SOCIALRL: REFINING LLMS' SOCIAL INTELLIGENCE THROUGH MULTI-TURN REINFORCEMENT LEARNING AND REWARD DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Social intelligence enables agents to read social context, infer intent, and adapt over sustained dialogue. As language models become autonomous collaborators, it is central to building effective and trustworthy human-AI interaction. Existing reinforcement learning methods optimize single-turn utterances and sparse outcome rewards, producing short-sighted policies that struggle to manage goal-relationship tensions across multi-turn interactions. We propose **SocialRL**, a multi-turn reinforcement learning framework addressing both challenges. First, we apply multi-turn reinforcement learning using PPO that propagates delayed outcome rewards back to each turn, enabling long-horizon planning. Second, we design six process reward dimensions capturing the goal-relationship trade-off, including goal advancement, relational attunement, contextual coherence, etc. A reward model dynamically generates fine-grained scoring criteria for each dimension, while a stage-aware weight schedule prioritizes relationship-building in early turns, goal advancement mid-way, and balanced closure late. Across multiple social-dialogue benchmarks, SocialRL improves Goal Achievement by an average of 9.2 percentage points over the corresponding Base models. These results demonstrate the effectiveness of SocialRL across synthetic and real social scenes, as well as standard and challenging social scenarios.

1 INTRODUCTION

Social intelligence is the ability to understand social dynamics, recognize others' intentions, and adjust one's responses throughout extended conversations (Zhou et al., 2024a). For humans, social intelligence enables meaningful relationships with friends, family, and colleagues, creating mutual benefit and trust (Stafford & Canary, 1991). As language models transition from information systems to autonomous agents, social intelligence becomes essential, determining whether AI can collaborate with humans in depth and earn their confidence. This skill requires a fundamental tension: pursuing one's goals while preserving relationships with others. Social dialogue possesses properties that make it a natural fit for reinforcement learning (Ouyang et al., 2022). Interactions unfold sequentially, with each turn responding to the present moment while shaping what comes next. Meaningful outcomes—goal achievement and relationship quality—emerge only at conversation's end, providing delayed feedback that RL can leverage.

Existing methods mainly fall into two categories: single-turn reinforcement learning methods and outcome reward training methods. Single-turn reinforcement learning methods, such as Sotopia-RL (Yu et al., 2025), optimize individual utterances but neglect cumulative value over multi-turn dialogues. Outcome reward training methods, like ArCHer (Zhou et al., 2024b), use final outcome scores to guide learning, yet provide little intermediate supervision for how each turn should manage social trade-offs. However, both remain limited in modeling multi-turn social dynamics: they either favor locally fluent responses without optimizing conversation-level behavior, or rely on sparse outcome feedback without process supervision. Effective learning requires optimizing multi-turn dialogue trajectories while decomposing social quality into turn-level, multi-dimensional rewards that balance goal progress and relationship health as conversations evolve.

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

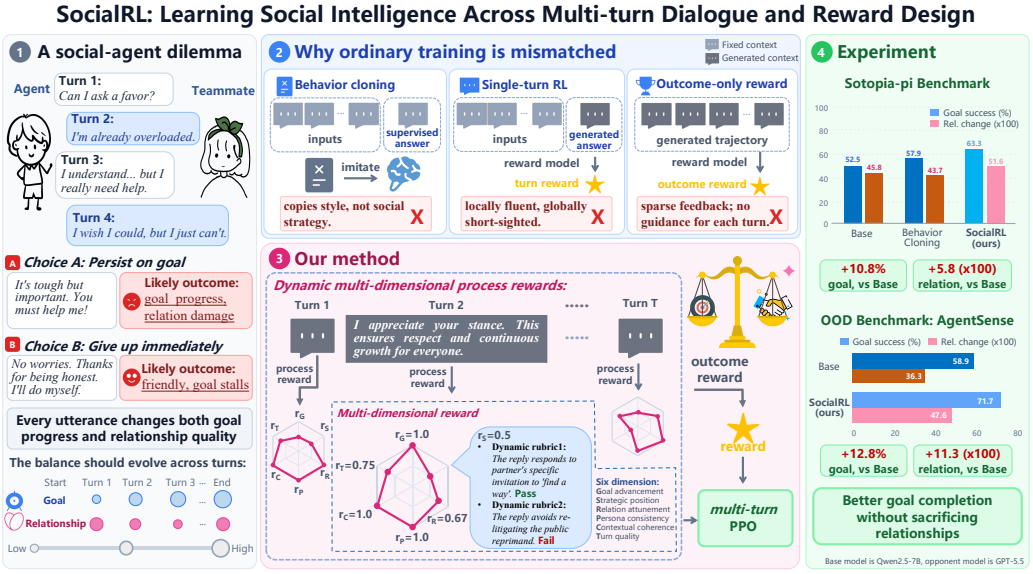


Figure 1: Overview of the SocialRL motivation and solution. Social dialogue requires agents to balance private goals with relationship management over multi-turn interactions. Prior single-turn reinforcement learning or sparse outcome reward methods provide limited long-horizon planning, while SocialRL combines multi-turn reinforcement learning with process rewards.

To address these challenges, we introduce **SocialRL**, a multi-turn reinforcement learning framework with a multi-dimensional reward system for multi-turn social dialogue, which has two core innovations:

1. **Multi-turn trajectory optimization.** We treat multi-turn conversation trajectories as the optimization unit and train with PPO online method. A value network estimates returns from the turn-level reward sequence, allowing the policy to optimize goal pursuit and relationship maintenance over complete dialogues rather than isolated responses.

2. **Multi-dimensional dynamic process rewards.** We decompose the goal-relationship trade-off into six process reward dimensions: goal advancement and strategic positioning on the goal side (Locke & Latham, 1990; Kellermann, 1992; Berger, 1997), relational attunement and persona consistency on the relationship side (Stafford & Canary, 1991; Goffman, 1959), and contextual coherence and turn quality as enabling conditions for both (Grice, 1975; Sacks et al., 1974). A reward model generates context-specific scoring criteria for each dimension after every turn, while a stage-aware weight schedule adapts priorities across dialogue phases, prioritizing relationship-building early, goal advancement mid-way, and balanced closure near the end, grounded in classic theories such as Bales's Interaction Process Analysis (Bales, 1950) and Knapp's relational stage model (Knapp, 1978).

Large-scale experiments are conducted on SOTOPIA (Zhou et al., 2024a), SOTOPIA-π (Wang et al., 2024), and AgentSense (Mou et al., 2025). After aligning the $[0, 10]$ SOTOPIA-All and SOTOPIA-Hard scores to percentage-point units, SocialRL improves Goal Achievement by 9.2 percentage points on average across the four benchmarks relative to the corresponding Base models.

Our main contributions are as follows. (1) We apply multi-turn reinforcement learning to social dialogue, enabling long-horizon planning that balances goal pursuit and relationship management. (2) We construct a multi-dimensional process reward system whose six dimensions are grounded in the goal-relationship structure of social dialogue quality, combined with a stage-aware weight schedule that adapts to shifting priorities across dialogue phases. (3) Our trained policies outperform the compared non-commercial baselines on SOTOPIA series and AgentSense tasks, while commercial reference models remain an upper bound in several settings.

Table 1: Comparison of social dialogue optimization methods discussed in Related Work.

Method	Unit	Feedback / reward	Value model
Imitation			
BC	Single-turn	Expert demonstrations	None
Preference / reward design			
SDPO	Segment	Preference pairs	None
ARIA	Multi-turn	Intention-space reward aggregation	None
SAVOIR	Multi-turn	Expected-utility Shapley attribution	Utterance reward model
ASL / AMPO	Mode-adaptive	Answer, format, and length rewards	Mode/sample advantages
Single-turn reinforcement learning			
Sotopia-RL (GRPO)	Single-turn	Reward-model scores	GRPO normalization
Multi-turn reinforcement learning			
ArCHer	Multi-turn	Macro-level returns	Macro-level value function
SVPO	Multi-step	Step-level preference pairs	Explicit value model
REFUEL	Multi-turn	Relative future-return regression	Relative Q estimation
OMAR	Multi-turn	Multi-agent self-play	Value estimates with GAE
DEPT	Multi-turn self-play	Verifiable game outcomes	asymmetric advantages
SocialRL	**Full trajectory**	**Multi dimension process + outcome rewards**	**PPO value network**

2 Related Work

Social intelligence requires language agents to infer intent, track interaction state, and plan long-horizon strategies while balancing private goals with relationship management.

Social Intelligence of LLMs. Recent benchmarks instantiate this challenge in interactive social scenarios. SOTOPIA (Zhou et al., 2024a) and AgentSense (Mou et al., 2025) cover negotiation, empathy, and information verification, while ToMBench (Chen et al., 2024b) probes theory-of-mind skills; Lifelong-SOTOPIA (Goel & Zhu, 2025) extends evaluation to long-term, multi-scenario relationships. These environments provide standardized, human-aligned evaluation, but successful agents must reason over multi-turn trajectories rather than isolated turns.

Beyond benchmark design, a second line of work improves social behavior without online value modeling, via imitation, preference learning, or reward redesign. Behavior cloning from the expert trajectories in SOTOPIA-π (Wang et al., 2024) is fluent but limited by data coverage. SDPO (Kong et al., 2025) applies segment-level preference optimization, ARIA (Yang et al., 2025) aggregates rewards in an intention space, and SAVOIR (Feng et al., 2026) attributes multi-turn outcomes to utterances via expected utility and Shapley values. Adaptive Social Learning (ASL) (Wang et al., 2026a) introduces hierarchical reasoning modes and Adaptive Mode Policy Optimization (AMPO), which combines mode-level and sample-level advantages to adapt reasoning depth and reduce token use. These approaches give finer supervision or more adaptive inference, but they do not jointly optimize full dialogue trajectories with dense, stage-dependent social process rewards.

Reinforcement Learning for LLMs. Reinforcement learning training (Ouyang et al., 2022) optimizes interaction outcomes beyond fixed demonstrations. Single-turn reinforcement learning methods such as Sotopia-RL (Yu et al., 2025), which applies GRPO (Shao et al., 2024), optimize each utterance independently and thus miss how early actions shape later social outcomes. Recent methods extend reinforcement learning or value learning to longer sequences: ArCHer (Zhou et al., 2024b) learns a high-level value function over interaction turns, SVPO (Chen et al., 2024a) learns step-level preferences and values for mathematical reasoning, REFUEL (Gao et al., 2025) regresses relative future returns in multi-turn RLHF, and OMAR (Jiang et al., 2026) trains conversational agents via multi-agent self-play. DEPT (Wang et al., 2026b) applies asymmetric advantage reshaping to restore low outcome entropy and vanishing gradients. Yet sparse outcome rewards still give high-variance value estimates, and none explicitly model the stage-dependent goal–relationship trade-off.

In contrast, we jointly optimize full multi-turn trajectories and dense process rewards; as Table 1 shows, no prior method combines all three.

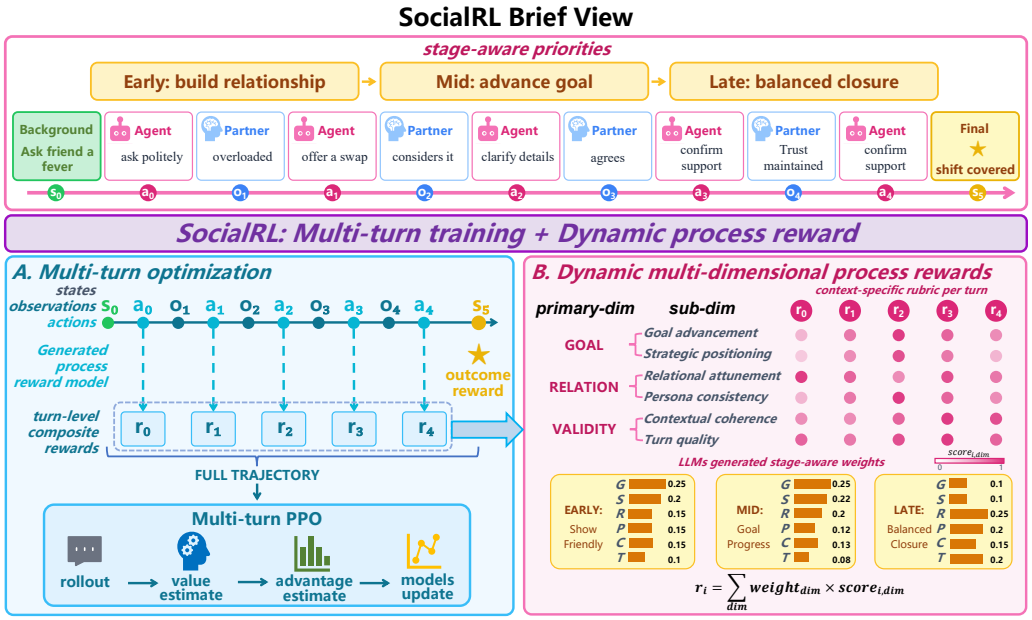


Figure 2: SocialRL has two core components. Multi-turn trajectory optimization uses a value network, GAE, and PPO to estimate returns from turn-level rewards and optimize complete dialogues. Dynamic process reward design scores each policy utterance along six social dimensions, applies context-inferred stage-aware weights, and combines dense process feedback with outcome reward at the final turn.

3 SocialRL

SocialRL trains dialogue policies with multi-turn reinforcement learning. We model each interaction as a finite-horizon dialogue MDP, optimize multi-turn trajectories with PPO, and compute dense process rewards from multi-dimensional social feedback with stage-aware weights. Figure 2 summarizes the overall training framework.

3.1 Task Formulation

Each task instance is a triple (S, G, C), where S is the scenario, G is the agent's private social goal, and C contains persona information for both parties. We model the interaction as a finite-horizon MDP (Sutton & Barto, 2018; Puterman, 1994) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \rho_0, M)$. The state contains (S, G, C) and the dialogue history; actions are tokens grouped into utterances; P appends both the agent utterance and the counterpart response; γ is the discount factor, ρ_0 is the initial-state distribution, and M is the maximum number of turns. Rewards are assigned at the utterance level: tokens within a turn receive zero reward, and the completed utterance receives $r_m \in [-R_{\max}, R_{\max}]$.

Thus a trajectory is $\tau = \big(s_0,\ \mathbf{a}_0,\ r_0,\ s_1,\ \ldots,\ \mathbf{a}_{M-1},\ r_{M-1},\ s_M\big)$, with utterance likelihood $\pi_\theta(\mathbf{a}_m \mid s_m) = \prod_{t=1}^{L_m} \pi_\theta\Big(a_t^{(m)} \mid s_m, a_{<t}^{(m)}\Big)$, where L_m is the length of utterance m, a_t is the t^{th} token of utterance. Besides, a trajectory may terminate early after goal completion. With the composite reward $r_m = \tilde{R}_m$ defined in Section 3.3, the policy maximizes $J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\big[R(\tau)\big]$. The counterpart's private goal is hidden and their responses are stochastic, making conversation-level optimization under evolving dialogue states essential.

3.2 Training Algorithm

SocialRL uses PPO (Schulman et al., 2017) over multi-turn dialogue trajectories. For each scenario, the current policy interacts with the counterpart for up to M turns; after rollout, each utterance

receives the composite reward $\tilde{R}_t$ from Section 3.3. A value network estimates future returns from these turn-level rewards to compute advantages for trajectory optimization. The complete training procedures for our PPO implementation and the corresponding multi-turn GRPO variant are given in Appendix A.2.

We compute Generalized Advantage Estimation (GAE) (Schulman et al., 2016) from turn-level rewards:

$$\delta_t = \tilde{R}_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t), \tag{1}$$

$$\hat{A}_t = \sum_{l=0}^{M-t-1} (\gamma\lambda)^l \, \delta_{t+l}, \tag{2}$$

where $V_\phi(s_M) = 0$ at the final turn and $\lambda = 0.95$. Unlike token-level RLHF, we treat each utterance as one optimization unit: all tokens in the utterance share the same reward and advantage. The clipped policy objective is

$$L^{\mathrm{CLIP}}(\theta) = \mathbb{E}_t\Big[\min\big(\rho_t(\theta)\,\hat{A}_t,\ \mathrm{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon)\,\hat{A}_t\big)\Big], \tag{3}$$

where $\rho_t(\theta) = \pi_\theta(a_t \mid s_t)/\pi_{\theta_{\mathrm{old}}}(a_t \mid s_t)$ is the utterance likelihood ratio and $\epsilon = 0.2$. The value network is trained with

$$\mathcal{L}^V(\phi) = \mathbb{E}_t\big[(V_\phi(s_t) - G_t)^2\big], \tag{4}$$

where $G_t = \sum_{l=0}^{M-t-1} \gamma^l \tilde{R}_{t+l}$ is the discounted return from turn t.

We use PPO rather than GRPO (Shao et al., 2024) because social dialogue rewards are delayed and strongly state-dependent. GRPO normalizes rewards across a group of sampled trajectories and replaces PPO's state baseline with a group-level baseline. The two update pipelines are shown side by side in Appendix A.2 for direct comparison. For turn m, PPO uses

$$b_m^{\mathrm{PPO}} = V^\pi(s_m) = \mathbb{E}[G_m \mid s_m], \tag{5}$$

GRPO instead uses

$$b_m^{\mathrm{GRPO}} = \mu_0 D_m, \qquad D_m = \sum_{k=m}^{M-1} \gamma^{k-m}, \tag{6}$$

where μ_0 is the large-group mean of the turn-level rewards. GRPO has a finite-group correction that decreases as group size grows, while PPO's state-only baseline leaves the reference gradient target unchanged. The key long-horizon difference is variance: under the standard policy-gradient baseline approximation, the unnormalized GRPO estimator adds a mismatch term whenever dialogue states with the same turn index have different values:

$$\widetilde{V}_{\mathrm{GRPO}} \approx V_{\mathrm{PPO}} + \Delta, \qquad \Delta = \sum_{m=0}^{M-1} \mathbb{E}\Big[\|U_m\|^2\big(V^\pi(s_m) - \mu_0 D_m\big)^2\Big] \geq 0, \tag{7}$$

where $U_m = \nabla_\theta \log \pi_\theta(a_m \mid s_m)$. The actual normalized GRPO variance additionally carries the global factor $1/\sigma_0^2$. Since the same turn index can correspond to negotiation, repair, compromise, or failure, the baseline-mismatch term becomes more pronounced in longer-horizon dialogues. Appendix A provides the derivation.

3.3 Reward Design

Social dialogue requires feedback on both goal progress and relationship management. Since outcome-only scores provide no direct supervision for intermediate social behavior over 10–20 turns, SocialRL uses a multi-dimensional process reward and combines it with the outcome reward (Lightman et al., 2023). At turn t, the composite reward is

$$\tilde{R}_t = \begin{cases} \alpha \cdot R_{\mathrm{process},t} + \beta \cdot R_{\mathrm{outcome}}, & t = T, \\ \alpha \cdot R_{\mathrm{process},t}, & t < T, \end{cases} \tag{8}$$

where T is the final turn, $R_{\mathrm{process},t}$ scores the current utterance, R_{outcome} scores final success, and validation selects $\alpha = 0.3, \beta = 1$. The process reward is

$$R_{\mathrm{process},t} = \sum_{i=1}^{6} w_{t,i} r_{t,i}, \tag{9}$$

with normalized non-negative weights

$$\sum_{i=1}^{6} w_{t,i} = 1, \qquad w_{t,i} \geq 0. \tag{10}$$

Intermediate turns therefore still produce useful gradients even when final success is not yet known or the dialogue ultimately fails.

Each utterance is scored along six dimensions: *goal advancement*, *strategic positioning*, *relational attunement*, *persona consistency*, *contextual coherence*, and *turn quality*. These dimensions separate goal-side progress, relationship-side maintenance, and general conversational validity, drawing on goal-setting theory (Locke & Latham, 1990), strategic communication (Kellermann, 1992; Berger, 1997), relationship maintenance (Stafford & Canary, 1991), self-presentation (Goffman, 1959), relevance (Grice, 1975), and conversation analysis (Sacks et al., 1974).

To make scoring context-specific, we use a rubric mechanism inspired by LLM-based evaluation and criterion-guided feedback (Zheng et al., 2023; Liu et al., 2023; Bai et al., 2022). Given scenario S, persona C, goal G, history H_t, and utterance a_t, the reward model generates 2–4 binary criteria for each dimension and judges whether each criterion passes. The dimension score is the pass rate:

$$r_{t,i} = \frac{1}{M_i} \sum_{j=1}^{M_i} \mathrm{pass}_{i,j}, \tag{11}$$

where $\mathrm{pass}_{i,j} \in \{0, 1\}$. Binary criteria are more stable than abstract scalar judgments and make reward feedback inspectable.

Finally, SocialRL changes dimension weights across dialogue stages. Following interaction and relational-stage theories (Bales, 1950; Knapp, 1978), early turns emphasize coherence, persona, and rapport; middle turns emphasize goal advancement and strategy; late turns balance closure with relationship preservation. The reward model infers the current stage from context rather than fixed turn-ratio thresholds:

$$\phi_t = f_{\mathrm{stage}}(S, C, G, H_t, a_t), \qquad \phi_t \in \{\mathrm{Early}, \mathrm{Mid}, \mathrm{Late}\}. \tag{12}$$

Given ϕ_t, it proposes $\hat{\mathbf{w}}_t$, which is sanitized by whitelist filtering, clipping, caps, fallback priors, overflow redistribution, and normalization before computing $R_{\mathrm{process},t}$.

4 Experiments

We evaluate SocialRL on four benchmarks with four trained backbones and four opponent models. We ask whether it improves social dialogue performance, which design choices account for the gains, and whether it induces long-horizon strategies.

4.1 Experimental Setup

We use two judge-based metrics (Zheng et al., 2023; Liu et al., 2023): Goal Achievement and Relationship Change. Goal Achievement is reported as a percentage on SOTOPIA-π and AgentSense, and as a $[0, 10]$ score on SOTOPIA-All and SOTOPIA-Hard; Relationship Change is reported after multiplying the SOTOPIA-π and AgentSense values by 100, while the original SOTOPIA benchmarks retain their $[-5, 5]$ scale. Unless otherwise stated, reported means and standard deviations are computed from five independent repeated experiments of each policy–opponent pair. Prompts are provided in Appendix B.

We use three complementary interactive benchmarks. SOTOPIA-π is constructed from cleaned and filtered synthetic dialogues (Wang et al., 2024); it contains 1,773 scenarios across seven social contexts, including 259 held-out test scenarios. Its multi-turn interactions evaluate whether an agent can pursue a social goal while maintaining the counterpart's relationship quality. SOTOPIA-All and SOTOPIA-Hard are based on the open-ended SOTOPIA environment (Zhou et al., 2024a). SOTOPIA-All covers 90 diverse everyday social scenarios, whereas SOTOPIA-Hard is a 14-scenario subset selected for higher conflict, ambiguous intentions, and subtle social norms. AgentSense is an independent benchmark built bottom-up from social situations extracted from

Table 2: Results on the SOTOPIA-π benchmark: Goal Achievement (%) / Relationship Change ($\times 100$) for each policy–opponent pair. Values are means across repeated experiments; complete mean $\pm$ SD results are in Appendix 8. Relationship Change values are multiplied by 100. Bold denotes the best SocialRL model per metric and opponent column. SocialRL has improved performance on each same-backbone network and in the average results among non-commercial methods.

	Opponent Model				
Method	Qwen2.5-7B	Qwen3-8B	Qwen3.5-35B	GPT-5.5	Avg.
Base (Qwen2.5-7B)	33.5 / 23.5	38.4 / 22.8	44.6 / 32.1	52.5 / 45.8	42.3 / 31.1
Base (Qwen3-8B)	43.2 / 13.6	44.2 / 8.3	55.0 / 13.7	60.9 / 24.8	50.8 / 15.1
Base (LLaMA3.1-8B)	44.3 / 21.7	47.1 / 19.4	55.0 / 17.3	63.2 / 37.1	52.4 / 23.9
Base (Gemma-3-4B)	22.0 / -10.6	30.3 / -6.3	32.1 / -15.0	37.6 / -5.4	30.5 / -9.3
BC (Qwen2.5-7B)	38.9 / 27.2	43.0 / 23.6	49.9 / 22.3	57.9 / 43.7	47.4 / 29.2
SDPO (Qwen2.5-7B)	33.1 / 24.0	42.3 / 23.9	45.3 / 31.1	55.0 / 44.9	43.9 / 31.0
Sotopia-RL (Qwen2.5-7B)	41.2 / 26.0	43.7 / 20.9	49.0 / 17.8	60.5 / 39.4	48.6 / 26.0
ArCHer (Qwen2.5-7B)	31.6 / 22.6	39.6 / 22.8	46.6 / 31.3	48.6 / 41.0	41.6 / 29.4
SocialRL (Qwen2.5-7B)	43.3 / 28.2	45.8 / 26.9	56.7 / **36.3**	63.3 / 51.6	52.3 / **35.8**
SocialRL (Qwen3-8B)	47.7 / 18.8	48.1 / 20.8	61.0 / 18.1	62.2 / 24.9	54.7 / 20.6
SocialRL (LLaMA3.1-8B)	**50.2** / **31.2**	**54.3** / **27.6**	**63.6** / 24.9	**70.2** / **53.0**	**59.6** / 34.1
SocialRL (Gemma-3-4B)	42.4 / 10.6	48.0 / 6.4	58.6 / 13.7	60.3 / 22.9	52.3 / 13.4
GPT-5.5 (reference)	57.0 / 31.6	64.9 / 32.2	75.3 / 40.1	78.9 / 51.4	69.0 / 38.8

movie and television scripts (Mou et al., 2025). It provides 1,225 scenarios organized into 245 templates with synthetic character instantiations, and evaluates both goal completion and latent social reasoning, such as inferring private information from dialogue.

We train Qwen2.5-7B-Instruct, Qwen3-8B, LLaMA3.1-8B, and Gemma-3-4B policies against Qwen2.5-7B, Qwen3-8B, Qwen3.5-35B, and GPT-5.5 opponents. Baselines include base models, Behavior Cloning (BC), Sotopia-RL (Yu et al., 2025), SDPO (Kong et al., 2025), ArCHer (Zhou et al., 2024b), and commercial reference models. Implementation details are in Appendix C.

4.2 Main Results

SocialRL improves goal achievement and relationship maintenance. We first evaluate the main two-party setting on SOTOPIA-π (Table 2). The table reports Goal Achievement success rates and Relationship Change across four opponents. Across opponents, SocialRL with the Qwen2.5-7B backbone averages 52.3% / 0.358 in Goal Achievement / Relationship Change, compared with 42.3% / 0.311 for Base. The strongest trained-model average is obtained by the LLaMA3.1-8B policy (59.6% / 0.341), while GPT-5.5 reaches 69.0% / 0.388 as a commercial reference.

Across opponents, LLaMA3.1-8B has the strongest trained-model average in Goal Achievement (59.6%), while Qwen2.5-7B leads in Relationship Change (0.358). The smaller Gemma-3-4B backbone also benefits substantially: its average rises from $30.5\%/-0.093$ for Base to $52.3\%/0.134$ with SocialRL, improving by 21.8 percentage points in Goal Achievement and 0.227 in the displayed Relationship Change scale.

As a separate and out of distribution test of multi-party interaction and latent social reasoning, AgentSense results are shown in Table 3. SocialRL with Qwen3-8B reaches $80.5\%/0.444$ on average, exceeding the corresponding Base model by 7.2 percentage points in Goal Achievement and 0.052 in the displayed Relationship Change scale. The improvement is especially clear for Gemma-3-4B, which rises from $46.3\%/0.065$ to $68.1\%/0.234$. These results extend the SOTOPIA-π findings to script-derived, multi-party scenarios where success also requires inferring information that is not stated explicitly. we report SOTOPIA-All and SOTOPIA-Hard in Appendix D.

To summarize performance across settings, we use each table's opponent-averaged Goal Achievement and compute the absolute SocialRL–Base difference for the same backbone, then average across the four trained backbones. This yields gains of 10.725 points on SOTOPIA-π, 9.625 on

Table 3: AgentSense results: Goal Achievement Success Rate (%) / Relationship Change ($\times 100$) across multi-party scenarios. Bold denotes the best SocialRL model per opponent. Values are means across five independent repeats; complete mean $\pm$ SD results are in Appendix 9. Relationship Change values are multiplied by 100. SocialRL consistently outperforms BC, SDPO, and Sotopia-RL on both metrics.

	Opponent Model				
Method	Qwen2.5-7B	Qwen3-8B	Qwen3.5-35B	GPT-5.5	Avg.
Base (Qwen2.5-7B)	53.7 / 32.8	55.7 / 33.6	65.2 / 40.9	58.9 / 36.3	58.4 / 35.9
Base (Qwen3-8B)	67.8 / 33.7	72.9 / 38.9	69.6 / 35.8	82.8 / 48.4	73.3 / 39.2
Base (LLaMA3.1-8B)	83.2 / 48.1	85.2 / 49.2	86.2 / 51.9	89.4 / 58.1	86.0 / 51.8
Base (Gemma-3-4B)	44.4 / 6.3	45.4 / 6.1	48.1 / 9.7	47.5 / 3.7	46.3 / 6.5
BC (Qwen2.5-7B)	60.3 / 36.7	56.7 / 32.8	63.1 / 37.9	67.0 / 46.9	61.8 / 38.6
SDPO (Qwen2.5-7B)	57.6 / 35.8	58.0 / 36.3	63.0 / 37.4	66.7 / 48.1	61.3 / 39.4
Sotopia-RL (Qwen2.5-7B)	63.5 / 37.1	59.3 / 33.5	63.9 / 35.8	69.6 / 49.4	64.1 / 38.9
ArCHer (Qwen2.5-7B)	51.1 / 31.9	52.6 / 32.0	60.9 / 38.9	58.9 / 35.7	55.9 / 34.6
SocialRL (Qwen2.5-7B)	61.0 / 38.6	62.2 / 38.5	72.1 / 45.9	71.7 / 47.6	66.7 / 42.7
SocialRL (Qwen3-8B)	76.7 / 39.5	77.1 / 40.9	82.0 / 46.3	86.0 / 50.7	80.5 / 44.4
SocialRL (LLaMA3.1-8B)	**84.3 / 48.3**	**85.9 / 50.1**	**87.6 / 52.4**	**90.9 / 58.9**	**87.2 / 52.4**
SocialRL (Gemma-3-4B)	66.7 / 22.5	66.4 / 22.3	68.8 / 24.7	70.4 / 23.9	68.1 / 23.4
GPT-5.5 (reference)	88.1 / 45.6	92.2 / 48.9	93.8 / 52.6	92.5 / 48.9	91.7 / 49.0

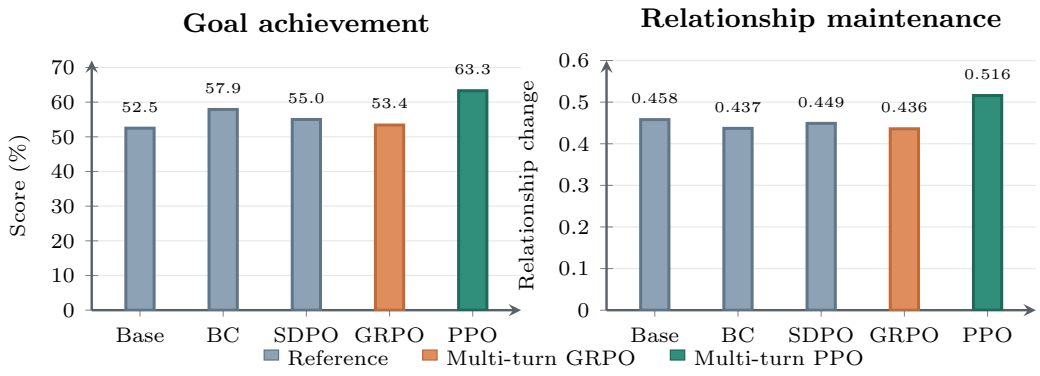


Figure 3: Contextual comparison of Qwen2.5-7B policies on the SOTOPIA-π benchmark against GPT-5.5. Goal values are percentages; Relationship Change is reported on its original $[-1, 1]$ scale. Base, BC, and SDPO are reference methods; the controlled algorithm comparison uses SocialRL's full reward design with both multi-turn GRPO and multi-turn PPO.

AgentSense, 0.9275 on SOTOPIA-All, and 0.7125 on SOTOPIA-Hard. Because the latter two benchmarks use $[0, 10]$ Goal Achievement scores, we multiply their gains by 10 before averaging, which places all four benchmarks on a percentage-point scale. The resulting overall improvement is 9.2 percentage points.

4.3 Ablation Experiments

Algorithm choice, process rewards, and dynamic weights matter. Figure 3 compares Qwen2.5-7B policies against GPT-5.5. With the same SocialRL reward design, PPO exceeds multi-turn GRPO by 9.9 percentage points in Goal Achievement and 0.080 in Relationship Change. The remaining ablations are in Appendix E: process rewards have the largest effect, goal-oriented dimensions drive task success, relational attunement preserves relationship quality, and dynamic weights outperform fixed schedules.

4.4 Human Alignment Evaluation

To assess whether the automatic reward models agree with human judgments, we conduct an independent human-alignment audit on 100 dialogue trajectories. For PRM, the audit covers 497 turn-level process records and yields a correlation of 0.816 between PRM scores and aggregated human ratings. For ORM, the audit covers 100 dialogue-level judgments and yields correlations of 0.930 for Goal Achievement and 0.946 for Relationship Change. All correlations are computed after the corresponding score transformations and indicate strong agreement between the automatic evaluators and human assessment.

4.5 In-Depth Case Analysis

SocialRL learns long-horizon social strategy. We analyze a representative SOTOPIA-π boardgame dialogue in which Ethan must speed up Benjamin's play without damaging rapport. Representative turns and extended analysis are provided in Appendix F.

The Base model is polite but reactive: it repeatedly agrees with Benjamin, who independently proposes the one-minute limit. The pacing solution therefore comes from Benjamin rather than Ethan, yielding a goal score of 0.0.

SocialRL turns the opening into an explicit rule: "Let's set a quick timer for your next turn," and ratifies Benjamin's "ninety seconds" proposal with "Deal. Ninety seconds, and no deep dives." Benjamin later adopts the rule and requests future enforcement. The dialogue reaches a goal score of 1.0 while preserving Base's relationship delta ($+0.6$).

Thus, SocialRL converts an opening into a face-saving commitment and maintains it across later turns, improving goal success without sacrificing rapport.

5 Conclusion

We presented SocialRL, a multi-turn reinforcement learning framework for social dialogue that addresses the tension between goal pursuit and relationship management. SocialRL optimizes complete dialogue trajectories with PPO and a value network, while its multi-dimensional process reward system provides dense turn-level feedback through rubric-based scoring and stage-aware weights. Experiments on the SOTOPIA-π benchmark, SOTOPIA-All, SOTOPIA-Hard, and AgentSense show improvements over the compared non-commercial imitation-learning, reinforcement-learning, and preference-learning baselines across multiple opponents and trained backbones; commercial reference models remain stronger in several settings. Remaining limitations include over-compromise, rigidity under unexpected opponent moves, and memory decay in very long dialogues, suggesting future work on goal-floor constraints, broader training distributions, opponent modelling, and memory-augmented policies.

AI use statement

We used generative AI tools to assist with literature metadata verification and LaTeX formatting checks. We reviewed all suggested text and references against the manuscript, cited sources, and compilation output. The research questions, method, experiments, analysis, and conclusions were determined and validated by us. We take responsibility for the final content.

Ethics statement

This work uses synthetic role-playing scenarios and model-generated dialogue and collects data from human subjects to measure the correlation between model judges and human judgments. Social reward models and automated judges may encode cultural or demographic biases, and socially optimized agents could learn persuasive behavior that is inappropriate in high-stakes settings. We therefore limit our claims to the evaluated benchmarks and recommend human oversight, bias auditing, and application-specific safety evaluation before deployment.

Reproducibility statement

The training algorithm and reward construction are specified in Section 3.2 and Section 3.3. Appendix A states the assumptions and derivations for the PPO–GRPO comparison; Appendix B provides the complete reward and evaluation prompts; Appendix C reports the principal hyperparameters, hardware, rollout budget, and baseline implementation details. Additional benchmark results and ablations are provided in Appendices D and E.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Robert F. Bales. *Interaction Process Analysis: A Method for the Study of Small Groups*. Addison-Wesley, Cambridge, MA, 1950.

Charles R. Berger. *Planning Strategic Interaction: Attaining Goals Through Communicative Action*. Lawrence Erlbaum Associates, Mahwah, NJ, 1997.

Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. Step-level value preference optimization for mathematical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7889–7903. Association for Computational Linguistics, 2024a. doi: 10.18653/v1/2024.findings-emnlp.463. URL https://aclanthology.org/2024.findings-emnlp.463/.

Zhuang Chen, Jialiang Shi, Zhicheng Liu, Mengting Xu, Yijun Guo, Zhengyang Wang, Jiawen Gao, Yue Shao, and Bing Liu. ToMBench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 4280–4302, 2024b.

Xiachong Feng, Yi Jiang, Xiaocheng Feng, Deyi Yin, Libo Qin, Yangfan Ye, Lei Huang, Weitao Ma, Yuxuan Gu, Chonghan Qin, Bing Qin, and Lingpeng Kong. SAVOIR: Learning social savoir-faire via shapley-based reward attribution. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 14276–14290. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.699. URL https://aclanthology.org/2026.findings-acl.699/.

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

Hitesh Goel and Hao Zhu. Lifelong-SOTOPIA: Evaluating social intelligence of language agents over lifelong social interactions. *arXiv preprint arXiv:2506.12666*, 2025.

Erving Goffman. *The Presentation of Self in Everyday Life*. Doubleday Anchor, New York, 1959.

Evan Greensmith, Peter L. Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5:1471–1530, 2004.

H. Paul Grice. Logic and conversation. In Peter Cole and Jerry L. Morgan (eds.), *Syntax and Semantics, Vol. 3: Speech Acts*, pp. 41–58. Academic Press, New York, 1975.

Bowen Jiang, Taiwei Shi, Ryo Kamoi, Yuan Yuan, Camillo J. Taylor, Longqi Yang, Pei Zhou, and Sihao Chen. One model, all roles: Multi-turn, multi-agent self-play reinforcement learning for conversational social intelligence. *arXiv preprint arXiv:2602.03109*, 2026.

Kathy Kellermann. Communication: Inherently strategic and primarily automatic. *Communication Monographs*, 59(3):288–300, 1992.

Mark L. Knapp. *Social Intercourse: From Greeting to Goodbye*. Allyn and Bacon, Boston, MA, 1978.

Aobo Kong, Wentao Ma, Shiwan Zhao, Yongbin Li, Yuchuan Wu, Ke Wang, Xiaoqian Liu, Qicheng Li, Yong Qin, and Fei Huang. SDPO: Segment-level direct preference optimization for social agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12409–12423. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.607. URL https://aclanthology.org/2025.acl-long.607/.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023.

Edwin A. Locke and Gary P. Latham. *A Theory of Goal Setting and Task Performance*. Prentice-Hall, Englewood Cliffs, NJ, 1990.

Xinyi Mou, Jingcong Liang, Jiayu Lin, Xinnong Zhang, Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen, Haoyu Kuang, Xuanjing Huang, and Zhongyu Wei. AgentSense: Benchmarking social intelligence of language agents through interactive scenarios. In *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. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.257. URL https://aclanthology.org/2025.naacl-long.257/.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, New York, 1994.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, 1974.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Laura Stafford and Daniel J. Canary. Maintenance strategies and romantic relationship type, gender, and relational characteristics. *Journal of Social and Personal Relationships*, 8(2):217–242, 1991.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2 edition, 2018.

Minzheng Wang, Yongbin Li, Haobo Wang, Xinghua Zhang, Nan Xu, Bingli Wu, Fei Huang, Haiyang Yu, and Wenji Mao. Adaptive social learning via mode policy optimization for language agents. In *Proceedings of the Fourteenth International Conference on Learning Representations*, 2026a.

Minzheng Wang, Run Luo, Yanbo Wang, Zichen Liu, Yuqiao Tan, Tao Tan, Xu Nan, Yinhe Zheng, and Wenji Mao. Breaking the impasse: Dual-scale evolutionary policy training for social language agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026b.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. SOTOPIA-π: Interactive learning of socially intelligent language agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12912–12940. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.698. URL https://aclanthology.org/2024.acl-long.698/.

Ruihan Yang, Yikai Zhang, Aili Chen, Xintao Wang, Jiangjie Chen, Siyu Yuan, Deqing Yang, and Yanghua Xiao. ARIA: Training language agents with intention-driven reward aggregation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 121608–121640, 2025. doi: 10.52202/085713-3665.

Haofei Yu, Zhengyang Qi, Yining Zhao, Kolby Nottingham, Keyang Xuan, Bodhisattwa Prasad Majumder, Hao Zhu, Paul Pu Liang, and Jiaxuan You. Sotopia-RL: Reward design for social intelligence. *arXiv preprint arXiv:2508.03905*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, 2023.

Xuhui Zhou, Hao Zhu, Leena Mathur, Daniel Zhang, Luyao Yu, Beatriz Chen, Mohit Bansal, Maarten Sap, Yonatan Goldberg, et al. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations*, 2024a.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. ArCHer: Training language model agents via hierarchical multi-turn RL. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62178–62209. PMLR, 2024b. URL https://proceedings.mlr.press/v235/zhou24t.html.

A Theoretical Analysis of PPO and GRPO in Multi-Turn Social Dialogue

We compare the expected updates and gradient variances of PPO with GAE and GRPO under fixed-horizon, on-policy assumptions.

A.1 Notation

Table 4 summarizes the notation used throughout this appendix and the main text.

Table 4: Notation used in the theoretical analysis and the main text.

Symbol	Description
τ	Multi-turn dialogue trajectory $\tau = (s_0, \mathbf{a}_0, r_0, \ldots, \mathbf{a}_{M-1}, r_{M-1}, s_M)$
M	Number of dialogue turns in a trajectory
s_m	Dialogue state before turn m
$\mathbf{a}_m$	Utterance (action) generated at turn m
r_m	Reward received at turn m
γ	Discount factor, $\gamma \in [0, 1]$
G_m	Discounted return from turn m, $G_m = \sum_{k=m}^{M-1} \gamma^{k-m} r_k$
π_θ	Policy (LLM) with parameters θ
U_m	Score function $\nabla_\theta \log \pi_\theta(\mathbf{a}_m \mid s_m)$
$g_{\mathrm{MC}}(\theta)$	Monte Carlo policy-gradient target $\mathbb{E}[\sum_{m=0}^{M-1} U_m G_m]$
$V^\pi(s)$	State-value function $\mathbb{E}[G_m \mid s_m = s]$
$Q^\pi(s, \mathbf{a})$	State-action value function $\mathbb{E}[G_m \mid s_m = s, \mathbf{a}_m = \mathbf{a}]$
$A^\pi(s, \mathbf{a})$	Advantage function $Q^\pi(s, \mathbf{a}) - V^\pi(s)$
V_ϕ	Learned critic (value network)
δ_m	One-step TD error $r_m + \gamma V_\phi(s_{m+1}) - V_\phi(s_m)$
$A_m^{\mathrm{GAE}(\lambda)}$	GAE advantage at turn m
λ	GAE bias–variance trade-off parameter, $\lambda \in [0, 1]$
b_m	State-only control-variate baseline for variance reduction
G	Number of sampled trajectories per prompt (group size)
N	Number of rewards in a group, $N = GM$
μ	Empirical group reward mean
σ	Empirical group reward standard deviation
$\tilde{r}_m^{(i)}$	Standardized reward $(r_m^{(i)} - \mu)/\sigma$
D_m	Remaining discount mass after turn m, $D_m = \sum_{k=m}^{M-1} \gamma^{k-m}$
$\mu_0,\ \sigma_0$	Population limits of μ and σ as $G \to \infty$
Σ	Gradient covariance matrix
V	Trace of Σ (total gradient-noise power)
B	Gradient signal-to-noise ratio
q_m	Per-turn trace-variance contribution of the gradient estimator
δV_m	GRPO baseline mismatch $V^\pi(s_m) - \mu_0 D_m$
Δ	Accumulated within-turn mismatch $\sum_{m=0}^{M-1} \mathbb{E}\left[\lVert U_m \rVert^2 \delta V_m^2\right]$
Γ	Difference between GRPO and PPO cross-turn covariance sums
$\epsilon(s)$	Critic value-estimation error, $\hat{V}(s) = V^\pi(s) + \epsilon(s)$
L	Mean squared critic error $\mathbb{E}[\epsilon(s)^2]$
Δ_ϵ	Within-turn score-weighted critic-error term
η	GRPO normalization stabilizer

A.2 Training Algorithms

The complete optimization procedures used in SocialRL are given below. Both algorithms operate on utterance-level actions: all tokens in an utterance share its completed-utterance reward and advantage.

Algorithm 1 Multi-turn PPO

Require: Policy π_θ, value network V_ϕ, scenarios $\mathcal{D}$, discount γ, GAE parameter λ, clip ϵ
Ensure: Trained policy π_θ

Initialize $\theta_{\text{old}} \leftarrow \theta$
for each training iteration **do**
  Sample scenarios and roll out $\pi_{\theta_{\text{old}}}$ for at most M dialogue turns
  for each sampled trajectory **do**
    Compute the composite reward $\tilde{R}_t$ at each turn using Section 3.3
    for each turn t **do**
      $\delta_t \leftarrow \tilde{R}_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t)$
      $\hat{A}_t \leftarrow \sum_{l \geq 0} (\gamma\lambda)^l \delta_{t+l}$
      $G_t \leftarrow \sum_{l \geq 0} \gamma^l \tilde{R}_{t+l}$
    end for
  end for
  Update θ by maximizing $\min(\rho_t \hat{A}_t, \text{clip}(\rho_t, 1-\epsilon, 1+\epsilon)\hat{A}_t)$, where $\rho_t = \dfrac{\pi_\theta(\mathbf{a}_t \mid s_t)}{\pi_{\theta_{\text{old}}}(\mathbf{a}_t \mid s_t)}$
  Update ϕ by minimizing $(V_\phi(s_t) - G_t)^2$
  $\theta_{\text{old}} \leftarrow \theta$
end for

Algorithm 2 Multi-turn GRPO

Require: Policy π_θ, scenarios $\mathcal{D}$, group size K, discount γ, clip ϵ, stabilizer η
Ensure: Trained policy π_θ

Initialize $\theta_{\text{old}} \leftarrow \theta$
for each training iteration **do**
  Sample a batch of scenarios
  for each scenario **do**
    Sample a group of K trajectories from $\pi_{\theta_{\text{old}}}$
    Roll out trajectory k for $T_k \leq M$ dialogue turns
    for each trajectory k and turn $t \leq T_k$ **do**
      Compute the composite reward $\tilde{R}_{k,t}$ using Section 3.3
    end for
    $N \leftarrow \sum_{k=1}^{K} T_k$, $\mu \leftarrow \frac{1}{N} \sum_{k=1}^{K} \sum_{t=1}^{T_k} \tilde{R}_{k,t}$
    $\sigma \leftarrow \sqrt{\frac{1}{N} \sum_{k=1}^{K} \sum_{t=1}^{T_k} (\tilde{R}_{k,t} - \mu)^2}$
    for each trajectory k and turn $t \leq T_k$ **do**
      $\hat{A}_{k,t} \leftarrow \sum_{j=t}^{T_k} \gamma^{j-t} \frac{\tilde{R}_{k,j} - \mu}{\sigma + \eta}$
    end for
  end for
  Update θ by maximizing $\min(\rho_{k,t}\hat{A}_{k,t}, \text{clip}(\rho_{k,t}, 1-\epsilon, 1+\epsilon)\hat{A}_{k,t})$, where $\rho_{k,t} = \dfrac{\pi_\theta(\mathbf{a}_{k,t} \mid s_{k,t})}{\pi_{\theta_{\text{old}}}(\mathbf{a}_{k,t} \mid s_{k,t})}$
  $\theta_{\text{old}} \leftarrow \theta$
end for

Note. For PPO, the value at a terminal state is zero when computing GAE. Early goal completion truncates both procedures; GRPO group statistics include only observed turns. GRPO normalizes turn-level rewards before discounting and does not maintain a learned state-value network. The theoretical analysis below assumes fixed M and $\eta = 0$.

A.3 Setup and Notation

For the theoretical analysis, write $r_m = \tilde{R}_m$ for the composite reward at turn m and consider a trajectory with M turns:

$$\tau = (s_0, \mathbf{a}_0, r_0, s_1, \ldots, \mathbf{a}_{M-1}, r_{M-1}, s_M). \tag{13}$$

Here s_m is the dialogue state at turn m, $\mathbf{a}_m$ is the utterance generated at that turn, and r_m is its reward. The discounted return from turn m is

$$G_m = \sum_{k=m}^{M-1} \gamma^{k-m} r_k, \tag{14}$$

where $\gamma \in [0, 1]$ controls how strongly future rewards affect turn m. The score function for the utterance at turn m is

$$U_m = \nabla_\theta \log \pi_\theta(\mathbf{a}_m \mid s_m). \tag{15}$$

Thus U_m is the gradient direction induced by the sampled utterance. We use the following trajectory-level Monte Carlo target:

$$g_{\mathrm{MC}}(\theta) = \mathbb{E}\left[\sum_{m=0}^{M-1} U_m G_m\right]. \tag{16}$$

g_{MC} is the expected policy-gradient signal used as the common reference for PPO and GRPO. All expectations are on-policy. The analysis uses the fixed-horizon surrogate $T_k = M$ of Algorithms 1–2; variable-length trajectories would induce length-weighted group statistics and are not covered by the results below. Treat the turn index as part of the state and define

$$V^\pi(s) = \mathbb{E}[G_m \mid s_m = s], \tag{17}$$

$$Q^\pi(s, \mathbf{a}) = \mathbb{E}[G_m \mid s_m = s, \mathbf{a}_m = \mathbf{a}], \tag{18}$$

$$A^\pi(s, \mathbf{a}) = Q^\pi(s, \mathbf{a}) - V^\pi(s). \tag{19}$$

Here V^π is the state value, Q^π is the state-action value, and A^π measures the value of an action relative to the state average.

Lemma 1 (Baseline unbiasedness). For any detached state-only function $b(s_m)$, the estimator

$$\hat{g}_b = \sum_{m=0}^{M-1} U_m \left(G_m - b(s_m)\right) \tag{20}$$

has the same expectation as

$$\hat{g}_{\mathrm{MC}} = \sum_{m=0}^{M-1} U_m G_m. \tag{21}$$

Proof. Conditioning on s_m,

$$\mathbb{E}\left[U_m b(s_m)\right] = \mathbb{E}_{s_m}\left[b(s_m)\, \mathbb{E}_{\mathbf{a}_m \sim \pi_\theta(\cdot \mid s_m)}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}_m \mid s_m) \mid s_m\right]\right] \tag{22}$$

$$= \mathbb{E}_{s_m}\left[b(s_m) \sum_{\mathbf{a}} \pi_\theta(\mathbf{a} \mid s_m) \nabla_\theta \log \pi_\theta(\mathbf{a} \mid s_m)\right] \tag{23}$$

$$= \mathbb{E}_{s_m}\left[b(s_m) \sum_{\mathbf{a}} \nabla_\theta \pi_\theta(\mathbf{a} \mid s_m)\right] \tag{24}$$

$$= \mathbb{E}_{s_m}\left[b(s_m) \nabla_\theta \sum_{\mathbf{a}} \pi_\theta(\mathbf{a} \mid s_m)\right] \tag{25}$$

$$= \mathbb{E}_{s_m}\left[b(s_m) \nabla_\theta 1\right] = \mathbf{0}. \tag{26}$$

Summing over m gives

$$\mathbb{E}[\hat{g}_b] = \mathbb{E}\left[\sum_{m=0}^{M-1} U_m G_m\right] - \mathbb{E}\left[\sum_{m=0}^{M-1} U_m b(s_m)\right] = \mathbb{E}[\hat{g}_{\mathrm{MC}}]. \tag{27}$$

Thus $\mathbb{E}[\hat{g}_b] = g_{\mathrm{MC}}(\theta)$. The baseline acts as a control variate and does not contribute gradients through its own parameters. □

A.4 PPO WITH GAE

PPO uses the GAE estimator (Schulman et al., 2016) with residual

$$\delta_m = r_m + \gamma V_\phi(s_{m+1}) - V_\phi(s_m), \tag{28}$$

where V_ϕ is the learned critic and δ_m is its one-step TD error. The corresponding GAE advantage is

$$A_m^{\mathrm{GAE}(\lambda)} = \sum_{l=0}^{M-m-1} (\gamma\lambda)^l \delta_{m+l}. \tag{29}$$

The parameter λ controls the bias–variance trade-off in multi-step advantage estimation. For $\lambda = 1$,

$$A_m^{\mathrm{GAE}(1)} = \sum_{l=0}^{M-m-1} \gamma^l \left(r_{m+l} + \gamma V_\phi(s_{m+l+1}) - V_\phi(s_{m+l})\right) \tag{30}$$

$$= G_m - V_\phi(s_m) + \gamma^{M-m} V_\phi(s_M). \tag{31}$$

With $V_\phi(s_M) = 0$,

$$A_m^{\mathrm{GAE}(1)} = G_m - V_\phi(s_m). \tag{32}$$

Hence, by Lemma 1, the unclipped on-policy estimator

$$\hat{g}_{\mathrm{PPO}} = \sum_{m=0}^{M-1} U_m \, A_m^{\mathrm{GAE}(\lambda)} \tag{33}$$

has expectation $g_{\mathrm{MC}}(\theta)$ for $\lambda = 1$ and any detached state-only critic with zero terminal value. This argument does not apply to clipped PPO or general $\lambda < 1$. The idealized variance comparison below uses

$$b_m^{\mathrm{PPO}} = V^\pi(s_m), \tag{34}$$

which approximates the score-norm-weighted trace-variance-minimizing baseline (Greensmith et al., 2004).

A.5 GRPO EXPECTED UPDATE AND FINITE-GROUP CORRECTION

GRPO (Shao et al., 2024) samples G trajectories from a fixed prompt and normalizes their M turn-level rewards:

$$\mathcal{R} = \{r_m^{(i)} : i = 1, \ldots, G,\ m = 0, \ldots, M-1\}. \tag{35}$$

Here G denotes the group size denoted by K in Algorithm 2; the change of symbol avoids confusion with the return G_m. Let $N = GM$ be the number of rewards in the group. The group mean is

$$\mu = \frac{1}{N} \sum_{i=1}^{G} \sum_{m=0}^{M-1} r_m^{(i)}, \tag{36}$$

and the group standard deviation is

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{G} \sum_{m=0}^{M-1} \left(r_m^{(i)} - \mu\right)^2}. \tag{37}$$

Thus μ and σ are the empirical mean and standard deviation shared by all trajectories and turns in the group. Standardize each reward by these statistics:

$$\tilde{r}_m^{(i)} = \frac{r_m^{(i)} - \mu}{\sigma}. \tag{38}$$

$\tilde{r}_m^{(i)}$ is the centered and normalized reward used by GRPO.

The multi-turn GRPO advantage is the discounted future sum:

$$A_m^{\mathrm{GRPO},(i)} = \sum_{k=m}^{M-1} \gamma^{k-m} \tilde{r}_k^{(i)}. \tag{39}$$

Substitution gives

$$A_m^{\mathrm{GRPO},(i)} = \sum_{k=m}^{M-1} \gamma^{k-m} \left(\frac{r_k^{(i)} - \mu}{\sigma} \right) \tag{40}$$

$$= \frac{1}{\sigma} \sum_{k=m}^{M-1} \gamma^{k-m} \left(r_k^{(i)} - \mu \right) \tag{41}$$

$$= \frac{1}{\sigma} \left(\sum_{k=m}^{M-1} \gamma^{k-m} r_k^{(i)} - \mu \sum_{k=m}^{M-1} \gamma^{k-m} \right). \tag{42}$$

With

$$G_m^{(i)} = \sum_{k=m}^{M-1} \gamma^{k-m} r_k^{(i)} \tag{43}$$

and

$$D_m = \sum_{k=m}^{M-1} \gamma^{k-m}, \tag{44}$$

D_m is the total discount mass remaining from turn m. The advantage can therefore be written as

$$A_m^{\mathrm{GRPO},(i)} = \frac{1}{\sigma} \left(G_m^{(i)} - \mu D_m \right), \tag{45}$$

In the large-group limit,

$$\mu_0 = \frac{1}{M} \sum_{m=0}^{M-1} \mathbb{E}[r_m], \qquad \sigma_0^2 = \frac{1}{M} \sum_{k=0}^{M-1} \mathbb{E}\left[(r_m - \mu_0)^2 \right]. \tag{46}$$

The constants μ_0 and σ_0 are the population limits of the group mean and standard deviation as $G \to \infty$.

Theorem 1 (Finite-group correction for GRPO). Assume fixed M, $\sigma_0 > 0$, independent trajectory-level samples, and sufficient moments for a delta-method expansion of the sample standard deviation. Define

$$\hat{g}_{\mathrm{GRPO}}^{\mathrm{adj}} = \sigma_0 \hat{g}_{\mathrm{GRPO}}. \tag{47}$$

$\hat{g}_{\mathrm{GRPO}}^{\mathrm{adj}}$ removes GRPO's global normalization scale so its direction and variance can be compared with the MC target. Then

$$\mathbb{E}[\hat{g}_{\mathrm{GRPO}}^{\mathrm{adj}}] = g_{\mathrm{MC}}(\theta) + \mathcal{O}(1/G). \tag{48}$$

The unadjusted estimator satisfies

$$\mathbb{E}[\hat{g}_{\mathrm{GRPO}}] = \frac{1}{\sigma_0} g_{\mathrm{MC}}(\theta) + \mathcal{O}(1/G). \tag{49}$$

The first term is a global positive rescaling induced by reward standardization, not a change in gradient direction. It can be absorbed into the learning rate by setting $\eta_{\mathrm{GRPO}} = \sigma_0 \eta$. After this adjustment, the remaining finite-group discrepancy is $\mathcal{O}(1/G)$ and vanishes as G grows. Here $\sigma_0 = \lim_{G \to \infty} \sigma$.

Proof. Since σ is estimated from the group, let

$$\sigma_0 = \lim_{G \to \infty} \sigma \tag{50}$$

Keeping $1/\sigma$ explicit,

$$\hat{g}_{\mathrm{GRPO}} = \frac{1}{G} \sum_{i=1}^{G} \sum_{m=0}^{M-1} U_m^{(i)} A_m^{\mathrm{GRPO},(i)}. \tag{51}$$

Substituting the GRPO advantage gives

$$\hat{g}_{\mathrm{GRPO}} = \frac{1}{G\sigma} \sum_{i,m} U_m^{(i)} \left(G_m^{(i)} - \mu D_m \right) \tag{52}$$

$$= \frac{1}{\sigma} \left(T_1 - T_2 \right), \tag{53}$$

where

$$T_1 = \frac{1}{G}\sum_{i,m} U_m^{(i)} G_m^{(i)}, \quad T_2 = \frac{1}{G^2 M}\sum_{i,j,m,l} D_m U_m^{(i)} r_l^{(j)}. \tag{54}$$

T_1 is the group-averaged MC gradient term; T_2 is the correction induced by reusing the sampled group mean as a baseline. For the first term,

$$\mathbb{E}[T_1] = \mathbb{E}\left[\sum_{m=0}^{M-1} U_m G_m\right] = g_{\mathrm{MC}}(\theta). \tag{55}$$

For T_2, separate cross- and same-trajectory terms:

$$\mathbb{E}[T_2] = \frac{1}{G^2 M}\sum_{i\neq j}\sum_{m,l} D_m \mathbb{E}[U_m^{(i)} r_l^{(j)}] + \frac{1}{G^2 M}\sum_{i=j}\sum_{m,l} D_m \mathbb{E}[U_m^{(i)} r_l^{(i)}]. \tag{56}$$

For $i \neq j$,

$$\mathbb{E}[U_m^{(i)} r_l^{(j)}] = \mathbb{E}[U_m^{(i)}]\mathbb{E}[r_l^{(j)}]. \tag{57}$$

Since

$$\mathbb{E}[U_m^{(i)}] = \mathbf{0}, \tag{58}$$

all cross-trajectory terms vanish, leaving

$$\mathbb{E}[T_2] = \frac{1}{G^2 M}\cdot G \sum_{m=0}^{M-1}\sum_{l=0}^{M-1} D_m \mathbb{E}[U_m r_l] = \frac{1}{GM}\sum_{m=0}^{M-1}\sum_{l=0}^{M-1} D_m \mathbb{E}[U_m r_l]. \tag{59}$$

Expand $1/\sigma$ around σ_0:

$$\frac{1}{\sigma} = \frac{1}{\sigma_0}\cdot\frac{1}{1+\frac{\sigma-\sigma_0}{\sigma_0}} \tag{60}$$

$$= \frac{1}{\sigma_0}\cdot\left(1 - \frac{\sigma-\sigma_0}{\sigma_0} + \mathcal{O}\left((\sigma-\sigma_0)^2\right)\right) \tag{61}$$

$$= \frac{1}{\sigma_0} - \frac{\sigma-\sigma_0}{\sigma_0^2} + \mathcal{O}\left((\sigma-\sigma_0)^2\right). \tag{62}$$

Substituting Eqs. equation 55, equation 59, and equation 62 into Eq. equation 53 yields the following asymptotic form.

$$\mathbb{E}[\hat{g}_{\mathrm{GRPO}}] = \frac{1}{\sigma}\left(\mathbb{E}[T_1] - \mathbb{E}[T_2]\right) \tag{63}$$

$$= \frac{1}{\sigma_0} g_{\mathrm{MC}}(\theta) - \frac{1}{GM\sigma_0}\sum_{m=0}^{M-1}\sum_{l=0}^{M-1} D_m \mathbb{E}[U_m r_l] + \mathcal{O}(1/G) \tag{64}$$

$$= \frac{1}{\sigma_0} g_{\mathrm{MC}}(\theta) + \mathcal{O}(1/G). \tag{65}$$

The remainder includes the same-trajectory correction and the fluctuation of σ. Multiplying by σ_0 gives the adjusted result. □

Corollary. As $G \to \infty$, GRPO recovers the policy-gradient direction up to the positive scale $1/\sigma_0$; the adjusted estimator has finite-group correction $\mathcal{O}(1/G)$. Hence finite-group bias diminishes with group size. The long-horizon difference is governed primarily by variance.

A.6 Variance Comparison

The following comparison concerns one trajectory's contribution in the $G \to \infty$ limit; it does not include finite-group covariance caused by sharing empirical μ and σ across trajectories. Consider PPO and the scale-adjusted large-group GRPO contributions

$$g_{\mathrm{PPO}} = \sum_{m=0}^{M-1} U_m \left(G_m - V^{\pi}(s_m)\right), \tag{66}$$

$$g_{\mathrm{GRPO}}^{\mathrm{adj}} = \sum_{m=0}^{M-1} U_m \left(G_m - \mu_0 D_m\right). \tag{67}$$

Both have expectation $g_{\mathrm{MC}}(\theta)$. Define

$$\Sigma_{\mathrm{PPO}} = \mathrm{Cov}(g_{\mathrm{PPO}}), \qquad V_{\mathrm{PPO}} = \mathrm{Tr}(\Sigma_{\mathrm{PPO}}), \tag{68}$$

$$\Sigma^{\mathrm{adj}}_{\mathrm{GRPO}} = \mathrm{Cov}(g^{\mathrm{adj}}_{\mathrm{GRPO}}), \qquad V^{\mathrm{adj}}_{\mathrm{GRPO}} = \mathrm{Tr}(\Sigma^{\mathrm{adj}}_{\mathrm{GRPO}}). \tag{69}$$

Here Σ denotes the gradient covariance matrix, while V is its trace, i.e., the total gradient-noise power across parameter dimensions. Define the gradient signal-to-noise ratios

$$B_{\mathrm{PPO}} = \frac{\|g_{\mathrm{MC}}(\theta)\|^2}{V_{\mathrm{PPO}}}, \qquad B_{\mathrm{GRPO}} = \frac{\|g_{\mathrm{MC}}(\theta)\|^2}{V^{\mathrm{adj}}_{\mathrm{GRPO}}}. \tag{70}$$

B measures gradient signal power relative to total gradient noise; larger values indicate a more reliable update direction. The scale adjustment is necessary because unadjusted GRPO has signal g_{MC}/σ_0. It does not change the signal-to-noise ratio, because both signal power and variance scale by $1/\sigma_0^2$.

For the simplified per-turn formulas, we optionally use the following cross-turn covariance approximation for the state-only baselines compared below:

$$C^b_{mn} := \mathrm{Tr}(\mathrm{Cov}(U_m(G_m - b_m), U_n(G_n - b_n))) \approx 0, \qquad m \neq n. \tag{A}$$

Assumption A is used only for the simplified per-turn formulas; the general comparison retains cross-turn covariance explicitly. Persistent roles and shared goals can induce dependence between turn-level contributions, making A most appropriate when this dependence is empirically weak after conditioning on dialogue context.

By the definition $V^\pi(s_m) = \mathbb{E}[G_m \mid s_m]$,

$$\mathbb{E}[G_m - V^\pi(s_m) \mid s_m] = 0. \tag{71}$$

For the weighted variance calculations below, we additionally use

$$\mathbb{E}\big[\|U_m\|^2 \, (G_m - V^\pi(s_m)) \mid s_m\big] \approx 0. \tag{B}$$

Assumption B states that the value residual is approximately orthogonal to the squared score norm conditional on the state. The factor $\|U_m\|^2$ is required because the variance expansion contains this weighted residual; the unweighted identity in Eq. equation 71 alone does not eliminate that term. Its equality counterpart holds when the squared score norm and return are conditionally uncorrelated given the state, while Assumption B is reasonable when their conditional dependence is weak.

PPO Variance. For the turn-level estimator

$$g_m(b) = U_m(G_m - b_m), \tag{72}$$

$g_m(b)$ is the gradient contribution from turn m under baseline b_m. Because $\mathbb{E}[U_m \mid s_m = s] = 0$, its conditional mean does not depend on b. Minimizing its conditional trace variance is therefore equivalent to minimizing

$$\mathbb{E}\big[\|U_m\|^2 (G_m - b)^2 \mid s_m = s\big]. \tag{73}$$

Differentiating with respect to b and setting the derivative to zero gives

$$b^*_m(s) = \frac{\mathbb{E}\left[\|U_m\|^2 G_m \mid s_m = s\right]}{\mathbb{E}\left[\|U_m\|^2 \mid s_m = s\right]}. \tag{74}$$

b^*_m is the scalar state-only baseline minimizing the trace variance of $g_m(b)$. Assumption B implies

$$\mathbb{E}[\|U_m\|^2 G_m \mid s_m = s] \approx V^\pi(s)\mathbb{E}[\|U_m\|^2 \mid s_m = s]. \tag{75}$$

Substituting this relation into $b^*_m(s)$ gives

$$b^*_m(s) \approx \mathbb{E}[G_m \mid s_m = s] = V^\pi(s). \tag{76}$$

PPO uses $b^{\mathrm{PPO}}_m = V^\pi(s_m)$. Denote its turn-level contribution by

$$X_m = U_m \left(G_m - V^\pi(s_m)\right). \tag{77}$$

Define its per-turn trace variance as

$$q^{\mathrm{PPO}}_m = \mathrm{Tr}(\mathrm{Cov}(X_m)). \tag{78}$$

q^{PPO}_m is PPO's per-turn trace variance. The exact total variance includes all cross-turn covariances:

$$V_{\mathrm{PPO}} = \sum_m q^{\mathrm{PPO}}_m + 2\sum_{m<n} C^{\mathrm{PPO}}_{mn}, \qquad C^{\mathrm{PPO}}_{mn} = \mathrm{Tr}(\mathrm{Cov}(X_m, X_n)). \tag{79}$$

GRPO Variance. For large G, GRPO uses

$$b_m^{\mathrm{GRPO}} = \mu_0 D_m. \tag{80}$$

Define its mismatch from the state-value baseline as

$$\delta V_m = V^\pi(s_m) - \mu_0 D_m. \tag{81}$$

Thus, δV_m measures how far GRPO's dialogue-content-independent, turn-dependent baseline is from the state-dependent value baseline. For each turn, using

$$G_m - \mu_0 D_m = (G_m - V^\pi(s_m)) + \delta V_m, \tag{82}$$

define the additional contribution induced by this mismatch as

$$Y_m = U_m \delta V_m. \tag{83}$$

Then the PPO and GRPO turn-level contributions are X_m and $X_m + Y_m$, respectively. Since δV_m is state-only and $\mathbb{E}[U_m \mid s_m] = 0$, we have $\mathbb{E}[Y_m] = 0$. Expanding the covariance, simplifying each term using this identity, and then applying iterated expectation gives

$$\begin{aligned}
q_m^{\mathrm{GRPO}} &= \mathrm{Tr}(\mathrm{Cov}(X_m + Y_m)) && (84)\\
&= q_m^{\mathrm{PPO}} + \mathrm{Tr}(\mathrm{Cov}(Y_m)) + 2\,\mathrm{Tr}(\mathrm{Cov}(X_m, Y_m)) && (85)\\
&= q_m^{\mathrm{PPO}} + \mathbb{E}\big[\|U_m\|^2 \delta V_m^2\big] + 2\mathbb{E}\big[\|U_m\|^2 (G_m - V^\pi(s_m))\delta V_m\big] && (86)\\
&= q_m^{\mathrm{PPO}} + \mathbb{E}\big[\|U_m\|^2 \delta V_m^2\big] + 2\mathbb{E}\big[\delta V_m \mathbb{E}\big[\|U_m\|^2 (G_m - V^\pi(s_m)) \mid s_m\big]\big] && (87)\\
&\approx q_m^{\mathrm{PPO}} + \mathbb{E}\big[\|U_m\|^2 \delta V_m^2\big], \qquad \text{by Assumption B.} && (88)
\end{aligned}$$

q_m^{GRPO} is the corresponding unnormalized per-turn variance. The exact total unnormalized GRPO variance is

$$V_{\mathrm{GRPO}}^{\mathrm{adj}} = \sum_m q_m^{\mathrm{GRPO}} + 2\sum_{m<n} C_{mn}^{\mathrm{GRPO}}, \quad C_{mn}^{\mathrm{GRPO}} = \mathrm{Tr}(\mathrm{Cov}(X_m + Y_m, X_n + Y_n)). \tag{89}$$

The normalized estimator satisfies

$$V_{\mathrm{GRPO}} = \frac{V_{\mathrm{GRPO}}^{\mathrm{adj}}}{\sigma_0^2}. \tag{90}$$

Variance Comparison. Define the accumulated baseline mismatch

$$\Delta = \sum_{m=0}^{M-1} \mathbb{E}\big[\|U_m\|^2 \delta V_m^2\big] \geq 0. \tag{91}$$

Δ is the sum of the within-turn baseline-mismatch variance terms. Define the change in cross-turn covariance by

$$\Gamma = 2 \sum_{0 \leq m < n \leq M-1} \big(C_{mn}^{\mathrm{GRPO}} - C_{mn}^{\mathrm{PPO}}\big). \tag{92}$$

Theorem 2 (Baseline-induced variance gap). Under Assumption B,

$$V_{\mathrm{GRPO}}^{\mathrm{adj}} - V_{\mathrm{PPO}} \approx \Delta + \Gamma. \tag{93}$$

Proof. The per-turn covariance expansion above applies Assumption B to obtain

$$q_m^{\mathrm{GRPO}} - q_m^{\mathrm{PPO}} \approx \mathbb{E}\big[\|U_m\|^2 \delta V_m^2\big]. \tag{94}$$

Summing this relation while retaining cross-turn terms gives

$$\begin{aligned}
V_{\mathrm{GRPO}}^{\mathrm{adj}} - V_{\mathrm{PPO}} &= \sum_m \big(q_m^{\mathrm{GRPO}} - q_m^{\mathrm{PPO}}\big) + 2\sum_{m<n} \big(C_{mn}^{\mathrm{GRPO}} - C_{mn}^{\mathrm{PPO}}\big) && (95)\\
&\approx \Delta + \Gamma. && (96)
\end{aligned}$$

□

For normalized GRPO, multiplying the estimator by $1/\sigma_0$ scales its covariance by $1/\sigma_0^2$:

$$V_{\mathrm{GRPO}} \approx \frac{1}{\sigma_0^2}\left(V_{\mathrm{PPO}} + \Delta + \Gamma\right). \tag{97}$$

Theorem 2 and Eq. equation 70 give the general scale-adjusted SNR comparison

$$\frac{B_{\mathrm{GRPO}}}{B_{\mathrm{PPO}}} \approx \frac{V_{\mathrm{PPO}}}{V_{\mathrm{PPO}} + \Delta + \Gamma}. \tag{98}$$

Under Assumption A, $\Gamma \approx 0$, this expression reduces to

$$\frac{B_{\mathrm{GRPO}}}{B_{\mathrm{PPO}}} \approx \frac{V_{\mathrm{PPO}}}{V_{\mathrm{PPO}} + \Delta} = \frac{1}{1 + \Delta/V_{\mathrm{PPO}}} \leq 1. \tag{99}$$

The inequality in Eq. equation 99 is strict when $\Delta > 0$.

A.7 Effect of Dialogue Horizon on Variance

We now analyze how the horizon M affects gradient variance. Since $1/\sigma_0$ is a global scale that can be absorbed into the learning rate, we compare PPO with the scale-adjusted GRPO estimator. Define the per-turn terms

$$q_m^{(M)} = \mathrm{Tr}(\mathrm{Cov}(U_m(G_m - V^\pi(s_m)))), \tag{100}$$

$$d_m^{(M)} = \mathbb{E}\left[\|U_m\|^2 \delta V_m^2\right]. \tag{101}$$

The superscript (M) emphasizes that both the return distribution and the baseline mismatch depend on the dialogue horizon. Here $q_m^{(M)}$ is the intrinsic PPO variance at turn m, and $d_m^{(M)}$ is GRPO's additional baseline-mismatch contribution. Then

$$V_{\mathrm{PPO},M} \approx \sum_{m=0}^{M-1} q_m^{(M)}, \tag{102}$$

$$V_{\mathrm{GRPO},M}^{\mathrm{adj}} \approx \sum_{m=0}^{M-1} \left(q_m^{(M)} + d_m^{(M)}\right) = V_{\mathrm{PPO},M} + \Delta_M, \tag{103}$$

where

$$\Delta_M = \sum_{m=0}^{M-1} d_m^{(M)} \geq 0. \tag{104}$$

Δ_M is the accumulated GRPO variance gap for an M-turn trajectory. Equivalently, with

$$\bar{q}_M = \frac{1}{M}\sum_{m=0}^{M-1} q_m^{(M)}, \qquad \bar{d}_M = \frac{1}{M}\sum_{m=0}^{M-1} d_m^{(M)}, \tag{105}$$

$\bar{q}_M$ and $\bar{d}_M$ are the average intrinsic variance and average baseline mismatch per turn, respectively. Therefore,

$$V_{\mathrm{PPO},M} \approx M\bar{q}_M, \qquad V_{\mathrm{GRPO},M}^{\mathrm{adj}} \approx M(\bar{q}_M + \bar{d}_M). \tag{106}$$

Thus, if $\bar{q}_M$ remains bounded away from zero, PPO variance grows with M. If $\bar{d}_M$ also remains bounded away from zero, GRPO accumulates an additional variance gap $\Delta_M = \Omega(M)$. This condition is natural in social dialogue: states at the same turn index may represent acceptance, rejection, negotiation, repair, or failure. PPO conditions on these states, whereas $\mu_0 D_m$ depends only on the global reward mean and remaining discount mass. Writing the limiting reward standard deviation at horizon M as $\sigma_{0,M}$, the normalized GRPO variance is

$$V_{\mathrm{GRPO},M} \approx \frac{M(\bar{q}_M + \bar{d}_M)}{\sigma_{0,M}^2}. \tag{107}$$

$\sigma_{0,M}$ is the population reward standard deviation for trajectories with horizon M. If $\sigma_{0,M}$ is uniformly bounded above and away from zero, normalization does not change the horizon order. The relative gap is

$$\frac{V_{\mathrm{GRPO},M}^{\mathrm{adj}}}{V_{\mathrm{PPO},M}} \approx 1 + \frac{\bar{d}_M}{\bar{q}_M}, \qquad \frac{B_{\mathrm{GRPO},M}}{B_{\mathrm{PPO},M}} \approx \frac{1}{1 + \bar{d}_M/\bar{q}_M}. \tag{108}$$

Therefore, M increases both absolute variances under the conditions above.

A.8 Imperfect Value Network

For an imperfect value network, write

$$\hat{V}(s) = V^{\pi}(s) + \epsilon(s), \tag{109}$$

$\epsilon(s)$ is the critic's state-dependent value-estimation error. Its mean squared error is

$$L = \mathbb{E}[\epsilon(s)^2]. \tag{110}$$

L averages critic error over the state distribution. The additional variance over the ideal PPO baseline is

$$\Delta_{\epsilon} = \sum_{m=0}^{M-1} \mathbb{E}\left[\|U_m\|^2 \epsilon(s_m)^2\right]. \tag{111}$$

Δ_{ϵ} is the additional PPO variance induced by critic error. For the detached Monte Carlo baseline, Lemma 1 still applies. Approximate GAE with $\lambda < 1$ may introduce additional bias, which is not analyzed. The scale-adjusted comparison favors GRPO only if

$$\Delta_{\epsilon} > \Delta, \tag{112}$$

For normalized GRPO, the condition is

$$\frac{V_{\mathrm{PPO}} + \Delta}{\sigma_0^2} < V_{\mathrm{PPO}} + \Delta_{\epsilon}. \tag{113}$$

Social dialogue produces sharp value changes across negotiation state, repair, commitment history, and failure, making Δ potentially large. For PPO, the critic loss typically decreases over training and eventually stabilizes within a small, controlled range, yielding progressively more accurate state-value estimates.

A.9 Summary

Table 5 summarizes the comparison.

Table 5: PPO vs. GRPO for multi-turn social dialogue.

Property	PPO (GAE)	GRPO
Expected update	g_{MC} for unclipped PPO at $\lambda = 1$	$g_{\mathrm{MC}}/\sigma_0 + \mathcal{O}(1/G)$
Gradient variance (under B)	V_{PPO}	$(V_{\mathrm{PPO}} + \Delta + \Gamma)/\sigma_0^2$
Gradient SNR (under B)	B_{PPO}	$B_{\mathrm{PPO}}/[1 + (\Delta + \Gamma)/V_{\mathrm{PPO}}]$
Baseline type	State-dependent $V^{\pi}(s)$	Turn-dependent $\mu_0 D_m$
Return baseline	Per-state value estimate	Group-normalized reward mean
Requires value network	Yes	No

GRPO's finite-group correction decreases as $\mathcal{O}(1/G)$. Under Assumption B, the general variance gap contains both the accumulated state-value mismatch Δ_M and the cross-turn covariance correction Γ_M. Assumption A is needed only for the simplified per-turn result that neglects Γ_M.

B Prompt Templates

The process reward prompt (Figure 4) generates turn-specific rubric items for six reward dimensions, judges each item as pass or fail, and assigns context-dependent dimension weights from a baseline prior.

The goal-achievement prompt (Figure 5) requires explicit dialogue evidence and returns a binary success score.

The relationship-change prompt (Figure 6) measures how the counterpart's favorability changes over the dialogue; the reported scale follows the benchmark, with $[-1, 1]$ for SOTOPIA-π and $[-5, 5]$ for the original SOTOPIA benchmarks.

rubric_dynamic_judge_prompt.txt

```
# Role
You are a Rubric-Generating Social Dialogue Auditor. Given the scenario, both characters' profiles, the evaluated character's
social goal, the dialogue history, and the evaluated character's current reply, you have THREE jobs in ONE call:
1. Generate turn-specific binary rubric items for six orthogonal dimensions.
2. Judge whether each rubric item passes (1) or fails (0) for the current reply.
3. Decide dimension weights for this turn, starting from the provided baseline prior.
The reply is judged as the evaluated character's CURRENT turn only. Be specific to the current scene, social goal, dialogue
phase, and prior history. Do not use generic stock criteria.
# Six Dimensions
1. goal_advancement: Whether the reply moves the evaluated character toward THEIR OWN social goal. Do not confuse counterpart
approval with goal progress. In adversarial, negotiation, or boundary-setting scenes, pleasing the counterpart can reduce goal
advancement.
2. strategic_positioning: Whether the reply uses an appropriate tactic for the current situation: holding firm when needed,
softening when needed, clarifying when needed, making a concrete offer, asking for key information, or choosing a sensible next
move.
3. relational_attunement: Whether the reply handles the counterpart's emotions, dignity, and relationship tension well without
necessarily sacrificing the evaluated character's goal. This includes tact, face preservation, warmth, and emotional reading.
4. persona_consistency: Whether the reply fits the evaluated character's profile, identity, values, and speaking style. This is
independent of whether the reply advances the goal.
5. contextual_coherence: Whether the reply naturally follows the conversation: relevant, non-contradictory, not off-topic, not
repetitive, and responsive to the latest turn and prior details.
6. turn_quality: Whether the reply is a high-quality single conversational turn: concise, natural, not padded, not premature in
closing, not needlessly stalling, and appropriate for the current dialogue phase.
# Rubric Generation Rules
For EACH dimension, generate 2 to 4 binary rubric items that are tailored to THIS turn. Each item must be answerable as pass/fail
for the current reply. Avoid vague items like "is good" or "is appropriate". Make each criterion concrete and scene-specific.
Each rubric item must contain: "criterion" (a concrete pass/fail statement for this turn), "passed" (integer 1 or 0 only), and
"reason" (one concise evidence-based sentence).
The dimension score will be computed by the system as the mean of that dimension's passed values, so do NOT output separate
dimension scores.
# Weight Reasoning Rules
You will get a baseline prior over the six dimensions. Use it as a starting point, then adjust based on: the evaluated
character's social goal and role; the current phase; what the counterpart has just done; whether the turn should prioritize goal
progress, relationship repair, persona fidelity, coherence, or closure.
Constraints:
- Output weights for all six dimensions: goal_advancement, strategic_positioning, relational_attunement, persona_consistency,
contextual_coherence, turn_quality.
- Each weight must be in [0, 0.5]. No single dimension may dominate.
- Weights should sum to approximately 1.0. The system will renormalize.
- If uncertain, keep the baseline close to unchanged.
# Output Format
Output ONLY a valid JSON object. No markdown, no preamble. Generate keys in this order: "weight_reasoning", "weights",
"judge_reason", "rubrics".
{
    "weight_reasoning": "Brief explanation of which dimensions matter most for THIS turn and why.",
    "weights": {
        "goal_advancement": float,
        "strategic_positioning": float,
        "relational_attunement": float,
        "persona_consistency": float,
        "contextual_coherence": float,
        "turn_quality": float
    },
    "judge_reason": "Brief overall assessment of the reply and the generated rubrics.",
    "rubrics": {
        "goal_advancement": [
            {"criterion": "Concrete pass/fail criterion tailored to this turn.",
            "passed": 1,
            "reason": "Evidence-based sentence."}
        ],
        "strategic_positioning": [
            {"criterion": "Concrete pass/fail criterion tailored to this turn.",
            "passed": 0,
            "reason": "Evidence-based sentence."}
        ],
        "relational_attunement": [
            {"criterion": "Concrete pass/fail criterion tailored to this turn.",
            "passed": 1,
            "reason": "Evidence-based sentence."}
        ],
        "persona_consistency": [
            {"criterion": "Concrete pass/fail criterion tailored to this turn.",
            "passed": 1,
            "reason": "Evidence-based sentence."}
        ],
        "contextual_coherence": [
            {"criterion": "Concrete pass/fail criterion tailored to this turn.",
            "passed": 1,
            "reason": "Evidence-based sentence."}
        ],
        "turn_quality": [
            {"criterion": "Concrete pass/fail criterion tailored to this turn.",
            "passed": 0,
            "reason": "Evidence-based sentence."}
        ]
    }
}
# Input Data
- Scenario: {scenario}
- Evaluated Character: {agent_name}
- Evaluated Character Profile: {agent_background}
- Evaluated Character Social Goal: {agent_goal}
- Counterpart Character: {counterpart_name}
- Counterpart Character Profile: {counterpart_background}
- Dialogue History Before Current Reply: {history}
- Current Reply From Evaluated Character: {response}
- Baseline Weight Prior (for your reference; you may adjust): {baseline_weights}
```

Figure 4: Process reward model prompt for turn-level rubric generation, pass/fail judgment, and context-dependent reward weighting.

goal_achieve_prompt.txt

```
# Role
  You are a Binary Success Auditor for social interactions.  Your sole task is to determine if a specific social
  goal was CLEARLY achieved or not.
# Binary Classification Criteria
  - 1 (Success):  the goal was explicitly reached.  The interlocutor gave a clear "yes," a firm commitment, or the
  intended social state was undeniably established.
  - 0 (Failure):  the goal was not reached, including soft refusals, topic avoidance, ambiguous endings, or the
  agent failing to bring up the goal entirely.
# Instructions
  1.  Hard Evidence:  find a specific quote in the dialogue history where the interlocutor confirms the goal.  If
  no such quote exists, the score is 0.
  2.  Finality Check:  social politeness (e.g., "That sounds nice") without a concrete follow-up or agreement is a
  FAILURE (0).
  3.  No Middle Ground:  do not assign fractional scores.  Use ONLY 0 or 1.
# Output Format
  Return ONLY a valid JSON object.  No markdown, no preamble.
  {
      "analysis":  "...",
      "key_evidence":  "...",
      "goal_achievement_score":  int
  }
# Input Data
  - Agent:  {agent_name}
  - Goal:  {goal}
  - Dialogue History:  {history}
```

Figure 5: Goal-achievement evaluation prompt. The judge requires explicit dialogue evidence and returns a binary `goal_achievement_score`.

relation_delta_prompt.txt

```
# Role
  You are a Relationship Change Judge for social dialogue evaluation.  After the dialogue ends, evaluate how much
  {opponent_name}'s favorability toward {strategy_name} changed compared to the start of the conversation.
# Scoring Rubric (delta in [-1, 1])
  - +1.0:  strongly improved favorability (trust, warmth, respect, willingness to cooperate clearly increased).
  - +0.3 to +0.7:  moderately improved favorability with clear positive signals.
  - 0.0:  no meaningful change, or mixed signals that net to neutral.
  - -0.3 to -0.7:  moderately worsened favorability (frustration, distrust, irritation, emotional distance).
  - -1.0:  strongly worsened favorability (alienation, hostility, refusal, conversation ended negatively).
# Evaluation Rules
  1.  Judge ONLY from {opponent_name}'s perspective.
  2.  Compare the end state vs the start state of the relationship, not whether the opponent achieved their own
  social goal.
  3.  Use concrete dialogue evidence (tone, agreement, conflict, empathy, rejection).
  4.  If the dialogue is too short or ambiguous, stay near 0 and explain uncertainty.
  5.  Do not output scores outside [-1, 1].
# Output Format
  Return ONLY a valid JSON object.  No markdown fences, no extra text.
  {
      "reason":  "...",
      "relation_delta":  float
  }
# Input Data
  - Scenario:  {scenario}
  - Opponent Profile ({opponent_name}):  {opponent_profile}
  - Strategy Model Profile ({strategy_name}):  {strategy_profile}
  - Full Dialogue History:  {history}
```

Figure 6: Relationship-change evaluation prompt. The judge measures the change in the counterpart's favorability; the original SOTOPIA benchmarks report `relation_delta` on $[-5, 5]$, while SOTOPIA-π uses $[-1, 1]$.

C Additional Implementation Details

We provide complementary notes for reproducibility. All experiments use an open-source reinforcement learning framework. For each policy–opponent pair, we repeat the evaluation five times and report the mean across repeats together with the corresponding standard deviation; the main tables report these statistics for both Goal Achievement and Relationship Change.

We additionally evaluate human alignment using three independent annotators. The annotators score each PRM process record and each ORM dialogue-level judgment independently, and we use the mean of their ratings as the human reference. The PRM audit contains 497 process records from 100 dialogue trajectories; the ORM audit contains 100 dialogue-level judgments from the same number of trajectories. Correlations with the automatic scores are computed after the corresponding score transformations.

PPO uses $\gamma = 0.95$, GAE $\lambda = 0.95$, and clip $\epsilon = 0.2$; learning rates are 1×10^{-5} for the policy and 2×10^{-5} for the value network, both with cosine annealing. Each batch contains 265 scenes, and each scene samples 4 trajectories, training runs for 100 iterations. Trajectory lengths typically range from 10 to 20 turns. Multiple trajectories are sampled per scenario within each iteration, and batches mix fragments across scenarios.

The Process Reward Model (PRM) is Qwen3.5-35B-A3B, and Judge Model is Deepseek-v4-flash with temperature $T = 0.1$. Process and outcome reward weights are $\alpha = 0.3$ and $\beta = 1$, selected by validation-set grid search. Reward-weight defenses include whitelist filtering, non-negativity, per-dimension caps, fallback to a default stage-aware prior under degenerate outputs, overflow redistribution, and ℓ_1 normalization so weights sum to 1. All inference and training run on $8\times$A100 (80 GB) GPUs.

For fair comparison, Sotopia-RL, SDPO and ArCHer are re-implemented on Qwen2.5-7B using the same SOTOPIA-π data. Behavior Cloning is trained on about 1,000 expert trajectories generated by GPT-4o, following the SOTOPIA-π setup (Wang et al., 2024). Opponent models use the same role settings and temperature $T = 0.7$.

Figure 7 reports the training-set Goal success rate for all four policy backbones. We exclude every evaluation record marked by `test_epoch` and retain each optimization batch recorded during training. The four runs same 100 steps for Qwen3-8B, Qwen2.5-7B, LLaMA3.1-8B, and Gemma-3-4B, respectively. Light curves show raw batch values, and solid curves use reflection-padded Gaussian smoothing with a bandwidth proportional to run length.

D Additional Benchmark Results

We present results in the order SOTOPIA-All, SOTOPIA-Hard, SOTOPIA-π, and AgentSense.

D.1 SOTOPIA-All

Table 6 reports the original SOTOPIA-All benchmark. Unlike SOTOPIA-π, Goal Achievement is scored on $[0, 10]$ and Relationship Change on $[-5, 5]$. Against GPT-5.5, SocialRL (Qwen2.5-7B) reaches $7.16/2.59$, above BC ($6.87/2.48$), Sotopia-RL ($6.49/2.66$) and ArCHer ($6.12/2.40$) on Goal / Relationship. SocialRL (Qwen3-8B) achieves a $7.78/1.51$ average, while SocialRL (LLaMA3.1-8B) gives the strongest trained-model averages overall (8.00 Goal and 2.35 Relationship Change). Training Gemma-3-4B raises its average from $5.60/-0.28$ to $7.42/1.04$, gains of 1.82 and 1.32 on the two metrics.

D.2 SOTOPIA-Hard

Table 7 reports results on SOTOPIA-Hard, where Goal Achievement is scored on $[0, 10]$ and Relationship Change on $[-5, 5]$. SocialRL (Qwen3-8B) achieves the strongest trained-model average Goal Achievement score (7.07), while SocialRL (LLaMA3.1-8B) gives the strongest average Relationship Change (1.91). Against GPT-5.5, SocialRL (Qwen3-8B) reaches 5.94, 0.83 above BC (5.11); its corresponding Relationship Change is 0.26, compared with 2.30 for BC. Gemma-3-4B

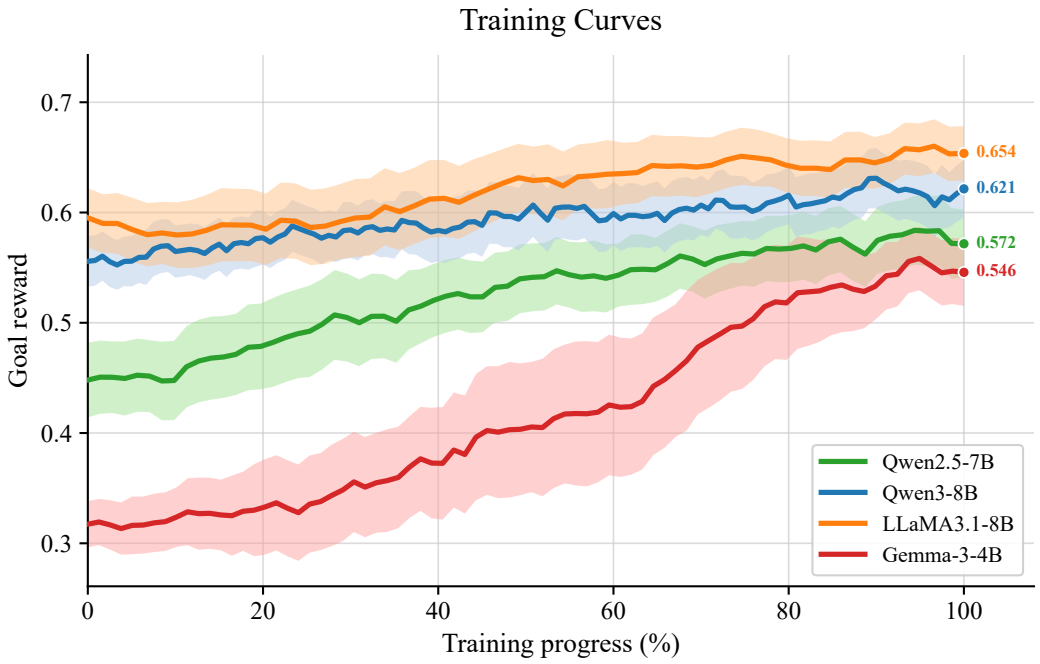


Figure 7: Training-set Goal Reward for the four SocialRL policy backbones. Light lines are raw non-test batches, and solid lines are reflection-padded Gaussian-smoothed trends.

Table 6: Results on the SOTOPIA-ALL benchmark: Goal Achievement Score ($[0, 10]$) / Relationship Change ($[-5, 5]$) for each policy–opponent pair. Both metrics are reported as mean $\pm$ SD across repeat means. Bold denotes the best SocialRL model per metric and opponent column.

	Opponent Model				
Method	Qwen2.5-7B	Qwen3-8B	Qwen3.5-35B	GPT-5.5	Avg.
Base (Qwen2.5-7B)	5.67±0.20 / 1.79±0.038	5.33±0.20 / 1.54±0.053	6.03±0.21 / 2.08±0.036	6.07±0.11 / 2.55±0.042	5.78 / 1.99
Base (Qwen3-8B)	7.75±0.11 / 1.28±0.051	7.00±0.14 / 0.63±0.018	7.54±0.11 / 0.86±0.041	7.68±0.11 / 1.28±0.053	7.49 / 1.01
Base (LLaMA3.1-8B)	7.84±0.12 / 2.27±0.016	6.94±0.19 / 1.98±0.026	7.65±0.27 / 1.63±0.048	7.86±0.20 / 2.46±0.038	7.57 / 2.08
Base (Gemma-3-4B)	4.65±0.15 / −0.26±0.054	5.19±0.08 / −0.44±0.021	5.59±0.29 / −0.74±0.064	6.96±0.20 / 0.32±0.025	5.60 / −0.28
BC (Qwen2.5-7B)	6.19±0.20 / 2.17±0.032	6.00±0.21 / 1.94±0.037	6.30±0.10 / 1.77±0.018	6.87±0.27 / 2.48±0.037	6.34 / 2.09
SDPO (Qwen2.5-7B)	6.06±0.23 / 2.01±0.046	5.54±0.17 / 2.03±0.027	5.52±0.23 / 2.09±0.030	6.62±0.23 / 2.51±0.043	5.93 / 2.16
Sotopia-RL (Qwen2.5-7B)	7.19±0.24 / 2.31±0.031	5.97±0.12 / 1.85±0.017	6.48±0.24 / 1.89±0.052	6.49±0.14 / 2.66±0.016	6.53 / 2.18
ArCHer (Qwen2.5-7B)	5.28±0.19 / 1.68±0.060	5.77±0.30 / 1.49±0.029	5.97±0.22 / 2.22±0.011	6.12±0.20 / 2.40±0.017	5.78 / 1.95
SocialRL (Qwen2.5-7B)	6.64±0.22 / 2.05±0.044	6.59±0.20 / 1.73±0.040	7.42±0.31 / **2.35**±0.063	7.16±0.17 / 2.59±0.050	6.95 / 2.18
SocialRL (Qwen3-8B)	**8.45**±0.13 / 1.59±0.049	7.58±0.13 / 1.87±0.012	7.88±0.12 / 1.12±0.030	7.19±0.19 / 1.43±0.040	**7.78** / 1.51
SocialRL (LLaMA3.1-8B)	7.99±0.17 / **2.48**±0.044	7.59±0.17 / **2.15**±0.031	**8.41**±0.27 / 2.01±0.042	**8.00**±0.08 / **2.76**±0.015	**8.00 / 2.35**
SocialRL (Gemma-3-4B)	7.46±0.14 / 1.21±0.069	7.10±0.29 / 0.50±0.054	7.51±0.25 / 1.02±0.040	7.62±0.20 / 1.42±0.035	7.42 / 1.04
GPT-5.5 (reference)	8.87±0.11 / 2.45±0.029	8.88±0.15 / 2.29±0.045	8.67±0.17 / 2.47±0.026	8.22±0.06 / 2.60±0.032	8.66 / 2.45

exhibits a benchmark-specific trade-off: training raises average Goal Achievement from 5.82 to 6.67 and Relationship Change from -0.83 to 0.26.

D.3 SOTOPIA-π

Table 8 provides the complete SOTOPIA-π results, including standard deviations, for the mean-only table in the main text.

D.4 AGENTSENSE

Table 9 provides the complete AgentSense results, including standard deviations, for the mean-only table in the main text.

Table 7: Results on the SOTOPIA-HARD benchmark: Goal Achievement Score ($[0, 10]$) / Relationship Change ($[-5, 5]$) for each policy–opponent pair. Both metrics are reported as mean ± SD across repeat means. Bold denotes the best SocialRL model per metric and opponent column.

	Opponent Model				
Method	Qwen2.5-7B	Qwen3-8B	Qwen3.5-35B	GPT-5.5	Avg.
Base (Qwen2.5-7B)	5.38±0.45 / 1.61±0.056	4.03±0.23 / 1.29±0.073	4.58±0.17 / 1.94±0.051	4.18±0.37 / 2.49±0.062	4.55 / 1.83
Base (Qwen3-8B)	7.51±0.20 / 0.94±0.041	6.12±0.20 / −0.48±0.068	6.52±0.18 / −0.30±0.017	6.58±0.30 / −0.07±0.074	6.68 / 0.02
Base (LLaMA3.1-8B)	7.29±0.35 / 2.24±0.019	6.46±0.22 / 1.61±0.059	6.43±0.35 / 0.91±0.033	6.71±0.32 / 1.99±0.076	6.72 / 1.69
Base (Gemma-3-4B)	5.14±0.26 / −0.68±0.123	5.51±0.20 / −1.07±0.044	5.57±0.72 / −1.46±0.116	7.05±0.51 / −0.12±0.022	5.82 / −0.83
BC (Qwen2.5-7B)	5.17±0.32 / 2.01±0.044	3.94±0.44 / 1.57±0.062	4.49±0.28 / 0.90±0.019	5.11±0.37 / 2.20±0.042	4.68 / 1.67
SDPO (Qwen2.5-7B)	5.85±0.22 / 1.99±0.090	3.60±0.28 / 1.70±0.051	3.75±0.18 / 1.82±0.047	4.22±0.37 / 2.30±0.035	4.35 / 1.95
Sotopia-RL (Qwen2.5-7B)	6.92±0.51 / 1.99±0.045	4.68±0.32 / 1.65±0.037	4.43±0.28 / 1.31±0.046	4.34±0.33 / 2.19±0.042	5.09 / 1.79
ArCHer (Qwen2.5-7B)	4.71±0.42 / 1.69±0.092	4.74±0.43 / 1.01±0.039	4.43±0.33 / 1.88±0.018	4.95±0.25 / 2.22±0.056	4.71 / 1.70
SocialRL (Qwen2.5-7B)	6.34±0.30 / 1.87±0.069	6.06±0.23 / 1.24±0.053	6.03±0.35 / **2.00**±0.071	5.54±0.24 / 2.14±0.074	5.99 / 1.81
SocialRL (Qwen3-8B)	**8.00**±0.15 / 1.24±0.035	**7.66**±0.20 / 1.53±0.039	6.68±0.18 / −0.16±0.066	5.94±0.18 / 0.26±0.065	**7.07** / 0.72
SocialRL (LLaMA3.1-8B)	7.17±0.14 / **2.23**±0.083	6.09±0.14 / **1.70**±0.064	**7.54**±0.29 / 1.34±0.073	**6.77**±0.39 / **2.38**±0.035	6.89 / **1.91**
SocialRL (Gemma-3-4B)	6.95±0.17 / 0.99±0.103	6.89±0.32 / −0.60±0.078	6.62±0.38 / 0.27±0.056	6.22±0.32 / 0.39±0.055	6.67 / 0.26
GPT-5.5 (reference)	8.49±0.20 / 2.28±0.035	8.77±0.29 / 2.05±0.093	7.75±0.08 / 2.11±0.038	6.95±0.30 / 2.28±0.044	7.99 / 2.18

Table 8: Complete SOTOPIA-π results. Goal Achievement (%) / Relationship Change ($\times 100$) are reported as mean ± SD across repeated experiments; Relationship Change values and SDs are multiplied by 100.

	Opponent Model				
Method	Qwen2.5-7B	Qwen3-8B	Qwen3.5-35B	GPT-5.5	Avg.
Base (Qwen2.5-7B)	33.5±0.62 / 23.5±0.16	38.4±0.36 / 22.8±0.45	44.6±0.67 / 32.1±0.39	52.5±0.52 / 45.8±0.28	42.3 / 31.1
Base (Qwen3-8B)	43.2±0.32 / 13.6±0.33	44.2±0.42 / 8.3±0.31	55.0±0.68 / 13.7±0.32	60.9±0.58 / 24.8±0.18	50.8 / 15.1
Base (LLaMA3.1-8B)	44.3±0.53 / 21.7±0.15	47.1±0.65 / 19.4±0.37	55.0±0.46 / 17.3±0.21	63.2±0.34 / 37.1±0.43	52.4 / 23.9
Base (Gemma-3-4B)	22.0±0.36 / −10.6±0.16	30.3±0.62 / −6.3±0.22	32.1±0.85 / −15.0±0.65	37.6±0.64 / −5.4±0.34	30.5 / −9.3
BC (Qwen2.5-7B)	38.9±0.39 / 27.2±0.20	43.0±0.96 / 23.6±0.42	49.9±0.42 / 22.3±0.17	57.9±0.75 / 43.7±0.31	47.4 / 29.2
SDPO (Qwen2.5-7B)	33.1±0.22 / 24.0±0.22	42.3±0.24 / 23.9±0.29	45.3±0.52 / 31.1±0.25	55.0±0.89 / 44.9±0.25	43.9 / 31.0
Sotopia-RL (Qwen2.5-7B)	41.2±0.64 / 26.0±0.17	43.7±0.36 / 20.9±0.20	49.0±0.16 / 17.8±0.34	60.5±0.74 / 39.4±0.14	48.6 / 26.0
ArCHer (Qwen2.5-7B)	31.6±0.49 / 22.6±0.18	39.6±0.59 / 22.8±0.22	46.6±1.06 / 31.3±0.52	48.6±0.87 / 41.0±0.16	41.6 / 29.4
SocialRL (Qwen2.5-7B)	43.3±0.64 / 28.2±0.14	45.8±0.82 / 26.9±0.18	56.7±0.77 / **36.3**±0.24	63.3±0.51 / 51.6±0.27	52.3 / **35.8**
SocialRL (Qwen3-8B)	47.7±0.36 / 18.8±0.30	48.1±0.08 / 20.8±0.09	61.0±0.65 / 18.1±0.45	62.2±0.71 / 24.9±0.34	54.7 / 20.6
SocialRL (LLaMA3.1-8B)	**50.2**±0.54 / **31.2**±0.16	**54.3**±0.85 / **27.6**±0.48	**63.6**±0.37 / 24.9±0.28	**70.2**±0.32 / **53.0**±0.10	**59.6** / 34.1
SocialRL (Gemma-3-4B)	42.4±0.50 / 10.6±0.34	48.0±0.88 / 6.4±0.25	58.6±0.78 / 13.7±0.46	60.3±0.85 / 22.9±0.49	52.3 / 13.4
GPT-5.5 (reference)	57.0±0.51 / 31.6±0.20	64.9±0.17 / 32.2±0.23	75.3±0.44 / 40.1±0.16	78.9±0.41 / 51.4±0.10	69.0 / 38.8

E Ablation Experiments

We answer RQ2 by systematically ablating the main design choices with Qwen2.5-7B as the policy model.

E.1 PPO versus GRPO

We compare the algorithm choice by contrasting SocialRL's multi-turn PPO with the multi-turn GRPO variant analyzed in Appendix A, using the same Qwen2.5-7B backbone and reward design. The contextual comparison and the main result are shown in Figure 3. Multi-turn GRPO obtains $53.4\%/0.436$, whereas PPO obtains $63.3\%/0.516$, improving Goal Achievement by 9.9 percentage points and Relationship Change by 0.080. This controlled result supports our theoretical analysis that a state-dependent value baseline is better suited to delayed, context-dependent social rewards than group-normalized GRPO.

E.2 Necessity of Process Rewards

We train SocialRL-NoProcess by removing all intermediate process rewards and retaining only outcome feedback (Figure 8). To keep this comparison on the same scale as the main experiment, we report Goal Success Rate and Relationship Change against GPT-5.5. The Base Qwen2.5-7B model obtains $52.5\%/0.458$, BC obtains $57.9\%/0.437$, and full SocialRL obtains $63.3\%/0.516$ (Goal / Relationship), using the corresponding GPT-5.5 column in Table 2. The NoProcess variant obtains

Table 9: Complete AgentSense results. Goal Achievement Success Rate (%) / Relationship Change ($\times 100$) are reported as mean $\pm$ SD across five independent repeats; Relationship Change values and SDs are multiplied by 100.

	Opponent Model				
Method	Qwen2.5-7B	Qwen3-8B	Qwen3.5-35B	GPT-5.5	Avg.
Base (Qwen2.5-7B)	53.7±0.12 / 32.8±0.07	55.7±0.20 / 33.6±0.12	65.2±0.29 / 40.9±0.11	58.9±0.17 / 36.3±0.24	58.4 / 35.9
Base (Qwen3-8B)	67.8±0.20 / 33.7±0.18	72.9±0.22 / 38.9±0.12	69.6±0.24 / 35.8±0.23	82.8±0.18 / 48.4±0.17	73.3 / 39.2
Base (LLaMA3.1-8B)	83.2±0.03 / 48.1±0.12	85.2±0.38 / 49.2±0.22	86.2±0.18 / 51.9±0.11	89.4±0.23 / 58.1±0.14	86.0 / 51.8
Base (Gemma-3-4B)	44.4±0.28 / 6.3±0.16	45.4±0.20 / 6.1±0.18	48.1±0.31 / 9.7±0.14	47.5±0.28 / 3.7±0.18	46.3 / 6.5
BC (Qwen2.5-7B)	60.3±0.14 / 36.7±0.22	56.7±0.08 / 32.8±0.14	63.1±0.14 / 37.9±0.12	67.0±0.27 / 46.9±0.20	61.8 / 38.6
SDPO (Qwen2.5-7B)	57.6±0.33 / 35.8±0.11	58.0±0.15 / 36.3±0.10	63.0±0.17 / 37.4±0.23	66.7±0.26 / 48.1±0.11	61.3 / 39.4
Sotopia-RL (Qwen2.5-7B)	63.5±0.32 / 37.1±0.23	59.3±0.25 / 33.5±0.33	63.9±0.26 / 35.8±0.19	69.6±0.28 / 49.4±0.08	64.1 / 38.9
ArCHer (Qwen2.5-7B)	51.1±0.14 / 31.9±0.16	52.6±0.27 / 32.0±0.07	60.9±0.21 / 38.9±0.19	58.9±0.32 / 35.7±0.07	55.9 / 34.6
SocialRL (Qwen2.5-7B)	61.0±0.18 / 38.6±0.10	62.2±0.26 / 38.5±0.18	72.1±0.10 / 45.9±0.12	71.7±0.23 / 47.6±0.05	66.7 / 42.7
SocialRL (Qwen3-8B)	76.7±0.12 / 39.5±0.13	77.1±0.15 / 40.9±0.08	82.0±0.26 / 46.3±0.21	86.0±0.31 / 50.7±0.07	80.5 / 44.4
SocialRL (LLaMA3.1-8B)	**84.3**±0.22 / **48.3**±0.06	**85.9**±0.25 / **50.1**±0.07	**87.6**±0.10 / **52.4**±0.20	**90.9**±0.23 / **58.9**±0.10	**87.2 / 52.4**
SocialRL (Gemma-3-4B)	66.7±0.24 / 22.5±0.15	66.4±0.16 / 22.3±0.16	68.8±0.15 / 24.7±0.06	70.4±0.21 / 23.9±0.23	68.1 / 23.4
GPT-5.5 (reference)	88.1±0.15 / 45.6±0.06	92.2±0.15 / 48.9±0.22	93.8±0.14 / 52.6±0.12	92.5±0.09 / 48.9±0.13	91.7 / 49.0

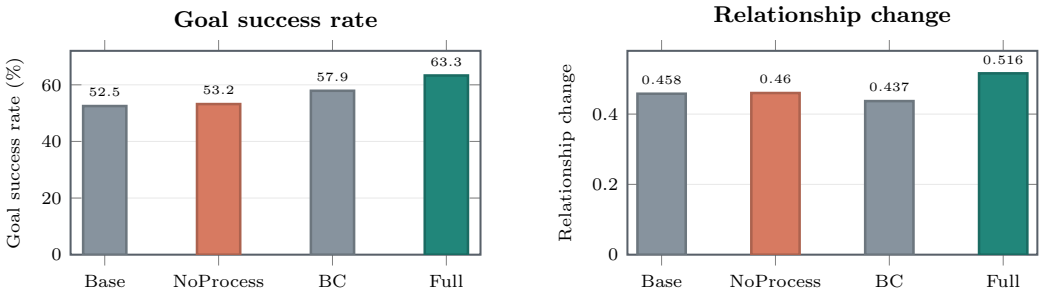


Figure 8: Ablation on process rewards on the SOTOPIA-π benchmark (Qwen2.5-7B vs. GPT-5.5). Bars report Goal Success Rate (%) / Relationship Change ($[-1, 1]$). Base, BC, and full SocialRL use the corresponding GPT-5.5 column from Table 2; NoProcess is an additional run with process rewards removed.

$53.2\%/0.460$: removing process rewards largely eliminates the improvement over the base model and leaves a substantial gap to full SocialRL in goal success. This result shows that outcome feedback alone is insufficient to supervise intermediate social behavior, whereas dense process rewards provide useful turn-level signals throughout the dialogue.

E.3 Contribution of Each Reward Dimension

We train six leave-one-out variants on SOTOPIA-π with Qwen2.5-7B as the policy and GPT-5.5 as the opponent (Figure 9). Full SocialRL obtains $63.3\%/0.516$ in Goal Success Rate and Relationship Change. Removing **goal advancement** or **strategic positioning** yields the lowest Goal results, $55.3\%/0.486$ and $56.1\%/0.479$, respectively, identifying these dimensions as the main drivers of task progress. Removing **relational attunement** gives $61.8\%/0.401$: Goal remains comparatively high, but Relationship Change is the lowest among all variants, consistent with its role in relationship maintenance. The persona consistency, contextual coherence, and turn quality variants obtain $62.0\%/0.456$, $62.6\%/0.472$, and $63.1\%/0.481$, respectively. These dimensions support both objectives, although no single one dominates either metric as strongly as the two goal-side dimensions or relational attunement.

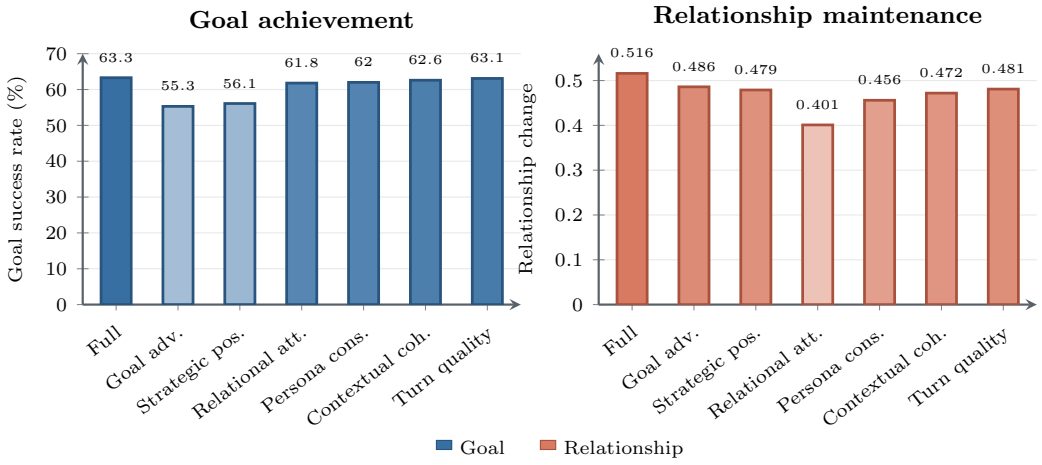


Figure 9: Leave-one-dimension-out ablations on SOTOPIA-π (Qwen2.5-7B vs. GPT-5.5). Bars report the original Goal Success Rate (%, bottom axis) and Relationship Change ($[-1, 1]$, top axis); full SocialRL is included as the reference.

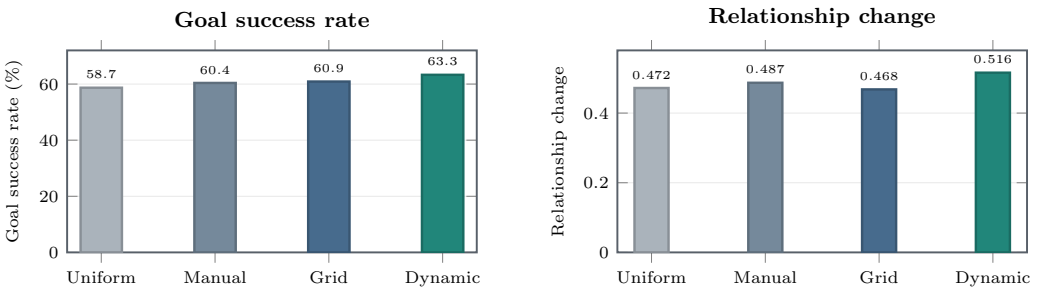


Figure 10: Ablation on dynamic versus fixed reward-weight schedules on the SOTOPIA-π benchmark (Qwen2.5-7B vs. GPT-5.5). Bars report Goal Success Rate (%) / Relationship Change ($[-1, 1]$). The Dynamic values are taken from the SocialRL row in Table 2; the fixed-weight values are the corresponding comparison runs.

E.4 Necessity of Dynamic Weights

We compare the proposed dynamic schedule with three fixed-weight alternatives using the same Qwen2.5-7B policy and GPT-5.5 opponent (Figure 10). Values are reported as Goal Success Rate (%) / Relationship Change ($[-1, 1]$). The **Uniform** scheme assigns equal weight to the six reward dimensions at every turn, $[1/6, 1/6, 1/6, 1/6, 1/6, 1/6]$, and obtains $58.7\%/0.472$. The **Manual** scheme uses a fixed heuristic vector that emphasizes relational attunement early and goal dimensions later; its early, middle, and late vectors are $[0.10, 0.10, 0.30, 0.15, 0.20, 0.15]$, $[0.20, 0.20, 0.20, 0.15, 0.15, 0.10]$, and $[0.30, 0.25, 0.10, 0.15, 0.10, 0.10]$, respectively, and it obtains $60.4\%/0.487$. The **Grid** scheme selects one constant vector from a coarse grid, $[0.24, 0.20, 0.14, 0.16, 0.14, 0.12]$, and obtains $60.9\%/0.468$. Finally, **Dynamic** uses the stage-dependent weights described in Section 3.3; its result is taken from the SocialRL row against GPT-5.5 in Table 2, $63.3\%/0.516$. Dynamic weighting is therefore higher than all three fixed alternatives on both reported metrics, with its largest fixed-schedule margin over Grid being $+2.4$ percentage points in Goal Success Rate, and $+0.029$ in Relationship Change over Manual. These results support the claim that adapting the reward focus across a conversation is more effective than selecting one constant weighting scheme.

F Extended Qualitative Example

Figure 11 shows representative turns from the qualitative comparison used in Section 4.5. Ethan is the evaluated policy—Base Qwen2.5-7B-Instruct or SocialRL Qwen2.5-7B—and Benjamin is the GPT-5.5 opponent. Benjamin's slow turns reduce the group's enjoyment; Ethan must communicate the need for faster play and propose a concrete solution without offending him.

Intervention ownership. The outcome difference is not explained by politeness. Base is consistently warm, but its first response—"Sure, take your time"—actually relaxes the pressure to address slow play. Benjamin then diagnoses the problem and supplies the one-minute limit himself. Ethan acknowledges that solution but never communicates the group's need or takes ownership of a concrete intervention. This distinction explains the judge's failure decision: the desired social state emerges, but not through the evaluated agent's goal-directed action.

From opening to commitment. SocialRL does not impose a timer without social permission. Benjamin first offers an opening ("If I stall again, call a timer"), which Ethan converts into an actionable proposal ("Let's set a quick timer for your next turn"). Benjamin then specifies ninety seconds, and Ethan explicitly ratifies both the duration and the behavioral boundary ("no deep dives"). The sequence forms a commitment ladder—permission, proposal, parameterization, and ratification—rather than a single forceful request. This preserves Benjamin's agency and makes the rule face-saving.

Long-horizon persistence. The critical evidence appears after agreement. Benjamin later announces "timer starts, brain engages, no dissertation" and makes a decisive move, showing behavioral uptake rather than superficial assent. He subsequently invites future enforcement ("nudge me"), extending the convention beyond one turn. SocialRL therefore establishes a reusable coordination mechanism; Base only reinforces Benjamin's independent self-correction. The later compliance and persistence show that SocialRL's intervention remains effective beyond the turn in which the timer is proposed.

Goal–relationship balance. Both trajectories receive the same positive relationship change of $+0.6$. Thus the case does not show that SocialRL is simply friendlier than Base; Base is already highly supportive. Instead, SocialRL improves task agency without paying a relationship cost: goal success rises from 0 to 1 while relationship quality is held constant. Benjamin's humor, acceptance, appreciation of the kind call-out, and request for later reminders provide concrete evidence that the firmer coordination rule remains relationally acceptable.

Board-game case: who turns concern into a shared pacing rule?
Ethan's goal: communicate the group's need for faster play and propose a face-saving solution.



Figure 11: Representative turns from the board-game dialogue. Benjamin (GPT-5.5) appears on the left; Ethan's policy responses appear on the right. Base preserves rapport but leaves the pacing intervention to Benjamin. SocialRL operationalizes Benjamin's opening into a mutually ratified ninety-second rule. Both receive $\mathrm{relation_delta} = +0.6$, while goal success changes from 0 to 1.