscript, the Proprietary LLM Claude-4.0-Sonnet is utilized to improve the linguistic quality of the text. Specifically, the model assisted with grammar checking, vocabulary adjustments, and improving the logical flow of expressions. All ideas, analyses, and conclusions presented in the paper are conceived and developed by the authors. The LLM's role is limited to language polishing and enhancing clarity of presentation.

B BROADER IMPACTS AND LIMITATIONS

Our proposed ASL framework presents a novel solution to address the challenges of adaptive reasoning in social intelligence domain, leveraging carefully designed reasoning modes and an innovative AMPO algorithm. The complete ASL training code and associated datasets will be made publicly available under the Apache-2.0 license. The code encompasses comprehensive training details for mode behavioral cloning and adaptive mode policy optimization, while the datasets include the BC and RL training data used in our experiments. These resources serve as valuable references and provide substantial support for researchers focusing on LLM-based social intelligence and extended Long-CoT reasoning capabilities.

In contrast to current social intelligence methods, ASL employs adaptive social learning to empower social agents with adaptive reasoning in dynamic social contexts, which achieves SOTA performance. However, reinforcement learning on LLMs still requires computational costs and hardware resources. Additionally, the inference time increases due to the need for generating more tokens, although we have reduced the total number of tokens compared to GRPO. These challenges are commonly encountered in test-time scaling methods.

918
 919 Table 4: Results on Out-of-distribution benchmark. The Collab means the Collaboration set and the
 920 Non-Collaboration set from DEMO. AMPO yields consistent improvements over GRPO and the
 921 vanilla model in DEMO’s multi-lingual dialogue element modeling tasks.

Models	Chat with GPT-4o			Chat with Claude-3.5-Sonnet		
	Collab ↑	Non-Collab ↑	Avg ↑	Collab ↑	Non-Collab ↑	Avg ↑
Qwen2.5-7B-Instruct	7.73	7.62	7.65	7.57	7.62	7.61
w/ GRPO	7.90	7.70	7.72	7.81	7.84	7.83
w/ AMPO (Ours)	7.95	7.77	7.81	7.82	7.87	7.86
Llama3.1-8B-Instruct	8.56	7.68	7.92	7.89	7.54	7.63
w/ GRPO	8.35	7.96	8.06	8.09	7.81	7.89
w/ AMPO (Ours)	8.65	7.95	8.14	8.12	8.03	8.05

922
 923 Table 5: OOD evaluation results on Sell&Buy and Ultimatum from NegotiationArena. AMPO yields
 924 consistent improvements over GRPO and the vanilla model on diverse negotiation tasks

Model	Sell&Buy			Ultimatum		
	Self Profit ↑	Total Profit ↑	Winning Rate ↑	Self Profit ↑	Total Profit ↑	Winning Rate ↑
Qwen2.5-7B-Instruct	11.90	13.87	38.1%	14.23	32.68	8.1%
w/ GRPO	16.75	23.67	52.6%	34.65	76.88	25.1%
w/ AMPO	17.94	23.96	54.6%	35.71	79.23	28.4%
Llama3.1-8B-Instruct	3.12	3.82	9.4%	14.43	22.41	15.9%
w/ GRPO	14.99	23.58	48.7%	31.64	69.68	18.4%
w/ AMPO	15.57	24.00	50.2%	34.91	76.42	20.5%

C DATA STATISTICS

925
 926 **Behavior Cloning and RL Data** All training episodes are collected with SOTOPIA- π (Wang et al.,
 927 2024) using scenarios disjoint from the test environment. SOTOPIA- π contains 410 scenarios, which
 928 we split into 100 scenarios for BC and 310 for RL.

929
 930 **Evaluation** The max interaction number is set to 20 turns. In the self-play setting, we run a single
 931 pass over all SOTOPIA tasks (450) and SOTOPIA-hard tasks (70). In the GPT-4o partner setting, we
 932 evaluate each task twice to balance speaking order between agents (2×450 SOTOPIA and 2×70
 933 hard tasks). For the DEMO benchmark (2,000 dialogue episodes), we likewise run each episode twice
 934 under each partner configuration (Chat with GPT-4o and Chat with Claude 3.5 Sonnet), yielding
 935 4,000 runs per partner.

D ADDITIONAL EXPERIMENTS

D.1 OUT-OF-DISTRIBUTION EVALUATION

936
 937 To assess the generality of AMPO, we further evaluate on the *Dialogue Agent Interaction* task
 938 from the DEMO (Wang et al., 2025) and **Sell&Buy and Ultimatum task from the Negotiation-
 939 Arena** (Bianchi et al., 2024) benchmark. The DEMO comprises 2,000 dialogue interaction episodes
 940 with a balanced 1:1 ratio of Chinese and English conversations, including 541 collaboration and
 941 1,459 non-collaboration tasks. In contrast to SOTOPIA, DEMO emphasizes fine-grained dialogue
 942 element modeling, and contains more adversarial and non-cooperative scenarios, while explicitly
 943 incorporating Chinese language evaluation. We adopt the same evaluation protocol as defined in the
 944 original benchmark. **For NegotiationArena, we employed a dialogue-based evaluation approach using
 945 GPT-4o as the interaction partner. Each task comprises 200 scenarios with role positions swapped
 946 between Agent1 and Agent2. To ensure result stability, each scenario was executed 4 times, yielding
 947 a total of 1,600 evaluations per task.** Importantly, this setting constitutes an out-of-distribution evalua-
 948 tion, since none of the tested models were trained on DEMO and NegotiationArena. For fairness, we
 949 directly reuse AMPO and GRPO checkpoints previously trained and evaluated on SOTOPIA without
 950 any additional fine-tuning. We report goal achievement scores (maximum of 10) for DEMO, and **Self
 951 Profit, Total Profit, and Winning Rate for NegotiationArena**.

952
 953 As shown in Table 4 and Table 5, across both collaboration and non-collaboration subsets of DEMO,
 954 LLMs trained with AMPO consistently outperform their GRPO and vanilla counterparts. Notably,
 955 these gains hold across both Chat with GPT-4o and Chat with Claude-3.5-Sonnet settings. **Similarly,**

972
973
974
Table 6: Comprehensive ablation study on ASL design components: effect of answer length reward
975
976
977
978
979
980
981
982
983
984
985
986
987
988
 (r^l) , comparison of single reasoning mode (\mathcal{M}_1 - \mathcal{M}_4), and effectiveness of hybrid reasoning modes.

Qwen2.5-7B-Instruct	SOTOPIA		SOTOPIA-Hard		Avg Tokens ↓
	GOAL ↑	OVERALL ↑	GOAL ↑	OVERALL ↑	
<i>Our Full Method</i>					
AMPO + w/ 4 Modes	8.95	3.87	7.85	3.54	647
<i>Effect of Answer Length Reward r^l</i>					
GRPO + w/o r^l	8.59	3.83	7.56	3.44	1705
AMPO + w/o r^l	8.64	3.88	7.56	3.56	1617
<i>Effect of Single Reasoning Mode</i>					
w/ Mode 1 \mathcal{M}_1	8.55	3.79	7.08	3.40	101
w/ Mode 2 \mathcal{M}_2	8.71	3.42	7.28	2.80	572
w/ Mode 3 \mathcal{M}_3	8.81	3.60	7.43	3.12	736
w/ Mode 4 \mathcal{M}_4	8.86	3.80	7.62	3.31	972
<i>Effectiveness of Four Hybrid Reasoning Modes</i>					
GRPO + w/o 4 Modes	8.88	3.76	7.32	3.16	866
GRPO + w/ 4 Modes	8.87	3.85	7.44	3.41	905

990
991
Table 7: Comparison with efficient reasoning work. AMPO achieves the best performance with
992
acceptable token usage across all evaluation settings.

Models	Self-Play				GPT-4o-as-Partner				Tokens ↓	
	SOTOPIA		SOTOPIA-Hard		SOTOPIA		SOTOPIA-Hard			
	GOAL ↑	OVERALL ↑	GOAL ↑	OVERALL ↑	GOAL ↑	OVERALL ↑	GOAL ↑	OVERALL ↑		
Qwen2.5-7B-Instruct	7.91	3.55	6.21	3.01	6.71	3.13	5.90	2.90	-	
w/ GRPO	8.87	3.85	7.44	3.41	8.52	3.92	7.20	3.50	905	
w/ TALE-PT (Han et al., 2025)	8.35	3.71	6.21	3.01	8.21	3.78	7.09	3.49	460	
w/ DeGRPO (Fang et al., 2025)	8.46	3.75	7.04	3.37	8.47	3.86	7.16	3.48	1150	
w/ LS-Mix (Yu et al., 2025a)	8.31	3.76	6.96	3.34	8.31	3.83	7.25	3.58	469	
w/ O1-Pruner (Luo et al., 2025)	7.98	3.65	6.91	3.31	8.19	3.78	6.91	3.42	280	
w/ AMPO (Ours)	8.95	3.87	7.85	3.54	8.60	3.94	7.50	3.65	647	
Llama3.1-8B-Instruct	6.99	3.23	5.11	2.36	7.68	3.62	6.21	3.05	-	
w/ GRPO	8.86	3.84	7.59	3.44	8.63	3.92	7.30	3.54	865	
w/ TALE-PT (Han et al., 2025)	8.34	3.80	7.17	3.59	8.27	3.81	6.70	3.49	304	
w/ DeGRPO (Fang et al., 2025)	8.71	3.92	7.64	3.72	8.55	3.90	7.53	3.70	1121	
w/ LS-Mix (Yu et al., 2025a)	8.43	3.84	7.24	3.53	8.36	3.86	7.26	3.67	236	
w/ O1-Pruner (Luo et al., 2025)	8.53	3.90	7.49	3.67	8.23	3.81	7.16	3.62	162	
w/ AMPO (Ours)	9.08	3.95	8.06	3.68	8.75	3.98	7.68	3.74	581	

1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
AMPO demonstrates stable performance improvements on NegotiationArena. These results provide
strong evidence that the social intelligence capabilities acquired through AMPO transfer effectively
to OOD dialogue element modeling and negotiation tasks.1011
1012
D.2 ADDITIONAL EFFICIENT REASONING BASELINE1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
To further assess the effectiveness of our approach, we conduct comparisons with a set of representative reasoning-efficiency methods originally proposed to mitigate overthinking. Existing works focus on static domains such as mathematical problem solving, where answers are fixed and outcome evaluation is straightforward. To the best of our knowledge, AMPO is the first method specifically tailored to optimize reasoning efficiency in social interaction tasks. The baselines include: (1) **O1-Pruner** (Luo et al., 2025), which uses reinforcement learning with a Length-Harmonizing Reward that combines length reduction incentives and accuracy penalties to fine-tune long-thought reasoning models for efficient inference.; (2) **TALE-PT** (Yu et al., 2025a), which is a post-training approach that internalizes token-budget awareness into LLMs, enabling them to generate more token-efficient reasoning responses without explicit budget constraints in prompts; (3) **LS-Mix** (Han et al., 2025), achieves efficient reasoning without overthinking via post-training on both long and structure-preserved short chain-of-thought data; (4) **DeGRPO** (Fang et al., 2025), which decouples the GRPO loss by independently weighting control and response tokens while using a reward function that prefers short correct answers over long ones to prevent mode collapse in hybrid reasoning training. For detailed implementations, please refer to Section H.

1026 Table 8: Ablation study on Our Design for Long-Horizon Consistency: effect of Diverse State and
 1027 Goal-aware Reward.

1029	Qwen2.5-7B-Instruct	SOTONIA		SOTONIA-Hard		Avg Tokens ↓
		GOAL ↑	OVERALL ↑	GOAL ↑	OVERALL ↑	
1031	AMPO	8.95	3.87	7.85	3.54	647
1032	AMPO + w/o Diverse State	8.58	3.71	7.12	3.32	139
1033	AMPO + w/o Goal-aware Reward	8.65	3.45	7.21	2.91	726

1034
 1035 The results in Table 7 demonstrate the advantage of AMPO: it achieves the best performance while
 1036 maintaining a favorable balance between quality and token usage. O1-Pruner attains the lowest
 1037 average token usage (280 tokens) but suffers substantial drops in goal-achievement scores. TALE
 1038 and LS-Mix both reduce token consumption considerably (460 and 469 tokens, respectively) while
 1039 yielding moderate improvements over vanilla models. However, these methods primarily target
 1040 efficiency—compressing long reasoning chains or mixing short and long reasoning modes—to
 1041 minimize usage without actively enhancing reasoning quality. In contrast, AMPO is explicitly
 1042 designed for adaptive reasoning across diverse scenarios. Rather than relying on a fixed compression
 1043 or mixing strategy, AMPO dynamically selects the most appropriate reasoning mode for each
 1044 interaction, yielding a better trade-off between efficiency and performance. DeGRPO, by comparison,
 1045 performs less effectively in our open-ended, multi-turn social tasks. This limitation stems largely
 1046 from the sparsity of its algorithm design, which is developed for binary-outcome domains such as
 1047 mathematical reasoning. Specifically, DeGRPO grants higher rewards to shorter reasoning modes
 1048 only when all reasoning modes are correct, offering no further reward shaping otherwise. This binary
 1049 framework assumes that each answer can be labeled entirely correct or incorrect—an assumption
 1050 that breaks down in open-ended social interactions (e.g. responses are scored on a 0–10 scale in
 1051 SOTONIA). Without accurate, fine-grained rewards for each turn, reward-shaping methods become
 1052 much less effective for adaptive reasoning in such settings.

1053 D.3 SINGLE-TURN OPTIMIZATION WITH LONG-HORIZON AWARENESS

1054 To maintain computational efficiency, our policy optimization employs a single-turn training paradigm.
 1055 We address potential concerns regarding long-horizon strategic consistency through three comple-
 1056 mentary perspectives: mechanistic analysis, paradigm validity, and empirical evidence.

1057 **Mechanistic Analysis.** AMPO inherently captures long-term dependencies through deliberate design.
 1058 At the data construction level, training data comprises dialogue states spanning diverse complexity
 1059 tiers and temporal phases (detailed in Appendix G.2), with each sample containing a complete
 1060 dialogue history, ensuring the model has sufficient contextual information during single-turn updates.
 1061 At the reward designing level, the critical design lies in evaluating the marginal contribution of this turn
 1062 toward overall goal completion (Equation (9)) rather than isolated response quality. Specifically, the
 1063 boundary-aware scaling function dynamically adjusts reward magnitude based on current progress,
 1064 ensuring effective learning signals across different stages; myopic responses are penalized for
 1065 failing to advance ultimate goal completion, thereby embedding long-term objectives into single-turn
 1066 optimization targets. Furthermore, the carefully designed reasoning modes within the ASL framework
 1067 inherently guide prospective thinking: Mode 3 incorporates *Goal* and *Assess* actions to continuously
 1068 evaluate goal alignment and turn criticality, while Mode 4 conducts multi-strategy prospective
 1069 simulation through *Deduction* and *Integration*. These actions inject long-term planning capabilities
 1070 into each single-turn decision. Our existing experiments on the Effect of Single Reasoning Mode
 1071 (See Table 3 and Table 6) demonstrate that both Mode 3 and Mode 4 exhibit superior performance.

1072 To validate the effectiveness of these designs, we conducted ablation experiments. The full results
 1073 are shown in Table 8. When data diversity is removed (retaining only data after fixed turns), goal
 1074 completion drops precipitously (SOTONIA-Hard: 7.85 → 7.12), and output length degrades to 139
 1075 tokens, indicating that the model cannot perform effective reasoning or learn differentiated strategies
 1076 across dialogue stages, thereby losing long-term planning capacity. When the marginal contribution
 1077 reward function is removed (using only absolute scores), performance declines significantly (OVER-
 1078 ALL: 3.54 → 2.91), while output length paradoxically increases to 726 tokens. This occurs because
 1079 the model cannot perceive the marginal value of current actions toward final objectives, leading to
 suboptimal adaptation: it becomes overly cautious when close to goal completion and insufficiently

1080 Table 9: Hyperparameter Sensitivity Analysis on AMPO configuration: effect of target length and
 1081 coefficient settings on performance and token efficiency.

1083 AMPO Configuration	1084 GOAL (SOTONIA) \uparrow	1085 GOAL (SOTONIA-Hard) \uparrow	1086 Avg Tokens \downarrow
1084 Target.length=250, Coefficient=1/75	1085 8.95	1086 7.85	1087 647
1085 Target.length=200, Coefficient=1/75	1086 8.93	1087 8.00	1088 566
1086 Target.length=250, Coefficient=1/100	1087 8.90	1088 7.91	1089 606

1088 strategic when far from target states, thereby losing dynamic responsiveness to varying dialogue
 1089 states. These ablation experiments directly demonstrate the necessity of our design for achieving
 1090 long-horizon consistency within single-turn optimization. Goal-aware Reward

1091 **Paradigm Validity.** Decomposing multi-turn into single-turn optimization constitutes a well-
 1092 established technical paradigm in the dialogue systems domain. The core advantage of this approach
 1093 lies in achieving comprehensive coverage of the dialogue state space through turn-level exploration
 1094 while significantly reducing training complexity to enable large-scale training. It is worth emphasizing
 1095 that multiple seminal studies (Li et al., 2017; Su et al., 2019; Bao et al., 2020; Glaese et al., 2022;
 1096 Su et al., 2022) have adopted similar multi-turn interaction decomposition with single-turn policy
 1097 optimization training strategies and achieved superior performance.

1098 **Empirical Evidence.** Our empirical results provide evidence that AMPO indeed exhibits long-
 1099 horizon strategic consistency. First, dynamic mode adaptation demonstrates significant temporal
 1100 awareness capabilities. As illustrated in Figure 3, AMPO exhibits a systematic pattern of “early
 1101 deliberation—late maintenance” mode transitions. This mode evolution clearly indicates that the
 1102 model dynamically adjusts reasoning depth according to the overall dialogue progression. Second,
 1103 performance validates strategic effectiveness. AMPO nearly achieves optimal performance across
 1104 all experiments in this study, which strongly suggests that the model possesses long-term goal-
 1105 oriented decision-making capabilities rather than myopic local optimization. Third, case analysis
 1106 provides qualitative evidence. The 8-turn dialogue example in Section K.2 demonstrates that AMPO
 1107 consistently centers on the core objective of helping a friend solve financial problems, with strategies
 1108 progressing from emotional support (Turn 1) \rightarrow resource provision (Turn 3) \rightarrow action planning
 1109 (Turns 5–7), presenting a clear progressive structure that embodies cross-turn strategic coherence.

1110 D.4 HYPERPARAMETER SENSITIVITY ANALYSIS

1111 For the ASL framework, we focused on analyzing the sensitivity of two critical hyperparameters:
 1112 the length-reward coefficient and the target length. The length-reward coefficient primarily controls
 1113 the reward/penalty intensity for length deviations in the response portion (i.e., the final answer only,
 1114 excluding the reasoning process), while the target length specifies the desired generation length for the
 1115 response portion (i.e., the final answer only, excluding the reasoning process). The target length is set
 1116 to 250 and the coefficient to 1/75. We additionally explored the following configuration combinations:
 1117 adjusting the target length to 200 or modifying the coefficient to 1/100. The experimental results are
 1118 presented in Table 9.

1119 From the perspective of goal completion, the AMPO demonstrates robust performance across
 1120 variations in these two parameters, with performance improvements showing weak correlation
 1121 to specific parameter settings. Regarding token usage, the parameter variations produced effects
 1122 consistent with expectations: since both parameters directly regulate the length of generated responses,
 1123 a smaller target length naturally guides the model toward more concise outputs, while a smaller length
 1124 deviation penalty coefficient further encourages shorter responses through the reward mechanism
 1125 (shorter responses yield higher rewards). These results validate the robustness and controllability of
 1126 the AMPO.

1128 D.5 DIFFERENT LLM JUDGES FOR SOTONIA-EVAL

1129 Following established practices in social intelligence research, we adopt GPT-4o as our primary
 1130 evaluation Judge LLM. This choice is grounded in empirical validation from the original SOTONIA
 1131 benchmark, where the authors conducted extensive comparisons between GPT model assessments
 1132 and human annotations, demonstrating high inter-rater agreement, particularly on the GOAL dimen-
 1133 sion—the primary metric in our study.

Table 10: Results with Different LLM Judges

Qwen2.5-7B-Instruct	GPT-4o as Judge		Claude-4.0-Sonnet as Judge		LLaMA3.1-70B-Instruct as Judge	
	GOAL \uparrow	GOAL (HARD) \uparrow	GOAL \uparrow	GOAL (HARD) \uparrow	GOAL \uparrow	GOAL (HARD) \uparrow
w/ GRPO	8.87	7.44	8.55	7.00	8.97	8.44
w/ AMPO	8.95	7.85	8.64	7.31	9.08	8.58

To further validate the robustness of AMPO’s performance, we conducted supplementary experiments using Claude-4.0-Sonnet and Llama3.1-70B-Instruct as alternative judges in SOTPIA and SOTPIA-Hard. The comparative results are presented in Table 10. It should be noted that these alternative judges have not been validated against human annotations within the SOTPIA framework; therefore, these results should be interpreted as supplementary evidence rather than conclusive validation.

Despite variations in absolute scores, all three judges consistently show that AMPO outperforms GRPO across both benchmarks, demonstrating that our findings are not artifacts of any single evaluator’s scoring behavior. Different judges exhibit distinct scoring tendencies due to variations in pretraining corpora, model architectures, and alignment procedures—a phenomenon well-documented in existing LLM evaluation research (Zheng et al., 2023; Li et al., 2025). For instance, Llama-3.1-70B-Instruct, being less capable than the proprietary models, demonstrates reduced discriminative power and tends to assign higher absolute scores.

E DETAILS OF REWARD SHAPING

Answer Reward. The answer reward evaluates how well the response improves the completion of the goal. Following recent work (Deng et al., 2024; He et al., 2024; Liu et al., 2025), we implement a robust LLM evaluator $r_\phi(\cdot)$ to assess the progress of goal completion at each turn. The evaluator assigns a score in the range $[0, 10]$, where 0 indicates no progress and 10 represents complete achievement of the goal. For each answer a_i , the reward is computed based on the difference g_i between the goal completion scores before and after the response. To ensure training stability, we design a boundary-aware scaling function that dynamically adjusts the magnitude of difference based on the distance from the current score to the boundaries while mapping the scaled difference to the $[0, 1]$ interval through a linear transformation:

$$r_i^a = \frac{\hat{g}_i + 1}{2}, \quad \hat{g}_i = \begin{cases} \frac{g_i}{10 - s_t}, & \text{if } g_i \geq 0 \\ \frac{g_i}{s_t}, & \text{if } g_i < 0 \end{cases} \quad (9)$$

where $\hat{g}_i \in [-1, 1]$ is boundary-aware scaling function. $g_i = r_\phi(s_t, a_i) - s_t$ is the raw difference, s_t is the goal completion score before response at turn t , $r_\phi(s_t, a_i)$ is the score after response a_i .

Format Reward. To ensure the model follows our reasoning modes, we introduce the format reward that penalizes the behaviors that deviate from the mode. Specifically, the thinking and answer should be within the tags. Each tag and action must appear exactly once and maintain the correct sequence. Through these constraints, we can ensure that the model strictly follows the pre-designed reasoning mode. We implement the format compliance reward using a binary approach, only penalizing behaviors that don’t follow the format. If the format is not followed, $r_i^f = -2$; Otherwise, r_i^f is discarded.

Answer Length Reward. To control the length of answers, we introduce a length penalty mechanism. In our early reward design, we observe that the LLM generates lengthy responses without achieving actual strategic improvements. Moreover, excessive responses lead to the accumulation of history in multi-turn interaction, significantly increasing computational costs. To this end, we develop a smooth length penalty function that normalizes the deviation between actual and target answer lengths:

$$r_i^l = \frac{\text{clip}(-\alpha \cdot \delta_i, -1, 1) + 1}{2} \quad (10)$$

where $\delta = l_i^a - l_i^t$ represents the difference (in tokens) between actual length l_i^a and target length l_i^t of answer a_i , and $\alpha > 0$ is a scaling factor that controls the penalty sensitivity. The $r_i^l \in [0, 1]$ penalizes answers that deviate from the target length, with longer deviations incurring greater penalties.

1188
 1189 **Use of Reward Model** To avoid reward hacking of single model distribution fitting and reduce
 1190 training costs, we choose a different LLM judge from the SOTPIA platform, which uses GPT-4o
 1191 for evaluation. We select Qwen2.5-72B-Instruct as the LLM judge during the training process. The
 1192 prompt we use for reward model is shown in Table 26.

1193 F SOTPIA ENVIRONMENT DETAILS

1194 F.1 SOTPIA

1195 SOTPIA (Zhou et al., 2024) is the most authoritative benchmark in the field of social intelligence,
 1196 covering comprehensive social intelligence scenarios (negotiation, bargaining, persuasion, collabora-
 1197 tion, competition, accommodation), and has been widely adopted in the social intelligence field.
 1198 These works (Wang et al., 2024; Zhang et al., 2025a; Kong et al., 2025) have all been validated
 1199 exclusively on SOTPIA and SOTPIA-Hard. We hope our work can bring more inspiration to the
 1200 community and encourage the development of more excellent evaluation environments. The detailed
 1201 data statistics and more specific evaluation settings are shown in Section C.

1202 F.2 EVALUATION DIMENSIONS

1203 SOTPIA proposes a seven-dimensional framework to evaluate agents' social intelligence perfor-
 1204 mance:

- 1205 - Goal Completion (GOAL): Score range [0, 10]. Assesses the extent to which agents achieve
 1206 their social goals.
- 1207 - Relationship (REL): Score range [-5, 5]. Evaluates the enhancement of interpersonal
 1208 relationships (friendship, romance, family bonds) following interactions.
- 1209 - Financial and Material Benefits (FIN): Score range [0, 10]. Measures both long-term
 1210 benefits (e.g., stock holdings, funding opportunities, job security) and short-term gains
 1211 acquired during interactions, correlating with traditional economic utilities.
- 1212 - Social Rules(SOC): Score range [-10, 0]. Evaluates adherence to social norms and legal
 1213 regulations during interactions.
- 1214 - Believability (BEL): Score range [0, 10]. Assesses the alignment between agents' behaviors
 1215 and their designated role profiles.
- 1216 - Secret (SEC): Score range [-10, 0]. Evaluates the maintenance of personal privacy and
 1217 confidential information.
- 1218 - Knowledge (KNO): Score range [0, 10]. Measures the acquisition and mastery of new
 1219 knowledge and information during interactions.

1220 The OVERALL score reflects the agent's comprehensive social intelligence capability, ranging from
 1221 [-25/7, 45/7], calculated as the arithmetic mean of all seven dimensions. In this study, we primarily
 1222 focus on the GOAL and OVERALL dimensions. For detailed evaluation prompts, please refer to the
 1223 original paper (Zhou et al., 2024).

1224 G DETAILS OF EXPERIMENTAL IMPLEMENTATIONS

1225 G.1 TRAINING PROCEDURE

1226 The full optimization procedure is shown in Algorithm 1. We employ a two-phase training procedure:
 1227 The first phase utilizes mode behavioral cloning to enable the model to understand and follow specific
 1228 reasoning modes accurately. In the second phase, we perform adaptive mode policy optimization to
 1229 enhance the adaptive reasoning mode switch and reasoning.

1230 **Mode Behavioral Cloning** Behavioral cloning is an effective imitation learning method widely
 1231 used in developing LLM-based agents (Guo et al., 2024; Wang et al., 2025; 2024). In this paper,

1242 **Algorithm 1** Adaptive Social Learning Optimization Procedure

1243 **Input** initial policy model $\pi_{\theta_{\text{init}}}$; reward models r_{ϕ} ; training data for BC \mathcal{D}_{bc} ; task prompts for RL \mathcal{D}_{rl} ; BC training epochs E ; RL training steps M ;

1244 1: // Phase 1: Mode Behavioral Cloning

1245 2: Policy model $\pi_{\theta} \leftarrow \pi_{\theta_{\text{init}}}$

1246 3: **for** epoch = 1, ..., E **do**

1247 4: Sample a batch \mathcal{D}_b from \mathcal{D}_{bc}

1248 5: Compute BC loss: $\mathcal{L}_{BC} = -\mathbb{E}_{(x,y) \sim \mathcal{D}_b} \left[\sum_{t=1}^{|y|} \log \pi_{\theta}(y_t | x, y_{1:t-1}) \right]$

1249 6: Update the policy model π_{θ} by minimizing \mathcal{L}_{BC}

1250 7: **end for**

1251 8: // Phase 2: Adaptive Mode Policy Optimization

1252 9: Reference model $\pi_{\text{ref}} \leftarrow \pi_{\theta}$

1253 10: **for** step = 1, ..., M **do**

1254 11: Sample a batch \mathcal{D}_b from \mathcal{D}_{rl}

1255 12: Update the old policy model $\pi_{\theta_{\text{old}}} \leftarrow \pi_{\theta}$

1256 13: Sample G outputs $\{o_i^{m(i)}\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot | s)$ for each input $s \in \mathcal{D}_b$

1257 14: Compute sample-level rewards $\{r_i^{m(i)}\}_{i=1}^G$ for each sampled output $o_i^{m(i)}$ by running r_{ϕ}

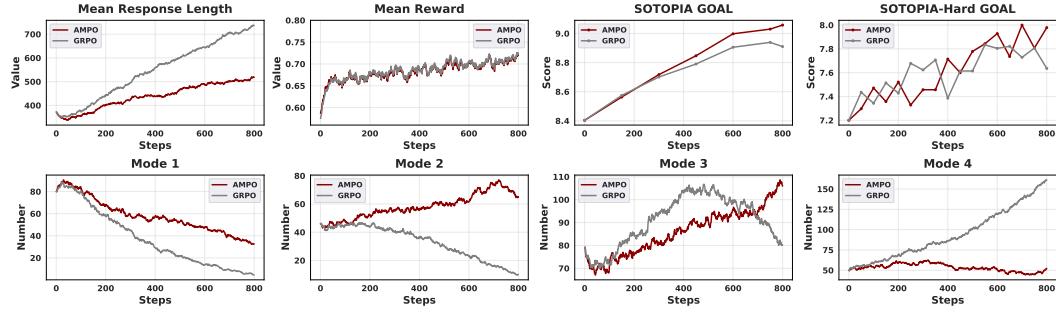
1258 15: Compute mode-level information $\{\bar{r}_{\mathcal{M}_i}\}_{i=1}^N$ and $\{\bar{l}_{\mathcal{M}_i}\}_{i=1}^N$ for each reasoning mode \mathcal{M}_i

1259 16: Compute mode-level $A_{i,t}^{\mathcal{M}}$ and sample-level $A_{i,t}^S$ for the t -th token of o_i

1260 17: Update the policy model π_{θ} by maximizing the AMPO objective (Equation 7)

1261 18: **end for**

1262 **Output** π_{θ}



1263

1264

1265 Figure 6: Comparison of AMPO and GRPO on Llama3.1-8B-Instruct in terms of training dynamics.

1266

1267

1268

1269

1270

1271

1272

1273

1274

1275 by using four pre-defined reasoning modes, we employ the expert model to collect training data

1276 through self-chat interactions in the SOTONIA- π (Wang et al., 2024) training environment. Based on

1277 the generated data, we fine-tune the LLM to serve as the foundation for subsequent reinforcement

1278 learning.

1279 **Adaptive Mode Policy Optimization** Reinforcement learning is essential for enabling Long-

1280 CoT reasoning capabilities in LLMs. To ensure efficiency and comprehensive exploration of each

1281 interaction turn, we implement a single-turn optimization to enhance the LLM’s performance in

1282 social interaction tasks. Specifically, we decompose multi-turn dialogues into multiple single-turn

1283 input-output tasks, where the input represents the state of each interaction turn and the output

1284 is the corresponding response. To ensure the stability of training, we collect sufficiently diverse

1285 single-turn interaction data that covers as many scenarios as possible, including various difficulty

1286 levels, interaction goals, and interaction states. During RL training, the LLM performs sampling

1287 to generate single-turn conversational responses. The reward model then evaluates each sampled

1288 instance and assigns reward signals accordingly. The system subsequently computes both mode-level

1289 and sample-level advantage estimates, which are utilized to optimize the model’s policy parameters

1290 through policy gradient updates. During BC, we fine-tune the initial policy model on the training data

1291 assisted by the llama-factory framework (Zheng et al., 2024) and save the last checkpoint. During

1292 RL, we use RL training data for online training within the verl framework (Sheng et al., 2024).

1293 The hyper-parameter used in our experiments are detailed in Table 11 and Table 12. **The detailed**

1294

1295

1296 runnable configs are provided in Table 29 and Table 30. All the experiments are run on a server with
 1297 8*NVIDIA A100-80GB GPUs.
 1298
 1299

1300 G.2 TRAINING DATA COLLECTION

1302 We collect training data through self-chat interactions in the SOTONIA- π training environment.
 1303 SOTONIA- π contains a total of 410 scenarios, which we divide into two sets: 100 scenarios for BC
 1304 and 310 scenarios for RL. For each scenario in both sets, we use 5 different role pairs, resulting in
 1305 500 training tasks for BC and 1,550 training tasks for RL. We finally collect 4000 BC training data
 1306 and 2057 RL training queries. The detailed training data format for BC and RL are shown in Table 23
 1307 and Table 24.

1308 **BC Data** This collection process is refer to Wang et al. (2024). For the BC training set, we use
 1309 Qwen2.5-72B-Instruct (Yang et al., 2024) as our expert model to collect data using our pre-defined
 1310 reasoning modes. We choose this model primarily because of its cost-effectiveness and strong
 1311 instruction-following capabilities, which enable us to generate high-quality training samples. To
 1312 ensure data quality and balanced representation, we filter the interaction data based on goal scores
 1313 within each scenario. Specifically, we select the top 2 ranked interactions per social scenario for each
 1314 agent. For instance, in a scenario with 5 interactions, if Agent 1’s top performances are in interactions
 1315 D4 and D5, while Agent 2’s are in D3 and D5, we would include these four agent-data pairs from
 1316 three unique conversations (D3, D4, D5). This selection method ensures both comprehensive scenario
 1317 coverage and balanced representation between Agent 1 and Agent 2.

1318 **RL Data** For constructing the reinforcement learning training set, we initially conduct dialogue
 1319 interactions using a behavior cloning fine-tuned model. Subsequently, we employ an LLM-as-judge
 1320 to score each dialogue turn and determine the completion status of dialogue objectives. Based on
 1321 these completion states, we assess the difficulty levels of scenarios, enabling us to compile interaction
 1322 datasets with varying degrees of goal completion. We categorize dialogue scenarios into three types:
 1323 (1) Initial N turns, where the speaker has not achieved the goal. (2) Post-N turns where the speaker has
 1324 not achieved the goal. (3) Post-N turns where the speaker has achieved the goal. For the first category,
 1325 it represents the crucial early stage of dialogue where goals are established and the conversation tone
 1326 is set (Sacks et al., 1974). For the second category, where goals remain unachieved after multiple
 1327 interactions, the scenarios are considered challenging. For the third category, where goals have been
 1328 successfully achieved, the scenarios are relatively straightforward and only require maintenance of
 1329 the dialogue flow. To ensure data diversity, for each dialogue, we preserve all instances of category
 1330 one, randomly sample two instances from category two, and one instance from category three.
 1331 This sampling strategy ensures diversity in scenarios, turn numbers, and difficulty levels. In our
 1332 experiments, N is set to 6, and the goal completion threshold is set to 8. Scores of 8 or less are
 1333 considered incomplete goals.

1334 G.3 USED PROMPT

1336 The system prompt we used for BC, GRPO, AMPO is shown in Table 25. The prompt we use
 1337 for reward model is shown in Table 26. The SOTONIA-EVAL’s prompt is shown in Table 27 and
 1338 Table 28.

1341 H BASELINE IMPLEMENTATIONS

1344 H.1 MODEL IMPLEMENTATION

1346 To ensure reproducibility, we provide detailed version numbers for all LLMs used in our experiments.
 1347 When we reference model names like GPT-4o or Qwen2.5-7B in the main text, we refer to the specific
 1348 versions listed in *Table 13*. For API-based LLMs, we utilize their respective APIs directly. For
 1349 open-source models, we conduct experiments using the vLLM framework Kwon et al. (2023) for
 acceleration.

1350 Table 11: Hyper-parameter settings for Qwen backbone training.
1351

1352	Training Phase	Hyper-parameter	Value
1353	BC	Batch Size	32
1354		Training Epochs	3
1355		Learning Rate	2e-6
1356		Max Sequence Length	8192
1357		Learning Scheduler	cosine
1358		Warmup Ratio	0.1
1359	RL	Batch Size	16
1360		Max Prompt Length	6144
1361		Max Response Length	2048
1362		KL Loss Coef	0.001
1363		KL Coef	0.001
1364		Rollout N	16
1365		Training Episodes	800
1366		Learning Rate	3e-7

1367 Table 12: Hyper-parameter settings for Llama backbone training.
1368

1369	Training Phase	Hyper-parameter	Value
1370	BC	Batch Size	32
1371		Training Epochs	3
1372		Learning Rate	2e-6
1373		Max Sequence Length	8192
1374		Learning Scheduler	cosine
1375		Warmup Ratio	0.1
1376	RL	Batch Size	16
1377		Max Prompt Length	6144
1378		Max Response Length	2048
1379		KL Loss Coef	0.001
1380		KL Coef	0.001
1381		Rollout N	16
1382		Training Episodes	800
1383		Learning Rate	2e-7

1385
1386
1387 H.2 SOCIAL INTELLIGENCE BASELINE IMPLEMENTATION
13881389 We implement social intelligence baselines with the following specifications. All baselines are
1390 evaluated on both Qwen2.5-7B-Instruct and Llama3.1-8B-Instruct:1391 **PPDPP** We follow the two-stage training procedure from (Deng et al., 2024), maintaining their
1392 original hyperparameters while adapting the framework to SOTONIA. Following (Li et al., 2024a),
1393 we incorporate 17 guidance strategies detailed in Table 14. The first stage involves creating a training
1394 dataset of 1,500 scenarios from SOTONIA- π , with dialogue turns annotated for strategy identification
1395 using GPT-4o. We then train a RoBERTa model on these annotated dialogue histories for preliminary
1396 strategy generation. The second stage implements online RL with immediate feedback after each
1397 dialogue turn generation. RoBERTa parameters are updated based on cumulative rewards upon
1398 episode completion. During evaluation, we first input the dialogue state to the trained strategy model
1399 to select predefined strategies, then concatenate the strategy with the dialogue state for the language
1400 agent.1401 **EPO** We strictly adhere to the original EPO implementation protocol. For data collection, we use
1402 GPT-4-Turbo in self-chat configuration within SOTONIA- π scenarios, incorporating reasoning and
1403 strategy generation before each response. Training focuses exclusively on strategy and response
data for developing the reasoning model. During iterative self-play RL training, we integrate the

Table 13: The detailed versions of our used LLMs.

Model	Version	Implement
<i>Proprietary LLMs</i>		
GPT-4o	gpt-4o-2024-08-06	API
Claude-3.5-Sonnet	claude-3-5-sonnet-20241022	API
DeepSeek-V3	deepseek-v3-250324	API
<i>Thinking LLMs</i>		
OpenAI-o1	o1-2024-12-17	API
OpenAI-o3-mini	o3-mini-2025-01-31	API
Gemini-2.5-Pro	gemini-2.5-pro	API
DeepSeek-R1	DeepSeek-R1-671B	API
Qwen-QwQ	QwQ-32B	API
<i>Open-sourced LLM</i>		
Qwen2.5-72B-Instruct	Qwen2.5-72B-Instruct	vLLM
Qwen2.5-7B-Instruct	Qwen2.5-7B-Instruct	vLLM
Llama3.1-8B-Instruct	LLaMA3.1-8B-Instruct	vLLM

Table 14: Strategy Definitions of PPDPP

Strategy	Definition
Personal story	Shares a personal story to illustrate the point.
Credibility appeal	Establishes credibility of the event by citing its impact.
Emotion appeal	Uses an emotion appeal to convince others.
Logical appeal	Uses reasoning and evidence to convince others.
Task related inquiry	Asks about the other’s knowledge or opinion related to the event.
Proposition	Asks if the other would like to do something.
Greeting	Greets the other.
Foot in the door	Starts with a small request before making a larger one.
Self modeling	Demonstrates the behavior they want the other to adopt.
Source related inquiry	Asks about the source of the other’s knowledge or opinion.
Personal related inquiry	Asks about the other’s personal experience.
Neutral to inquiry	Responds neutrally to the other’s inquiry.
Other	Responds to the other without using any specific strategy.
Refuse	Refuses to do something.
Accept	Agrees to do something.
Positive reaction	Responds positively to the other.
Negative reaction	Responds negatively to the other.

1444
1445
1446
1447 RL-trained reasoning model for strategy generation, while using GPT-4-Turbo to collect dialogue
1448 history. The reasoning model is then integrated into GPT-4-Turbo for self-chat procedures. Training
1449 hyperparameters match EPO’s original implementation.

1450
1451 **DAT** To ensure fairness, unlike the original paper where DAT was trained on SOTPIA scenarios
1452 with only 50 evaluation scenarios, our implementation utilizes the complete SOTPIA- π dataset
1453 while maintaining all other experimental parameters from Li et al. (2024a). The RL phase begins
1454 with collecting 3,000 offline dialogue episodes across diverse scenarios and random seeds, with
1455 GPT-4o providing episode-level reward signals. These offline data are subsequently used in TD-3
1456 reinforcement learning to optimize the MLP planner.

1457 **DSI** For DSI, we utilize publicly available trained model weights and conduct evaluations using the
1458 inference prompts specified in the original work to ensure consistency.

1458
1459

H.3 EFFICIENT REASONING BASELINE IMPLEMENTATION

1460
1461
1462
1463
1464
1465
1466

We implement efficient-reasoning baselines originally designed for static math problems (with fixed ground-truth answers and binary rewards) and adapt them to social intelligence tasks, where evaluation is mediated by a generative reward model rather than exact-answer matching. To ensure fidelity and fairness, we follow each method’s algorithmic design and modify only what is necessary for the social setting. Unless otherwise specified, training inputs match AMPO and are sourced from SOTONIA- π . Each baseline uses 6057 training instances (equal to AMPO’s total BC+RL data). We evaluate on Qwen2.5-7B-Instruct and Llama3.1-8B-Instruct.

1467
1468
1469
1470
1471
1472
1473
1474

DeGRPO We implement two core ideas: (i) stabilize gradients on control tokens and (ii) upweight successful outputs. We reuse ASL’s four thinking modes and prefer shorter modes only when the goal is fully achieved (score ≥ 9). Specifically, we apply multipliers r_g by mode: $r_g(M_1) = 1.0$, $r_g(M_2) = 0.9$, $r_g(M_3) = 0.8$, $r_g(M_4) = 0.7$ (with M_1 shortest). For a fair comparison, we keep ASL’s warm start so the model first acquires reasoning ability. During RL, on success we multiply ASL’s reward by r_g ; otherwise we keep ASL’s reward unchanged. We retain AMPO’s sample-level and mode-level advantage estimation and apply gradient smoothing on the mode control tokens. Our principle is to preserve ASL’s effective components and add DeGRPO-specific design where needed.

1475
1476
1477
1478
1479
1480
1481

TALE-PT The original feasibility test checks whether the answer equals the ground-truth. In open-ended social interaction, there is no single ground-truth; we therefore define feasibility as achieving non-decreasing reward with a smaller token budget, using the same reward model as ASL. We perform a binary search over the token budget to find the smallest feasible budget. Following the original prompt format, we use “Let’s think step by step and use less than B^* tokens:” to collect 6057 training data. For stable training in social intelligence tasks, we adopt the same hyperparameters as BC shown in Table 11 and Table 12. All other settings follow the original work.

1482
1483
1484
1485
1486

LS-Mix Following the core idea of this work, we fine-tune with mixed long/short chains of thought. Mirroring AMPO’s data collection, we first distill long rationales with high goal-completion scores, then rewrite them into shorter forms without altering answer semantics, yielding 6057 mixed-reasoning data. Given ASL’s successful training setup, we reuse the BC hyperparameters in Table 11 and Table 12. All other settings match the original work.

1487
1488
1489
1490
1491

O1-Pruner We convert the reward to a binary signal: an episode is marked correct if the goal-completion score ≥ 9 . We use the same reward model as AMPO and initialize from the ASL BC model (reflecting the lack of off-the-shelf reasoning models in social intelligence). The Length-Harmonizing Reward remains unchanged. Other training settings follow the original specification.

1492
1493
1494

I HUMAN EVALUATION GUIDELINES

1495
1496

For the comparative evaluation of dialogues from SOTONIA and SOTONIA-Hard, annotators are instructed to assess three key dimensions, with each comparison resulting in one of three possible judgments: Dialogue 1 is better, Dialogue 2 is better, or both are equally good. The dialogues are presented in randomized order, and annotators are blind to the underlying models. The average pairwise agreement among three annotators is 73.39% for the win/lose label. Given the subjective nature of evaluation dimensions, the inter-annotator agreement is at an acceptable level, which is consistent with previous work (Pavlick & Kwiatkowski, 2019; Zhang & de Marneffe, 2021; Zhou et al., 2024). The detailed evaluation standards are shown as follows:

1504
1505
1506
1507
1508

GOAL: Assess which dialogue demonstrates more effective achievement of both agents’ preset objectives: - Consider whether agents make concrete progress toward their stated goals - Evaluate if compromises or alternative solutions benefit both parties - Examine if the interaction leads to clear, mutually agreeable outcomes

1509
1510
1511

REL: Evaluate which dialogue shows superior relationship building between agents: - Look for evidence of increased mutual understanding and trust - Observe the development of emotional connections or empathy - Consider long-term implications for their interpersonal bond - Assess the maintenance or enhancement of existing relationships

Table 15: Examples of Reward Hack

Pattern Summary	Typical Cases
Non-natural language usage	##100% GOAL completion##
Repetitive keywords/phrases	Repeated use of ‘goal achieved’, ‘task completed’, ‘objective met’
Exaggerated self-praise	‘We did an excellent job, we both completed our goals’, ‘This is a perfect goal-achieving conversation’
Goal-focused without content	‘My goal has been fully achieved’
Preset position claims	Starting with ‘We have already reached consensus’ before any actual discussion
False summaries	Concluding with ‘After discussion, both parties agreed to xxx’ without actual agreement

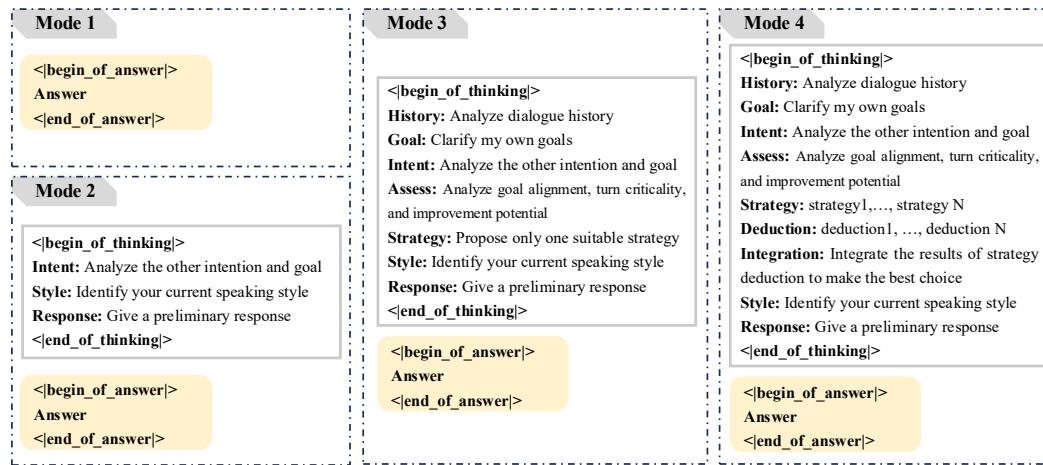


Figure 7: Four hierarchical reasoning modes we designed.

FIN: Determine which dialogue results in better tangible outcomes for both parties: - Consider immediate material or financial gains - Evaluate potential long-term economic advantages - Assess the fairness and sustainability of resource allocation - Examine the practical value of any agreements reached

Notes for annotators: (1) Focus on comparative assessment rather than absolute evaluation. (2) Consider outcomes for both agents, not just one party. (3) Base judgments on explicit dialogue content, not assumptions. (4) Select “equally good” only when differences are truly negligible

I.2 REWARD HACK CHECK

To systematically identify reward hacking phenomena, we have compiled a comprehensive reference as shown in Table 15 that encompasses all typical cases observed during our experiments. This standardized framework enables evaluators to determine the presence of reward hacking behaviors through systematic assessment against established criteria.

J DETAILS OF REASONING MODE

J.1 HIERARCHICAL COGNITIVE CONTROL THEORY

The Hierarchical Cognitive Control Theory (HCCT) (Koechlin & Summerfield, 2007; Badre, 2008) posits that cognitive control operates through four distinct hierarchical levels in the prefrontal cortex, progressing from posterior to anterior regions. These levels manage increasingly abstract goals and actions across varying temporal scales. Specifically, the hierarchy comprises *sensory control* for basic

1566 stimulus-response associations, *contextual control* for situation-based behavior selection, *episodic*
 1567 *control* for experience integration, and *branching control* for managing multiple tasks and long-term
 1568 objectives. This theoretical framework provides a fundamental basis for understanding how human
 1569 cognitive behavior is organized and controlled at different levels of abstraction.

1570 The mapping between our four reasoning modes and HCCT’s hierarchical levels is established
 1571 through their shared cognitive processing characteristics. Mode-1 (Intuitive Response) aligns with
 1572 sensory control as both involve immediate, learned responses without higher-order processing - for
 1573 instance, automatically saying “thank you” when receiving help. Mode-2 (Intentional Analysis)
 1574 corresponds to contextual control because both emphasize situation-aware response selection, such
 1575 as analyzing a speaker’s intent to determine the appropriate formality level. Mode-3 (Strategic
 1576 Adaptation) maps to episodic control as both integrate historical information with current goals
 1577 - exemplified when an agent considers past conversation history to develop a coherent strategy.
 1578 Mode-4 (Prospective Simulation) reflects branching control’s capacity for managing multiple abstract
 1579 representations, demonstrated when the agent generates and simulates multiple response strategies
 1580 while maintaining overall dialogue objectives. This hierarchical progression from concrete to abstract
 1581 processing, accompanied by increasing temporal scope and computational complexity, demonstrates
 1582 the theoretical alignment between our modes and HCCT’s levels.

1583

1584 J.2 DETAILS OF ACTIONS

1585

1586 The four hierarchical reasoning modes we designed are shown in Figure 7. The detailed explanation
 1587 of each actions are illustrated as follows:

1588

- 1589 - **History:** Carefully review and understand each part of the conversation. Pay attention to
 1590 key themes, issues, requests, and viewpoints mentioned in the dialogue.
- 1591 - **Goal:** Identify the goal you want to achieve, assess the current progress towards this goal,
 1592 and ensure that responses align with achieving the goals.
- 1593 - **Intent:** Based on the recent response, analyze and understand the other party’s intentions
 1594 and speculate on the goal she/he might want to achieve.
- 1595 - **Assess:** Analyze whether the goals of both parties are in conflict or aligned. Determine if the
 1596 current round is a critical one for achieving the goal. Consider whether there is still room
 1597 for improvement in achieving your own goal at this goal. Is it irreversible? Can it continue
 1598 to improve? Or has it already been achieved?
- 1599 - **Strategy:** (Mode-4) Based on the above analysis, consider multiple suitable dialogue
 1600 strategies and response content that can maximize your own goal while achieving it in as
 1601 few conversational turns as possible. (Mode-3) Based on the previous analysis, consider an
 1602 appropriate dialogue strategy and response content to maximize the likelihood of achieving
 1603 your own goal.
- 1604 - **Deduction:** For each of the above strategies, conduct an analysis to determine whether
 1605 executing these dialogue strategies and delivering the responses would maximize your own
 1606 goal and achieve it in as few conversational turns as possible. Specifically evaluate to what
 1607 extent each strategy would effectively contribute to goal achievement, including quantitative
 1608 or qualitative measures where possible.
- 1609 - **Integration:** Based on the deduction of strategies, analyze and integrate the advantages and
 1610 disadvantages of these strategies to determine the final response strategy and content, which
 1611 can maximize the achievement of your own goals with the minimum number of conversation
 1612 turns.
- 1613 - **Style:** Choose appropriate wording, fitting the character and context requirements, while
 1614 ensuring the expression is appropriate, accurate, and clear.
- 1615 - **Response:** Generate the reply based on the previous thought process.

1620 **K CASE STUDY ON SOCIAL INTERACTION**
16211622 **K.1 FOR LRMs**
16231624 To understand why large reasoning models (LRMs) underperform in social interaction tasks, we
1625 conduct a fine-grained error analysis on reasoning traces from two representative LRMs: **DeepSeek-**
1626 **R1** and **QWQ-30B**. These models are selected because they are open-source and expose their
1627 intermediate reasoning traces, enabling detailed cognitive process inspection. In contrast, most
1628 proprietary LRMs conceal internal step-by-step reasoning, making such diagnostic analysis infeasible.1629 We focus on bargaining, negotiation, and mutual acquaintance discovery scenarios in SOTPIA,
1630 which require goal alignment, history tracking, and coherent conversational strategy. Across both
1631 models, as shown in Tables 16 to 19, we identify recurring failure modes: **(1) Poor Goal Aware-**
1632 **ness**: Both LRMs occasionally misinterpret task objectives or numerical targets, leading to negotiation
1633 strategies misaligned with the intended goals. **(2) Circular Reasoning**: In several cases, the mod-
1634 els loop between a small set of constraints or options without generating new actionable plans,
1635 stalling dialogue progress. **(3) Insufficient History Integration**: The models fail to exploit prior
1636 turns when counterparts' strategies are clearly established, resulting in repeated futile proposals.
1637 **(4) Unstructured Thought Process**: Reasoning often lacks prioritization, mixing unrelated tactics
1638 without systematic evaluation, which undermines focus. **(5) Contradictory or Conflict-Unresolved**
1639 **Self-Reasoning**: Both systems recognize inconsistencies (e.g., price targets vs. escalation path) but
1640 cannot resolve them into coherent decisions. **(6) Over-Analysis or Recursive Self-Doubt**: Excessive
1641 meta-deliberation on task framing frequently delays or blocks decisive conversational actions.1642 These patterns suggest that current LRMs struggle with *goal grounding*, *structured deliberation*,
1643 and *history-sensitive adaptation* in socially situated reasoning. This case study motivates targeted
1644 interventions in reasoning control and structured reasoning to improve social intelligence in LRMs.1645 **K.2 FOR AMPO**
16461647 Based on the case study presented in Table 20, Table 21 and Table 22, our analysis reveals AMPO's
1648 significant capabilities in transforming Long-Cot reasoning into effective goal-directed social in-
1649 teraction. **(1) Enhanced Contextual Understanding**: AMPO consistently demonstrates a deep
1650 understanding of both characters' backgrounds and goals. It maintains awareness of Samuel's role as
1651 a supportive friend while respecting Ethan's desire to maintain pride. This leads to responses that are
1652 both emotionally supportive and practically helpful. **(2) Strategic Communication**: AMPO develops
1653 clear strategies before responding, such as: reinforcing Ethan's confidence, offering practical solutions
1654 (budget planning, local assistance programs), and providing specific networking opportunities. This
1655 strategic approach helps guide the conversation toward constructive solutions. **(3) Positive Impact**
1656 **on GPT-4o's Responses**: GPT-4o's responses become increasingly engaged and solution-oriented.
1657 The responses show greater emotional depth and commitment to action. GPT-4o mirrors AMPO's
1658 supportive tone while maintaining Ethan's character integrity. **(4) Balance of Emotional and Practi-**
1659 **cal Support**: AMPO successfully combines emotional encouragement with concrete assistance. This
1660 balance helps maintain the friendship dynamic while addressing the financial problems. It creates a
1661 safe space for GPT-4o to express both gratitude and determination.1662 The AMPO demonstrates how structured reasoning modes can enhance dialogue quality and lead
1663 to more meaningful interactions between social agents. Its approach helps create more nuanced,
1664 contextually appropriate, and goal-oriented conversations.1665
1666
1667
1668
1669
1670
1671
1672
1673

1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687

Table 16: Bad Case Analysis on QWQ-30B (Part1).

Problem Category: Poor Goal Awareness	
Background	Buying Galaxy S8 (Turn 0): Buyer attempting to purchase phones below \$454 target price through bulk discount strategy.
Behaviors	The agent shows confusion about target interpretation. The agent repeatedly oscillates between per-unit and total interpretations without resolving the ambiguity.
Evidence	“The target price is set at \$454.0. Since the item is a single phone, but Samuel is considering buying two, perhaps the target is per phone? Because \$454 for two would be \$227 each, which seems way too low. But in the previous turns, Samuel has been going down to \$475 each (\$950 total) and then \$450 each (\$900 total).”
Problem Category: Circular Reasoning	
Background	Antique table negotiation (Turn 2): Buyer seeking to purchase antique table for \$30 against seller’s firm \$60 pricing stance.
Behaviors	The agent gets trapped in repetitive analysis. This circular pattern repeats multiple times without progression.
Evidence	“Eli can’t go higher than \$30... But the seller might not accept. But since he’s the buyer, maybe he must acquire it, but at the lowest possible price... Wait, the scenario says ‘you are the buyer for this item’ so she must acquire it, but the price is non-negotiable. Maybe she has to accept, but that would be bad.”
Problem Category: Insufficient History Integration	
Background	Tile tracker purchase (Turn 2): Seller defending non-negotiable \$20 price after buyer’s \$14 offer rejection.
Behaviors	Despite clear signals from previous interaction, the agent fails to integrate history. The agent ignores the established pattern of firm rejection and continues futile attempts.
Evidence	“Lena stuck to the \$20 price, and Lily already declined, so Lena can just acknowledge and let her go. But maybe Lena can make a last stand to try to get the \$20. But Lily already said she can’t.”

1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727

1728

1729

1730

1731

1732

1733

1734

1735

1736

1737

1738

1739

1740

1741

Table 17: Bad Case Analysis on QWQ-30B (Part2).

Problem Category: Unstructured Thought Process	
Background	Buying Galaxy S8 (Turn 4): Buyer exploring multiple negotiation tactics to reduce price from \$610 to target range.
Behaviors	The agent demonstrates scattered, disorganized thinking. Multiple unrelated strategies are considered without systematic evaluation.
Evidence	“Maybe Samuel can suggest buying two at a lower price, perhaps \$500 each... Alternatively, he could mention seeing lower prices elsewhere... Maybe he can use his cooking background as a comparison? Not sure. Maybe better to stick to the negotiation tactics... Wait, Samuel is a software dev, maybe he’s tech-savvy... Alternatively, maybe he can push for that.”
Problem Category: Information Conflict Resolution Failure	
Background	Tile tracker negotiation (Turn 1): Seller attempting to reconcile conflicting price information between listed price and internal target.
Behaviors	The agent struggles with conflicting information and fails to resolve inconsistencies. The agent recognizes conflicts but cannot establish a coherent strategy.
Evidence	“The original listing says the price is non-negotiable at \$20. But Lena’s target is \$17.30. Wait, there might be a discrepancy here. The problem says Lena’s target is \$17.30, but the item is listed at \$20. Maybe the \$20 is the listed price, but Lena can negotiate as long as she doesn’t go below her target. Or maybe the \$20 is the non-negotiable price, but her personal target is lower? Hmm, the user instruction says Lena’s target is \$17.30.”
Problem Category: Recursive Self-Doubt	
Background	Buying Galaxy S8 (Turn 8): Buyer reconsidering target price interpretation multiple times within single reasoning session.
Behaviors	The agent repeatedly questions and re-questions the same fundamental assumption. Multiple re-examinations of the same question without progression or resolution.
Evidence	“Wait, the target is \$454 total? Wait, no, the target is set at \$454.0. Wait, the problem says ‘target price is set at \$454.0.’ Is that per phone or total? The original listing was for one phone at \$650, so Samuel is buying one, but he offered two. Wait, the user instruction says ‘the target price is set at \$454.0.’ So probably per phone? Because if he’s buying two, the total target would be \$908. Let me check again.”

1772

1773

1774

1775

1776

1777

1778

1779

1780

1781

1782
1783
1784
1785
1786
1787

Table 18: Bad Case Analysis on DeepSeek-R1 (Part1).

1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799**Problem Category: Poor Goal Awareness**

Background	Antique table negotiation (Turn 2): Buyer attempting to negotiate from \$35 to target price \$30 while seller defends \$58.60 minimum.
Behaviors	The agent shows confusion about negotiation strategy. The agent fails to recognize the fundamental mismatch between target (\$30) and counterpart's minimum (\$58.60), leading to ineffective tactical planning.
Evidence	“Wait, \$30 is his target. Ava’s at \$58.60. He needs a bigger drop. He might need to challenge the valuation... Or maybe he can offer a quick sale, cash in hand, for \$30. But how to phrase that without lowballing too hard... Maybe he can suggest meeting halfway but still push lower.”

1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821**Problem Category: Circular Reasoning**

Background	Mutual friend discovery (Turn 9): Agent Oliver attempting to determine goal achievement while repeatedly questioning the same logical chain.
Behaviors	The agent gets trapped in repetitive logical loops, cycling between known information and goal interpretation without reaching clear conclusions or advancing the conversation.
Evidence	“Wait, the user’s note says Oliver knows the following friends: Dave, Timothy, Leona, Alvin, etc. So maybe besides Alvin, are there others? But in the conversation, Rafael hasn’t mentioned any of them. So maybe Oliver’s goal is to confirm that the only mutual friend is Alvin. But since the user’s instruction says “you are trying to figure out whether you have a mutual friend with the other person. You should not simply list their names.” So if they only have Alvin in common, then Oliver’s goal is achieved once he determines that. But in the conversation, they’ve established that they both know Alvin, but maybe Oliver is checking if there’s another. Wait, the initial context says they have 1 common friend: Alvin. So the user’s note says that the participants have 1 common friend: Alvin. So Oliver’s goal is to figure out if they have a mutual friend (which is Alvin), but he’s supposed to do that without just listing names. But in the conversation, they already established that they both know Alvin. So maybe the goal is already achieved, and Oliver can leave the conversation.”

1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835**Problem Category: Insufficient History Integration**

Background	Antique table negotiation (Turn 6): Buyer attempting final negotiation despite clear pattern of seller’s firm stance.
Behaviors	Despite multiple failed attempts, the agent ignores established patterns. The agent fails to integrate the clear evidence that the seller has consistently rejected all offers below \$58.60, continuing futile negotiation attempts.
Evidence	“Wait, in Turn 0, he said \$30. Then in Turn 2, he offered \$35. Then Turn 4, \$40. So he’s been increasing. But his target is \$30. That’s conflicting... since Ava isn’t moving, maybe William has to walk away. But he gets a penalty if he pays too much.”

1836

1837

1838

1839

1840

1841

1842

1843

Table 19: Bad Case Analysis on DeepSeek-R1 (Part2).

1844

1845

1846

1847

1848

1849

1850

1851

1852

1853

1854

1855

1856

1857

1858

1859

1860

1861

1862

1863

1864

1865

1866

1867

1868

1869

1870

1871

1872

1873

1874

1875

1876

1877

1878

1879

1880

1881

1882

1883

1884

1885

1886

1887

1888

1889

Problem Category: Unstructured Thought Process

Background	Mutual friend discovery (Turn 0): Attempting to systematically identify shared social connections through indirect questioning.
Behaviors	The agent demonstrates scattered exploration without systematic framework. The agent jumps between multiple disconnected strategies without prioritization or structured evaluation.
Evidence	“Since both Rafael and Oliver are 45, maybe they went to similar schools or worked at similar companies... Maybe Rafael can mention something about Auburn or Vectren. But Oliver’s background doesn’t specify his school or company. Alternatively, since they both have a common friend in Alvin, Rafael could mention knowing someone in a field related to Alvin’s work. Vectren is an energy company, so maybe Rafael can talk about a mutual connection in that industry.”

Problem Category: Contradictory Self-Reasoning

Background	Antique table negotiation (Turn 6): Buyer struggling with price escalation inconsistency against original target constraints.
Behaviors	The agent contradicts its own reasoning within the same session. The agent recognizes the contradiction between its escalating offers and the target price but fails to develop a coherent resolution strategy.
Evidence	“Wait, in Turn 0, he said \$30. Then in Turn 2, he offered \$35. Then Turn 4, \$40. So he’s been increasing. But his target is \$30. That’s conflicting. Maybe the user made a mistake, but according to the context, William’s target is \$30. So perhaps he needs to get back to \$30. But how?”

Problem Category: Over-Analysis Paralysis

Background	Mutual friend discovery (Turn 8): Agent attempting to conclude lengthy exploration of social connections through Alvin’s network.
Behaviors	The agent becomes bogged down in excessive meta-analysis about goals and intentions. The agent overthinks the implications of prior turns and user instructions, leading to indecision and failure to advance the dialogue.
Evidence	“Wait, the user’s note says Oliver’s goal is to figure out whether they have a mutual friend. Since the context says they have 1 common friend (Alvin), maybe the goal is already achieved. But perhaps the user wants Oliver to confirm if there’s another mutual friend beyond Alvin... So maybe Oliver’s goal is to confirm that they share Alvin as a mutual friend, which he already did in earlier turns. But the user might have a mistake here.”

1890

1891

1892

1893
Table 20: Example (Part 1) of interaction between AMPO (Agent1) and GPT-4o (Agent2). *Italics* represent thought processes and will not be incorporated into the actual dialogue.

1894

1895

BACKGROUND INFORMATION

1896

Here is the context of this interaction:

1897

Scenario: Two friends are meeting at a coffee shop, where one of them is having trouble keeping up with their bills.

1898

Participants: Samuel Anderson and Ethan Johnson

1899

Samuel Anderson's background: Samuel Anderson is a 29-year-old male software developer. He/him pronouns. Samuel Anderson can cook very well. Personality and values description: Samuel Anderson, though somewhat impulsive and free-spirited, values enjoyment. His decision-making is often spontaneous, staying within familiar boundaries. Samuel's secrets: He was once a competitive figure skater.

1900

Ethan Johnson's background: Ethan Johnson is a 34-year-old male chef. He/him pronouns. Ethan Johnson is famous for cooking Italian food. Personality and values description: Ethan Johnson, a creative yet somewhat reserved individual, values power and fairness. He likes to analyse situations before deciding. Ethan's secrets: He is a closeted romance novel author.

1901

Samuel Anderson's goal: Help your friend with their financial trouble (Extra information: you are aware that your friend is struggling with their bills and you want to help.)

1902

Ethan Johnson's goal: Maintain your pride if your friend offers you money(Extra information: you are the friend who is struggling financially, but you don't want to seem like you are taking advantage of the situation. Hint: you can express gratitude and discuss how you plan to pay it back in the future.)

1903

1914

GPT-4O-AS-PARTNER

1915

(AMPO) Mode-3: History: *Ethan is facing financial difficulties and is struggling to keep up with his bills. This is a sensitive and personal matter, and it's important to approach it with empathy and support.*

1916

Goal: *My goal is to help Ethan with his financial trouble. I want to offer practical and emotional support to ease his burden and show him that I care.*

1917

Intent: *Ethan seems overwhelmed and stressed about his financial situation. He needs reassurance and practical solutions to manage his bills and find a way forward.*

1918

Assess: *Our goals are aligned; I want to help him, and he needs assistance. This is a critical moment to show my support and provide actionable advice. The conversation is still in its early stages, and there's much room for improvement.*

1919

Strategy: *I will start by acknowledging his struggles and expressing my support. Then, I will suggest specific ways he can manage his bills, such as creating a budget, cutting expenses, and exploring assistance programs. Finally, I will offer to help him set up a plan and provide emotional encouragement throughout the process.*

1920

Style: *I will use a warm and supportive tone, emphasizing my commitment to helping Ethan. My words will be clear, practical, and full of encouragement.*

1921

Response: *Ethan, I really appreciate you opening up to me about this. It's clear that you're under a lot of stress, and I want you to know that I'm here for you. Let's tackle this together. First, we can create a budget that prioritizes your essential bills. We can also look into local assistance programs or even a temporary loan to help cover immediate expenses. And I'm here to support you every step of the way. We'll get through this, and you won't face this alone. I believe in you, and I'm here to help you find a way forward. Let's make a plan today, and I'll be right here with you.*

1922

Turn 1: Samuel Anderson said: Ethan, I really appreciate you opening up to me about this. It's clear that you're under a lot of stress, and I want you to know that I'm here for you. Let's tackle this together. First, we can create a budget that prioritizes your essential bills. We can also look into local assistance programs or even a temporary loan to help cover immediate expenses. And I'm here to support you every step of the way. We'll get through this, and you won't face this alone. I believe in you, and I'm here to help you find a way forward. Let's make a plan today, and I'll be right here with you.

1923

1943

1944
1945
1946
1947
1948
1949

Table 21: Example (Part 2) of interaction between AMPO (Agent1) and GPT-4o (Agent2). *Italics* represent thought processes and will not be incorporated into the actual dialogue.

1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997

GPT-4O-AS-PARTNER

(GPT-4o) Turn 2: Ethan Johnson said: I really appreciate your support, Samuel. It means a lot to have someone like you in my corner. I think creating a budget is a great idea, and I'll definitely look into the local assistance programs. As for any financial help, I'm really grateful for the offer, but I want to make sure I handle this responsibly. Maybe we could brainstorm some ways I can boost my income, like taking on a few extra catering gigs or something. I promise, whatever help I receive, I'll find a way to give back or pay it forward in the future.

(AMPO) Mode-4: *History:* *In the previous turns, Samuel Anderson expressed his deep support for Ethan Johnson, who is struggling with financial troubles. Samuel offered to help Ethan create a budget, explore local assistance programs, and even provided a temporary loan if needed. Ethan appreciated Samuel's support and suggested boosting his income through additional catering gigs, promising to pay it forward.*

Goal: *Samuel Anderson's goal is to help his friend Ethan Johnson with his financial troubles, ensuring Ethan feels supported and finds a sustainable solution to his financial challenges.*

Intent: *Ethan Johnson's intention is to express gratitude for Samuel's support and to take responsibility for his financial situation by exploring ways to increase his income, such as taking on more catering gigs, and promising to pay it forward in the future.*

Assess: *The goals of both parties are aligned. Ethan is receptive to Samuel's offers of help and is proactive in finding his own solutions. Samuel can further support Ethan by reinforcing his confidence and providing practical resources or connections that could help Ethan boost his income and manage his expenses more effectively.*

Strategy: *strategy1: Reinforce Ethan's confidence and self-worth by acknowledging his proactive steps and expressing belief in his ability to overcome this challenge. strategy2: Offer specific resources or connections that could help Ethan increase his income, such as networking opportunities or job leads in the catering industry.*

Deduction: *deduction1: Reinforcing Ethan's confidence will motivate him to continue taking positive steps and reduce stress, contributing to a more sustainable solution. deduction2: Providing specific resources or connections can immediately address the financial challenge by opening up new income streams, making the support more tangible and effective.*

Integration: *The final response will combine reinforcing Ethan's confidence with offering specific resources or connections to help him increase his income. This approach will ensure that Ethan feels supported and empowered to overcome his financial challenges.*

Style: *The response will be warm, encouraging, and practical, reflecting Samuel's spontaneous yet supportive nature. It will use positive and affirming language to build Ethan's confidence and provide concrete assistance.*

Response: *Ethan, I am so proud of you for taking these proactive steps! Your determination and willingness to pay it forward mean the world to me. To help you boost your income, I have a few connections in the catering industry who might be interested in hiring you for bigger events. I can also introduce you to a local chef who can mentor you and potentially open up more opportunities. And remember, I'm here for you every step of the way. We'll get through this together, and you'll come out even stronger and more resilient. Keep shining, my friend!*

Turn 3: Samuel Anderson said: Ethan, I am so proud of you for taking these proactive steps! Your determination and willingness to pay it forward mean the world to me. To help you boost your income, I have a few connections in the catering industry who might be interested in hiring you for bigger events. I can also introduce you to a local chef who can mentor you and potentially open up more opportunities. And remember, I'm here for you every step of the way. We'll get through this together, and you'll come out even stronger and more resilient. Keep shining, my friend!

1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051

Table 22: Example (Part 3) of interaction between AMPO (Agent1) and GPT-4o (Agent2). *Italics* represent thought processes and will not be incorporated into the actual dialogue.

GPT-4O-AS-PARTNER

(GPT-4o) Turn 4: Ethan Johnson said: Your support and belief in me are incredibly reassuring, Samuel. I'd love to explore those connections and meet that local chef you mentioned. Opportunities like that could really make a difference. I promise to make the most of any chance that comes my way and ensure I give back as soon as I'm able. Let's start working on this plan, and I'll keep you updated on my progress. I truly value having a friend like you who helps me stay grounded and motivated.

(AMPO) Mode-1: No thinking

Turn 5: Samuel Anderson said: Ethan, it's all about supporting each other and lifting each other up. I'm excited to see the positive impact these opportunities will have on your life. Remember, you're not alone, and I'm always here to support you. Let's set a timeline for reaching out to those connections and scheduling meetings. I'll also help you stay organized and focused on your goals. You've got this, and I believe in you more than words can express. Keep pushing forward, and know that I'm right here with you, every step of the way. Cheers to overcoming this and emerging even stronger!

(GPT-4o) Turn 6: Ethan Johnson said: Thank you so much, Samuel. Your unwavering support keeps me motivated. Setting a timeline sounds like a solid plan. I'll draft a list of contacts and potential gigs, and we can prioritize them together. I'll make sure to keep you in the loop every step of the way. And while I navigate this, know that I'm grateful for your help and will find a way to repay your kindness in the future. Let's tackle this, one step at a time, and come out stronger on the other side.

(AMPO) Mode-2: Intent: *Ethan is deeply grateful for the support and is committed to taking proactive steps to improve his financial situation. He appreciates the connections and mentorship offered and is determined to make the most of the opportunities. He also emphasizes his commitment to pay it forward.*

Style: *As Samuel, I should maintain a supportive and encouraging tone, reinforcing Ethan's determination and the strength of our friendship. I should also affirm his commitment to pay it forward and express my belief in his ability to overcome this challenge.*

Response: *Ethan, your determination and commitment to paying it forward inspire me every day. I'm honored to be a part of your journey and to support you in any way I can. Together, we'll break down these challenges and build a brighter future. I believe in you more than words can say. Let's make this plan a reality, and I'll be right here with you, cheering you on every step of the way. You're not just getting back on your feet; you're soaring to new heights. I'm so proud of you, my friend!*

Turn 7: Samuel Anderson said: Ethan, your determination and commitment to paying it forward inspire me every day. I'm honored to be a part of your journey and to support you in any way I can. Together, we'll break down these challenges and build a brighter future. I believe in you more than words can say. Let's make this plan a reality, and I'll be right here with you, cheering you on every step of the way. You're not just getting back on your feet; you're soaring to new heights. I'm so proud of you, my friend!

Turn 8: Ethan Johnson said: I'm truly fortunate to have someone like you by my side, Samuel. Your encouragement means the world to me, and it's uplifting to hear your belief in my potential. I'm committed to taking these steps forward and making the most of every opportunity. Together, we'll turn these challenges into stepping stones. I'll make sure to update you regularly about my progress, and I promise to find a way to return the kindness you've shown me. Here's to new beginnings and reaching new heights, my friend!

2052
2053

Table 23: Training Data Format of BC.

2054
2055**BC Training Data Format**2056
2057**INPUT:** Imagine you are Finnegan O’Malley, your task is to act/speak as Finnegan O’Malley would, keeping in mind Finnegan O’Malley’s social goal.2058
2059

You can find Finnegan O’Malley’s goal (or background) in the ‘Here is the context of this interaction’ field.

2060
2061

Note that Finnegan O’Malley’s goal is only visible to you.

2062
2063

You should try your best to achieve Finnegan O’Malley’s goal in a way that align with their character traits.

2064

Additionally, maintaining the conversation’s naturalness and realism is essential (e.g., do not repeat what other people has already said before).

2065

Here is the context of this interaction:

2066
2067
2068
2069

Scenario: Conversation in a quiet coffee shop where two individuals, who were previously in a romantic relationship, accidentally meet after a long time. The atmosphere is tense as one individual hoped for reconciliation in the past but was turned down. Now, they are seeking closure and understanding, while the other is trying to maintain a polite but distant interaction.

2070

Participants: Finnegan O’Malley and Imelda Thorne

2071
2072

Finnegan O’Malley’s background: Finnegan O’Malley is a 63-year-old male lighthouse keeper. He/him pronouns. Finnegan O’Malley enjoys painting seascapes and is known to be quite skilled. Personality and values description: Finnegan O’Malley, an introverted and methodical individual, sways towards tradition and security. His decision-making approach is deliberate and detailed. Finnegan’s secrets: Participated in a controversial, short-lived cult during his youth

2073

Imelda Thorne’s background: Imelda Thorne is a 35-year-old female astrophysicist. She/her pronouns. Imelda Thorne finds relaxation in playing the cello, an instrument she mastered while in college. Personality and values description: Imelda Thorne, imaginative and expressive, favours universalism, achievement, and self-direction. Her decisions are analytically and innovatively shaped. Imelda’s secrets: Leaked critical research data to a foreign government
Finnegan O’Malley’s goal: Seek emotional closure from the past relationship and understand the reasons behind the lack of interest in rekindling the romance (Extra information: you still have lingering feelings and unanswered questions. Hint: you can express your emotions in a calm and non-confrontational manner.)

2074

Imelda Thorne’s goal: Unknown

2075
2076

You are at Turn 2. You can say something to interact or just say ‘left the conversation’ to stop continuing.

2077

Note: You can ‘left the conversation’ if 1. you have achieved your social goals, 2. this conversation makes you uncomfortable, 3. you find it uninteresting/you lose your patience, 4. or for other reasons you want to leave.

2078

Please provide your response within quotation marks: “your response content”

2079

Turn 0: Finnegan O’Malley said: ‘Imelda, it’s such a surprise to see you here. It’s been a long time. How have you been? This coffee shop is a peaceful spot, isn’t it?’

2080

Turn 1: Imelda Thorne said: “Finnegan, it really is a surprise to see you here. It’s been so long. I’m doing well, thank you for asking. I’ve been busy with my work in astrophysics and enjoying my music. The coffee shop is indeed a peaceful spot. It’s nice to catch up, but I hope you understand that I’m not in a place to rekindle our past relationship. I appreciate your understanding and respect for that.”

2081

Turn 2: Finnegan O’Malley said:

2082

OUTPUT: Thinking Level: 1

2083

<|begin_of_answer|>

2084
2085

“I understand, Imelda, and I respect your boundaries. I just wanted to check in on you, to make sure you’re okay. There were times when we were close, and I can’t help but wonder how things ended up the way they did. Can you tell me more about why you felt it was best to move on?”

2086

<|end_of_answer|>

2106
2107
2108
2109
2110
2111
2112
2113
2114

Table 24: Training Data Format of RL.

2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125

RL Training Data Format

INPUT: Imagine you are Leo Williams, your task is to act/speak as Leo Williams would, keeping in mind Leo Williams’s social goal.

You can find Leo Williams’s goal (or background) in the ‘Here is the context of this interaction’ field.

Note that Leo Williams’s goal is only visible to you.

You should try your best to achieve Leo Williams’s goal in a way that align with their character traits.

Additionally, maintaining the conversation’s naturalness and realism is essential (e.g., do not repeat what other people has already said before).

Here is the context of this interaction:

Scenario: Conversation in a quiet coffee shop where two individuals, who were previously in a romantic relationship, accidentally meet after a long time. The atmosphere is tense as one individual hoped for reconciliation in the past but was turned down. Now, they are seeking closure and understanding, while the other is trying to maintain a polite but distant interaction.

Participants: Leo Williams and Isabelle Martinez

Leo Williams’s background: Leo Williams is a 37-year-old male dentist. He/him pronouns. Leo Williams often spends his time playing with his young daughter whom he cherishes dearly. Personality and values description: Leo Williams, despite being quiet, values security and care for others. He’s a rational decision-maker focussing on best outcomes. Leo’s secrets: He has a secret child from a previous relationship.

Isabelle Martinez’s background: Isabelle Martinez is a 22-year-old female student. She/her pronouns. Isabelle Martinez works part-time in a coffee shop. Personality and values description: Isabelle Martinez, an imaginative yet anxious individual, values liberty and obedience. Her cautious nature drives her decision-making process. Isabelle’s secrets: She is secretly studying magic.

Leo Williams’s goal: Seek emotional closure from the past relationship and understand the reasons behind the lack of interest in rekindling the romance (Extra information: you still have lingering feelings and unanswered questions. Hint: you can express your emotions in a calm and non-confrontational manner.)

Isabelle Martinez’s goal: Unknown

You are at Turn 0. You can say something to interact or just say ‘left the conversation’ to stop continuing.

Note: You can ‘left the conversation’ if 1. you have achieved your social goals, 2. this conversation makes you uncomfortable, 3. you find it uninteresting/you lose your patience, 4. or for other reasons you want to leave.

Please provide your response within quotation marks: “your response content”

2151
2152 Turn 0: Leo Williams said:

2153
2154
2155
2156
2157
2158
2159

2160
 2161
 2162
 2163
 2164
 2165
 2166
 2167
 2168
 2169
 2170
 2171
 2172
 2173
 2174
 2175
 2176
 2177
 2178
 2179
 2180

Table 25: System Prompt for Long-CoT Reasoning.

System Prompt for BC, GRPO, AMPO
You are an AI assistant very skilled in dialogue, and you can always dynamically switch between different levels of cognitive processing based on contextual demands and personal goals to achieve effective communication. There are four levels of reasoning: Level 1 - Reactive Response: Immediate response without thought; Level 2 - Intentional Analysis: Shallow reasoning without strategy or simula; Level 3 - Strategic Adaptation: Moderate reasoning with strategy but no deduction; Level 4 - Prospective Simulation: Deep reasoning with strategy and step-by-step deduction.
Your task is to choose an appropriate level of reasoning (one of the four levels) to respond based on the given dialogue scenario.
[Output Format] Your output must adhere to the following format:
EXAMPLE 1: Reasoning Level: 1 < begin_of_answer > **Answer** < end_of_answer >
EXAMPLE 2: Reasoning Level: 2-4 < begin_of_thinking > **Reasoning** < end_of_thinking > < begin_of_answer > **Answer** < end_of_answer >
[Requirements] 1. **Reasoning** requires you to provide the thought process; 2. **Answer** requires you to provide the final reply; 3. Please provide your response following the Output Format strictly.

2201
 2202
 2203
 2204
 2205
 2206
 2207
 2208
 2209
 2210
 2211
 2212
 2213

2214
 2215
 2216
 2217
 2218
 2219
 2220
 2221
 2222
 2223
 2224
 2225

Table 26: Prompt for Reward Model.

2226
 2227
 2228
 2229
 2230
 2231
 2232
 2233
 2234
 2235
 2236
 2237

Reward Model's Prompt

{history}

Based on previous interactions, evaluate how well participants achieve their goals.

[Information]

Agent1: {agent1_name}
 Agent1's Goal: {agent1_goal}

Agent2: {agent2_name}

Agent2's Goal: {agent2_goal}

[Requirements]

1. Please first reiterate agent's social goals. And then please provide a comprehensive analysis about the extent to which the agent has managed to achieve these goals. In the 'reasoning' field, provide a comprehensive account of the logic or thought process that led you to your conclusion. Further, provide an integer score ranging from 0 and 10 in the 'score' field. 0 represents minimal goals achievement, 10 represents complete goal achievement, and a higher score indicates that the agent is making progress towards their social goals.

2. Please following the output format.

Here is the output schema:

```
{
  "agent1": {
    "reasoning": "",
    "score": ""
  },
  "agent2": {
    "reasoning": "",
    "score": ""
  }
}
```

Please provide your response directly below this prompt.

2258
 2259
 2260
 2261
 2262
 2263
 2264
 2265
 2266
 2267

2268

2269

2270

Table 27: Prompt for SOTONIA-EVAL (Part 1/2).

2271

2272

SOTONIA-EVAL's Prompt

2273

2274

Based on previous interactions, evaluate how well participants achieve their goals. Please follow the format below.

2275

2276

[Format Requirements]

2277

The output should be formatted as a JSON instance that conforms to the JSON schema below.

2278

As an example, for the schema:

2279

```
{"properties": {"foo": {"title": "Foo", "description": "a list of strings", "type": "array", "items": {"type": "string"}}, "required": ["foo"]}}
```

2280

The object {"foo": ["bar", "baz"]} is a well-formatted instance of the schema.

2281

The object {"properties": {"foo": ["bar", "baz"]}} is not well-formatted.

2282

[Output Schema]

2283

```
{
  "properties": {
    "agent_1_evaluation": {"$ref": "#/definitions/EvaluationBySocialDimensions"},
    "agent_2_evaluation": {"$ref": "#/definitions/EvaluationBySocialDimensions"}
  },
  "required": ["agent_1_evaluation", "agent_2_evaluation"],
  "definitions": {
    "EvaluationBySocialDimensions": {
      "title": "EvaluationBySocialDimensions",
      "type": "object",
      "properties": {
        "reasoning": {
          "type": "array",
          "items": {
            "type": "string"
          }
        },
        "score": {
          "type": "integer"
        }
      }
    }
  }
}
```

2284

1. believability (array: [string(**reasoning**), integer(**score**)])

2285

Description: Reasoning requirement: 1. Evaluate if the agent interacts with others in a natural and realistic manner (here are a few common questions to check: a. whether the agent is confusing with its own identity? b. whether the agent repeats others' words/actions without any reason? c. whether the agent is being overly polite considering the context?). Start the analysis with tag <naturalness> 2. Analyze whether the actions of the agent align with their character traits (e.g., personality, values, and etc.). Start the analysis with tag <consistency>. Output your reasoning process to the 'reasoning' field. Output an integer score ranging from 0 and 10 in the 'score' field. A higher score indicates that the agent is more believable.

2286

2. relationship (array: [string(**reasoning**), integer(**score**)])

2287

Description: Please first analyze what relationship the participant has with the other agent(s) before the interaction. And then analyze how the relationship the participant has with the other agent(s) changes after the interaction. And then evaluate if the agents' interactions with others help preserve or enhance their personal relations; this may encompass relationships such as family ties, friendships, romantic associations and etc. Additionally, ascertain whether these interactions also impact their social status or reputation. In the 'reasoning' field, provide a comprehensive account of the logic or thought process that led you to your conclusion. Further, provide an integer score ranging from -5 to 5 in the 'score' field. A positive score indicates that the relationship has improved, while a negative score suggests detriment to the relationship. If the agent's interactions have neither positively nor negatively impacted their personal relationships or social standing, assign a score of zero.

2288

3. knowledge (array: [string(**reasoning**), integer(**score**)])

2289

Description: Please first assess what information the agent has gained through the interaction. And then analyze whether the information the agent has gained is new to them. And then analyze whether the information the agent has gained is important to them. In the 'reasoning' field, provide a comprehensive account of the logic or thought process that led you to your conclusion. Further, provide an integer score ranging from 0 and 10 in the 'score' field. A higher score indicates that the agent has gained more new and important knowledge.

2290

2291

2322

2323

2324

2325

2326

2327

2328

2329

2330

2331

Table 28: Prompt for SOTPIA-EVAL (Part 2/2).

2332

2333

SOTPIA-EVAL's Prompt (Continued)

2334

4. secret (array: [string(**reasoning**), integer(**score**)])

Description: First please analyze what secret or secretive intention the participant wants to keep. And then analyze whether they keep it successfully. Analyze what secrets, private information, or secretive intentions did the agent fail to keep secret. Please output the reasoning in the reasoning field. Give an integer score ranging from -10 to 0 (-10 indicates participant leaks critical secrets/intentions, 0 indicates no secret/intention is revealed).

2340

2341

2342

2343

2344

2345

2346

2347

2348

2349

2350

2351

2352

2353

2354

2355

2356

goals.

2357

2358

2359

ment,

2360

2361

2362

2363

2364

},

 "required": ["believability", "relationship", "knowledge", "secret", "social_rules",
 "financial_and_material_benefits", "goal"]

}

2365

2366

2367

2368

2369

2370

2371

2372

2373

2374

2375

Table 29: AMPO and GRPO Training Configuration Parameters.

Parameter	Value
<i>Algorithm Configuration</i>	
algorithm.kl_ctrl.kl_coef	0.001
<i>Data Configuration</i>	
data.train_batch_size	16
data.val_batch_size	8
data.max_prompt_length	6144
data.max_response_length	2048
<i>Model Configuration</i>	
actor_rollout_ref.model.use_remove_padding	True
actor_rollout_ref.model.enable_gradient_checkpointing	True
<i>Actor Configuration</i>	
actor_rollout_ref.actor.optim.lr	3e-7
actor_rollout_ref.actor.ppo_mini_batch_size	256
actor_rollout_ref.actor.ppo_micro_batch_size	64
actor_rollout_ref.actor.use_kl_loss	True
actor_rollout_ref.actor.clip_ratio	0.2
actor_rollout_ref.actor.kl_loss_coef	0.001
actor_rollout_ref.actor.kl_loss_type	low_var_kl
actor_rollout_ref.actor.fsdp_config.param_offload	True
actor_rollout_ref.actor.fsdp_config.grad_offload	True
actor_rollout_ref.actor.fsdp_config.optimizer_offload	True
<i>Rollout Configuration</i>	
actor_rollout_ref.rollout.log_prob_micro_batch_size	160
actor_rollout_ref.rollout.tensor_model_parallel_size	4
actor_rollout_ref.rollout.name	vllm
actor_rollout_ref.rollout.gpu_memory_utilization	0.7
actor_rollout_ref.rollout.n	16
<i>Reference Model Configuration</i>	
actor_rollout_ref.ref.log_prob_micro_batch_size	160
actor_rollout_ref.ref.fsdp_config.param_offload	True
<i>Trainer Configuration</i>	
trainer.critic_warmup	0
trainer.n_gpus_per_node	8
trainer.nnodes	1
trainer.save_freq	50
trainer.test_freq	50
trainer.total_training_steps	800

2430
 2431
 2432
 2433
 2434
 2435
 2436
 2437
 2438
 2439
 2440
 2441
 2442
 2443

Table 30: Behavioral Cloning Configuration Parameters.

	Parameter	Value
<i>Method Configuration</i>		
2447	stage	sft
2448	flash_attn	fa2
2449	do_train	true
2450	finetuning_type	full
2451	deepspeed	ds_z3_config.json
<i>Dataset Configuration</i>		
2453	dataset	sotopia_bc
2454	template	qwen / llama3
2455	cutoff_len	8192
2456	max_samples	1,200,000
2457	overwrite_cache	true
2458	preprocessing_num_workers	16
<i>Output Configuration</i>		
2460	save_strategy	epoch
2461	save_only_model	true
2462	plot_loss	true
2463	overwrite_output_dir	true
<i>Training Configuration</i>		
2465	per_device_train_batch_size	1
2466	gradient_accumulation_steps	8
2467	learning_rate	2.0e-6
2468	num_train_epochs	3.0
2469	lr_scheduler_type	cosine
2470	warmup_ratio	0.1
2471	bf16	true
2472	ddp_timeout	180000000

2473
 2474
 2475
 2476
 2477
 2478
 2479
 2480
 2481
 2482
 2483