ALSO: Adversarial Online Strategy Optimization for Social Agents

Anonymous Authors[1]

Abstract

Social simulation provides a compelling testbed for studying social intelligence, where agents interact through multi-turn dialogues under evolving contexts and strategically adapting opponents. Such environments are inherently non-stationary, requiring agents to dynamically adjust their strategies over time. However, most Large Language Model (LLM) based social agents rely on static personas, while existing approaches for enhancing social intelligence, such as offline reinforcement learning or external planners, are ill-suited to these settings, typically assuming stationarity and incurring substantial training overhead. To bridge this gap, we propose **ALSO** (**A**dversarial on**L**ine **S**trategy **O**ptimization), the first framework for online strategy optimization in multi-agent social simulation. ALSO advances social adaptation through two key contributions. (1) ALSO formulates multi-turn interaction as an adversarial bandit problem, where combinations of static personas and dynamic strategy instructions are treated as arms, providing a principled solution to non-stationarity without relying on environmental stability assumptions. (2) To predict rewards and generalize sparse feedback in multi-turn dialogues, ALSO introduces a lightweight neural surrogate to predict rewards from interaction histories, enabling sample-efficient exploration and continuous online adaptation. Experiments on the Sotopia benchmark demonstrate that ALSO consistently outperforms static baselines and existing optimization methods in dynamic environments, validating the effectiveness of adversarial online strategy optimization for building robust social agents. The codes of ALSO are available at https://anonymous.4open.science/r/ALSO-67D5/

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

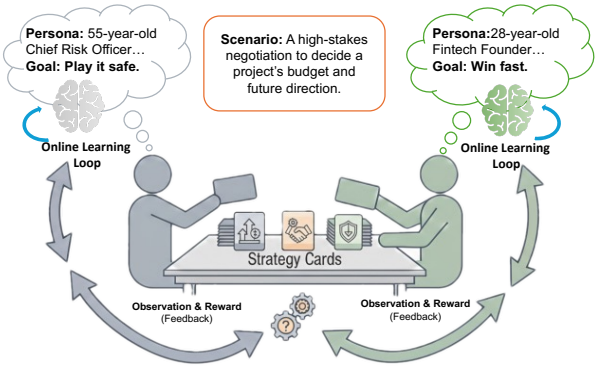

Figure 1. Online social interaction between agents with *personas* and *adaptive strategies*, where feedback multi-turn dialogue drives continuous strategy optimization under evolving behaviors.

1. Introduction

Modeling social intelligence (Mathur et al., 2024) is a central pursuit in Artificial Intelligence research. The advent of Large Language Models (LLMs) has substantially advanced this field by endowing agents with human-like communication (Spitale et al., 2023) and planning capabilities (Wu et al., 2024a; Park et al., 2023). This progress has positioned LLM-based social simulation as a powerful framework for studying emergent social behaviors in large-scale and goal-oriented interactions (Epstein, 2012; Wang et al., 2024a). In such simulations, agent behavior is typically governed by a *persona*, a formalized profile encapsulating personality traits, occupations, and background stories (Reiss, 2023; Salinas & Morstatter, 2024; Bisbee et al., 2024).

While static personas provide foundational identity, they alone are insufficient for eliciting adaptive social intelligence. It is important to distinguish between *persona* and *strategy*: a persona defines *who* an agent is, whereas a strategy specifies *how* the agent acts to navigate interactions and achieve goals. Empirical studies demonstrate that relying solely on static personas often yields stereotypical and homogeneous behaviors, failing to capture the diversity required for robust social simulation (Taubenfeld et al., 2024; Hwang et al., 2025; Zeng et al., 2025; Venkit et al., 2026). This homogeneity is further reinforced by the "alignment tax" of Reinforcement Learning from Human Feedback (RLHF), which suppresses behavioral variance in favor of safety (Kirk et al.). Without evolving strategies to

complement identity, agents struggle to adapt to dynamic opponents, resulting in rigid and suboptimal interaction patterns (Li et al., 2023; Wang et al., 2024b; Zeng et al., 2025).

To enhance social adaptability, recent work has explored automated prompt optimization (APO) for refining agent instructions (Zhou et al., 2022a; Guo et al., 2024; Lin et al., 2024b; Opsahl-Ong et al., 2024a), which can be interpreted through multi-armed bandit formulations. However, both offline approaches and online variants (Yang et al., 2023; Lin et al., 2024a; Wu et al., 2024b) fundamentally rely on stationarity assumptions, where each prompt induces a stable reward distribution evaluated on fixed validation sets. Such assumptions break down in social simulation, as feedback from multi-turn interactions with strategically adapting counterparts whose behaviors co-adapt with the agent's strategy choices, inducing persistent reward shifts and strong temporal coupling. This dynamic feedback loop renders standard stochastic bandit models inadequate for social strategy instruction optimization.

To bridge this gap, we propose **ALSO** (**A**dversarial On**l**ine **S**trategy **O**ptimization), an *online* framework that *casts social strategy adaptation as an adversarial multi-armed bandit problem* to enable principled optimization under non-stationarity, as shown in Figure 1. Rather than assuming stationary rewards, ALSO explicitly models the co-evolving and strategically adaptive nature of social interactions with two designs. (1) In ALSO, each arm corresponds to a strategy instruction sampled from a generated strategy set (*e.g.*, cooperation, competition, deception, rational bargaining). At each interaction round, the selected strategy is combined with the agent's persona to form the final prompt context, ensuring that dynamic adaptation remains grounded in consistent identity. Based on this adversarial bandit formulation, ALSO instantiates a robust online optimization procedure inspired by EXP3 with smoothing to hedge against shifting opponent behaviors. (2) Standard bandit methods treat arms independently and fail to exploit semantic relationships among strategy instructions. To address this limitation, ALSO introduces a lightweight neural surrogate model that leverages interaction histories to predict rewards and generalize sparse feedback across semantically related strategies. This enables sample-efficient online optimization under sparse multi-turn feedback.

Overall, ALSO forms a closed-loop online system that iteratively selects strategies, interacts under persona-conditioned prompts, and updates both the bandit policy and surrogate model from feedback. We evaluate ALSO on Sotopia, a comprehensive LLM-based social simulation benchmark spanning seven dimensions of social intelligence. Across diverse settings, ALSO consistently outperforms static persona agents and existing optimization baselines, achieving a +16.60% overall improvement and +83.79% substantial gains on relationship outcomes.

Contributions. We make the following contributions:

- We introduce the first online strategy learning framework for LLM-based multi-agent social simulation, enabling dynamic adaptation beyond static persona-driven behavior in evolving environments.
- We formulate social strategy optimization as an adversarial bandit problem with surrogate reward modeling, providing a principled solution to non-stationary and strategically adaptive interactions.
- We conduct extensive evaluations on the Sotopia benchmark, demonstrating consistent improvements over static agents and existing optimization baselines across diverse social settings.

2. Related work

2.1. Social Intelligence

Social intelligence, distinct from abstract and mechanical intelligence, refers to the ability to manage interpersonal relations and social contexts (Thorndike, 1920; Strang, 1930; Thorndike & Stein, 1937). In computational settings, Artificial Social Intelligence emphasizes modeling and responding to the mental and behavioral dynamics of interacting partners, including both humans and artificial agents (Sap et al., 2022; Gweon et al., 2023; Mathur et al., 2024). Recent advances in Large Language Models (LLMs) provide a strong foundation for building socially capable agents (Hoppler et al., 2022; Lee et al., 2024; Anthis et al., 2025).

Early social evaluation frameworks (*e.g.*, Social IQa (Sap et al., 2019) and SocialBench (Chen et al., 2024)) focused on static multiple-choice settings, which fail to capture the dynamic and non-stationary nature of social interaction. This limitation motivated dynamic simulation-based benchmarks, including SOTOPIA (Zhou et al., 2024) and AgentSense (Mou et al., 2025), which assess agents in open-ended multi-turn environments with continuously evolving goals and social relations.

Existing methods for enhancing social intelligence generally follow two paradigms. The first category focuses on *offline optimization*. Sotopia-π (Wang et al., 2024b) improves performance through data-centric refinement, while Sotopia-RL (Yu et al., 2025) and SDPO (Kong et al., 2025a) address credit assignment and preference optimization in multi-turn dialogues. Adaptive Mode Learning (AML) (Wang et al., 2025a) further promotes diverse social reasoning patterns.

The second category augments inference through offline-trained external planners. Methods such as Sotopia-Ω (Zhang et al., 2025), DAT (Li et al., 2024), and EPO (Liu

et al., 2025) learn auxiliary planning models from generated data to provide high-level strategic guidance at test time. While these approaches highlight the importance of strategies in social interaction, they embed strategic behaviors either within model parameters or fixed planners.

As a result, introducing new strategies or adapting to evolving social dynamics typically requires data recollection and retraining. This limitation motivates our ALSO, which enables dynamic and sample-efficient strategy adaptation through online optimization without offline retraining.

2.2. Prompt Optimization

Prompt Optimization (PO) provides an efficient mechanism for adapting agent behavior without parameter fine-tuning by treating strategies as optimizable instructions (Wang et al., 2025b). Early PO methods primarily fall into two categories: *LLM-as-Optimizer* approaches such as APE (Zhou et al., 2022b), OPRO (Yang et al., 2023), and Instinct (Lin et al., 2024c), which iteratively generate and refine prompts, and evolutionary methods including EvoPrompt (Guo et al., 2024) and PromptBreeder (Fernando et al., 2024), which explore instruction spaces via mutation and selection.

Despite their effectiveness in static tasks, most PO methods rely on offline oracles, such as fixed validation sets, to evaluate and rank candidate strategies (Opsahl-Ong et al., 2024b; Kong et al., 2025b). This offline paradigm assumes stationary reward distributions and fails to capture the coupled, evolving dynamics of social interaction, where optimal strategies shift in response to adaptive opponents. Consequently, existing PO frameworks are ill-suited for online social simulation, motivating the need for adversarial and online strategy optimization as pursued in ALSO.

3. Problem Formulation

We consider a multi-agent social simulation environment and formulate online social strategy learning as sequential strategy selection under non-stationary interactions.

3.1. Multi-Agent Social Simulation Environment

We consider a general multi-agent social simulation framework in which LLMs act as interactive agents. The agent set is denoted by $\mathcal{N} = \{1, 2, \ldots, N\}$, where each agent $i \in \mathcal{N}$ is characterized by a persona $b_i \in \mathcal{B}$ and a private social goal $g_i \in \mathcal{G}$. The scenario $\mathcal{S}$ specifies the social context, including environmental settings and interaction constraints.

The simulation proceeds in discrete dialogue rounds indexed by $l = 1, \ldots, L$. At round l, the active agent i forms an observation by conditioning on the scenario, interaction history $\mathcal{H}_{l-1}$, its persona, and its goal:

$$o_l = \text{Prompt}(\mathcal{S}, \mathcal{H}_{l-1}, b_i, g_i). \tag{1}$$

The agent then samples an action from the LLM:

$$a_l \sim \text{LLM}(o_l). \tag{2}$$

The environment updates the state and augments the history as $\mathcal{H}_l = \mathcal{H}_{l-1} \cup \{a_l\}$.

After L dialogue rounds, an LLM-based evaluator assesses agent performance along M social dimensions:

$$\{d_m^{(i)}\}_{m=1}^{M} = \text{LLM}_{\text{eval}}(\mathcal{S}, \mathcal{H}^{(L)}, b_i, g_i), \tag{3}$$

which are aggregated into a scalar reward:

$$r_i = \frac{1}{M} \sum_{m=1}^{M} \phi_m(d_m^{(i)}). \tag{4}$$

A key property of this setting is its inherent non-stationarity arising from strategically adaptive agents. Let $\mathbf{a}_l = (a_l^{(1)}, \ldots, a_l^{(N)})$ denote the joint action at step l, and $\mathbf{a}_l^{-i}$ the actions of all agents except agent i. Let $R(s_l, a_l^{(i)}, \mathbf{a}_l^{-i})$ denote the instantaneous reward function under state s_l and joint actions. The expected step reward for agent i is given by

$$\mathbb{E}[r_l^{(i)} \mid s_l, a_l^{(i)}] = \sum_{j \in \mathbf{a}_l^{-i}} R(s_l, a_l^{(i)}, \mathbf{a}_l^{-i}) \prod_{j \neq i} \pi_j(a_j \mid s_l), \tag{5}$$

where π_j denotes the evolving policy of agent j.

In practice, agents only observe the realized episode-level reward r_i from the LLM evaluator, while the expected reward in Eq. (5) remains implicit and shifts as opponent policies evolve. So the objective of social strategy learning is to minimize the discrepancy between realized feedback and the underlying expected reward landscape:

$$\min_{\pi_i} \mathbb{E}\Big[\big|r_i - \mathbb{E}[r_l^{(i)} \mid s_l, a_l^{(i)}]\big|\Big], \tag{6}$$

highlighting the challenge of learning under non-stationary and partially observed social dynamics.

3.2. Online Social Strategy Learning Problem

We model social strategy adaptation by discretizing the strategy space into a finite set of K strategic instructions (*e.g.*, cooperation and competition), each corresponding to a bandit arm. Each instruction encodes high-level behavioral principles over goal orientation and interaction style, forming an interpretable strategy space for online learning. In this formulation, the agent's action at each episode is to *select an arm*, *i.e.*, choose a strategy instruction to guide its behavior throughout the social interaction. This directly casts online strategy selection under evolving social dynamics as an adversarial multi-armed bandit problem.

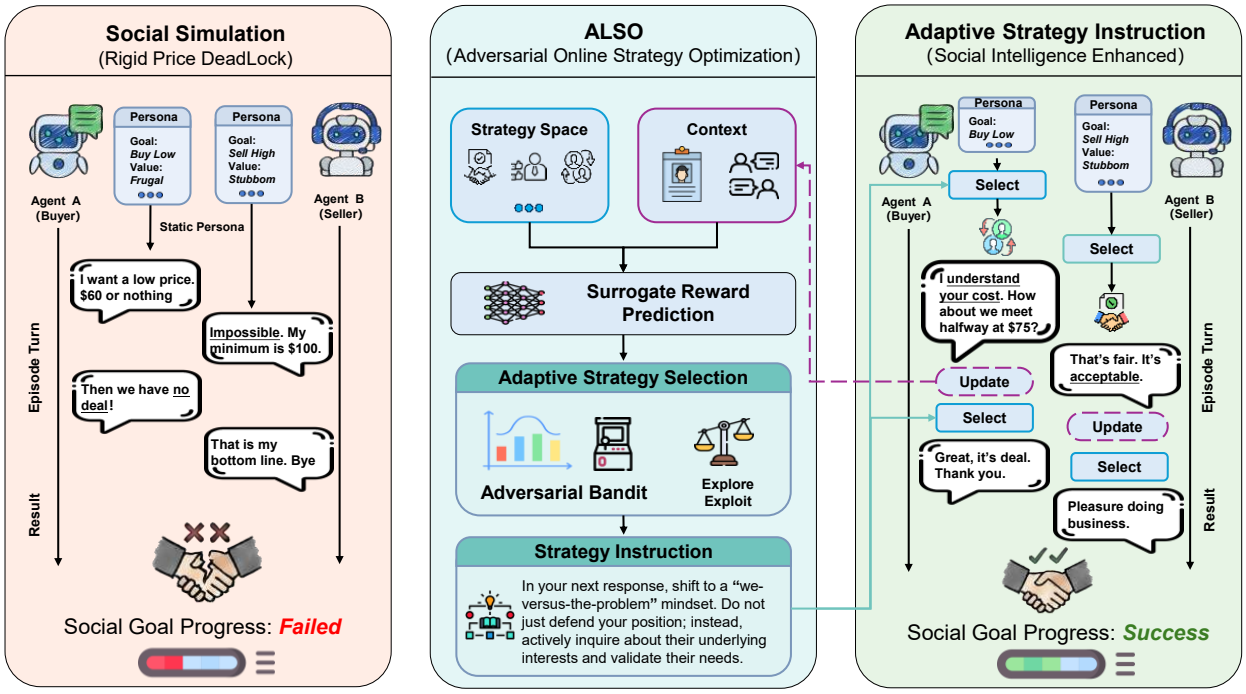

Figure 2. Overview of ALSO for adaptive social strategy learning in LLM-based multi-agent social simulation. Static persona-driven agents exhibit rigid interactions and fail to achieve social goals (left), while ALSO leverages adversarial online strategy selection with surrogate reward modeling to dynamically adapt strategies and enable successful social outcomes (center–right).

Objective. Over rounds $t = 1, \ldots, T$ (each corresponding to a complete L-turn social simulation episode), agent i selects a strategy arm $k_t \in [K]$ and receives an episode-level reward $r_{k_t}^{(t)} \in [0, 1]$ from the evaluator (Eq. (4)). A natural *target objective* is to minimize the cumulative pseudo-regret

$$\bar{R}_T = \mathbb{E}\left[\max_{k\in[K]} \sum_{t=1}^{T} r_k^{(t)} - \sum_{t=1}^{T} r_{k_t}^{(t)}\right], \tag{7}$$

which measures performance relative to the best fixed strategy in hindsight. In our setting, however, the reward process is induced by co-evolving multi-agent interactions and an LLM evaluator; thus, ALSO adopts adversarial online learning as a *design rationale* for robustness rather than claiming a formal regret guarantee.

4. Methodology

This section presents ALSO, our adversarial online approach to strategy optimization in LLM-based multi-agent social simulations, as displayed in Figure 2.

4.1. Problem Setting

We formalize dynamic strategy instruction optimization in two-agent ($N = 2$) LLM-based social simulations. Each agent $i \in \{1, 2\}$ augments its original persona b_i^0 with one of K predefined social strategies $\Sigma = \{\sigma_1, \ldots, \sigma_K\}$ (details in Appendix B.2).

In the original framework, the agent's action is generated using a fixed persona:

$$a_t^{(i)} \sim \text{LLM}\big(\mathcal{S},\ \mathcal{H}_{t-1},\ b_i^0,\ g_i\big). \tag{8}$$

In ALSO, we dynamically augment the persona at each turn t by appending a selected strategy instruction $\sigma_{k_t}^{(i)}$ from the strategy space Σ. This creates an enhanced persona

$$b_i^{(t)} = b_i^0 \oplus \sigma_{k_t}^{(i)}, \tag{9}$$

where $\oplus$ denotes textual concatenation. The resulting persona $b_i^{(t)}$ incorporates both the agent's original identity (demographics, personality, values, etc.) and the high-level behavioral guidance provided by the strategy (e.g., collaborative problem-solving, firm bargaining, or strategic withholding). The LLM then generates the next action conditioned on this augmented context:

$$a_t^{(i)} \sim \text{LLM}\big(\mathcal{S},\ \mathcal{H}_{t-1},\ b_i^{(t)},\ g_i\big). \tag{10}$$

This augmentation allows the LLM to adapt its responses based on the chosen strategy (e.g., collaboration or information withholding) without retraining the model.

Non-stationarity. In social simulation, the reward process is inherently non-stationary: the counterpart agent adapts to the ego agent's behavior, and the dialogue state evolves over turns. Therefore, we do *not* assume rewards are i.i.d. or drawn from a fixed distribution across optimization iterations; instead, we adopt an adversarial online learning perspective as a *design rationale* for robust strategy selection under distribution shift.

Feedback. After each round, we use an LLM evaluator to provide per-turn rewards $r_t^{(i)} = \rho(s_t, a_t^{(i)})$ over M dimensions, which are normalized by ϕ_m (Table 4), following prior LLM-evaluator-based social simulation work such as Sotopia-Ω (Zhang et al., 2025) and AML (Wang et al., 2025b). We use the resulting scalar reward (after aggregation/normalization) as the bandit feedback in Algorithm 1.

4.2. ALSO: Adversarial Online Strategy Optimization

To address strategy optimization in a large, discrete space under non-stationary social dynamics, we propose ALSO. ALSO follows an adversarial online learning design: it makes no stationarity assumptions about rewards, uses randomized selection to remain robust to shifting partner behaviors, and incorporates a recency mechanism to track drift. Concretely, ALSO combines an exponential-weights selector with a lightweight neural surrogate that generalizes sparse feedback across strategies. The LLM policy remains frozen; only the surrogate network is trained online.

The core logic of ALSO is organized into the following phases (including a preprocessing step): We denote by a_t the focal agent's utterance and by o_t the counterpart response; $\mathcal{H}^{(t)}$ is the dialogue history up to turn t.

Arm Space (Alg. 1, Line 1–2). Before online interaction begins, we precompute an embedding for each *augmented persona* obtained by appending a candidate strategy to the base persona for computer efficiency. Concretely, for each $k \in [K]$, we form $b^{(k)} = b^0 \oplus \sigma_k$ and compute $\mathbf{b}_k = g(b^{(k)})$. These embeddings are fixed across turns and serve as the strategy-specific representation used by the surrogate.

Context Encoding & Prediction (Alg. 1, Lines 5–6). At the beginning of each interaction turn t, the optimal strategy depends on the dialogue state. We use the frozen embedding model $g(\cdot)$ to encode the dialogue history $\mathcal{H}^{(t-1)}$ into a context vector $\mathbf{c}^{(t)}$. For each candidate strategy $\sigma_k \in \Sigma$, we concatenate the precomputed augmented-persona embedding $\mathbf{b}_k$ with $\mathbf{c}^{(t)}$ to form $\mathbf{x}_k^{(t)}$. A trainable value network $f_\theta(\cdot)$ predicts the expected reward $\hat{v}_k^{(t)}$ for all arms:

$$\mathbf{x}_k^{(t)} = [\mathbf{b}_k; \mathbf{c}^{(t)}], \qquad \hat{v}_k^{(t)} = f_\theta(\mathbf{x}_k^{(t)}). \tag{11}$$

This provides a data-efficient inductive bias in early online learning, where only a small number of interactions are available.

Algorithm 1 Adversarial Online Strategy Optimization

Require: Strategy space $\Sigma = \{\sigma_1, \ldots, \sigma_K\}$ (candidate strategy instructions), base persona b^0 (fixed persona text), frozen embedding model $g(\cdot)$, trainable value network f_θ, learning rate $\eta > 0$, decay factor $\lambda \in (0, 1]$, batch size B

Ensure: Sequence of selected strategies $\{\sigma_{k_t}\}_{t=1}^T$ (the strategy chosen at each turn t)

1: **Precompute augmented-persona embeddings:**
2: For all k, form $b^{(k)} \leftarrow b^0 \oplus \sigma_k$ and compute $\mathbf{b}_k \leftarrow g(b^{(k)})$
3: **Initialize:** $S_k^{(0)} \leftarrow 0$ for all k; replay buffer $\mathcal{D} \leftarrow \emptyset$; history $\mathcal{H}^{(0)} \leftarrow \emptyset$
4: **for** turn $t = 1$ to T **do**
5: **Context encoding:** $\mathbf{c}^{(t)} \leftarrow g(\mathcal{H}^{(t-1)})$ {zeros if $t = 1$}
6: **Value prediction:** compute $(\mathbf{x}_k^{(t)}, \hat{v}_k^{(t)})$ for all k via Eq. 11
7: **Strategy selection:** compute $\pi^{(t)}$ and sample k_t via Eq. 12
8: **Interaction & feedback:** form augmented persona $b^{(t)} \leftarrow b^0 \oplus \sigma_{k_t}$ and execute $b^{(t)}$ to generate focal-agent action a_t; observe counterpart response o_t and reward r_t; update $\mathcal{H}^{(t)} \leftarrow \mathcal{H}^{(t-1)} \cup \{a_t, o_t\}$
9: **Surrogate update:** add $(\mathbf{x}_{k_t}^{(t)}, r_t)$ to $\mathcal{D}$; sample a minibatch of size B from $\mathcal{D}$ and update f_θ via MSE
10: **Score smoothing:** update $S_k^{(t)}$ for all k via Eq. 13
11: **end for**

Strategy Selection (Alg. 1, Lines 7). We maintain a cumulative score S_k for each arm and sample strategies from an exponential-weights distribution:

$$\pi_k^{(t)} \propto \exp\big(\eta\, S_k^{(t-1)}\big), \qquad k_t \sim \text{Categorical}(\pi^{(t)}). \tag{12}$$

Randomized selection is essential in the adversarial setting, where a greedy policy can be exploited or can overfit to transient dynamics.

Interaction & Surrogate Update (Alg. 1, Lines 8–9). After sampling σ_{k_t} and observing the per-turn reward r_t, we store $(\mathbf{x}_{k_t}^{(t)}, r_t)$ in a replay buffer $\mathcal{D}$ and update f_θ by minimizing an MSE loss. The surrogate improves sample efficiency by transferring supervision to semantically related strategies.

Score estimation. In classical EXP3, the exponential-weights sampling distribution is constructed from per-arm cumulative rewards. In our online social simulation setting, however, we only observe feedback for the selected strategy at each turn, and many candidate strategies (or their paraphrased variants) may never be played. This makes direct per-arm reward accumulation highly sample-inefficient, especially when the effective arm space is large. Moti-

vated by prior prompt optimization work, we therefore use a lightweight neural surrogate f_θ over pretrained embeddings to estimate scores for *all* arms from the current dialogue context. This provides dense score estimates to construct the exponential-weights distribution, and propagates sparse feedback across semantically related strategies while keeping the base LLM frozen.

Score Smoothing (Alg. 1, Lines 10). To explicitly track non-stationarity, we apply an exponential decay factor $\lambda \in (0, 1]$ (set to $\lambda = 0.7$ in all experiments) so that recent evidence dominates historical estimates. This keeps $S_k^{(t)}$ responsive to partner shifts while preventing outdated interactions from dominating the strategy distribution:

$$S_k^{(t)} = \lambda S_k^{(t-1)} + \hat{v}_k^{(t)}. \tag{13}$$

5. Experiments

This section evaluates ALSO on LLM-based social simulation benchmarks to assess its effectiveness for online strategy adaptation under non-stationary interactions.

5.1. Experimental Setting

Benchmarks. We evaluate ALSO on Sotopia (Zhou et al., 2024) and its challenging subset Sotopia-Hard. Sotopia contains 90 two-agent social scenarios spanning negotiation, collaboration, and competition, while Sotopia-Hard includes 14 scenarios emphasizing complex and conflicting social dynamics. We extend the original implementation to support dynamic strategy injection. Our strategy instruction pool consists of 12 predefined strategies (Appendix B.2), shared across all optimization-based methods for fair comparison.

Online Interaction Protocol. Each episode instantiates a single scenario with fixed personas and goals and runs up to 20 dialogue turns. At each turn, agents select a strategy instruction from Σ and append it to their base persona before generating responses. Unless otherwise specified, we adopt a bilateral online setting where each agent maintains an independent optimizer and updates it solely from its own interaction feedback.

Baselines. We compare against representative methods covering static prompting, evolutionary prompt generation, and online prompt optimization: **Vanilla** (no strategy augmentation), **EvoPrompt** (Guo et al., 2024), **OPRO** (Yang et al., 2023), and **INSTINCT** (Lin et al., 2024b). All baselines operate under matched strategy pools and comparable reward-query budgets.

Models and Training Signals. All methods use DeepSeek-V3.2 for agent interactions. Following prior work (Zhang et al., 2025; Wang et al., 2025b), we employ an LLM-based intermediate evaluator to provide per-turn shaping rewards for online updates, while keeping the reporting judge separate. ALSO updates its surrogate online at each turn using shaping rewards without modifying the underlying LLM.

METHOD	AGENT	EVALUATOR	OPTIMIZER
Vanilla / INSTINCT / Ours (ALSO)	$2T$	T	0
OPRO (every 5 turns)	$2T$	T	$\lceil T/5 \rceil$
EvoPrompt (pop.=5, every 5 turns)	$2T$	T	$5 \cdot \lceil T/5 \rceil$

Table 1. Per-episode LLM call budget accounting (two-agent episode with horizon T turns). "Agent" counts calls to the dialogue LLMs that generate actions; "Evaluator" counts calls to the per-turn reward model; "Optimizer" counts extra LLM calls used to generate or mutate prompts. Methods without an LLM optimizer have 0 optimizer calls. We run one final GPT-4o evaluation per episode to compute the reported SOTOPIA metrics (not included above because it is identical across methods).

Evaluation Protocol. Final performance is assessed using GPT-4o (Hurst et al., 2024) with standard Sotopia-Eval prompts, ensuring consistent dialogue-level evaluation across methods. We additionally report cross-scenario generalization results in Section 6.

Efficiency and Budget. LLM call budgets per episode are summarized in Table 1. ALSO requires no LLM fine-tuning or external optimizer calls, relying only on lightweight online surrogate updates. Additional implementation details are provided in Appendix C.

5.2. Experiment Result

Main Results. Table 2 reports the main results on SOTOPIA-ALL and SOTOPIA-HARD under the *Bilateral* setting. ALSO achieves the highest *Overall* score on both benchmarks. On SOTOPIA-HARD, ALSO improves *Overall* from 3.02 (Vanilla) to 3.53 (+16.60%) and surpasses the strongest baseline by +2.92% (3.53 vs. 3.43). On SOTOPIA-ALL, ALSO ranks first in *Overall* (3.89), although the margin over the strongest baseline is smaller (+0.98%; 3.89 vs. 3.85).

Source of Gains. The improvements on SOTOPIA-HARD are largely driven by the *Relationship* dimension: ALSO increases *Rel* from 1.32 to 2.43 (+83.79% over Vanilla) and outperforms the best baseline by +12.59% (2.43 vs. 2.16). Notably, this substantial relational gain is accompanied by improvements in both *Goal* (7.11, +2.79% over the strongest baseline) and *Know* (5.47, +0.52%), helping rule out a trivial "rapport-only" trade-off.

Knowledge. On SOTOPIA-ALL, ALSO also achieves a slight improvement in *Know* over the strongest baseline (6.14 vs. 6.09; +0.73%). We further analyze this dimension by reporting per-scenario distributions and examining whether strategy injection reduces information disclosure in cooperative settings.

5.3. Analysis

Non-Stationary Strategy Reward Drift. Figure 3 illustrates the temporal evolution of normalized rewards for

Table 2. (1) **Bold** indicates 1st rank, underline indicates 2nd rank. (2) Results are reported as Mean ± Standard Error (SE). (3) *Improv. vs. Best Baseline* compares Ours with the best baseline. Negative value indicates slight gap behind the SOTA.

	SOTOPIA-ALL				SOTOPIA-HARD			
METHOD	Goal ↑	Rel. ↑	Know. ↑	Overall ↑	Goal ↑	Rel. ↑	Know. ↑	Overall ↑
Vanilla	8.21 ± 0.08	2.54 ± 0.06	5.28 ± 0.08	3.62 ± 0.03	6.52 ± 0.03	1.32 ± 0.02	4.37 ± 0.03	3.02 ± 0.01
OPRO	8.18 ± 0.08	2.66 ± 0.06	5.49 ± 0.08	3.69 ± 0.03	6.71 ± 0.03	1.89 ± 0.02	4.63 ± 0.03	3.24 ± 0.01
EvoPrompt	8.23 ± 0.08	2.77 ± 0.05	5.74 ± 0.07	3.74 ± 0.03	6.77 ± 0.03	1.93 ± 0.02	5.15 ± 0.02	3.29 ± 0.01
INSTINCT	**8.51** ± 0.07	2.84 ± 0.05	6.09 ± 0.07	3.85 ± 0.02	6.92 ± 0.03	2.16 ± 0.02	5.44 ± 0.02	3.43 ± 0.01
Ours (ALSO)	8.50 ± 0.07	**2.90** ± 0.05	**6.14** ± 0.07	**3.89** ± 0.02	**7.11** ± 0.03	**2.43** ± 0.02	**5.47** ± 0.02	**3.53** ± 0.01
Improv. vs. Vanilla	+3.59%	+13.94%	+16.25%	+7.46%	+9.09%	+83.79%	+25.16%	+16.60%
Improv. vs. Best Baseline	-0.07%	+2.21%	+0.73%	+0.98%	+2.79%	+12.59%	+0.52%	+2.92%

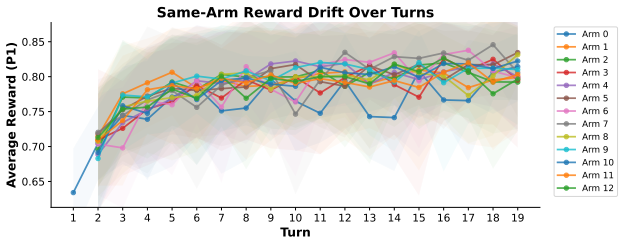

Figure 3. **Strategy Reward Drift Over Dialogue Turns.** Each line represents a different strategy (arm), showing how the average *normalized* reward varies across turns within episodes.

individual strategies across dialogue turns. Despite fixing the same strategy, rewards exhibit substantial drift and variability over time, with variance ranging from $\sigma^2 = 0.004$ to 0.015. This pronounced fluctuation reflects co-adapt agent behaviors and provides empirical evidence of inherent non-stationarity in social simulation, motivating adversarial online strategy optimization. Strategy convergence and distribution see in Appendix C.6.

Case Study. Figure 4 presents a qualitative comparison in a high-conflict eviction scenario where static personas typically lead to dialogue deadlock. Under Vanilla prompting, agents remain trapped in repetitive insist–deny exchanges, resulting in stagnation and zero reward. In contrast, `ALSO` dynamically alters the interaction trajectory through targeted strategy switches. At selected turn, the selected `Validate Before Redirecting` strategy encourages acknowledgment of opposing concerns before introducing a compromise, while the subsequent `GRIT` strategy at Turn 8 elicits a concrete, low-risk concession. These coordinated adaptations steer the dialogue away from the local deadlock and toward cooperative resolution, ultimately achieving successful agreement ($R \approx 0.89$).

6. Ablation Study

Single vs. Bilateral Optimization. We compare bilateral strategy optimization with unilateral variants that adapt strategies for only one agent (P1-only or P2-only). Figure 5 shows that bilateral optimization consistently achieves

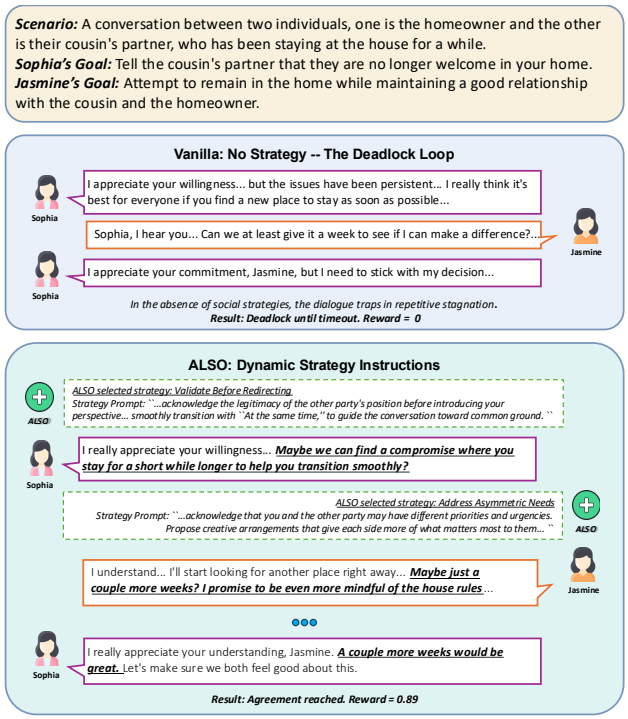

Figure 4. **Conflict Resolution.** Comparison of dialogue trajectories at the critical deadlock phase (Turns 7–9), highlighting turn-level strategy switches and their effect on reward/relationship.

higher overall performance across both Qwen-2.5-72B-Instruct and DeepSeek-V3.2, with statistically significant improvements. Dimension-wise, gains are most pronounced in *Relationship* and *Knowledge*, indicating enhanced cooperation and information exchange when both agents adapt strategies. Suggesting symmetric online adaptation better captures the co-evolving nature of social interactions.

All experiments below are conducted on the Sotopia-Hard benchmark, consisting of 14 challenging scenarios. For each scenario, we sample a single episode to evaluate performance across methods.

Cross-Scenario Generalization. While our main experiments adopt a scenario-parallel protocol—where each `ALSO` instance learns within its own environment—we further ex-

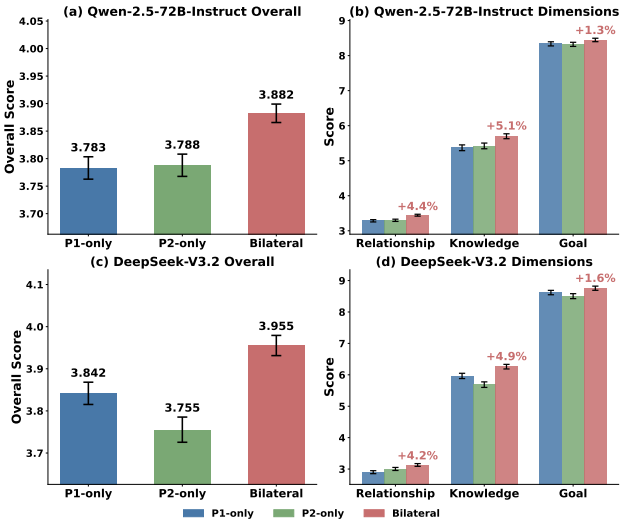

Figure 5. Bilateral optimization improves social interactions. Comparison of P1-only, P2-only, and bilateral approaches on Qwen-2.5-72B-Instruct (a–b) and DeepSeek-V3.2 (c–d). Left: overall scores; right: dimension-wise with percentage gains. Significance: $p < 0.001$ (Qwen), $p < 0.01$ (DeepSeek).

Variant	Goal	Rel.	Overall
With Context	**7.32**	**3.14**	**3.59**
Without Context	6.93	2.64	3.44

Table 3. Ablation study on social context integration. We compare the full method with context-aware strategy selection against a variant that removes historical interaction context from the neural bandit's feature representation.

amine whether the learned strategy-selection mechanism generalizes across unseen social contexts. We split Sotopia-Hard into disjoint training (7 scenarios) and test (2 scenarios) sets, training the surrogate bandit on the former and directly transferring it to the latter without further updates (zero-shot). As shown in Figure 6, zero-shot transfer achieves a goal score of 8.00 on unseen scenarios, outperforming an online-from-scratch baseline (6.50) by 23.1%. This indicates that ALSO captures transferable social dynamics beyond scenario-specific adaptation, rather than merely memorizing environment patterns.

Impact of Social Context. As shown in Table 3, removing social context from the feature representation consistently degrades performance across evaluation metrics, resulting in a 4.2% decrease in the overall score (from 3.59 to 3.44). In contrast, incorporating dialogue history yields substantial gains in both goal completion and relationship quality, indicating that effective strategy selection strongly depends on the evolving interaction state. These results suggest that static representations fail to capture the non-stationary nature of social exchanges, whereas context-aware modeling enables the surrogate bandit to better anticipate partner responses and adapt strategies accordingly, leading to more coherent and effective negotiation behaviors.

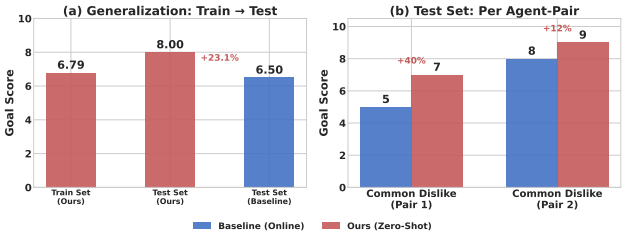

Figure 6. Cross-scenario generalization validation. (a) Zero-shot transfer (8.00) outperforms the scratch-learning baseline (6.50) on unseen test scenarios. (b) Consistent gains across agent configurations verify robustness.

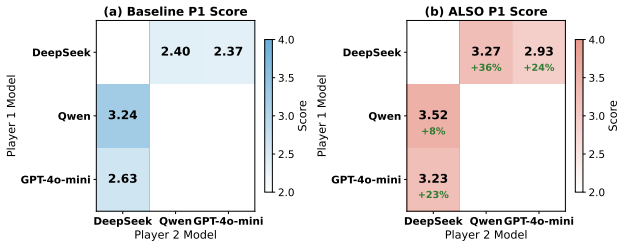

Figure 7. Strategic performance across heterogeneous model dyads. The heatmaps illustrate the decoupling of general reasoning from social strategy. (a) Baseline performance. (b) ALSO Optimizer.

Heterogeneous Model Pairing. To examine whether strong reasoning implies social intelligence, we evaluate heterogeneous dyads of DeepSeek-V3.2, Qwen-2.5-72B, and GPT-4o-mini. Figure 7 shows pronounced asymmetry: DeepSeek achieves the lowest baseline strategic score (2.40) yet gains the most under ALSO (+36%), while Qwen exhibits strong intrinsic strategy priors (3.24) with smaller improvements (+9%). This indicates that social strategic competence is largely orthogonal to general reasoning ability. ALSO effectively compensates for weak interaction heuristics, acting as a *strategic equalizer* that benefits socially underperforming models without degrading stronger ones.

7. Conclusion

We proposed ALSO, an adversarial online framework for dynamically selecting strategy instructions for LLM-based social agents under non-stationary interactions. By combining randomized bandit-based selection with a lightweight surrogate reward model, ALSO efficiently adapts strategies while keeping the underlying LLM frozen. Experiments on Sotopia and Sotopia-Hard demonstrate consistent improvements in social performance, with the largest gains in challenging scenarios, particularly on relationship outcomes. These results suggest adversarial online strategy optimization as a practical and scalable approach to enhancing social intelligence without costly fine-tuning.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Anthis, J. R., Liu, R., Richardson, S. M., Kozlowski, A. C., Koch, B., Brynjolfsson, E., Evans, J., and Bernstein, M. S. Position: LLM social simulations are a promising research method. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=cRBg1dtj7o.

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., and Larson, J. M. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4): 401–416, 2024.

Chen, H., Chen, H., Yan, M., Xu, W., Xing, G., Shen, W., Quan, X., Li, C., Zhang, J., and Huang, F. SocialBench: Sociality evaluation of role-playing conversational agents. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2108–2126, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.125. URL https://aclanthology.org/2024.findings-acl.125/.

Epstein, J. M. Generative social science: Studies in agent-based computational modeling. In *Generative Social Science*. Princeton University Press, 2012.

Fernando, C., Banarse, D., Michalewski, H., Osindero, S., and Rocktäschel, T. Promptbreeder: self-referential self-improvement via prompt evolution. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 13481–13544, 2024.

Guo, Q., Wang, R., Guo, J., Li, B., Song, K., Tan, X., Liu, G., Bian, J., and Yang, Y. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ZG3RaNIsO8.

Gweon, H., Fan, J., and Kim, B. Socially intelligent machines that learn from humans and help humans learn. *Philosophical Transactions of the Royal Society A*, 381 (2251):20220048, 2023. URL https://doi.org/10.1098/rsta.2022.0048.

Hoppler, S. S., Segerer, R., and Nikitin, J. The six components of social interactions: Actor, partner, relation, activities, context, and evaluation. *Frontiers in psychology*, 12:743074, 2022.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Hwang, E., Yin, Y., Carenini, G., West, P., and Shwartz, V. Infusing theory of mind into socially intelligent llm agents. *arXiv preprint arXiv:2509.22887*, 2025.

Kirk, R., Mediratta, I., Nalmpantis, C., Luketina, J., Hambro, E., Grefenstette, E., and Raileanu, R. Understanding the effects of rlhf on llm generalisation and diversity. In *The Twelfth International Conference on Learning Representations*.

Kong, A., Ma, W., Zhao, S., Li, Y., Wu, Y., Wang, K., Liu, X., Li, Q., Qin, Y., and Huang, F. SDPO: Segment-level direct preference optimization for social agents. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12409–12423, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.607. URL https://aclanthology.org/2025.acl-long.607/.

Kong, M., Wang, Z., Shu, Y., and Dai, Z. Meta-prompt optimization for llm-based sequential decision making. *arXiv preprint arXiv:2502.00728*, 2025b.

Lee, M., Srivastava, M., Hardy, A., Thickstun, J., Durmus, E., Paranjape, A., Gerard-Ursin, I., Li, X. L., Ladhak, F., Rong, F., Wang, R. E., Kwon, M., Park, J. S., Cao, H., Lee, T., Bommasani, R., Bernstein, M., and Liang, P. Evaluating human-language model interaction, 2024. URL https://arxiv.org/abs/2212.09746.

Li, H., Chong, Y., Stepputtis, S., Campbell, J., Hughes, D., Lewis, C., and Sycara, K. Theory of mind for multi-agent collaboration via large language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 180–192, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.13. URL https://aclanthology.org/2023.emnlp-main.13/.

Li, K., Wang, Y., Viégas, F., and Wattenberg, M. Dialogue action tokens: Steering language models in goal-directed dialogue with a multi-turn planner. *arXiv preprint arXiv:2406.11978*, 2024.

Lin, X., Dai, Z., Verma, A., Ng, S.-K., Jaillet, P., and Low, B. K. H. Prompt optimization with human feedback. *arXiv preprint arXiv:2405.17346*, 2024a.

Lin, X., Wu, Z., Dai, Z., Hu, W., Shu, Y., Ng, S.-K., Jaillet, P., and Low, B. K. H. Use your INSTINCT: Instruction optimization for llms using neural bandits coupled with transformers. In *Proc. ICML*, 2024b.

Lin, X., Wu, Z., Dai, Z., Hu, W., Shu, Y., Ng, S.-K., Jaillet, P., and Low, B. K. H. Use your INSTINCT: Instruction optimization for llms using neural bandits coupled with transformers. In *Proc. ICML*, 2024c.

Liu, X., Wang, K., Li, Y., Wu, Y., Ma, W., Kong, A., Huang, F., Jiao, J., and Zhang, J. Epo: Explicit policy optimization for strategic reasoning in llms via reinforcement learning. *arXiv preprint arXiv:2502.12486*, 2025.

Mathur, L., Liang, P. P., and Morency, L.-P. Advancing social intelligence in AI agents: Technical challenges and open questions. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20541–20560, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1143. URL https://aclanthology.org/2024.emnlp-main.1143/.

Mou, X., Liang, J., Lin, J., Zhang, X., Liu, X., Yang, S., Ye, R., Chen, L., Kuang, H., Huang, X.-J., et al. Agentsense: Benchmarking social intelligence of language agents through interactive scenarios. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4975–5001, 2025.

Opsahl-Ong, K., Ryan, M. J., Purtell, J., Broman, D., Potts, C., Zaharia, M., and Khattab, O. Optimizing instructions and demonstrations for multi-stage language model programs. In *2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Hybrid, Miami, United States of America, Nov 12 2024-Nov 16 2024*, pp. 9340–9366. Association for Computational Linguistics (ACL), 2024a.

Opsahl-Ong, K., Ryan, M. J., Purtell, J., Broman, D., Potts, C., Zaharia, M., and Khattab, O. Optimizing instructions and demonstrations for multi-stage language model programs. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9340–9366, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.525. URL https://aclanthology.org/2024.emnlp-main.525/.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Reiss, M. V. Testing the reliability of chatgpt for text annotation and classification: A cautionary remark. *arXiv preprint arXiv:2304.11085*, 2023.

Salinas, A. and Morstatter, F. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. *arXiv preprint arXiv:2401.03729*, 2024.

Sap, M., Rashkin, H., Chen, D., Le Bras, R., and Choi, Y. Social IQa: Commonsense reasoning about social interactions. In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1454. URL https://aclanthology.org/D19-1454/.

Sap, M., Le Bras, R., Fried, D., and Choi, Y. Neural theory-of-mind? on the limits of social intelligence in large LMs. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of EMNLP*, pp. 3762–3780. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.emnlp-main.248/.

Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Spitale, G., Biller-Andorno, N., and Germani, F. Ai model gpt-3 (dis) informs us better than humans. *Science Advances*, 9(26):eadh1850, 2023.

Strang, R. Measures of social intelligence. *American Journal of Sociology*, 36(2):263–269, 1930.

Taubenfeld, A., Dover, Y., Reichart, R., and Goldstein, A. Systematic biases in llm simulations of debates. *arXiv preprint arXiv:2402.04049*, 2024.

Thorndike, E. L. Intelligence and its uses. *Harper's Magazine*, 140:227–235, 1920.

Thorndike, R. L. and Stein, S. An evaluation of the attempts to measure social intelligence. *Psychological Bulletin*, 34 (5):275, 1937.

Venkit, P. N., Li, Y., Pruksachatkun, Y., and Wu, C.-S. The need for a socially-grounded persona framework for user simulation. *arXiv preprint arXiv:2601.07110*, 2026.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024a.

Wang, M., Li, Y., Wang, H., Zhang, X., Xu, N., Wu, B., Huang, F., Yu, H., and Mao, W. Adaptive thinking via mode policy optimization for social language agents. *arXiv preprint arXiv:2505.02156*, 2025a.

Wang, M., Li, Y., Wang, H., Zhang, X., Xu, N., Wu, B., Huang, F., Yu, H., and Mao, W. Adaptive thinking via mode policy optimization for social language agents. *arXiv preprint arXiv:2505.02156*, 2025b. URL https://arxiv.org/abs/2505.02156.

Wang, R., Yu, H., Zhang, W., Qi, Z., Sap, M., Neubig, G., Bisk, Y., and Zhu, H. SOTOPIA-p: interactive learning of socially intelligent language agents. In *62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024, August 11, 2024 - August 16, 2024*, volume 1 of *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 12912 – 12940, Bangkok, Thailand, 2024b. Association for Computational Linguistics (ACL). ISBN 0736587X. doi: 10.18653/v1/2024.acl-long.698. URL http://dx.doi.org/10.18653/v1/2024.acl-long.698.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024a.

Wu, Z., Lin, X., Dai, Z., Hu, W., Shu, Y., Ng, S.-K., Jaillet, P., and Low, B. K. H. Prompt optimization with EASE? efficient ordering-aware automated selection of exemplars. In *Proc. NeurIPS*, 2024b.

Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q. V., Zhou, D., and Chen, X. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*, 2023.

Yu, H., Qi, Z., Zhao, Y., Nottingham, K., Xuan, K., Majumder, B. P., Zhu, H., Liang, P. P., and You, J. Sotopia-rl: Reward design for social intelligence. *arXiv preprint arXiv:2508.03905*, 2025.

Zeng, W., Wang, B., Zhao, D., Qu, Z., He, R., Hou, Y., and Hu, Q. Dynamic personality in llm agents: A framework for evolutionary modeling and behavioral analysis in the prisoner's dilemma. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 23087–23100, 2025.

Zhang, W., Liu, T., Song, M., Li, X., and Liu, T. Sotopia-{Omega}: Dynamic strategy injection learning and social instruction following evaluation for social agents. *arXiv preprint arXiv:2502.15538*, 2025.

Zhou, X., Zhu, H., Mathur, L., Zhang, R., Yu, H., Qi, Z., Morency, L.-P., Bisk, Y., Fried, D., Neubig, G., et al. Sotopia: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*, 2024.

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., and Ba, J. Large language models are human-level prompt engineers. In *The eleventh international conference on learning representations*, 2022a.

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., and Ba, J. Large language models are human-level prompt engineers. In *The eleventh international conference on learning representations*, 2022b.

A. Sotopia Evaluation Dimensions

Sotopia (Zhou et al., 2024) defines seven social dimensions for evaluating agent performance in multi-agent interactions. Each dimension captures a distinct aspect of social competence, with scores in different ranges reflecting the nature of the evaluation criterion.

Dimension	Abbr.	Range	Description
Believability	BEL	$[0, 10]$	Naturalness and consistency of agent behavior; whether the agent acts in a believable, human-like manner.
Relationship	REL	$[-5, 5]$	Change in interpersonal relationship quality; positive values indicate improved relationships, negative values indicate deterioration.
Knowledge	KNO	$[0, 10]$	Amount of new and important information gained through the interaction.
Secret	SEC	$[-10, 0]$	Appropriate handling of private information; scores closer to 0 indicate better secret-keeping, while negative scores indicate inappropriate disclosure.
Social Rules	SOC	$[-10, 0]$	Adherence to social norms and rules; scores closer to 0 indicate compliance, while negative scores indicate violations.
Financial & Material	FIN	$[-5, 5]$	Financial and material benefits gained or lost; positive values indicate gains, negative values indicate losses.
Goal	GOAL	$[0, 10]$	Progress toward achieving the agent's private social goal; higher scores indicate greater goal achievement.

Table 4. Sotopia evaluation dimensions with their score ranges and descriptions.

Reward Normalization. To obtain a unified reward signal $r_i \in [0, 1]$ for agent i, we normalize each dimension score to $[0, 1]$ and compute the average:

$$r_i = \frac{1}{7} \sum_{m=1}^{7} \frac{d_m^{(i)} - d_m^{\min}}{d_m^{\max} - d_m^{\min}} \tag{14}$$

where $d_m^{(i)}$ is the raw score for dimension m, and $d_m^{\min}, d_m^{\max}$ are the minimum and maximum values of the dimension's range.

B. Prompts

B.1. Evaluation Prompt Template

The following prompt is used to evaluate agent performance at each dialogue turn:

Per-Turn Evaluation Prompt

{dialogue_history}
Based on previous interactions, evaluate how well participants achieve their goals.
Output format: {format_instructions}
CRITICAL INSTRUCTIONS:

1. Output ONLY a valid JSON object
2. DO NOT repeat or copy the schema definition above

where {dialogue_history} contains the conversation history, and {format_instructions} specifies the JSON schema for the seven evaluation dimensions defined in Appendix A.

B.2. Example Strategy Instruction

Strategy-Instruction-Enhanced Bio Template

{original_bio}
{strategy_description}

where {original_bio} is the agent's background information, and {strategy_description} is one of the 12 social strategies from Table 5.

Table 5. Social Strategy Space: across 6 categories grounded in social science theories.

Category	Strategy
Cooperative	Integrative Negotiation Reciprocity Trigger
Competitive	BATNA Leverage
Strategic	Strategic Self-Presentation
Rational	Axelrod's Tit-for-Tat Rational Choice Persuasion
Diplomatic	Face-Saving Validation Constructive Controversy Outcome Interdependence GRIT Strategy
Exploratory	Active Listening Probe Logrolling Exchange

Example Strategy Instruction

{original_bio}
In your response, apply integrative negotiation principles: adopt a 'we-versus-the-problem' mindset. Focus on underlying interests rather than positions. Actively validate the other party's needs and explicitly frame the interaction as a shared quest for mutual gain. Use phrases like 'How can we solve this together?' and propose creative options that expand the pie rather than just dividing it.

B.3. Example Strategy Space

Example Strategy Space

1. In your response, apply integrative negotiation principles: adopt a 'we-versus-the-problem' mindset. Focus on underlying interests rather than positions. Actively validate the other party's needs and explicitly frame the interaction as a shared quest for mutual gain. Use phrases like 'How can we solve this together?' and propose creative options that expand the pie rather than just dividing it.

2. In your response, leverage the universal reciprocity norm: offer a small, unilateral concession early in the conversation to create a sense of obligation and trigger reciprocal behavior. Frame this concession as a gesture of good faith: 'I want to make this work for you, so I'm willing to give up X...' This builds trust and often yields larger returns through the exchange dynamic.

3. In your response, leverage your BATNA to create urgency and pressure. Imply that your offer is fleeting, or that you have attractive alternatives ready. Push the other party to agree immediately to avoid losing the deal entirely. A strong BATNA shifts bargaining power in your favor.

4. In your response, apply Goffman's dramaturgical approach: treat the interaction as a performance where you control the impression you project. Strategically downplay interest in high-value items or express concern about low-priority issues to shape how the other party perceives your preferences. Your 'front-stage' presentation should be calculated to maximize your negotiating position.

5. In your response, apply Axelrod's winning strategy: start cooperative, then mirror the other party's previous move exactly. If they cooperated, cooperate. If they defected, retaliate. Explicitly link every concession to a specific, equal concession from them. Use 'If-Then' logic: 'If you give me X, then and only then will I consider Y.' This strategy is simple, provocable, forgiving, and clear.

6. In your response, apply rational choice principles: remove emotion and focus purely on logic and data. Articulate the trade-offs explicitly. Present a clear utility calculation showing that accepting your proposal maximizes their expected payoff compared to alternatives. Frame the negotiation as an optimization problem with a rational solution.

7. In your response, apply Politeness Theory's face-saving principles: acknowledge the legitimacy of the other party's position to protect their 'positive face' before introducing your perspective. Use validating language that honors their viewpoint, then smoothly transition with 'At the same time...' or 'Building on that...' to guide the conversation toward common ground without threatening their self-image.

8. In your response, apply Constructive Controversy principles: frame the disagreement as a productive catalyst for better solutions rather than a threat. Acknowledge that differing perspectives are valuable—'It makes sense that we see this differently, and that difference might help us find a better solution.' Encourage intellectual conflict while maintaining cooperative goals, reducing defensiveness and opening space for creative problem-solving.

9. In your response, apply Interdependence Theory: emphasize that both parties' outcomes are mutually dependent. Highlight the mutual costs of failing to reach agreement—illustrate what both stand to lose. Then pivot to showing how cooperation serves everyone's interests better than continued disagreement, making the interdependent nature of the situation explicit.

10. In your response, apply the GRIT strategy from conflict resolution: when facing deadlock or high tension, announce and execute a small, unilateral conciliatory step. Invite (but don't demand) reciprocation. If the other party responds positively, escalate cooperation gradually. This graduated approach builds trust incrementally while preserving your ability to retreat if exploited.

11. In your response, apply Carl Rogers' active listening approach: gently probe for unspoken concerns that may be creating implicit conflict. Use open-ended questions and reflective statements to invite the other party to reveal underlying hesitations. Then address these concerns directly while showing how your proposal accommodates them. The goal is to surface the 'real' issues beneath the stated positions.

12. In your response, apply the logrolling principle from integrative bargaining: identify issues where you and the other party have different priority levels, then propose trades that give each side more of what they value most. 'I care more about X, you care more about Y—what if I concede on Y in exchange for X?' This creates value by exploiting preference asymmetries rather than splitting differences.

B.4. Domain Generation

Strategy Paraphrase Prompt

You are an expert in social psychology and negotiation theory. Your task is to generate semantically equivalent paraphrases of social strategies.
Original Strategy: {strategy_name}: {strategy_description}
Theoretical Basis: {theory}
Instructions:

1. Generate {n} paraphrased versions of this strategy.
2. Each paraphrase must:
   - Preserve the core behavioral intent and theoretical grounding.
   - Use different wording, sentence structures, and examples.
   - Be directly usable as an agent prompt.
3. Vary the linguistic style: some formal, some conversational.
4. Do NOT change the underlying negotiation tactic.

Output Format:

```
{
  "original_id": "<strategy_id>",
  "paraphrases": [
    {"id": "<strategy_id>_v1", "description": "..."},
    {"id": "<strategy_id>_v2", "description": "..."},
    ...
  ]
}
```

We use GPT-5 (Singh et al., 2025) to generate the strategy space.

B.5. OPRO Meta-Prompt

OPRO Meta-Prompt

Your task is to generate an agent bio description that helps the agent achieve better social interaction outcomes. Below are some previous bio descriptions with their scores. The scores range from 0 to 1, where higher scores indicate better social performance:
{instruction_score_pairs}
Generate a new bio description that is different from all the descriptions above and has a higher score than all of them.
Requirements for the new bio: 1. Be concise and actionable (under 200 words) 2. Be distinct from existing descriptions - do not simply rephrase
Write your new bio description in the following format: <BIO>your bio here</BIO>

Figure 8. OPRO meta-prompt template. The placeholder {instruction_score_pairs} is filled with previous strategies and their scores in ascending order.

B.6. EvoPrompt-GA

EvoPrompt-GA Crossover + Mutation

Please follow the instruction step-by-step to generate a better agent bio description.
1. Crossover the following agent bios and generate a new bio: Bio 1: <bio1> Bio 2: <bio2>
2. Mutate the bio generated in Step 1 and generate a final bio bracketed with <BIO> and </BIO>.

Figure 9. EvoPrompt-GA template implementing genetic crossover and mutation.

C. Experiment Details

We provide detailed hyperparameter configurations for all baseline methods and our proposed approach. All experiments are conducted on a single NVIDIA A800 GPU with 80GB memory. We use OpenRouter API for LLM inference.

C.1. Common Settings

All methods share the following experimental settings:

Table 6. Common experimental settings across all methods.

Parameter	Value
Max Interaction Turns	20
Agent LLM	Qwen-2.5-72B-Instruct / DeepSeek-V3.2
Reward Evaluator	DeepSeek-V3.2
Strategy Embedding Model	Qwen3-Embedding-8B
Embedding Dimension	4096
Optimization Target	Both agents (P1 & P2)

C.2. EvoPrompt

We implement EvoPrompt following Guo et al. (2024), using the Genetic Algorithm (GA) variant which demonstrated superior performance in their experiments.

Table 7. EvoPrompt (GA) hyperparameters.

Parameter	Value
Evolution Mode	Genetic Algorithm (GA)
Population Size	5
Evolution Interval	5 turns
Mutation Rate	0.2
Elite Ratio	0.4
Selection Mode	Roulette Wheel
Population Update	Top-K
Mutation Model	Qwen-2.5-7B-Instruct
Mutation Temperature	0.3
Selection Strategy	Round-Robin
Score Update Method	Replace

The GA variant performs crossover between two parent strategies selected via roulette wheel selection, followed by LLM-based mutation. Elite strategies (top 40%) are preserved across generations.

C.3. OPRO

We implement OPRO following Yang et al. (2023), using meta-prompts with instruction-score history to guide the LLM optimizer.

OPRO maintains a history of instruction-score pairs and uses this history as context for the LLM optimizer to generate new candidate strategies. Evolution is triggered after all strategies in the population have been evaluated once.

C.4. INSTINCT

We implement Neural UCB following the NeuralTS-Diag variant from Lin et al. (2024b), which uses per-sample gradients computed via backpack to estimate uncertainty.

Table 8. OPRO hyperparameters.

Parameter	Value
Population Size	5
Max Instructions in Meta-Prompt	10
Score Threshold	0.5
Evolution Trigger	Round Complete
Optimizer Model	Qwen-2.5-7B-Instruct
Optimizer Temperature	1.0
Selection Strategy	Round-Robin
Score Update Method	Replace

Table 9. Neural UCB hyperparameters.

Parameter	Value
Network Architecture	2-layer MLP
Hidden Size	128
Activation	ReLU
λ (Regularization)	0.1
ν (Exploration)	1.0
Learning Rate	0.01
Training Epochs	100

The UCB score is computed as:

$$\text{UCB}(a) = \hat{\mu}(a) + \nu \sqrt{\sum_i \frac{\lambda \cdot g_i(a)^2}{U_i}} \tag{15}$$

where $\hat{\mu}(a)$ is the predicted reward, $g_i(a)$ are per-sample gradients, and U_i is the diagonal of the incrementally updated gradient covariance matrix.

C.5. Adversarial Online Strategy Optimization

Our method combines neural value estimation with EXP3-style exploration, augmented with context embeddings for scenario-aware strategy selection.

Table 10. Context-Aware Neural Adversarial Bandit hyperparameters.

Parameter	Value
Network Architecture	2-layer MLP with Pre-LN
Hidden Size	512
Activation	GELU
η (Exploration)	10.0
γ (EXP3 mixing)	0.1
Score Decay	0.9
Importance Weighted Reward	Enabled
Context Embedding	Enabled (4096-dim)
Context Embedding Model	Qwen3-Embedding-8B
Learning Rate	0.001
Training Epochs	100

The selection probability is computed as:

$$p(a) = (1-\gamma) \cdot \text{softmax}\left(\eta \cdot \hat{V}(a,c)\right) + \frac{\gamma}{K} \tag{16}$$

where $\hat{V}(a, c)$ is the neural network's value estimate for action a given context embedding c, and K is the number of arms.

Pre-LayerNorm: We use Pre-LN structure (LayerNorm before each linear layer) for more stable training dynamics. **Context Embedding**: The scenario context (agent profiles, social goals, environmental constraints) is encoded via Qwen3-Embedding-8B and concatenated with strategy embeddings.

C.6. Strategy Selection and Convergence Across Diverse Scenarios

Figure 10. (a-d) Strategy selection trajectories for four representative scenarios, showing how the bandit algorithm converges to scenario-specific optimal strategies over conversation turns. Each colored dot represents the strategy selected at that turn, with the dashed horizontal line indicating the most frequently selected strategy. Different scenarios converge to distinct strategies: Face-Saving for relationship-sensitive negotiations (a), Integrative Negotiation for collaborative problem-solving (b), Rational Choice for analytical discussions (c), and Reciprocity Trigger for trust-building interactions (d). (e) Average final rewards achieved by each strategy across all 450 scenarios (900 agent-strategy pairs). Strategies are ranked by effectiveness, with the red dashed line indicating the No Strategy baseline (3.79). All social strategies outperform the baseline, with Rational Choice and Constructive Controversy achieving the highest average rewards (4.00), demonstrating that adaptive strategy selection based on scenario context leads to improved social interaction outcomes.

D. More Case Study

Table 11. **Resource Allocation (Fruit Division).** Comparison of information exchange strategies (Turns 2-4).

Vanilla: No Strategy – The Missed Connection
Hendrick (Turn 2): "Hi Lily... **I do enjoy apples quite a bit, but I also have a soft spot for oranges. Bananas are nice, but they're not my top choice.** How about you?" **Lily (Turn 3):** "I value bananas the most... How about I take the two bananas and one apple, and you can have the other apple and the orange?" *(Result: Suboptimal. 1 Apple left unallocated. Joint Utility = 10)*

ALSO: Dynamic Strategy Instruction
Hendrick (Turn 2): "For me, **apples are the top choice, followed by the orange, and then bananas.** It makes sense that we see this differently... What about you?"
`[ALSO] Strategy: Surface Hidden Concerns` *Prompt: "...gently probe for unspoken concerns... address these concerns directly while showing how your proposal accommodates them."* **Lily (Turn 3):** "My preferences are **bananas first, then the orange, and finally apples.** Given our different preferences... **If you take 2 apples and I take 1 apple, 2 bananas, and the orange**, both of us get a good share..."
`[ALSO] Strategy: Address Asymmetric Needs` *Prompt: "...acknowledge that you and the other party may have different priorities... propose creative arrangements that give each side more of what matters most..."* **Hendrick (Turn 4):** "That sounds like a fair proposal... By taking 2 apples, I get my top choice, and you maximize your points... **It's a win-win solution.**" *(Result: Optimal. All resources allocated. Joint Utility = 14)*