LARGE LANGUAGE MODELS LEARN TO PERSUADE MULTIPLE RECEIVERS THROUGH SOCIAL CONTEXT

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) enable open-ended persuasion, yet existing LLM-based approaches largely focus on isolated, short-horizon interactions, overlooking how target selection and social influence jointly shape collective outcomes. To address this gap, we formulate *LLM-based multi-receiver persuasion* as a natural-language sequential decision problem, in which an LLM persuader decides whom to address and what to say to heterogeneous receivers connected by a social graph. We instantiate this setting in three domains: academic rebuttal, political decision-making, and charitable campaigns. Our key insight is that evolving social context is central to both persuasion evaluation and policy learning. For evaluation, we find that incorporating social context generally improves correlation and reduces prediction error with human judgments across all three domains and accordingly develop Persuade-RM, which outperforms the task-specific Rebuttal-RM in academic rebuttal. For policy learning, we propose *Graph-Contextual Group Policy Optimization* (GCGPO), a critic-free algorithm that groups socially comparable steps and combines step-level advantages with episode-level relative advantages. Across all three domains, GCGPO achieves strong quality and stance scores while avoiding the degradation observed in several baselines. We further develop a coupled actor-reward generation mechanism that reduces training time per iteration by 44.7% on average while maintaining comparable evaluation performance. Together, these results establish social context as a key ingredient for effective evaluation and policy learning in LLM-based multi-receiver persuasion.

1 INTRODUCTION

Persuasion, as a fundamental form of human communication that seeks to shape others' beliefs, attitudes, or opinions, has been widely studied in the social sciences (O'keefe, 2016; Cialdini, 2001; Simons & Jones, 2011). In many real-world settings, however, persuasion involves multiple heterogeneous receivers and evolving social interactions rather than isolated individuals. For example, firms target heterogeneous consumers with personalized advertisements (Arieli & Babichenko, 2019), political campaigns and lobbyists seek to influence multiple voters or committee members (Alonso & Câmara, 2016; Kerman & Tenev, 2025), and charitable organizations tailor appeals to different potential donors (Wang et al., 2019). These scenarios share a common structure: persuasion depends not only on message quality, but also on strategic interactions among multiple receivers (Arieli & Babichenko, 2019).

Existing research on persuasion has explored this problem from two complementary perspectives. First, *Bayesian Persuasion (BP)* studies strategic information disclosure, where an informed sender designs signals to influence receivers' decisions (Kamenica & Gentzkow, 2011). Extensions to multi-receiver settings consider personalized signaling strategies and collective objectives over multiple receivers (Arieli & Babichenko, 2019; Kerman & Tenev, 2025). However, these approaches rely on predefined signals and simplified receiver models, limiting their applicability to open-ended natural language persuasion. Sec-

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.

ond, *LLM-based persuasion* utilizes LLMs to extend persuasion from predefined signals to open-ended natural language. Recent studies explore verbalized Bayesian persuasion (Li et al., 2025), grounding persuasion in realistic dialogues (He et al., 2025), strategic persuasion evaluation and training (Cheng & You, 2026), and domain-specific persuasive agents such as academic rebuttal systems (He et al., 2026). Nevertheless, existing LLM-based persuasion primarily focuses on short interactions and limited strategic planning, leaving multi-receiver persuasion under evolving social contexts largely unexplored. As illustrated in Figure 1, existing paradigms capture only partial aspects of realistic persuasion.

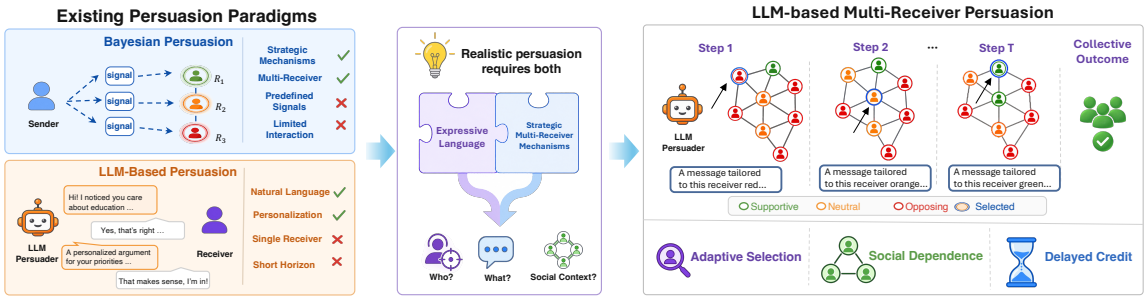

Figure 1: Motivation for LLM-based multi-receiver persuasion. Realistic persuasion requires combining expressive language with strategic multi-receiver reasoning under evolving social context.

To bridge this gap, we formulate **LLM-based multi-receiver persuasion**, a natural-language persuasion setting where an LLM-based persuader sequentially interacts with multiple heterogeneous receivers in a social network. Unlike existing settings, we require the persuader to jointly determine *which receiver to approach* and *what message to generate* under an evolving multi-step persuasion process. Each receiver has a unique profile and stance, while neighbors' stances create social dependencies that shape collective outcomes. We instantiate this setting across three representative domains, including academic rebuttal (He et al., 2026), political (Kerman & Tenev, 2025), and charitable campaigns (Arieli & Babichenko, 2019).

This formulation introduces three fundamental challenges. First, **adaptive receiver selection**: Sequential information design demonstrates the benefits of optimizing disclosure order (Wu et al., 2022). The persuader must decide which receiver to target sequentially, as different receivers have different potential influence on future outcomes. Second, **socially dependent persuasion**: Receivers' decisions depend on peer information, as illustrated by conformity patterns in academic rebuttal (Gao et al., 2019; Kargaran et al., 2025). Therefore, effective evaluation and decision-making must incorporate social context rather than treating each receiver in isolation. Third, **long-horizon credit assignment**: The collective outcome emerges from receiver selections and messages, where early interventions reshape future social states. This creates a delayed-reward optimization challenge, making action contributions difficult to locate under evolving dynamics.

Our main contributions are summarized as follows: (i) We formulate LLM-based multi-receiver persuasion as a natural-language sequential decision problem with heterogeneous receivers, adaptive target selection, social influence, and collective objectives, and instantiate it in academic, political, and charitable domains. (ii) Motivated by observed conformity patterns, we develop Persuade-RM, which incorporates receiver information and visible neighbors' stances. Including peer stances improves correlation with human judgments across all domains. (iii) We propose Graph-Contextual Group Policy Optimization (GCGPO), which combines episode-level and graph-contextual step advantages for fine-grained credit assignment, together with coupled actor-reward generation for single-pass self-assessment. (iv) Across all three domains, GCGPO consistently achieves the highest quality and stance scores among the compared baselines steadily without degradation. Coupled actor-reward architecture reduces training time per iteration by 44.7% while maintaining comparable evaluation performance to the decoupled variant.

2 Related Work

Multi-receiver Bayesian Persuasion. Multi-receiver Bayesian persuasion generalizes the classical Bayesian persuasion framework of Kamenica & Gentzkow (2011) to multiple receivers. The sender designs a signaling scheme to induce a desired collective outcome. This literature follows two main directions, signal design and social influence. First, signal design examines how receiver heterogeneity (Shimoji, 2022), private communication (Arieli & Babichenko, 2019), and signal leakage (Haghtalab et al., 2025) shape signaling schemes. Second, social influence studies how receivers update their opinions through repeated peer interactions. These models capture both attachment to initial beliefs (Friedkin & Johnsen, 1990) and the effects of others' opinions (De et al., 2016; Zhu et al., 2021; Li et al., 2026a). However, these approaches typically represent persuasion through signaling schemes over discrete states and actions (Kamenica & Gentzkow, 2011; Arieli & Babichenko, 2019), or through scalar opinion updates under fixed influence structures (Friedkin & Johnsen, 1990; Zhu et al., 2021), limiting their applicability to open-ended language persuasion.

LLM-Based Persuasion. Recent work extends persuasion from discrete signals to open-ended natural-language communication. In the single-receiver setting, prior studies verbalize Bayesian persuasion in theory-grounded environments (Li et al., 2025), construct persuasive dialogue data with multi-LLM pipelines (Ma et al., 2025), develop domain-specific agents for academic rebuttal and misinformation (He et al., 2026; Han et al., 2026). These methods mainly consider isolated, short-horizon dyadic interactions and do not model social influence among receivers. By contrast, research on multi-receiver persuasion remains preliminary. Although Verbalized Bayesian Persuasion discusses multi-receiver applications, its formulation and experiments are primarily dyadic (Li et al., 2025). Overall, existing approaches remain task-specific and lack a unified framework for LLM-based persuasion with multi receivers (Rogiers et al., 2024; Bozdag et al., 2026).

3 Task Formulation

For a persuasion task with context c, we consider a single persuader and n receivers indexed by $i \in \mathcal{V} = [n]$. We model a persuasion task as a finite-horizon MDP with $T = 2n$ steps. At each pair of steps, the persuader selects a receiver and then attempts to persuade them with an open-ended natural-language message. Receivers form a directed social graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where an edge $(j, i) \in \mathcal{E}$ indicates that receiver i observes receiver j's stance. Each receiver i has a textual profile p_i, an initial stance s_i^0 and a stance score s_i^t after step $t \in [T]$, with higher values indicating greater persuasiveness. Let $\mathbf{p} = (p_i)_{i\in\mathcal{V}}$ and $\mathbf{s}^t = (s_i^t)_{i\in\mathcal{V}}$ denote the collections of profiles and stances. Accordingly, each task instance is represented as a tuple $(c, \mathcal{G}, \mathbf{p}, \mathbf{s}^0)$. Following He et al. (2026), we assess message quality across four dimensions $\mathcal{L}_Q = \{\mathrm{att}, \mathrm{cla}, \mathrm{con}, \mathrm{per}\}$ (Attitude, Clarity, Constructiveness, Persuasiveness). At each step t:

(1) **State (x_t):** The state x_t contains the full interaction context available before a_t, including c, $\mathcal{G}$, $\mathbf{p}$, $\mathbf{s}^{t-1}$.

(2) **Action (a_t):** At a selection step, the persuader selects a receiver $i \in \mathcal{V}$ that has not been selected before; at the subsequent persuasion step, i denotes the selected target receiver and the persuader generates an open-ended natural-language persuasion message for that receiver i.

(3) **Transition:** At a selection step, the environment validates the output and returns a success signal η_t without altering stances; at a persuasion step, the environment returns a message-quality vector $\mathbf{q}_t = \big(q_{t,\ell}\big)_{\ell\in\mathcal{L}_Q}$, a post-intervention stance s_i^t for receiver i, and a success signal η_t.

(4) **Reward (r_t):** The reward is computed directly from these returned signals. At a selection step, $r_t = U_s(\eta_t)$ evaluates format compliance and receiver validity; at a persuasion step, each receiver can be selected only once. We therefore define a single receiver-level stance uplift as $\Delta s_i := s_i^t - s_i^0$, and compute $r_t = U_p(\eta_t, \mathbf{q}_t, \Delta s_i)$. U_s and U_p are the respective reward functions.

In all, the persuader learns a policy $\pi_\theta(a_t \mid x_t)$ that maximizes the expected quality scores and stance uplift.

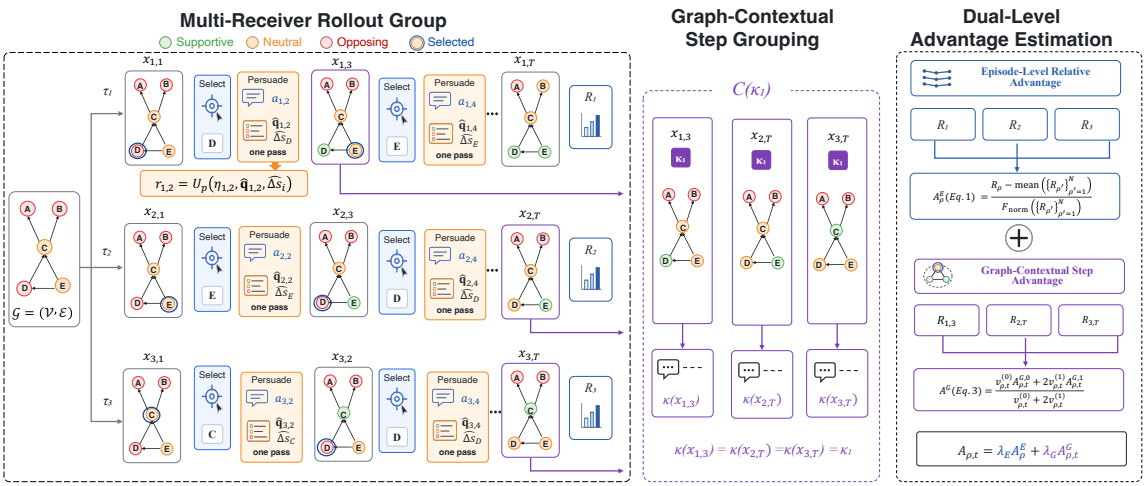

Figure 2: Overview of GCGPO. During LLM rollouts, a coupled actor-reward architecture jointly generates each persuasive action and its self-assessment. GCGPO then groups steps by identical graph-contextual keys and combines trajectory-level and graph-contextual advantages for fine-grained credit assignment.

4 METHOD

As shown in Figure 2, we propose *Graph-Contextual Group Policy Optimization* (GCGPO), a critic-free method for LLM-based multi-receiver persuasion, with two key improvements. First, we combine an *episode-level relative advantage*, which compares complete trajectories from the same task instance, with a *graph-contextual step advantage*, which compares selection and persuasion actions under similar receiver and social contexts. Second, we introduce a *coupled actor-reward* architecture that generates both the action and a self-assessment in a single pass, eliminating separate reward-model inference.

4.1 EPISODE-LEVEL RELATIVE ADVANTAGE

For each task instance, we sample N complete trajectories from the old policy $\pi_{\theta_{\text{old}}}$. The ρ-th trajectory is $\tau_\rho = \{(x_{\rho,t}, a_{\rho,t}, r_{\rho,t})\}_{t=1}^{T}, \rho \in [N]$, where $x_{\rho,t}$ is the persuasion state, $a_{\rho,t}$ is the action (either receiver selection or a natural-language message), and $r_{\rho,t}$ is the reward. Its trajectory return is $R_\rho := \sum_{t=1}^{T} r_{\rho,t}$. The episode-level relative advantage is computed by normalizing this return within the trajectory group:

$$A_\rho^E = \frac{R_\rho - \text{mean}\left(\{R_{\rho'}\}_{\rho'=1}^{N}\right)}{F_{\text{norm}}\left(\{R_{\rho'}\}_{\rho'=1}^{N}\right)}. \tag{1}$$

Following Feng et al. (2026); He et al., the default normalization factor is defined as the standard deviation, i.e., $F_{\text{norm}} = std$. A_ρ^E measures the performance of τ_ρ relative to rollouts from the same task instance; however, it cannot identify which specific step drives the final performance gain.

4.2 GRAPH-CONTEXTUAL STEP ADVANTAGE

To enable finer-grained credit assignment, GCGPO constructs step-level comparison groups from collected rollouts. Unlike prior state-based grouping methods (Feng et al., 2026; He et al.), exact persuasion states rarely recur: textually different observations may represent the same social decision role. We therefore define

graph-contextual keys to compare decisions under similar social conditions while abstracting away textual variation.

First, we define a receiver-level context summary that captures the local social context of each receiver. Let $\mathcal{V}_t^{\text{done}} \subseteq \mathcal{V}$ denote the receivers whose persuasion steps have been completed before step t. We distinguish the incoming visible-neighbor set $\mathcal{N}_{\text{in}}(i) := \{j \in \mathcal{V} : (j,i) \in \mathcal{E}\}$ from the incident-neighbor set used for grouping, $\mathcal{N}(i) := \{j \in \mathcal{V} : (j,i) \in \mathcal{E} \text{ or } (i,j) \in \mathcal{E}\}$. Receiver observations and Persuade-RM use $\mathcal{N}_{\text{in}}(i)$, whereas GCGPO constructs its grouping keys using $\mathcal{N}(i)$. We use two fixed context levels: $\psi_t^{(0)}(i) = \left(p_i, \Delta s_i^{t-1}\right), \psi_t^{(1)}(i) \quad = \left(\psi_t^{(0)}(i), \text{Agg}\left(\{s_j^{t-1} : j \in \mathcal{N}(i)\}\right)\right)$. Here, p_i is the receiver's profile and Δs_i^{t-1} is its stance change before step t (defaulting to 0). Agg records the neighborhood size and mean current stance, with the latter rounded to two decimal places; for $\mathcal{N}(i) = \varnothing$, it returns $(0, \text{NA})$.

Next, GCGPO uses different graph-contextual keys for the two action types at each level $\ell \in \{0,1\}$. For a selection step, the action determines which receiver will be addressed next. The key $\kappa^{(\ell)}(x_t)$ summarizes the current progress of the persuasion process, including the contexts of both completed and remaining receivers: $\kappa^{(\ell)}(x_t) = \left(|\mathcal{V}_t^{\text{done}}|, |\mathcal{V}|, \{\{\psi_t^{(\ell)}(i) : i \in \mathcal{V} \setminus \mathcal{V}_t^{\text{done}}\}\}, \{\{\psi_t^{(\ell)}(i) : i \in \mathcal{V}_t^{\text{done}}\}\}\right)$. For a persuasion step, the selected receiver i_t is already fixed, so the key is the corresponding context of the target receiver: $\kappa^{(\ell)}(x_t) = \left(\psi_t^{(\ell)}(i_t)\right)$. We then group steps with identical keys at each level: $\mathcal{C}^{(\ell)}(\kappa) = \left\{(\rho,u) : \kappa^{(\ell)}(x_{\rho,u}) = \kappa,\ \rho \in [N],\ 1 \le u \le T\right\}$. To account for the downstream effects of each action, we define the discounted return-to-go from step t in trajectory τ_ρ as $R_{\rho,t} = \sum_{u=t}^{T} \gamma^{u-t} r_{\rho,u}$, where $\gamma \in (0,1]$ is the discount factor. In particular, $R_{\rho,1} = R_\rho$ when $\gamma = 1$. GCGPO computes a step advantage at each level by normalizing discounted returns within the corresponding group:

$$A^{G,\ell}(x_{\rho,t}, a_{\rho,t}) = \frac{R_{\rho,t} - \text{mean}\left(\{R_{\rho',u} : (\rho',u) \in \mathcal{C}^{(\ell)}(\kappa^{(\ell)}(x_{\rho,t}))\}\right)}{F_{\text{norm}}\left(\{R_{\rho',u} : (\rho',u) \in \mathcal{C}^{(\ell)}(\kappa^{(\ell)}(x_{\rho,t}))\}\right)}. \tag{2}$$

Let $A_{\rho,t}^{G,\ell} := A^{G,\ell}(x_{\rho,t}, a_{\rho,t})$, and $v_{\rho,t}^{(\ell)}$ indicate whether its group contains at least $m_{\min}$ samples. We aggregate valid levels with fixed weights and set $A_{\rho,t}^{G} = 0$ when the denominator is zero, excluding undersized groups to mitigate variance (Li et al., 2026c):

$$A_{\rho,t}^{G} = \frac{v_{\rho,t}^{(0)} A_{\rho,t}^{G,0} + 2 v_{\rho,t}^{(1)} A_{\rho,t}^{G,1}}{v_{\rho,t}^{(0)} + 2 v_{\rho,t}^{(1)}}. \tag{3}$$

4.3 GCGPO Objective

As shown in Alg. 1, episode-level and graph-contextual advantages are combined into a step-level advantage:

$$A_{\rho,t} = \lambda_E A_\rho^E + \lambda_G A_{\rho,t}^G, \rho \in [N], t \in [T]. \tag{4}$$

where λ_E and λ_G control their relative contributions. We set both to 1.0 in all experiments. We optimize this advantage using a token-level PPO clipped objective with KL regularization toward a fixed reference policy $\pi_{\theta_{\text{ref}}} := \pi_{\theta_0}$, where π_{θ_0} denotes the initial pretrained policy. Let $\mathcal{M}$ denote the index set of all policy-action tokens across $\{\tau_\rho\}_{\rho=1}^N$; self-assessment-token indices are excluded. Let $w_{\rho,t,h}(\theta) = \frac{\pi_\theta(a_{\rho,t,h}|x_{\rho,t},a_{\rho,t,<h})}{\pi_{\theta_{\text{old}}}(a_{\rho,t,h}|x_{\rho,t},a_{\rho,t,<h})}$ be the token-level importance weight. We write its clipped counterpart as $w_{\rho,t,h}^{\text{clip}}(\theta) := \text{clip}(w_{\rho,t,h}(\theta), 1-\epsilon, 1+\epsilon)$, where ϵ is the clipping threshold. The GCGPO objective is then

$$J(\theta) = \mathbb{E}_{\{\tau_\rho\}_{\rho=1}^N \sim \pi_{\theta_{\text{old}}}}\left[\frac{1}{|\mathcal{M}|} \sum_{(\rho,t,h)\in\mathcal{M}} \left(\min\left\{w_{\rho,t,h}(\theta)A_{\rho,t}, w_{\rho,t,h}^{\text{clip}}(\theta)A_{\rho,t}\right\} - \beta d_{\rho,t,h}(\theta)\right)\right]. \tag{5}$$

Algorithm 1: Graph-Contextual Group Policy Optimization

Input: Initial policy π_{θ_0}, task instance $(c, \mathcal{G}, \mathbf{p}, \mathbf{s}^0)$, rollout size N, discount γ, context-group size $m_{\min}$

Initialize $\pi_\theta \leftarrow \pi_{\theta_0}$, $\pi_{\theta_{\text{ref}}} \leftarrow \pi_{\theta_0}$
for *each training iteration* **do**
  Set $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
  Roll out N trajectories $\{\tau_\rho\}_{\rho=1}^N$ from the same task instance using $\pi_{\theta_{\text{old}}}$
  Store $r_{\rho,t}$ and $\{\kappa^{(\ell)}(x_{\rho,t})\}_{\ell=0}^1$ for every $(\rho, t) \in [N] \times [T]$
  Compute R_ρ and A_ρ^E for every $\rho \in [N]$ using Eq. (1)
  for $(\rho, t) \in [N] \times [T]$ **do**
    Compute $R_{\rho,t}$ for each step
  Build $\mathcal{C}^{(0)}$ and $\mathcal{C}^{(1)}$ by hashing their keys over $(\rho, u) \in [N] \times [T]$
  Compute and aggregate the level-wise advantages using Eqs. (2)–(3)
  Combine advantages with Eq. (4)
  Update π_θ with the clipped objective in Eq. (5)

where $d_{\rho,t,h}(\theta)$ is the token-level KL regularizer toward $\pi_{\theta_{\text{ref}}}$, and $\beta \geq 0$ controls the strength of KL regularization. The same step-level advantage $A_{\rho,t}$ is assigned to all action tokens generated at step t.

4.4 Coupled Actor-Reward Generation

To avoid separate online reward-model inference (Liu et al., 2025; 2026), $\pi_{\theta_{\text{old}}}$ jointly generates each action and structured evaluation signals required by the environment in one pass. Specifically, at a persuasion step, the actor produces an actor-assessed quality vector $\widehat{\mathbf{q}}_t = (\widehat{q}_{t,\ell})_{\ell \in \mathcal{L}_Q}$, and an actor-assessed post-intervention stance $\widehat{s}_i^t$. We define $\widehat{\Delta s}_i := \widehat{s}_i^t - s_i^0$ and use $r_t = U_p(\eta_t, \widehat{\mathbf{q}}_t, \widehat{\Delta s}_i)$ for training. During evaluation, the frozen Persuade-RM independently produces $\mathbf{q}_t$ and s_i^t, yielding $\Delta s_i := s_i^t - s_i^0$; these outputs are not used for training. The actor and Persuade-RM share only the assessment rubric; the actor receives no Persuade-RM supervision, distillation, or calibration.

5 Experiment

5.1 Datasets and Evaluation

LLM-based multi-receiver persuasion remains underexplored. Existing approaches are often confined to narrow domains (He et al., 2026; Han et al., 2026), overlook social influence among receivers, or lack a unified task and evaluation framework (Rogiers et al., 2024; Bozdag et al., 2026). To address these limitations, we make two contributions: (1) we introduce three multi-receiver persuasion datasets; and (2) we identify limitations in existing evaluation criteria and develop a unified evaluation framework.

Datasets. Beyond the academic rebuttal dataset built from real ICLR 2024-2025 peer-review data (Kargaran et al., 2025; Tan et al., 2024), we collect data in two additional domains: political persuasion (Hackenburg & Margetts, 2024) and charity persuasion (Wang et al., 2019). Since existing data are limited to scalar survey responses without large-scale dialogue texts, we combine LLM-simulated augmentation with human-annotated calibration, where the test set is exclusively human-annotated.

Social-Context-Aware Evaluator. Because human evaluation does not scale to open-ended responses, we propose Persuade-RM as an evaluator. We denote human-annotated stance uplift as $\Delta s_{i,\text{H}}$ and LLM-

predicted stance uplift as $\Delta s_{i,\mathrm{RM}}$. Rebuttal-RM (He et al., 2026) has been proposed to evaluate persuasion

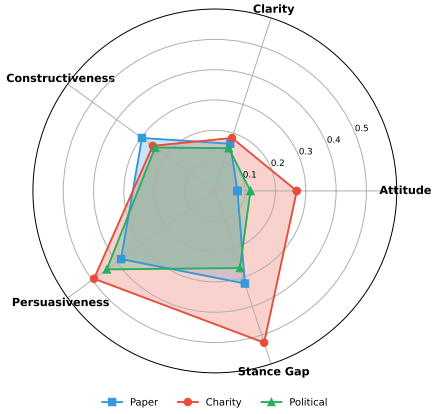

Figure 3: Pearson correlations of the four quality scores and the initial neighbor-stance gap g_i^0 with ground-truth stance uplift $\Delta s_{i,\mathrm{H}}$ across the three domains.

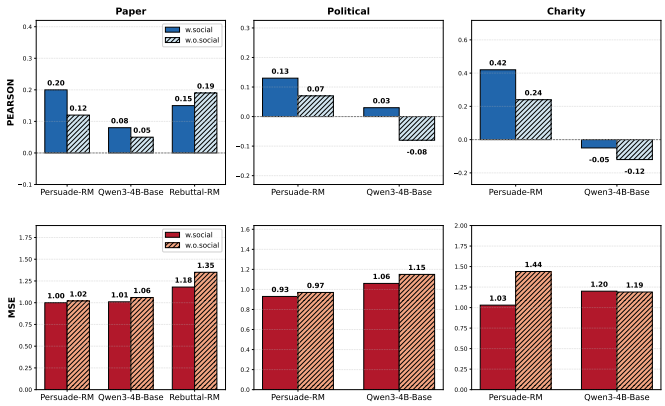

Figure 4: Comparison of LLM-generated stance-change predictions across the three domains. The upper and lower rows report Pearson correlation and MSE between $\Delta s_{i,\mathrm{RM}}$ and $\Delta s_{i,\mathrm{H}}$, respectively. Adding social context improves prediction accuracy.

rebuttal responses while it ignores social context. We first define receiver i's initial neighbor-stance gap as $g_i^0 := \bar{s}^0_{\mathcal{N}_{\mathrm{in}}(i)} - s_i^0$, $\bar{s}^0_{\mathcal{N}_{\mathrm{in}}(i)} := \frac{1}{|\mathcal{N}_{\mathrm{in}}(i)|}\sum_{j\in\mathcal{N}_{\mathrm{in}}(i)} s_j^0$, where $\mathcal{N}_{\mathrm{in}}(i)$ denotes receiver i's incoming visible neighbors (we omit receivers with $\mathcal{N}_{\mathrm{in}}(i) = \emptyset$). Figure 3 reports Pearson correlations between ground-truth uplift $\Delta s_{i,\mathrm{H}}$ and four message-quality scores plus g_i^0. The neighbor-stance gap correlates positively with uplift in Charity ($r = 0.526$), Paper ($r = 0.321$), and Political ($r = 0.268$), with $p < 0.01$ throughout. These results highlight the importance of peer influence in persuasion evaluation and that social context contains substantial predictive signal for persuasion effectiveness. Motivated by this finding, Persuade-RM incorporates social context into its evaluation prompt. We then evaluate its predicted uplift, $\Delta s_{i,\mathrm{RM}}$ against human annotated uplift $\Delta s_{i,\mathrm{H}}$ using MSE and Pearson correlation with and without social context. As shown in Figure 4, social context improves Pearson correlation and reduces MSE. Utilizing Persuade-RM as evaluator, we report four dimension-wise message quality (Q_ℓ): Attitude, Clarity, Constructiveness, and Persuasiveness; their equally weighted mean (Quality); mean final stance (Stance); and the proportion of successful persuasion rollouts (Success). Detailed definitions are provided in Appendix C.1.

Implementation Details. We compare two categories of baselines. *Prompting* includes zero-shot prompting and reflection prompting, where the latter elicits a strategy reflection before each action. *RL baselines* include GiGPO (Feng et al., 2026), HGPO (He et al.), G2PO (Wang et al., 2026), and SALT (Li et al., 2026b), representing trajectory-, hierarchical-, graph-, and state-based advantage grouping, respectively. *GCGPO* is our graph-contextual grouping method described in Section 4. We adopt Qwen3-4B-Instruct and Llama-3.2-3B-Instruct as backbones, with Llama results deferred to Appendix E.1. For each backbone-domain pair, all RL methods share the same training and optimization settings, differing only in method-specific components and hyperparameters. Full configurations are provided in Appendix D.

5.2 Results

Overall Performance Across Domains. As shown in Table 1, GCGPO achieves the highest mean across all 21 dataset-metric combinations over the three domains. Compared with the strongest baseline in each domain, GCGPO improves the Quality score by 1.4% across the three domains. GCGPO also achieves

Table 1: Evaluation results for LLM-based multi-receiver persuasion across the Paper, Political, and Charity datasets using Qwen3-4B-Instruct as the backbone model. Results are reported as mean ± standard deviation, with higher values indicating better performance for all metrics. **Bold**: best result; underline: second best.

Dataset	Baseline	Attitude	Clarity	Constructiveness	Persuasiveness	Quality	Stance	Success
Paper	Prompting	8.96±0.02	8.64±0.03	7.99±0.11	7.67±0.03	8.32±0.09	6.70±0.00	0.97±0.01
	Reflection	8.86±0.02	8.77±0.03	7.94±0.03	7.74±0.03	8.33±0.03	6.70±0.01	0.94±0.01
	G2PO	8.81±0.02	8.46±0.01	7.92±0.00	7.45±0.02	8.16±0.01	6.65±0.01	0.93±0.01
	HGPO	5.25±1.61	5.03±1.62	4.57±1.60	4.24±1.48	4.77±1.57	6.07±0.23	0.25±0.43
	GIGPO	6.43±1.21	6.06±1.24	5.63±1.17	5.07±1.16	5.80±1.19	6.17±0.17	0.46±0.42
	SALT	8.82±0.02	8.80±0.02	8.25±0.01	7.77±0.03	8.41±0.02	6.62±0.01	0.98±0.01
	GCGPO	**8.99±0.01**	**8.85±0.01**	**8.41±0.02**	**7.85±0.02**	**8.53±0.02**	**6.76±0.01**	**0.99±0.01**
Political	Reflection	8.09±0.06	8.09±0.06	7.82±0.05	7.03±0.05	7.76±0.05	5.65±0.08	0.78±0.01
	Prompting	7.99±0.08	7.99±0.08	7.71±0.07	6.93±0.06	7.66±0.07	5.64±0.07	0.74±0.02
	G2PO	1.77±0.03	1.29±0.00	0.85±0.01	0.84±0.01	1.19±0.00	4.90±0.01	0.13±0.02
	HGPO	7.32±0.71	7.32±0.71	7.12±0.68	6.35±0.61	7.03±0.68	5.75±0.09	0.53±0.20
	GIGPO	8.67±0.03	8.59±0.04	7.75±0.29	7.03±0.21	8.01±0.14	5.43±0.16	0.92±0.01
	SALT	8.69±0.02	8.69±0.02	8.44±0.02	7.52±0.03	8.34±0.02	5.80±0.09	0.92±0.01
	GCGPO	**8.85±0.05**	**8.85±0.05**	**8.63±0.06**	**7.69±0.04**	**8.51±0.05**	**5.84±0.09**	**0.95±0.01**
Charity	Reflection	7.85±0.17	7.85±0.17	7.62±0.16	7.04±0.16	7.59±0.17	6.59±0.02	0.64±0.05
	Prompting	7.71±0.03	7.71±0.03	7.47±0.02	6.90±0.03	7.45±0.03	6.58±0.02	0.61±0.01
	G2PO	1.68±0.02	1.13±0.04	0.99±0.05	1.03±0.03	1.21±0.04	5.35±0.01	0.09±0.00
	HGPO	6.26±0.01	6.24±0.01	6.10±0.01	5.45±0.02	6.01±0.01	6.39±0.00	0.05±0.00
	GIGPO	8.20±0.37	8.19±0.37	7.97±0.30	7.26±0.33	7.91±0.34	6.56±0.03	0.71±0.13
	SALT	8.84±0.01	8.84±0.01	8.64±0.01	7.88±0.00	8.55±0.01	6.62±0.00	0.94±0.00
	GCGPO	**8.91±0.06**	**8.90±0.06**	**8.70±0.08**	**7.93±0.02**	**8.61±0.02**	**6.65±0.03**	**0.96±0.02**

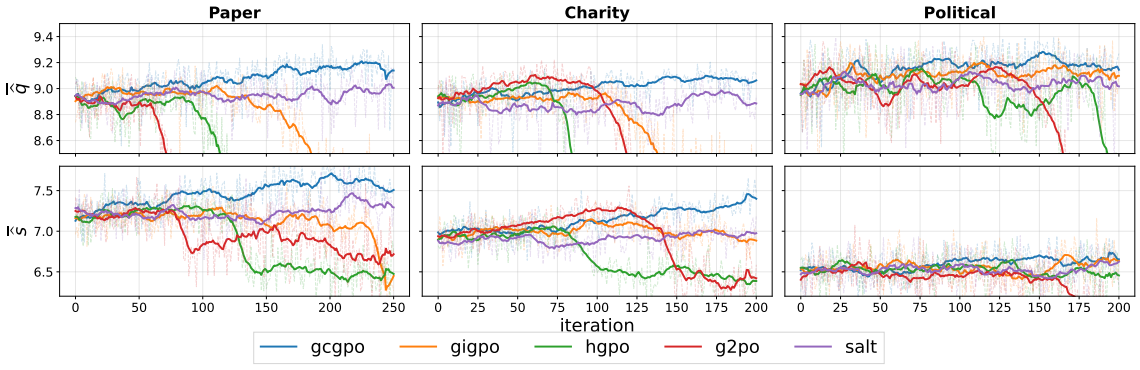

Figure 5: Training dynamics across three domains. At each training update, the upper and lower rows show batch means of actor-assessed message quality $\bar{\hat{q}}$ and post-intervention stance $\bar{\hat{s}}$, respectively. Solid lines show smoothed trends; dashed lines show unsmoothed batch means.

the highest Success rates across all domains, with relative improvements of 1.0%, 3.3%, and 2.1% over the strongest baselines, respectively. Several baselines show domain-specific degradation, including G2PO on Political and GiGPO on Paper. Results with a second backbone further support GCGPO's robustness (Appendix E.1). The contrast is particularly informative for credit assignment: G2PO and HGPO produce lower aggregate Quality than both prompting methods on Paper and Political, while GiGPO degrades sharply on Paper. This pattern is consistent with noisy or sparse step-level comparisons as trajectories diverge; for GiGPO, temporal discounting further attenuates early consequential actions. SALT is more robust because

it shares trajectory-level advantages across trajectories for steps with the same action type, and this coarse grouping conflates decisions involving different social states. For example, selecting a receiver before and after neighboring stances change represents two distinct social contexts, despite sharing the same action type. GCGPO addresses this ambiguity through two complementary graph-contextual levels: receiver-state matching and a neighborhood-augmented social context. By distinguishing steps according to their social context, this design provides more informative step-level credit, consistent with the observed Quality gains over SALT of 0.12, 0.17, and 0.06 points on Paper, Political, and Charity, respectively. Results across domains and backbones further support its robust final performance.

Training Stability. At each training update, we track mean actor-assessed message quality $\bar{\hat{q}}$ and mean actor-assessed post-intervention stance $\bar{\hat{s}}$. Detailed definitions are provided in Appendix C.2. Figure 5 shows that both metrics increase more steadily under GCGPO across domains, while G2PO and HGPO exhibit abrupt degradation and GiGPO and SALT generally plateau earlier. This instability is consistent with unreliable step-level comparisons in groups that are too small or mix steps under different social conditions. GCGPO alleviates this by incorporating receiver and neighborhood context, restricting step-level advantages to sufficiently large groups, and retaining episode-level advantages. Together, these mechanisms filter unreliable comparisons while preserving a robust learning signal, leading to the more stable and sustained optimization observed across domains.

5.3 Efficiency Analysis

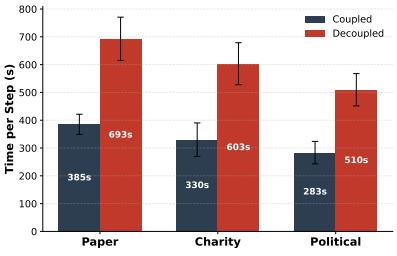

Figure 6: Average training iteration time for the coupled and decoupled actor-reward architectures across three domains. The former reduces iteration time by 44.7% on average.

We compare the efficiency of the coupled actor-reward architecture with a vanilla decoupled self-reward variant. In the decoupled baseline, the actor generates an action, and a separate frozen model assesses it in a second pass. Both models are initialized from the same policy SFT checkpoint. This training-time self-assessor is distinct from Persuade-RM, which is used only for evaluation. Figure 6 shows consistent time savings across all three domains. Coupling reduces average training iteration time from 693 to 385 seconds on Paper, from 510 to 283 seconds on Political, and from 603 to 330 seconds on Charity. The average reduction across domains is 44.7%. These savings come from removing the separate assessment pass after each persuasion step. Coupling also eliminates the need to store separate self-assessor weights. Appendix E.3 further shows that coupled actor-reward architecture achieves substantial efficiency gains while maintaining comparable performance.

6 Conclusion

In this paper, we introduce *LLM-based multi-receiver persuasion*, where an LLM-based persuader sequentially interacts with multiple socially connected heterogeneous receivers. We instantiate this setting in academic rebuttal, political decision-making, and charitable campaigns. Our key finding is that social context is critical for both evaluation and policy learning. For evaluation, we show that incorporating social context improves alignment with human judgments and develop Persuade-RM. For policy learning, we propose GCGPO, which performs graph-contextual credit assignment by grouping socially comparable steps, together with coupled actor-reward generation for efficient training. Across all three domains, GCGPO achieves the best quality and stance performance among the compared methods. Coupled actor-reward generation reduces training time per iteration by 44.7% on average across domains while maintaining comparable performance. Overall, these results highlight the importance of social context in LLM-based multi-receiver persuasion.

7 Ethics Statement

This work studies multi-receiver persuasion with LLMs in socially contextualized environments. While our goal is to better understand and model how social context affects persuasion, such techniques could also be misused for manipulation, targeted influence, or other forms of undesirable behavioral intervention. We therefore restrict our study to controlled experimental settings. We emphasize that improvements in persuasion effectiveness should not be interpreted as endorsing manipulative applications. Real-world deployment would require additional safeguards concerning user consent, transparency, privacy, fairness, and the potential amplification of social influence or bias.

8 Reproducibility Statement.

To support reproducibility, we provide detailed descriptions of the environment, GCGPO training objectives and evaluation protocols. Additional implementation details are included in the appendix. Code and data are available at `https://anonymous.4open.science/r/Persuasion/README.md`.

9 Generative AI Disclosure.

In this work, we used generative AI tools for language polishing, data augmentation, and base-model training. All AI-assisted content and outputs were reviewed by the authors. We take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the assistance of generative AI.

References

Ricardo Alonso and Odilon Câmara. Persuading voters. *American Economic Review*, 106(11):3590–3605, 2016.

Itai Arieli and Yakov Babichenko. Private bayesian persuasion. *Journal of Economic Theory*, 182:185–217, 2019.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439): 509–512, 1999.

Nimet Beyza Bozdag, Shuhaib Mehri, Xiaocheng Yang, Hyeonjeong Ha, Zirui Cheng, Esin Durmus, Jiaxuan You, Heng Ji, Gokhan Tur, and Dilek Hakkani-Tür. Must read: A comprehensive survey of computational persuasion. *ACM Computing Surveys*, 58(12):1–39, 2026.

Zirui Cheng and Jiaxuan You. Towards strategic persuasion with language models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *Proceedings of the International Conference on Learning Representations (ICLR)*, volume 2026, pp. 58190–58229, 2026.

Robert B Cialdini. The science of persuasion. *Scientific American*, 284(2):76–81, 2001.

Abir De, Isabel Valera, Niloy Ganguly, Sourangshu Bhattacharya, and Manuel Gomez Rodriguez. Learning and forecasting opinion dynamics in social networks. *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 29, 2016.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025. doi: 10.48550/ARXIV.2501.12948.

Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, Guorui Zhou, Yutao Zhu, Ji-Rong Wen, and Zhicheng Dou. Agentic reinforced policy optimization. *CoRR*, abs/2507.19849, 2025. doi: 10.48550/ARXIV.2507.19849.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 46375–46408, 2026.

Noah E Friedkin and Eugene C Johnsen. Social influence and opinions. *Journal of mathematical sociology*, 15(3-4):193–206, 1990.

Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. Does my rebuttal matter? insights from a major NLP conference. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 1274–1290. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1129.

Robert H Gass and John S Seiter. *Persuasion: Social influence and compliance gaining*. Routledge, 2022.

Kobi Hackenburg and Helen Margetts. Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences (PNAS)*, 121(24):e2403116121, 2024.

Nika Haghtalab, Mingda Qiao, and Kunhe Yang. Leakage-robust bayesian persuasion. In *Proceedings of the 26th ACM Conference on Economics and Computation (EC)*, pp. 1018–1018, 2025.

Chen Han, Yijia Ma, Jin Tan, Wenzhen Zheng, and Xijin Tang. Beyond detection: Exploring evidence-based multi-agent debate for misinformation intervention and persuasion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 38542–38550, 2026.

Buwei He, Yang Liu, Zhaowei Zhang, Zixia Jia, Huijia Wu, Zhaofeng He, Zilong Zheng, and Yipeng Kang. Make an offer they can't refuse: Grounding bayesian persuasion in real-world dialogues without pre-commitment. *CoRR*, abs/2510.13387, October 2025.

Shuo He, Lang Feng, Qi Wei, Xin Cheng, Lei Feng, and Bo An. Hierarchy-of-groups policy optimization for long-horizon agentic tasks. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*.

Zhitao He, Zongwei LYU, and Yi R. Fung. Dancing in chains: Strategic persuasion in academic rebuttal via theory of mind. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

Amir Hossein Kargaran, Nafiseh Nikeghbal, Jing Yang, and Nedjma Ousidhoum. Insights from the ICLR peer review and rebuttal process. *CoRR*, abs/2511.15462, 2025. doi: 10.48550/ARXIV.2511.15462.

Toygar T Kerman and Anastas P Tenev. Information design for weighted voting. *Economic Theory*, 79(3): 809–852, 2025.

Paul F Lazarsfeld, Bernard Berelson, and Hazel Gaudet. *The people's choice: How the voter makes up his mind in a presidential campaign*. Columbia University Press, 1968.

Jiayi Li, Xiao Liu, and Yansong Feng. From single to societal: Analyzing persona-induced bias in multi-agent interactions. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 31609–31617, 2026a.

Jiazheng Li, Yawei Wang, Qiaojing Yan, Yijun Tian, Zhichao Xu, Huan Song, Panpan Xu, and Lin Lee Cheong. Salt: Step-level advantage assignment for long-horizon agents via trajectory graph. In *Findings of the Association for Computational Linguistics (EACL)*, pp. 4709–4725, 2026b.

Wenhao Li, Yue Lin, Xiangfeng Wang, Bo Jin, Hongyuan Zha, and Baoxiang Wang. Verbalized bayesian persuasion. *arXiv preprint arXiv:2502.01587*, 2025.

Yuanfan Li, Qi Zhou, Wenjing Duan, and Lu Chen. When denser credit is not enough: Evidence-calibrated policy optimization for long-horizon llm agent training. *arXiv preprint arXiv:2606.05885*, 2026c.

Wei Liu, Siya Qi, Xinyu Wang, Chen Qian, Yali Du, and Yulan He. Nover: Incentive training for language models via verifier-free reinforcement learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7450–7469, 2025.

Xiaoyuan Liu, Tian Liang, Zhiwei He, Jiahao Xu, Wenxuan Wang, Pinjia He, Zhaopeng Tu, Haitao Mi, and Dong Yu. Trust, but verify: A self-verification approach to reinforcement learning with verifiable rewards. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 130475–130501, 2026.

Weicheng Ma, Hefan Zhang, Ivory Yang, Shiyu Ji, Joice Chen, Farnoosh Hashemi, Shubham Mohole, Ethan Gearey, Michael Macy, Saeed Hassanpour, et al. Communication makes perfect: Persuasion dataset construction via multi-llm communication. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4017–4045, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Daniel J O'keefe. *Persuasion: Theory and research*. SAGE Publications, Inc., 2016.

Alexander Rogiers, Sander Noels, Maarten Buyl, and Tijl De Bie. Persuasion with large language models: a survey. *arXiv preprint arXiv:2411.06837*, 2024.

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

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

Makoto Shimoji. Bayesian persuasion in unlinked games. *International Journal of Game Theory*, 51(3): 451–481, 2022.

Herbert W Simons and Jean Jones. *Persuasion in society*. Taylor & Francis, 2011.

Saksham Sahai Srivastava and Vaneet Aggarwal. A technical survey of reinforcement learning techniques for large language models. *CoRR*, abs/2507.04136, 2025. doi: 10.48550/ARXIV.2507.04136.

Cheng Tan, Dongxin Lyu, Siyuan Li, Zhangyang Gao, Jingxuan Wei, Siqi Ma, Zicheng Liu, and Stan Z Li. Peer review as a multi-turn and long-context dialogue with role-based interactions. *arXiv preprint arXiv:2406.05688*, 2024.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th annual meeting of the association for computational linguistics (ACL)*, pp. 5635–5649, 2019.

Yunan Wang, Minghui Song, Zihan Zhang, Shaohan Huang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. Group-graph policy optimization for long-horizon agentic reinforcement learning. *arXiv preprint arXiv:2606.22995*, 2026.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, Eli Gottlieb, Yiping Lu, Kyunghyun Cho, Jiajun Wu, Li Fei-Fei, Lijuan Wang, Yejin Choi, and Manling Li. RAGEN: understanding self-evolution in LLM agents via multi-turn reinforcement learning. *CoRR*, abs/2504.20073, 2025. doi: 10.48550/ARXIV.2504.20073.

Jibang Wu, Zixuan Zhang, Zhe Feng, Zhaoran Wang, Zhuoran Yang, Michael I. Jordan, and Haifeng Xu. Sequential information design: Markov persuasion process and its efficient reinforcement learning. *Proceedings of the 23rd ACM Conference on Economics and Computation (EC)*, pp. 471–472, 2022. doi: 10.1145/3490486.3538313.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: an open-source LLM reinforcement learning system at scale. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *CoRR*, abs/2507.18071, 2025. doi: 10.48550/ARXIV.2507.18071.

Liwang Zhu, Qi Bao, and Zhongzhi Zhang. Minimizing polarization and disagreement in social networks via link recommendation. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 2072–2084, 2021.

Appendix: Contents **14**

A Additional Related Work **15**

B Dataset **15**

B.1 Dataset Format 15

B.2 Simulation Pipeline 16

B.3 Human Annotation Calibration 17

C Evaluation **17**

C.1 Evaluation Metrics 17

C.2 Training Diagnostic Metrics 18

C.3 Persuade-RM Training Methodology 18

C.4 Social-Context Analysis and Ablation 19

C.5 Persuade-RM Performance Evaluation 20

D Experiment Details **21**

D.1 Compared Methods and RL Implementation 21

D.2 SFT Data Collection and Fine-tuning 23

E Ablation Experiments **23**

E.1 Ablation on the Backbone Model 23

E.2 Ablation on Component Contributions 24

E.3 Ablation on Reward Mechanisms 27

F Interaction and Evaluation Prompts **28**

F.1 Policy Interaction and Output Formats 28

F.2 Reflection-Prompting Instruction 29

F.3 Persuade-RM Evaluation Prompt 29

G Future Work **30**

A Additional Related Work

Reinforcement Learning for LLM Agents. Reinforcement learning has become a central paradigm for LLM post-training, with reinforcement learning from human feedback (RLHF) aligning model outputs with human preferences through learned reward models (Ouyang et al., 2022). A widely adopted optimization algorithm is Proximal Policy Optimization (PPO), which combines actor-critic learning with clipped policy updates (Schulman et al., 2017). More recently, reinforcement learning with verifiable rewards (RLVR) has advanced reasoning-oriented post-training, using automatically checkable outcomes as training signals (DeepSeek-AI, 2025; Yu et al., 2025). Alongside this development, critic-free group-wise policy optimization methods, including GRPO (Shao et al., 2024), DAPO (Yu et al., 2025), and GSPO (Zheng et al., 2025), estimate advantages through relative reward comparisons among sampled responses, avoiding a separately trained value model.

Credit Assignment in Reinforcement Learning. Building on these advances in reinforcement learning, agentic RL extends response-level optimization to multi-step interactions with dynamic environments, enabling agents to refine planning, tool use, and decision-making through environment feedback (Srivastava & Aggarwal, 2025). Long-horizon agentic RL extends LLMs from single-turn generation to multi-turn decision-making, making effective credit assignment increasingly important as interaction horizons grow (Wang et al., 2025; Dong et al., 2025). Recent studies show that incorporating step-level supervision or intermediate signals can substantially improve the effectiveness and efficiency of agent training over extended interactions. Accordingly, recent methods have moved beyond trajectory-level optimization toward finer-grained credit assignment across interaction steps (Feng et al., 2026; Wang et al., 2026; Li et al., 2026b; He et al.). However, existing methods primarily organize credit assignment around trajectories, interaction histories, or intermediate states, without explicitly accounting for social dependencies among interacting individuals. This limitation is particularly relevant to AI4SS tasks, where an action can affect not only its immediate target but also the social context that shapes subsequent decisions.

B Dataset

Below we detail the data format, the simulated data generation pipeline, and the human annotation calibration process.

B.1 Dataset Format

For a persuasion task with context c, we consider a single persuader and n receivers indexed by $i \in \mathcal{V} = [n]$. Receivers form a directed social graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where an edge $(j, i) \in \mathcal{E}$ indicates that receiver i observes receiver j's stance. The receiver set is taken directly from the task's `social_network` field, and `network_structure` is restricted to the subgraph induced by that set. A receiver with no incoming edge observes no neighbor and is represented by an empty social context. Each receiver i has a textual profile p_i and an initial stance score $s_i^0 \in \{1, \ldots, 10\}$. Let $\mathbf{p} = (p_i)_{i \in \mathcal{V}}$ and $\mathbf{s}^0 = (s_i^0)_{i \in \mathcal{V}}$ denote the collections of profiles and initial stances, respectively. Accordingly, each task instance is represented as a tuple $(c, \mathcal{G}, \mathbf{p}, \mathbf{s}^0)$.

The persuasion process unfolds over a finite horizon $T = 2n$. At each step t, the persuader either selects a target receiver $i \in \mathcal{V}$ or generates a persuasion message. The policy model π_θ observes the full state x_t, which includes the context c, the graph $\mathcal{G}$, all profiles $\mathbf{p}$, the current stance configuration $\mathbf{s}^{t-1}$, and the conversation history. Each rollout produces a complete trajectory $\tau = \{(x_t, a_t, r_t)\}_{t=1}^{T}$. Because each receiver can be selected only once, $s_i^{t-1} = s_i^0$ immediately before its persuasion action at step t.

During training rollouts, our *Coupled Actor-Reward Generation* scheme allows the actor to jointly generate each persuasion message and its structured self-assessment in a single inference pass. The scalar training reward depends on the action type. At a selection step, $r_t = U_s(\eta_t)$, where U_s evaluates format compliance

and receiver validity; At a persuasion step t, the self-assessment for the target receiver i comprises an actor-assessed message-quality vector $\widehat{\mathbf{q}}_t$ and a predicted post-intervention stance $\widehat{s}_i^t$. The corresponding receiver-level stance uplift is $\widehat{\Delta s}_i := \widehat{s}_i^t - s_i^0$, and the scalar reward is $r_t = U_p\Big(\eta_t, \widehat{\mathbf{q}}_t, \widehat{\Delta s}_i\Big)$. Below we detail the domain-specific instantiations of profiles, stances, and graph construction. Dataset statistics are summarized in Table 2.

Table 2: Dataset statistics across three multi-receiver persuasion domains.

			Dataset		
Domain	Source	$\lvert\mathcal{V}\rvert$	Train	Val	Test
Academic Rebuttal	ICLR 2024-25	3-5	7,000	1,000	860
Political Persuasion	Human+Simulated	2-4	2,640	694	368
Charity Persuasion	Human+Simulated	3-5	2,627	690	367

Academic Rebuttal. Receivers are paper reviewers and context is the submitted manuscript. Data is collected from real peer review at ICLR 2024-2025 conferences (Kargaran et al., 2025; Tan et al., 2024). Each reviewer i has profile $p_i = (\text{soundness}_i,\ \text{contribution}_i,\ \text{presentation}_i,\ \text{confidence}_i)$ with each dimension rated $1-5$. Following ICLR conference guidelines, the initial stance $s_i^0 \in \{1, 3, 5, 6, 8, 10\}$ corresponds to the overall rating. Edges in $\mathcal{E}$ connect reviewers who can observe each other's reviews. The persuader (author) generates rebuttals as persuasion messages, and Δs_i denotes reviewer i's resulting rating change.

Political Persuasion. Receivers are council members and context is a proposed policy. Each receiver i has profile $p_i = (\text{leaning}_i,\ \text{consistency}_i)$ rated $1-5$. Initial stance $s_i^0 \in \{1, \ldots, 10\}$ represents support level for the policy. Edges in $\mathcal{E}$ connect voters who can observe each other's positions. The persuader presents a policy pitch followed by responses to audience questions. Data is constructed via human-annotated seed scenarios augmented by LLM.

Charity Persuasion. Receivers are potential donors and context is a charity campaign. Each receiver i has profile $p_i = (\text{income}_i,\ \text{generosity}_i)$ rated $1-5$. Initial stance $s_i^0 \in \{1, \ldots, 10\}$ represents willingness to donate. Edges in $\mathcal{E}$ connect donors who can observe each other's responses. The persuader makes a campaign appeal followed by responses to donor inquiries. Data is constructed via human-annotated seed scenarios augmented by LLM.

B.2 Simulation Pipeline

For each domain, we construct a multi-receiver persuasion environment with one persuader and n receivers connected through a directed social graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Each trial samples a domain-admissible number of receivers: $2-4$ for Political and $3-5$ for Paper and Charity, and a topic from the domain's topic set. The simulation proceeds through five stages: (1) **Proposal**: the persuader generates a public proposal for the topic, observable by all receivers, denoted as c; (2) **Initial stance formation**: each receiver i independently forms an initial stance s_i^0 based on p_i and c and states a core concern; (3) **Social pressure observation**: when the conformity mechanism is enabled, receiver i observes its visible neighbors' initial stances and computes the initial neighbor-stance gap $g_i^0 = \bar{s}^0_{\mathcal{N}_{\text{in}}(i)} - s_i^0$ when its visible-neighbor set is nonempty. Otherwise, the social-context list is empty and no neighbor-stance gap is assigned; (4) **Persuader response**: the persuader generates a persuasion message to each receiver's concern; and (5) **Final stance update**: receiver i updates its stance by integrating response quality, social pressure, and its profile, recording the receiver-level stance uplift $\widetilde{\Delta s}_i$. We adopt Qwen3.5-27B-FP8 as the underlying language model. Receiver sampling temperature is set to 0.9 for individual heterogeneity; persuader temperature is 0.1 for argumentative stability.

Social graph design. The two domains adopt different network structures. The political domain uses a layered graph, partitioning receivers into three layers, with edges directed from upper to lower layers, reflecting top-down influence from firm to flexible individuals (Gass & Seiter, 2022; Lazarsfeld et al., 1968). The Charity domain uses a BA network (Barabási & Albert, 1999). Both domains employ stratified sampling to ensure layer coverage, with 5 political topics and 6 charity topics.

B.3 Human Annotation Calibration

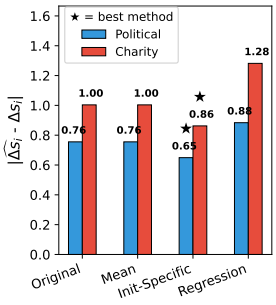

Figure 7: Delta MAE across correction strategies and init-specific correction performs best.

To validate the reliability of LLM-simulated data, we use human-annotated stances as ground truth stance. Annotators observe the full interaction for each receiver: the persuader's proposal, the receiver's initial stance s_i^0 and stated concern, any visible neighbors' initial stances, and the persuader's persuasion message. The annotator's task is to assign a final stance to the receiver based on the observed interaction. Let $\Delta s_{i,\mathrm{H}}$ denote the human-annotated stance change, and let $\widetilde{\Delta s}_i$ denote the corresponding change predicted by the LLM simulator.

We quantitatively assess the fidelity of LLM simulations using human annotations on the held-out annotated subset. As shown in Figure 7, we compare four correction strategies: (1) no correction, (2) subtracting a constant offset, (3) subtracting an offset stratified by the receiver's initial stance s_i^0, and (4) linear regression on $\widetilde{\Delta s}_i$ and s_i^0. Evaluated by the absolute error in the between-group stance-uplift difference (higher-lower delta MAE) against human annotations, the init-specific correction achieves the best calibration on both domains. We therefore adopt it to adjust the LLM-augmented samples in both the training and validation sets. We use human-annotated samples for the test set.

C Evaluation

In this section, we outline the evaluation and training diagnostic metrics. Additionally, we provide further details on the evaluation model Persuade-RM, including its training methodology, accuracy assessment, and ablation study results.

C.1 Evaluation Metrics

At evaluation time, the test-set contains M task instances and one rollout per instance, i.e., $N = 1$. For instance $m \in [M]$, let $\mathcal{V}_m$ be its receiver set, $n_m = |\mathcal{V}_m|$, and $T_m = 2n_m$. At each executed persuasion step, Persuade-RM independently returns $\mathbf{q}_{m,t} = (q_{m,t,\ell})_{\ell \in \mathcal{L}_Q}$ and the corresponding post-intervention stance. For aggregation, we set $\mathbf{q}_{m,t} = \mathbf{0}$ at selection and unexecuted steps, and $\Delta s_{m,i} = 0$ when receiver i does not complete a persuasion step.

We report four types of evaluation metrics: (1) Q_ℓ measures message quality along each dimension $\ell \in \mathcal{L}_Q$ (Attitude, Clarity, Constructiveness, and Persuasiveness, following (He et al., 2026)); (2) $\mathrm{Quality}$ is the equally weighted mean of the four dimension-wise scores; (3) Stance is the mean final stance over all

receivers; and (4) Success is the fraction of rollouts in which every receiver's persuasion step is completed:

$$Q_\ell = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{n_m}\sum_{t=1}^{T_m} q_{m,t,\ell}, \qquad \ell \in \mathcal{L}_Q, \tag{6}$$

$$\text{Quality} = \frac{1}{|\mathcal{L}_Q|}\sum_{\ell\in\mathcal{L}_Q} Q_\ell, \tag{7}$$

$$\text{Stance} = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{n_m}\sum_{i\in\mathcal{V}_m}(s^0_{m,i} + \Delta s_{m,i}), \tag{8}$$

$$\text{Success} = \frac{1}{M}\sum_{m=1}^{M}\mathbb{I}\big[\big|\mathcal{V}^{\text{done}}_{m,T_m+1}\big| = n_m\big]\,. \tag{9}$$

Here, $\mathcal{V}^{\text{done}}_{m,T_m+1}$ denotes the receivers whose persuasion steps are completed by the end of instance m's evaluation rollout. Each task instance receives equal weight.

C.2 Training Diagnostic Metrics

At each training update, we track two actor-assessed diagnostic metrics: mean message quality $\bar{\widehat{q}}$ and mean post-intervention stance $\bar{\widehat{s}}$. Each training batch contains M task instances and N rollouts per instance. For instance $m \in [M]$, let $\mathcal{V}_m$ be its receiver set, $n_m = |\mathcal{V}_m|$, and $T_m = 2n_m$. We set $\widehat{\mathbf{q}}_{m,\rho,t} = \mathbf{0}$ at selection and unexecuted steps, and $\widehat{\Delta s}_{m,\rho,i} = 0$ if receiver i does not complete a persuasion step, retaining its initial stance. The diagnostics are

$$\bar{\widehat{q}} = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{N}\sum_{\rho=1}^{N}\frac{1}{n_m}\sum_{t=1}^{T_m}\frac{1}{|\mathcal{L}_Q|}\sum_{\ell\in\mathcal{L}_Q}\widehat{q}_{m,\rho,t,\ell}, \tag{10}$$

$$\bar{\widehat{s}} = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{N}\sum_{\rho=1}^{N}\frac{1}{n_m}\sum_{i\in\mathcal{V}_m}\Big(s^0_{m,i} + \widehat{\Delta s}_{m,\rho,i}\Big)\,. \tag{11}$$

Here, $\widehat{\mathbf{q}}_{m,\rho,t}$ is the message-quality vector, $\widehat{q}_{m,\rho,t,\ell}$ is its ℓ-th component, and $\bar{\widehat{q}}$ is the scalar mean defined above. The stance diagnostic first averages over all receivers in each rollout, then over rollouts and task instances in the training batch. Receivers whose persuasion steps are not completed retain their initial stances. Here, t indexes interaction steps within a rollout, not training updates. These diagnostics are used only to visualize training dynamics and are excluded from the final Persuade-RM evaluation.

C.3 Persuade-RM Training Methodology

The Persuade-RM training methodology consists of three components:

Output Response Generation. We employ Qwen3.5-27B-FP8[1] for response generation. Following He et al. (2026), Persuade-RM assigns ratings on a 0-10 scale from four message-quality perspectives indexed by

$$\mathcal{L}_Q = \{\text{att}, \text{cla}, \text{con}, \text{per}\},$$

corresponding to *Attitude*, *Clarity*, *Constructiveness*, and *Persuasiveness*. Additionally, the output includes *OverallStance* on the same scale as the receiver's final stance rating and *ScoreExplanation* as a brief justification.

[1] `https://huggingface.co/Qwen/Qwen3.5-27B-FP8`

Prompt Design. For every Persuade-RM SFT and evaluation example, we take the task receiver set directly from `social_network` and restrict `network_structure` to the induced graph on this set. For Paper, the induced graph is a complete directed graph, so the social-context field contains the stance of every non-target receiver. For Political and Charity, it contains only the stances of receivers with an incoming edge to the target receiver. If the target has no visible in-neighbor, the stance list is empty and its mean is represented as `Unknown`. Task-provided related information is included only as optional background evidence and is not treated as a reference answer. The same context construction is used for SFT data collection and test-time evaluation. The complete evaluation template and output schema are provided in Appendix F.

Supervised Fine-Tuning. We initialize Persuade-RM from Qwen3-4B-Instruct-2507 and fine-tune it for three epochs. Each record uses the `instruction`, `input`, and `output` fields as the prompt, query, and response, respectively, and is formatted with the `qwen2_5` chat template. Training uses `bfloat16`, a maximum sequence length of 8,192 tokens, and sequence packing. We use two GPUs with a per-device batch size of 2 and accumulate gradients for 32 steps, yielding an effective batch size of 128 examples per optimizer update. The learning rate is fixed at 1×10^{-6} throughout training. We set the random seed to 42 and use four workers for data preprocessing.

C.4 Social-Context Analysis and Ablation

We assess the role of social context from two complementary perspectives. First, we examine whether a pre-interaction measure of local social context is associated with the subsequently observed receiver-level stance uplift. Second, we conduct an input ablation to determine whether providing visible-neighbor stances improves stance-uplift prediction. The former establishes predictive relevance, whereas the latter evaluates incremental predictive utility.

Association with Human-Annotated Stance Uplift. For receiver i, we define the initial neighbor-stance gap as $g_i^0 = \bar{s}^0_{\mathcal{N}_{\text{in}}(i)} - s_i^0$, where $\bar{s}^0_{\mathcal{N}_{\text{in}}(i)}$ denotes the mean initial stance of i's visible incoming neighbors. Table 3 reports its Pearson correlation with the human-annotated stance uplift $\Delta s_{i,\text{H}}$, together with the correlations of four message-quality scores. The neighbor-stance gap is positively correlated with stance uplift in Paper ($r = 0.321$), Charity ($r = 0.526$), and Political ($r = 0.268$), with $p < 0.01$ in all three domains. Among the five examined predictors, it exhibits the largest correlation in Charity and the second-largest correlation in Paper and Political. These results show that the pre-interaction social configuration contains predictive signal comparable in magnitude to commonly used message-quality measures. Because both g_i^0 and $\Delta s_{i,\text{H}}$ depend on the receiver's initial stance, we interpret this correlation descriptively and rely on the controlled input ablation below as more direct evidence of incremental predictive utility.

Table 3: Pearson correlations between quality scores, the initial neighbor-stance gap ($g_i^0 = \bar{s}^0_{\mathcal{N}_{\text{in}}(i)} - s_i^0$), and actual receiver-level stance uplift (Δs_i). **Bold**: best result; underline: second best. $^{**}p < 0.01$, $^{*}p < 0.05$.

Metric	Paper	Charity	Political
Attitude	0.075**	0.270**	0.118*
Clarity	0.164**	0.184**	0.148**
Constructiveness	0.297**	0.253**	0.244**
Persuasiveness	**0.383****	0.493**	**0.440****
Initial Neighbor-Stance Gap (g_i^0)	0.321**	**0.526****	0.268**

Incremental Utility for Stance Prediction. We next compare matched evaluator variants with and without visible-neighbor stances while holding the model and all other prompt inputs fixed. Prediction quality is measured using Pearson correlation and MSE between $\Delta s_{i,\text{RM}}$ and $\Delta s_{i,\text{H}}$.

For Persuade-RM, adding social context increases the overall Pearson correlation in every domain: from 0.12 to 0.20 on Paper, from 0.24 to 0.42 on Charity, and from 0.07 to 0.13 on Political. It also reduces MSE on Paper ($1.02 \rightarrow 1.00$), Charity ($1.44 \rightarrow 1.03$), and Political ($0.97 \rightarrow 0.93$). More broadly, across the seven model-domain configurations in Tables 4, 5, and 6, social context yields a higher Pearson and a lower MSE correlation.

The bucket-level results further localize this benefit. In the lower bucket, where the receiver's initial stance is below the visible-neighbor mean, social context consistently increases Pearson correlation by 0.06, 0.12, and 0.11, while reducing MSE by 0.06, 0.87, and 0.24 in Paper, Charity, and Political, respectively. By contrast, the higher-bucket changes are negligible or mixed. Thus, the benefit of social context is most consistent for receivers who initially hold a less favorable stance than their visible neighbors.

C.5 Persuade-RM Performance Evaluation

Having established the incremental value of social context, we evaluate the full, context-aware Persuade-RM against the available baselines. For all three domains, we compare the predicted stance uplift $\Delta s_{i,\mathrm{RM}}$ with the human-annotated uplift $\Delta s_{i,\mathrm{H}}$, where higher Pearson correlation and lower MSE indicate better prediction performance. Tables 4, 5, and 6 report both bucket-level and overall results.

Table 4: Bucket-level and overall stance-uplift prediction for the Charity domain. Higher Pearson and lower MSE are better. **Bold**: best; underline: second best.

Bucket	Metric	Persuade-RM		Qwen3-4B-Base	
		w.social	w.o.social	w.social	w.o.social
Lower	Pearson	**0.27**	0.15	0.07	-0.10
	MSE	**1.26**	2.13	1.50	1.52
Equal	Pearson	**0.28**	0.09	-0.08	-0.16
	MSE	0.97	1.22	1.02	**0.95**
Higher	Pearson	0.03	0.05	0.03	**0.05**
	MSE	0.79	**0.77**	1.02	1.01
Overall	Pearson	**0.42**	0.24	-0.05	-0.12
	MSE	**1.03**	1.44	1.20	1.19

Table 5: Bucket-level and overall stance-uplift prediction for the Political domain. Higher Pearson and lower MSE are better. **Bold**: best; underline: second best.

Bucket	Metric	Persuade-RM		Qwen3-4B-Base	
		w.social	w.o.social	w.social	w.o.social
Lower	Pearson	**0.02**	-0.09	-0.08	-0.16
	MSE	1.41	1.65	**1.20**	1.35
Equal	Pearson	**0.28**	0.20	0.12	0.07
	MSE	**0.61**	0.62	0.70	0.87
Higher	Pearson	**0.21**	0.20	0.09	0.18
	MSE	0.59	**0.58**	1.01	1.05
Overall	Pearson	**0.13**	0.07	0.03	-0.08
	MSE	**0.93**	0.97	1.06	1.15

Across all evaluated configurations, Persuade-RM with social context achieves the highest overall Pearson correlation in each domain: 0.20 on Paper, 0.42 on Charity, and 0.13 on Political. Under the matched social-

context condition, the corresponding correlations of Qwen3-4B-Base are 0.08, -0.05, and 0.03, respectively. Persuade-RM also yields lower MSE than Qwen3-4B-Base on Charity (1.03 vs. 1.20), Political (0.93 vs. 1.06) and Paper (1.00 vs. 1.01).

For the academic-rebuttal domain, where a task-specific evaluator is available, we additionally compare against Rebuttal-RM (He et al., 2026). As shown in Table 6, Persuade-RM with social context achieves an overall Pearson correlation of 0.20 and an MSE of 1.00, compared with 0.15 and 1.18, respectively, for Rebuttal-RM under the matched prompt condition. Even when selecting Rebuttal-RM's best value for each metric across its two prompt variants, Persuade-RM retains a modest Pearson advantage (0.20 vs. 0.19) and reduces MSE from 1.18 to 1.00. Overall, Persuade-RM provides the strongest correlation with human judgments across all three domains, while its MSE advantage is domain dependent.

Table 6: Bucket-level and overall stance-uplift prediction for the Paper domain. Higher Pearson and lower MSE are better. **Bold**: best; underline: second best.

Bucket	Metric	Persuade-RM		Qwen3-4B-Base		Rebuttal-RM	
		w.social	w.o.social	w.social	w.o.social	w.social	w.o.social
Lower	Pearson	**0.17**	0.11	0.08	0.02	0.08	0.10
	MSE	**1.46**	1.52	1.54	1.71	1.55	1.80
Equal	Pearson	-0.02	0.01	0.04	0.03	**0.04**	0.03
	MSE	**0.72**	0.78	0.82	0.79	1.00	1.34
Higher	Pearson	0.12	0.11	0.12	0.13	0.21	**0.22**
	MSE	**0.65**	**0.65**	0.70	0.70	0.90	0.90
Overall	Pearson	**0.20**	0.12	0.08	0.05	0.15	0.19
	MSE	**1.00**	1.02	1.01	1.06	1.18	1.35

D Experiment Details

The task representation, interaction process, and domain-specific environments are defined in Appendices B.1 and B.2. Here, we provide only the additional details needed to reproduce the compared methods. Evaluation metrics, aggregation, and training diagnostics follow Appendices C.1 and C.2. The interaction and evaluation prompt templates are reported separately in Appendix F.

D.1 Compared Methods and RL Implementation

Compared methods. We compare two post-training paradigms. *Reflection Prompting* applies a fixed strategy-reflection instruction to the unmodified instruction-tuned model before every action. It does not use environment feedback to revise a previous action, iterative reflection, external memory, or retries.

We compare two categories of baselines. *Prompting* includes zero-shot prompting and reflection prompting, where the latter elicits a strategy reflection before each action; *RL baselines* include GiGPO, HGPO, G2PO, and SALT. GiGPO (Feng et al., 2026) combines trajectory-level relative advantages with step-level advantages computed over repeated states, whereas HGPO (He et al.) constructs comparison groups from hierarchical interaction contexts. G2PO (Wang et al., 2026) forms comparison groups from a global state-transition graph. SALT (Li et al., 2026b) aggregates trajectory advantages at matching states while retaining the original trajectory advantage for singleton states. GCGPO replaces these estimators with the graph-contextual grouping mechanism introduced in Section 4. All remaining training components are held fixed within each backbone-domain setting.

Table 7: Shared RL implementation settings. Each update contains $M = 16$ task groups and $N = 4$ trajectories per group.

Item	Setting
Initialization	Final SFT checkpoint for the corresponding backbone and domain
Rollout grouping	16 task groups $\times$ 4 trajectories, totaling 64 trajectories per update
Interaction and length	At most 10 agent actions; sequence length 8192; a 512-token action budget; up to 1024 generated tokens for the joint action-self-assessment output
Sampling	Temperature 0.99; top-p 0.99; top-k 100
Optimization	Learning rate 5×10^{-7}; constant schedule; actor micro-batch size 1; gradient accumulation 16; warmup 10 optimization updates; one PPO epoch per update; BF16 precision
Clipping	Policy-ratio clip 0.1; advantage clip 0.2; reward clip 2.0; dual clipping enabled
Regularization	KL coefficient 0.01; entropy coefficient 0; maximum gradient norm 1
Advantage processing	Groupwise mean-standard-deviation normalization and advantage whitening
Final checkpoint	Iteration 250 for Paper; iteration 200 for Political and Charity
Compute and randomness	8 GPUs for Paper; 4 GPUs for Political and Charity; random seed 42

Table 8: Method-specific advantage-estimation settings.

Method	Setting
GiGPO	Episode/step weights 1/1; discount factor 0.95; separate normalization within trajectory and repeated-state groups
HGPO	Maximum context hierarchy 4; `hgpo_alpha`= 1; minimum comparison-group size 2
G2PO	Episode/step weights 1/1; discount factor 0.95; `g2po_w`= 1; `g2po_eps`= 10^{-6}
SALT	Aggregation within each `(traj_group_id, state_hash)` group; singleton states retain their trajectory advantage
GCGPO	Fixed context levels 0 and 1 with relative weights 1:2; no higher levels; minimum comparison-group size 2

Shared RL protocol. We use `Qwen3-4B-Instruct-2507` and `Llama-3.2-3B-Instruct` as policy backbones. For a fixed backbone and domain, all RL methods use the same final domain-specific SFT checkpoint, training split, environment, reward function, rollout protocol, optimizer, and training horizon. Thus, the advantage estimator and its necessary method-specific hyperparameters are the only intended differences among RL methods. Each RL update samples $M = 16$ task groups with $N = 4$ trajectories per group, with at most 10 agent actions per trajectory. For both backbones, we train SFT/RL for 400/250, 200/200, and 300/200 optimization updates on Paper, Political, and Charity, respectively. Table 7 summarizes the shared settings.

We always evaluate the final saved checkpoint and do not use validation-based early stopping or best-checkpoint selection. The GPU allocation is matched across RL methods within each backbone-domain setting; Paper experiments use 8 GPUs, while Political and Charity experiments use 4 GPUs. For each setting, we run the complete experiment twice using exactly the same configuration. The mean and standard deviation reported in the two full backbone result tables (Qwen3-4B-Instruct and Llama-3.2-3B-Instruct) are computed across these two runs.

Reward implementation. Appendix B.1 defines the coupled actor-reward interaction. For a persuasion action, the quality reward is the mean of Attitude, Clarity, Constructiveness, and Persuasiveness divided by 10. Its reward weight is 1, whereas the direct per-step stance component has weight 0. A malformed action receives a reward of -0.1. Once all receivers have been completed, we additionally assign a terminal reward equal to the mean current receiver stance divided by 10. A selection action outputs only its selected target, whereas a persuasion action outputs both its response and structured self-assessment during RL rollout.

Method-specific settings. Table 8 lists the only settings that differ across the RL advantage estimators.

D.2 SFT Data Collection and Fine-tuning

Collection and processing. We generate demonstrations for all three domains with `Qwen3-4B-Instruct-2507`. From the rollouts, we extract three types of single-step supervision: *select* records for target selection, *persuade-infer* records containing a persuasion response and its structured self-assessment, and *persuade-train* records containing the persuasion response alone. We convert them to instruction-input-output examples. The rollout configurations define collection caps rather than the final dataset sizes; Table 9 reports the actual post-processing counts.

Collection uses BF16 inference on two GPUs with seed 42 and rollout batches of 128 single trajectories. Each trajectory permits at most 10 agent actions, with a sequence length of 8192, a 512-token action budget, and up to 1024 generated tokens for the joint action-self-assessment output; the sampling settings match those in Table 7.

Processing performs parsing and format validation only, without filtering by task success or reward value. A select record must contain a nonempty <p> field, and a persuade-train record must contain a nonempty <r> field. A persuade-infer record must additionally contain a parseable reward object with Attitude, Clarity, Constructiveness, Persuasiveness, and OverallStance. All successfully parsed and validated records are retained and shuffled with seed 42. For Paper, 628 persuade-infer records are removed because their reward objects cannot be parsed; no post-categorization format failures are recorded for Political or Charity.

Table 9: Processed SFT examples. Counts are single-step supervision examples, not complete rollout trajectories.

Domain	Select	Persuade-infer	Persuade-train	Total
Paper	19,624	18,228	18,856	56,708
Political	9,808	9,394	9,394	28,596
Charity	15,590	15,186	15,186	45,962
Total	45,022	42,808	43,436	131,266

Fine-tuning. Both policy backbones are fine-tuned on the same domain-specific, Qwen-generated examples for one epoch on two GPUs, using seed 42 and sequence packing. We use a per-device batch size of 2, gradient accumulation of 32, a learning rate of 10^{-6} with a constant schedule, BF16 precision, and a maximum sequence length of 8192. We use the final saved checkpoints after 400, 200, and 300 optimization updates for Paper, Political, and Charity, respectively. These checkpoints serve as the initial policies for all corresponding RL runs; no validation-based early stopping or best-checkpoint selection is performed.

E Ablation Experiments

E.1 Ablation on the Backbone Model

Cross-Base-Model Stability. GCGPO maintains consistent superiority across both backbone models. When utilizing Llama-3.2-3B-Instruct, it achieves the highest performance in all 21 dataset-metric combinations, mirroring its performance with Qwen3-4B-Instruct. More importantly, GCGPO's Quality scores remain nearly identical across backbones: 8.25 compared to 8.53 on Paper, 8.40 compared to 8.51 on Political, and 8.55 compared to 8.61 on Charity, with a maximum gap of only 0.28 points. These results show that GCGPO maintains strong performance across the two tested backbones. SALT, the strongest alternative baseline,

Table 10: Evaluation results for LLM-based multi-receiver persuasion across the Paper, Political, and Charity datasets using Llama-3.2-3B-Instruct as the backbone model. Results are reported as mean ± standard deviation, with higher values indicating better performance for all metrics. **Bold**: best result; underline: second best.

Dataset	Baseline	Attitude	Clarity	Constructiveness	Persuasiveness	Quality	Stance	Success
Paper	Prompting	6.79±0.04	6.04±0.04	4.67±0.04	4.85±0.03	5.59±0.04	5.70±0.01	0.24±0.01
	Reflection	6.84±0.02	6.09±0.01	4.74±0.02	4.90±0.01	5.64±0.02	5.70±0.00	0.24±0.01
	G2PO	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	5.22±0.00	0.00±0.00
	HGPO	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	5.22±0.00	0.00±0.00
	GIGPO	2.17±0.02	2.00±0.02	1.80±0.02	1.60±0.02	1.89±0.02	5.59±0.00	0.00±0.00
	SALT	**8.98±0.00**	8.59±0.00	7.93±0.00	7.39±0.00	8.22±0.00	6.48±0.01	0.99±0.00
	GCGPO	**8.98±0.00**	**8.61±0.01**	**7.94±0.01**	**7.45±0.02**	**8.25±0.01**	**6.50±0.02**	**1.00±0.00**
Political	Prompting	7.82±0.07	7.15±0.05	6.03±0.03	5.36±0.03	6.59±0.04	5.04±0.00	0.67±0.03
	Reflection	7.22±0.02	6.61±0.01	5.63±0.02	4.96±0.02	6.10±0.02	5.03±0.00	0.49±0.00
	G2PO	8.75±0.03	8.74±0.03	8.39±0.03	7.44±0.02	8.33±0.03	5.52±0.01	0.93±0.01
	HGPO	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	4.98±0.00	0.00±0.00
	GIGPO	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	4.98±0.00	0.00±0.00
	SALT	7.69±0.12	6.83±0.11	5.15±0.07	4.69±0.06	6.09±0.09	4.99±0.00	0.66±0.03
	GCGPO	**8.95±0.02**	**8.82±0.02**	**8.40±0.02**	**7.45±0.02**	**8.40±0.02**	**5.55±0.01**	**0.98±0.01**
Charity	Prompting	8.25±0.03	7.66±0.02	6.57±0.02	6.09±0.01	7.14±0.02	5.89±0.00	0.74±0.01
	Reflection	7.07±0.05	6.57±0.03	5.68±0.02	5.21±0.03	6.13±0.03	5.87±0.01	0.36±0.03
	G2PO	4.96±0.19	4.93±0.19	4.74±0.19	4.32±0.17	4.74±0.19	6.23±0.02	0.18±0.01
	HGPO	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	5.59±0.00	0.00±0.00
	GIGPO	6.83±0.04	6.83±0.04	6.63±0.03	6.05±0.03	6.59±0.04	6.47±0.01	0.43±0.00
	SALT	8.77±0.04	8.77±0.04	8.56±0.04	7.79±0.04	8.47±0.04	6.61±0.01	0.91±0.01
	GCGPO	**8.85±0.03**	**8.85±0.03**	**8.64±0.01**	**7.87±0.02**	**8.55±0.02**	**6.62±0.01**	**0.94±0.01**

demonstrates reasonable cross-model transfer; however, its Political Quality score degrades more substantially (from 8.34 to 6.09). Prompting and Reflection also show lower Quality scores with Llama across all three domains, with the largest decreases on Paper (2.73 and 2.69 points, respectively).

Cross-Domain Stability. G2PO and HGPO experience a complete collapse in Quality and Success on the Paper dataset, which means that no persuasion rollout is successfully completed. Since quality scores are assigned only to successful actions, all quality dimensions are also zero. G2PO partially recovers on Political (Quality 8.33, Success 0.93) and Charity (Quality 4.74, Success 0.18), suggesting that its global state-transition graph can form usable comparison groups in some domains, but its performance is highly domain-dependent. HGPO's complete failure across all domains suggests that its hierarchical context grouping is more sensitive to the base model's instruction-following ability and trajectory coherence. With Llama, GiGPO records zero rollout Success on Paper and Political, while retaining partial functionality on Charity (Quality 6.59, Success 0.43). In contrast, GCGPO maintains high Success rates across domains, supporting its robustness in completing persuasion rollouts.

Figure 8 shows that GCGPO maintains stable actor-assessed message quality and improves mean post-intervention stance across all three domains. Several baselines exhibit sharp declines in one or both diagnostics. These trends complement the final evaluation results and support the robustness of GCGPO with the Llama backbone.

E.2 Ablation on Component Contributions

Ablation setup. To assess the contributions of episode-level optimization and graph-contextual credit assignment, we compare GCGPO with four baselines. *SFT* uses the supervised fine-tuned policy without

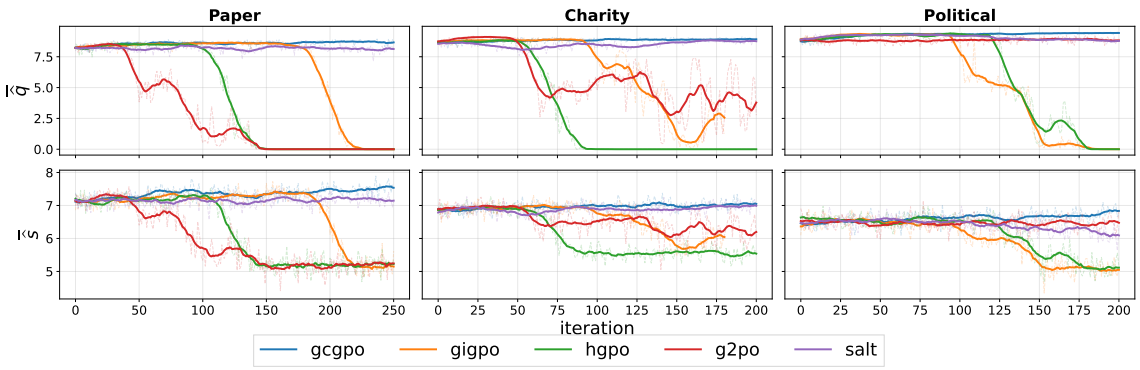

Figure 8: Training dynamics across three domains. At each training update, the upper and lower rows show batch means of actor-assessed message quality $\bar{\bar{q}}$ and post-intervention stance $\bar{\bar{s}}$, respectively. Solid lines show smoothed trends; dashed lines show unsmoothed batch means.

further reinforcement learning. *GCGPO (Episode-only)* retains only the episode-level relative advantage, removing graph-contextual step-level credit assignment. *GCGPO (Graph-only)* retains only the graph-contextual step advantage, removing the episode-level advantage. *GCGPO (Edge-free)* retains the receiver nodes but removes all graph edges, eliminating neighborhood information and testing the contribution of social dependencies. The full *GCGPO* model combines episode-level and graph-contextual advantages, incorporating both receiver states and neighborhood context. The full GCGPO model uses $\lambda_E = \lambda_G = 1$. Episode-only sets $\lambda_G = 0$, while Graph-only sets $\lambda_E = 0$.

We examine training dynamics across all three domains and report detailed evaluation results on Paper, which is grounded in real peer-review interactions from ICLR 2024–2025. Its observed reviewer ratings provide a human-derived basis for studying stance changes in a realistic persuasion setting. This setting allows us to examine component contributions using authentic interaction data, while retaining the model-based evaluation protocol described above for generated rollouts.

Training Stability. Figure 9 compares GCGPO with three ablated variants. The top panel tracks mean actor-assessed message quality $\bar{\bar{q}}$ and post-intervention stance $\bar{\bar{s}}$ across training iterations. GCGPO maintains stable quality across all three domains and improves post-intervention stance on Paper and Charity. Graph-only remains competitive and achieves higher stance scores on Charity, while the stance curves on Political remain close. Edge-free exhibits late declines in message quality across domains, with lower post-intervention stance on Paper and Charity. These trends support retaining neighborhood information for robust training, without implying that the full model leads on every diagnostic.

The bottom panel summarizes the diagnostics over training. GCGPO attains mean quality of 9.07 and post-intervention stance of 7.47, compared with 8.99/7.24 for Episode-only and 9.02/7.26 for Graph-only. Edge-free records lower means of 8.75 and 7.16. These averages complement the final evaluation in Table 11, while remaining distinct from the independent Persuade-RM scores.

Evaluation Performance with Persuade-RM. Table 11 confirms that GCGPO achieves the best scores across all seven evaluation metrics. Relative to *SFT*, the full model raises Quality by 5.5% (8.088 to 8.532), Persuasiveness by 8.0% (7.278 to 7.859), and Success from 0.935 to 0.999. Both *Episode-only* and *Graph-only* surpass SFT on every metric, demonstrating that either advantage signal alone provides useful supervision for policy optimization. Their combination yields additional gains: the full model attains Quality of 8.532 and Stance of 6.761, exceeding both variants (8.465/6.644 for Episode-only and 8.486/6.668 for Graph-only).

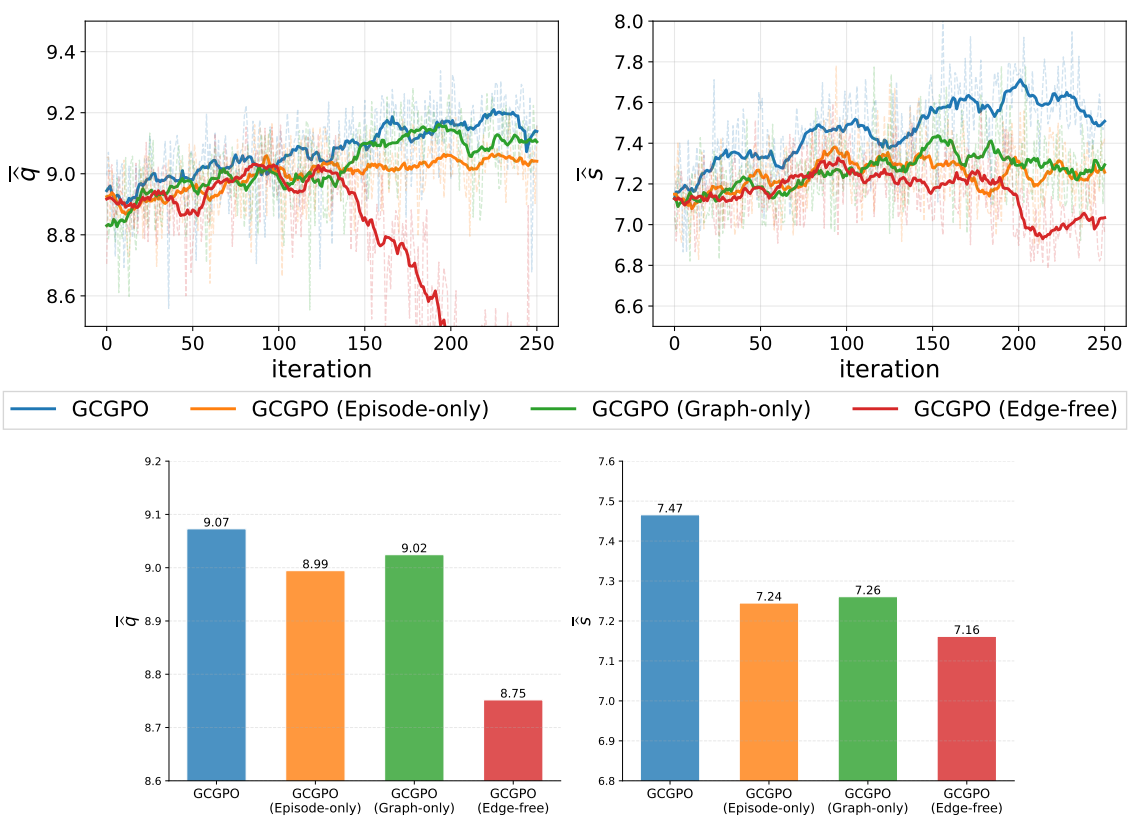

Figure 9: Training diagnostics for GCGPO and its ablation variants. Top: batch means of actor-assessed message quality $\overline{\hat{q}}$ and post-intervention stance $\overline{\hat{s}}$ for paper domain. Bottom: time-averaged diagnostic scores.

Table 11: Component ablation results on the Paper domain, comparing GCGPO with SFT and three ablated variants, all models are trained for 250 iterations. **Bold**: best result; underline: second best.

Variant	Attitude	Clarity	Constr.	Persuas.	Quality	Stance	Success
SFT	8.857	8.449	7.770	7.278	8.088	6.439	0.935
GCGPO (Episode-only)	8.972	8.800	8.383	7.704	8.465	6.644	0.987
GCGPO (Graph-only)	8.983	8.838	8.373	7.750	8.486	6.668	0.992
GCGPO (Edge-free)	7.357	7.133	6.696	6.112	6.824	6.411	0.305
GCGPO	**8.991**	**8.852**	**8.410**	**7.859**	**8.532**	**6.761**	**0.999**

These margins corroborate the complementary roles of episode-level and graph-contextual credit assignment. Removing graph edges causes the most severe degradation: *Edge-free* drops Quality to 6.824, Persuasiveness to 6.112, and Success to 0.305, underperforming SFT across all metrics. This degradation supports the importance of preserving social dependencies for both effective persuasion and reliable rollout completion. In all, these results collectively validate the design of GCGPO as a robust framework for multi-receiver persuasion.

E.3 Ablation on Reward Mechanisms

Table 12: Evaluation results for Coupled and Decoupled RM variants across the Paper, Political, and Charity datasets after 100 training iterations. Higher values indicate better performance for all metrics. **Bold**: best result; underline: second best.

Dataset	Baseline	Attitude	Clarity	Constructiveness	Persuasiveness	Quality	Stance	Success
Paper	Coupled	8.976	8.816	8.310	7.710	8.453	6.651	0.988
	Decoupled	**8.983**	**8.850**	**8.345**	**7.762**	**8.485**	**6.669**	**0.993**
Political	Coupled	**8.937**	**8.934**	**8.669**	**7.755**	**8.574**	**5.832**	**0.981**
	Decoupled	8.796	8.795	8.602	7.683	8.469	5.676	0.938
Charity	Coupled	8.912	8.917	8.704	7.926	8.619	6.642	0.955
	Decoupled	**8.923**	**8.924**	**8.758**	**7.987**	**8.648**	**6.666**	**0.970**

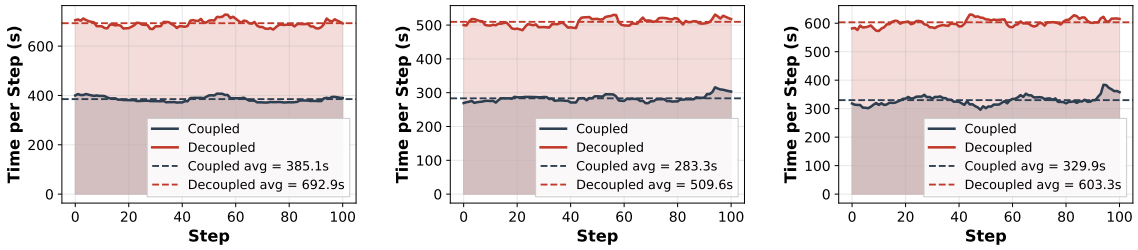

Figure 10: Training time per iteration for Coupled and Decoupled reward generation on Paper (left), Political (middle), and Charity (right).

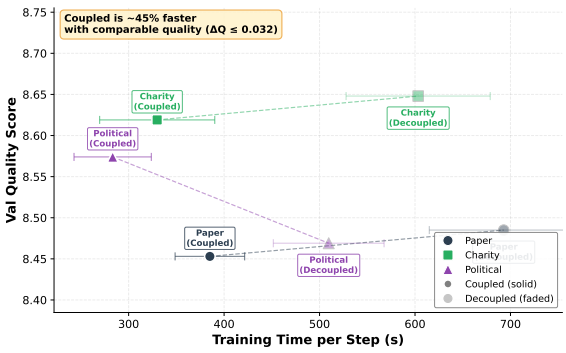

Figure 11: Pareto comparison of validation quality and training time for Coupled and Decoupled reward generation across the three domains.

Evaluation Performance Analysis. Table 12 shows comparable performance between Coupled and Decoupled after 100 training iterations. Decoupled scores slightly higher on all seven metrics on Paper and Charity. Its Quality advantages are only 0.032 and 0.029, respectively. Its Success advantages are also small, at 0.005 and 0.015. On Political, Coupled leads on all seven metrics. It improves Quality by 0.105, Stance by 0.156, and Success by 0.043. These results show that the additional reward-generation pass in Decoupled does not

consistently improve performance across domains. Figure 10 shows a clear efficiency advantage for Coupled. Average training iteration time decreases from 693 to 385 seconds on Paper, from 510 to 283 seconds on Political, and from 603 to 330 seconds on Charity. The average time reduction across domains is 44.7%. Figure 11 further illustrates the trade-off between quality and training time. Together, these results support Coupled as an efficient reward-generation strategy. It nearly halves training time for the same number of updates. Performance decreases slightly on Paper and Charity but improves on Political.

F Interaction and Evaluation Prompts

We report task-independent prompt skeletons shared across all three domains. Fields in square brackets are instantiated at run time; domain-specific roles, receiver profiles, related information, and stance rubrics follow the definitions in Appendix B.1.

F.1 Policy Interaction and Output Formats

At every interaction step, the policy receives a domain-specific system role and the following shared user-message wrapper:

```
## Environment Instructions:
[DOMAIN ROLE, PERSUASION OBJECTIVE, AND STANCE RUBRIC]

## Recent History:
You have completed [STEP COUNT] prior steps.
Your recent [HISTORY LENGTH] steps are:
[RECENT VALID ACTION TYPES AND DESCRIPTIONS]

## Current State:
You are at current step [CURRENT STEP].
[STATE-DEPENDENT OBSERVATION]

Max Answer length: [RESPONSE BUDGET] words (tokens).

Answer:
```

At a *selection* step, the current state lists the trajectory progress and, in randomized order, each remaining receiver's identifier, current stance, and profile. The required output is

```
<answer>
  <t>[2-3 SENTENCE SELECTION RATIONALE]</t>
  <p>[ONE VALID RECEIVER ID]</p>
</answer>
```

At a *persuasion* step, the state contains the selected receiver's identifier and profile, concern or problem, any available related information, current stance, and the current stances of receivers visible to that receiver in the directed social graph. The required output is

```
<answer>
  <t>[2-3 SENTENCE REBUTTAL RATIONALE]</t>
  <r>[POINT-BY-POINT PERSUASION RESPONSE]</r>
</answer>
```

During coupled RL rollout generation, a persuasion step additionally asks the same policy to assess its response. It scores Attitude, Clarity, Constructiveness, and Persuasiveness on a 0-10 scale and predicts the

receiver's new OverallStance from the domain-specific stance set. The prompt instructs the policy to use the current stance and confidence, the response's relevance to the receiver's concern, any useful related information, and social context; it also caps ordinary stance movement at two rating levels and prevents larger changes for high-confidence receivers. The combined output is

```
<answer>
  <t>[RATIONALE]</t>
  <r>[PERSUASION RESPONSE]</r>
</answer>
<reward>
{
  "Attitude": [0-10],
  "Clarity": [0-10],
  "Constructiveness": [0-10],
  "Persuasiveness": [0-10],
  "OverallStance": [ONE DOMAIN-SPECIFIC STANCE VALUE]
}
</reward>
```

The environment uses the reward block to score the interaction and update the simulated receiver stance. The optimization sample retains the ordinary interaction prompt and the `<answer>` block; the self-assessment instruction and reward tokens are excluded from the policy response used for optimization.

F.2 Reflection-Prompting Instruction

Reflection Prompting uses the same observations and output schemas, but inserts the following fixed, single-pass instruction immediately before the answer cue at every step:

```
## Strategy Reflection (think before answering):
Before generating your answer, carefully reflect on the following:
First determine whether you need to select a persuadee or write a
rebuttal, consider the persuadee's concerns, and interaction history
to choose the most effective strategy while avoiding ineffective or
alienating arguments used earlier.

After your reflection, provide your answer in the required format.
```

The model does not produce a separate reflection field and receives no feedback-driven retry or external reflection memory.

F.3 Persuade-RM Evaluation Prompt

At test time, the policy emits only the interaction answer. Following a valid persuasion response, an independent Persuade-RM judge receives a receiver-side domain role and the following evaluation skeleton:

The social-context fields follow the same domain-specific induced-graph construction used for Persuade-RM SFT; an empty visible-neighbor set is represented by an empty social context, with its peer mean treated as `Unknown`.

```
Your task is to give ratings for the rebuttal response.

Target receiver: [RECEIVER ID]
Receiver profile: [TEXTUAL PROFILE]
```

```
Receiver concern/problem: [FIRST RECEIVER MESSAGE]
Related information: [RELATED INFORMATION, IF PRESENT]

[IF VISIBLE PEERS ARE PRESENT:]
Others' current stances: [VISIBLE-PEER STANCES]
Average current stance: [VISIBLE-PEER MEAN]
[END OPTIONAL SOCIAL CONTEXT]
Target receiver's current stance: [CURRENT STANCE]

Rebuttal:
[POLICY RESPONSE]

Evaluate without overestimating:
- Attitude [0-10]: tone and professionalism;
- Clarity [0-10]: logic, focus, and structure;
- Constructiveness [0-10]: specific fixes and commitment;
- Persuasiveness [0-10]: argument strength, evidence, and
  coverage of the receiver's concerns;
- OverallStance [DOMAIN STANCE SET]: predicted updated stance,
  with peer average [VISIBLE-PEER MEAN OR UNKNOWN].

Use the current stance, receiver confidence, the response's relevance
to the receiver's concern, any useful related information, and the
social-context mean. If the mean is Unknown, do not infer peer influence.
Apply the same two-level and high-confidence movement constraints.
```

The requested evaluator output is

```
<reward>
{
  "Attitude": [0-10],
  "Clarity": [0-10],
  "Constructiveness": [0-10],
  "Persuasiveness": [0-10],
  "OverallStance": [ONE DOMAIN-SPECIFIC STANCE VALUE],
  "ScoreExplanation": [BRIEF JUSTIFICATION]
}
</reward>
```

The parser extracts the first valid balanced JSON object and accepts either the requested `<reward>` wrapper or a bare JSON response.

G FUTURE WORK

Our study is limited to small, fixed networks, a fixed horizon, and one persuasion intervention per receiver. Future work should examine larger and evolving networks, repeated interactions, adaptive stopping, and unseen graph topologies. Moreover, the Political and Charity training sets are partly LLM-simulated, online rewards rely on actor self-assessment, and final outcomes are measured by a frozen model-based evaluator. Larger-scale human interaction studies, causal tests of social influence, and externally calibrated reward signals would strengthen real-world validity and mitigate self-evaluation bias.