EvoCorps: An Evolutionary Multi-Agent Framework for Depolarizing Online Discourse

Anonymous ACL submission

Abstract

Polarization in online discourse erodes social trust and accelerates misinformation, yet technical responses remain largely diagnostic and post-hoc. Current governance approaches suffer from inherent latency and static policies, struggling to counter coordinated adversarial amplification that evolves in real-time. We present **EvoCorps**, an evolutionary multi-agent framework for proactive depolarization. EvoCorps frames discourse governance as a dynamic social game and coordinates roles for monitoring, planning, grounded generation, and multi-identity diffusion. A retrieval-augmented collective cognition core provides factual grounding and action–outcome memory, while closed-loop evolutionary learning adapts strategies as the environment and attackers change. We implement EvoCorps on the MOSAIC social-AI simulation platform for controlled evaluation in a multi-source news stream with adversarial injection and amplification. Across emotional polarization, viewpoint extremity, and argumentative rationality, EvoCorps improves discourse outcomes over an adversarial baseline under controlled simulation conditions, pointing to a promising direction from detection and post-hoc mitigation toward in-process, closed-loop intervention. The code is available at https://anonymous.4open.science/r/EvoCorps-DE08.

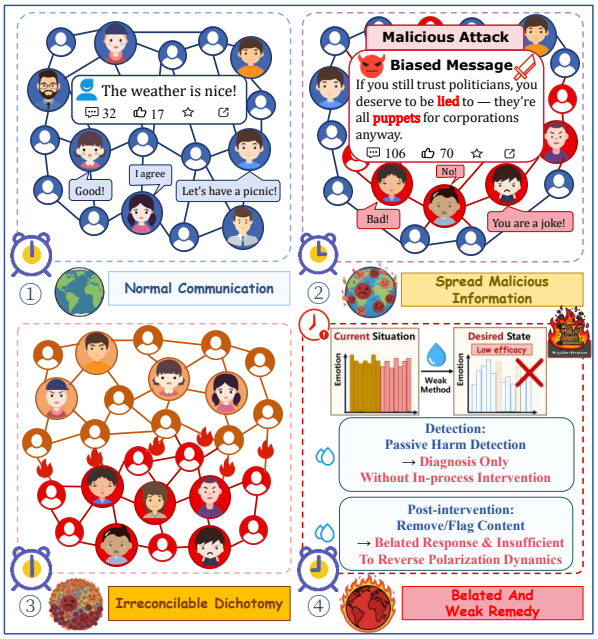

Figure 1: Online interaction evolves from normal user communication to malicious agents spreading harmful information, leading to amplified negative emotions and polarization. Existing harm detection and post-hoc intervention (e.g., content removal or labeling) occur only after diffusion and are insufficient to mitigate polarization.

1 Introduction

Online social media platforms have become a central infrastructure for public communication, shaping how opinions, emotions, and narratives are formed at scale. Through algorithmic curation and homophilous network interactions, these platforms influence information exposure and social alignment. A consequence of these dynamics is the emergence of polarization in online discourse. Here, polarization refers to the segregation of users, content, and interactions into opposing camps with limited cross-cutting exposure; emotionally charged and one-sided narratives can become salient within homophilous communities, reinforcing group identities while marginalizing moderate viewpoints. (Loru et al., 2025) This is a system-level outcome that can emerge from the interaction between homophily and engagement-optimized curation (Cinelli et al., 2021). Prior work has shown that recommender systems can intensify echo chambers under strong homophily (Cinus et al., 2021). Once entrenched, emotional reactions tend to dominate factual reasoning, rendering subsequent correction less effective.

Beyond these dynamics, contemporary online

discourse is no longer shaped solely by user interactions. Accumulating evidence indicates that malicious and highly coordinated agents have been actively deployed in real-world platforms to influence public opinion and steer discourse trajectories. Recent studies demonstrate that such adversarial collectives can strategically exploit emotional contagion, algorithmic reinforcement, and coordinated multi-identity behaviors to reshape discourse dynamics at an early stage, often before existing governance mechanisms can respond (Feng et al., 2024; Ma et al., 2024). Under these conditions, polarization is significantly accelerated through adversarial amplification, where coordinated agents inject and amplify emotionally provocative narratives across multiple identities, leveraging feedback loops in recommendation systems to rapidly diffuse content within homophilous communities and push discussions toward antagonistic camps (Figure 1). In such settings, emotional signals frequently outpace factual corrections, allowing early frames to lock in and reducing the leverage of later moderation or fact-checking.

In many platforms, the dominant technical response to harmful or polarizing discourse remains diagnostic: models can detect risks, but they do not by themselves execute timely, coordinated actions that change the trajectory of a live discussion. Building on detection, prior work has explored interventions such as fact-checking, content labeling, content removal or downranking, moderation workflows, and recommendation adjustments (Nakov et al., 2024; Vladika and Matthes, 2023; Habibi et al., 2024). However, these mechanisms are typically triggered only after diffusion has begun or polarization signals become salient, leaving governance structurally behind early adversarial amplification.

Under algorithmic feedback and homophilous interaction, emotionally charged narratives become path dependent: once amplified, they entrench echo chambers (Ferraz de Arruda et al., 2024), and later corrections or isolated moderation actions have limited leverage. Moreover, episodic or static policies lack closed-loop adaptation to evolving adversarial behavior. These gaps motivate in-process governance mechanisms that operate throughout the interaction, coordinate complementary roles, and adapt.

To this end, we propose EvoCorps, an evolutionary multi-agent framework for proactive discourse depolarization under adversarial amplification. Inspired by the Stackelberg–Mean-Field Control (SMFC) paradigm (Huang and Yang, 2020; Dayanikli and Laurière, 2023), EvoCorps models discourse governance as a dynamic social game in which an organized intervention team plans and executes strategies, while the broader user population responds at the population level. Operationally, EvoCorps instantiates this view as a closed-loop multi-agent system with an MMDP formulation and evolutionary learning that adapts strategies as discourse dynamics and adversarial tactics evolve.

Our main contributions are summarized as follows:

- We formulate proactive depolarization as an in-process, closed-loop intervention problem under adversarial amplification, targeting the latency and non-adaptive nature of prevalent post-hoc interventions.
- We propose **EvoCorps**, combining role-specialized coordination with retrieval-augmented collective cognition and closed-loop evolutionary learning to adapt intervention strategies over time.
- We implement EvoCorps on MOSAIC and evaluate it under adversarial injection and coordinated amplification in a multi-source news stream. EvoCorps improves emotional polarization, viewpoint extremity, and argumentative rationality over adversarial and post-hoc baselines.

2 Related Work

2.1 Agent-Based Malicious Information Diffusion

Recent studies have revealed that malicious or coordinated agents can do more than spread falsehoods; by amplifying emotionally charged narratives, they can mobilize audiences into antagonistic camps, erode trust in institutions, and normalize hostile or conspiratorial framings. Recent work shows that generative agents can propagate misinformation through coordinated liking and sharing (Liu et al., 2025), that AI–human collaboration can substantially extend the reach and persistence of sensational narratives (Ng et al., 2025), and that adaptive multi-agent coordination enables campaigns to evade moderation and sustain influence over time (Ren et al., 2025). Together, these findings suggest that adversarial collectives possess a systematic advantage in shaping discourse dynamics, as

they exploit emotional contagion, algorithmic homophily, and coordinated amplification to drive fast, adaptive, and highly organized processes that often outpace reactive governance mechanisms.

2.2 Research on Discourse Depolarization

Most existing approaches to mitigating polarization in online discourse emphasize diagnostic capabilities, including measuring and detecting toxicity, misinformation, stance extremity, or biased narratives (Nakov et al., 2024; Vladika and Matthes, 2023; Ma et al., 2024). Recent studies explore polarization detection using self-supervised objectives and interaction features (Cui and Abdelzaher, 2024), as well as large-scale analyses of polarized language and toxicity patterns in social media debates (Xu et al., 2025b; Arora et al., 2022). While effective for identifying harmful content and auditing polarization indicators, such methods primarily provide diagnosis and do not directly specify how to steer the unfolding interaction process.

Building on these diagnostic signals, a range of interventions have been explored, most commonly in a post-hoc manner. Representative actions include account suspension or content takedown, content downranking, warning labels or interstitials, and attaching fact-checks or clarifications (Habibi et al., 2024; Nakov et al., 2024). These mechanisms can reduce further spread and exposure, yet they are typically triggered only after polarization signals become salient and are often executed as isolated actions.

As a result, these approaches remain largely reactive, failing to establish sustained feedback and strategy evolution over time, and thus fall short of effectively modeling and intervening in discourse dynamics as they unfold through interaction.

3 Method

3.1 Problem Formulation and Environment Modeling

We consider an online discourse environment in which ordinary users interact over an underlying network and generate a time-indexed discourse stream over an intervention horizon T. At each round t, the environment reveals an observable discourse snapshot o_t, comprising the posts $\mathcal{P}_t$, comments $\mathcal{C}_t$, and their associated engagement signals observed at that time step. The intervention team then executes a joint action a_t following the closed-loop workflow in Figure 2.

Rather than modeling individual user states directly, EvoCorps uses a compact aggregate summary for high-level control, while each role acts on the relevant local context in o_t according to its function.

To formalize this process, we adopt a **multi-agent Markov decision process (MMDP)-style control abstraction** with explicit observations:

$$\mathcal{M} = (\mathcal{N}, \mathcal{O}, \mathcal{S}, \{\mathcal{A}^i\}_{i\in\mathcal{N}}, P, R, g), \qquad (1)$$

where $\mathcal{O}$ denotes the space of observable discourse snapshots and $g : \mathcal{O} \to \mathcal{S}$ maps each snapshot to the control state. This abstraction captures the perception-evaluation-decision-execution loop across heterogeneous roles.

Agent Set. We define the set of agents as $\mathcal{N} = \{\text{Analyst}, \text{Strategist}, \text{Leader}, \text{Amplifier}\}$ with role-specific action spaces $\mathcal{A}^i$. Their functional roles are described in Sec. 3.3.

State Space. We use the control state $s_t = (v_t, e_t)$, where v_t and e_t denote the average viewpoint extremity and sentiment computed from o_t. This state serves as a compact control summary rather than the full discourse representation.

Joint Action Space. We distinguish between the full agent set $\mathcal{N}$ and the intervention subset $\mathcal{N}_c = \{\text{Strategist}, \text{Leader}, \text{Amplifier}\}$ that directly executes interventions; the Analyst performs observation and evaluation but is not included in the intervention action. Concretely, a_t^{stg} specifies the target narrative, rhetorical style, and deployment scale; a_t^{ldr} specifies the final response and evidence subset; and a_t^{amp} specifies the selected personas, targets, and posting schedule. The joint intervention action is:

$$\begin{aligned} a_t &= (a_t^{\text{stg}}, a_t^{\text{ldr}}, a_t^{\text{amp}}) \\ &\in \mathcal{A}^{\text{stg}} \times \mathcal{A}^{\text{ldr}} \times \mathcal{A}^{\text{amp}} = \prod_{i\in\mathcal{N}_c} \mathcal{A}^i \end{aligned} \qquad (2)$$

Reward Function. The global cooperative reward is designed to reduce average viewpoint extremity while increasing average sentiment:

$$R(s_t, a_t) = -\lambda_1 \cdot \Delta v_t + \lambda_2 \cdot \Delta e_t, \qquad (3)$$

where $\Delta v_t = v_{t+1} - v_t$ and $\Delta e_t = e_{t+1} - e_t$ measure the change between the observed discourse metrics at consecutive time steps. Here, $\lambda_1, \lambda_2 > 0$ balance viewpoint moderation against sentiment

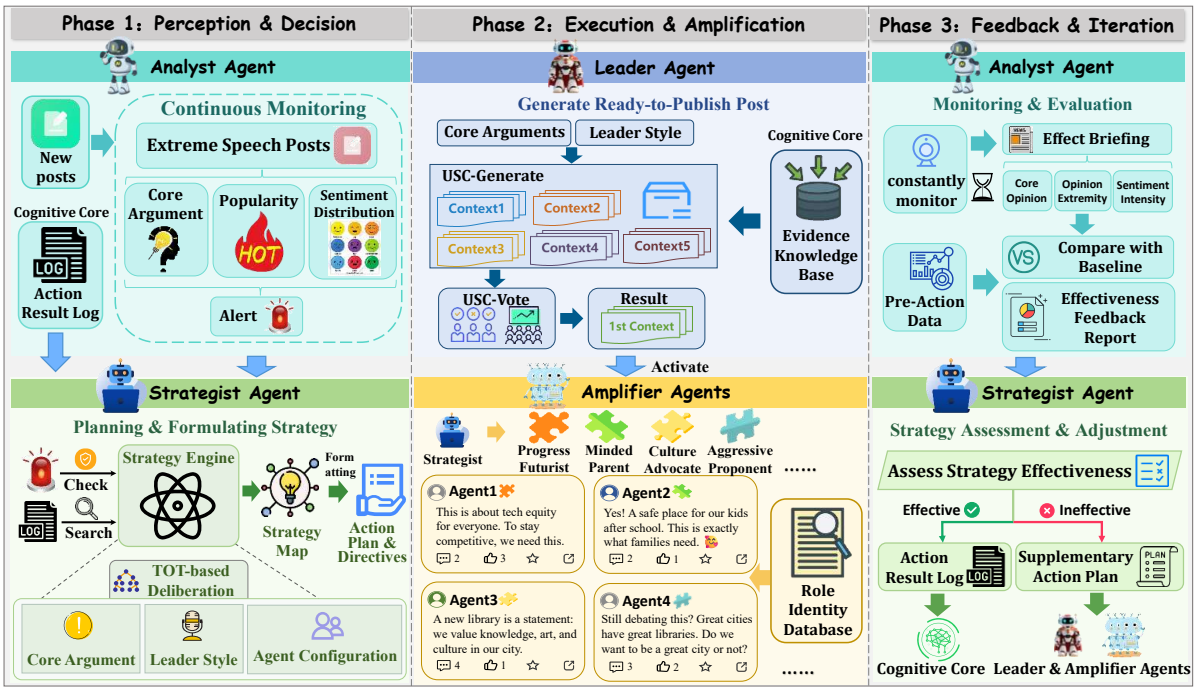

Figure 2: **Workflow of the closed-loop evolutionary system.** In the perception and decision phase, an analyst agent monitors incoming posts and evaluates opinion stance, extremity, popularity, and sentiment to trigger warnings. A strategist agent then formulates core arguments, leader styles, and amplification scales. In the execution and amplification phase, the leader agent, supported by the USC mechanism, retrieves evidence, generates candidate contents, and selects optimal outputs via voting, while amplifier agents coordinate role-based identities to expand and propagate the content. In the feedback and iteration phase, an analyst agent evaluates intervention effects against baselines and produces structured reports, which are used by the strategist agent to guide strategy adjustment and support subsequent decision making.

improvement. We use this reward for online adaptation because sentiment and extremity provide dense round-level signals; toxicity, AQS, fallacy, and evidence usage remain offline evaluation metrics in Sec. 4.2.

System Dynamics. Under the joint intervention, the discourse evolves according to

$$o_{t+1} \sim P(\cdot \mid o_t, a_t), \qquad s_{t+1} = g(o_{t+1}), \quad (4)$$

where P denotes the simulator-induced transition distribution over discourse snapshots, and g is the deterministic snapshot-to-state mapping defined above. Adversarial injections and other external perturbations are realized through the simulator dynamics.

3.2 EvoCorps Framework Overview

Figure 2 summarizes EvoCorps as a three-phase closed-loop workflow grounded in retrieval-augmented cognition and role-specialized coordination. In **Perception and Decision**, the Analyst monitors discourse signals and the Strategist formulates an intervention plan from current observations and historical memory. In **Execution and Amplification**, the Leader generates grounded responses and the Amplifiers disseminate them through diverse role identities. In **Feedback and Iteration**, the Analyst evaluates intervention outcomes and the Strategist updates subsequent decisions accordingly.

3.3 Dynamic Team with Role Coordination

EvoCorps models intervention as coordinated team behavior, reflecting governance settings in which specialized roles act jointly rather than in isolation.

EvoCorps defines four functional roles with complementary responsibilities:

Analyst. The Analyst monitors discourse and extracts core claims, viewpoint extremity, popularity, and sentiment polarity. It provides the controller summary and structured alerts used by downstream roles, and evaluates intervention outcomes to support subsequent rounds.

Strategist. The Strategist formulates intervention plans by combining current alerts with historical experience stored in the Action–Outcome

Memory. It determines target narratives, rhetorical style, and amplifier composition for coordinated response.

Leader. The Leader generates persuasive, evidence-grounded content aligned with the strategist's plan. It retrieves supporting arguments from the Evidence Knowledge Base, produces candidate drafts, and selects the most effective output through reflection and voting.

Amplifier. The Amplifiers disseminate leader outputs through diverse role identities such as ordinary users, topical experts, and community figures. By operating in parallel, they expand reach while maintaining coherence with the intervention strategy.

Within each round, decisions follow the sequence Analyst → Strategist → Leader → Amplifier, and their outputs jointly determine the intervention applied to the discourse state.

3.4 Retrieval-Augmented Coordination Core and Adaptive Updates

LLM-driven multi-agent systems often suffer from inconsistency and memory loss over long horizons. EvoCorps addresses this with a *retrieval-augmented coordination core* that couples an evidence store with action–outcome memory. The intervention process is formalized as a closed-loop system: at time step t, the joint action is produced by a prompt-driven coordination procedure conditioned on the current discourse context and the coordination core,

$$a_t = \Pi(o_t, s_t, K_t, M_t), \tag{5}$$

where o_t provides the current local discourse context, $s_t = (v_t, e_t)$ is the controller summary, K_t is the current evidence store, and M_t is the action–outcome memory. Operationally, Π unfolds as sequential prompting over Analyst → Strategist → Leader → Amplifier, with later roles conditioning on upstream outputs. The Strategist conditions on retrieved high-reward action–outcome tuples from M_t, while the Leader preferentially cites higher-scored evidence from K_t.

The adaptation objective is cumulative reward,

$$J = \mathbb{E}\left[\sum_{t=1}^{T} R(s_t, a_t)\right], \tag{6}$$

where $R(s_t, a_t)$ is defined in Eq. 3 and T is the intervention horizon. Unlike gradient-based RL, adaptation updates evidence scores and retained memories rather than model parameters.

Evidence Knowledge Base. The repository stores evidence items with persuasiveness scores, initialized from a fixed offline evidence pool (instantiation details are given in Sec. 4.1).

$$K_t = \{(k_i, p_i)\}_{i=1}^{n}, \quad p_i \in [0, 1], \tag{7}$$

where k_i is a candidate evidence item and p_i its persuasiveness score. Let q_t denote the query extracted from the disputed claim. A retrieved item (k', p') is incorporated if

$$\mathrm{rel}(k', q_t) = \cos(h_{k'}, h_{q_t}) > \delta, \tag{8}$$

where $h_{k'}$ and h_{q_t} are retrieval embeddings and $\delta \in (0, 1)$ is the acceptance threshold, yielding $K_{t+1} = K_t \cup \{(k', p')\}$.

After each round, cited evidence is reweighted by the global reward:

$$p_i \leftarrow p_i + \eta \cdot R(s_t, a_t) \cdot \mathbb{I}[k_i \in \mathcal{K}(a_t)], \tag{9}$$

where $\eta > 0$ is the learning rate and $\mathcal{K}(a_t) \subseteq K_t$ is the cited evidence subset.

Action–Outcome Memory. The cognition core records each action and its observed effect:

$$M_t = \{(a_i, o_{i+1}, r_i)\}_{i=1}^{t-1}, \tag{10}$$

where a_i is a past joint intervention, o_{i+1} the resulting snapshot, and $r_i = R(s_i, a_i)$ the associated reward. The log supports strategic evaluation through explicit action–outcome tuples.

After each round, the updated memory becomes

$$M_{t+1} = M_t \cup \{(a_t, o_{t+1}, r_t)\}. \tag{11}$$

Here $\mathrm{r}_t = R(s_t, a_t)$.

Adaptive Update Dynamics. Reward signals and memory updates form a closed-loop adaptive system: evidence scores are reweighted by Eq. 9, impactful trajectories are retained in M_t, and future actions follow the updated procedure $\Pi(o_t, s_t, K_t, M_t)$ rather than a parameterized policy $\pi_\theta(a_t \mid s_t)$.

This design treats depolarization as strategy-level adaptation, relying on evidence reweighting and memory-guided planning rather than LLM parameter updates, enabling long-horizon coordination without gradient-based training. Unlike RL-style policy evolution, "evolutionary learning" here refers to iterative refinement of intervention strategies across timesteps, not to updates of underlying model parameters.

4 Experiments

4.1 Experimental Setup

We run our experiments on top of the MOSAIC social-AI simulation platform (Liu et al., 2025), extending it with a multi-source news stream, adversarial injection and amplification, and a coordinated intervention team.

News stream and adversarial setting. We build a multi-source news stream from NELA-GT-2021 and the COVID-19 Fake News Dataset (Gruppi et al., 2023; Patwa et al., 2021). For a subset of real news items, we use an LLM to generate extremized variants as adversarial stimuli and inject the corresponding factual clarifications after a predetermined delay (4 time steps in our experiments). Malicious agents preferentially amplify adversarial sources to sustain polarization pressure.

Argumentation knowledge base. We build an Evidence Knowledge Base by integrating IBM Debater arguments (Mirkin et al., 2018) with relevant materials retrieved via the Wikipedia API.

Agents and evaluation configuration. All users and agents are instantiated from LLM-enriched Nemotron-Personas (Meyer and Corneil, 2025). We simulate 50 ordinary users for T=30 time steps under a fixed news stream; to increase heterogeneity, ordinary users sample between `gpt-4.1-nano` and `gemini-2.0-flash` (OpenAI, 2025; Cloud, 2025) (Appendix Fig. 7). For adversarial settings, we introduce malicious agents that preferentially target and amplify adversarial sources; we keep the news stream, population, horizon, and adversary fixed across Case 2/3/4 and vary only the defense mechanism. All reported values are mean±std over three independent simulation runs.

Experimental conditions. Based on this setup, we define four conditions (Case 1/2/3/4) for the main comparison:

- **Case 1 (Benign Baseline).** No adversary and no intervention, with discourse evolving organically from ordinary users.
- **Case 2 (Adversarial Propagation).** Malicious agents amplify adversarial sources, with no protection.
- **Case 3 (Post-hoc Moderation).** On top of Case 2, an LLM-based reviewer triggers fact-checking and takedown (Nakov et al., 2024; Vladika and Matthes, 2023; Habibi et al., 2024; Li et al., 2024) (full implementation details in Appendix B.7).
- **Case 4 (EvoCorps Intervention).** On top of Case 2, EvoCorps performs proactive, role-coordinated intervention.

For each condition, we select posts with more than 50 comments and compute all metrics on the *ordinary-user subset* to assess user-level outcomes.

4.2 Evaluation Metrics

To assess EvoCorps' depolarization capability, we evaluate emotional polarization, viewpoint extremity, and argumentative rationality. Following prior work, we use LLM-based evaluation (Liu et al., 2023; Lin and Chen, 2023; Li et al., 2024) and, under identical graders and settings across conditions, treat the resulting scores as comparative signals and report relative differences. Appendix A provides the evaluation protocol, prompt templates, and the justification and validation of LLM-based scoring. For sentiment, it also reports a sampled human comparison with strong aggregate agreement, supporting relative comparison (Appendix Table 3).

Emotional Polarization. We score ordinary-user utterances on a five-level sentiment scale (linearly mapped to $[0, 1]$) and compute toxicity via the Perspective API (Luong et al., 2024; Plaza-del Arco et al., 2024). Higher sentiment and lower toxicity indicate better emotional regulation.

Viewpoint Extremity. We label viewpoint extremity into five ordered levels (linearly mapped to $[0, 1]$) with deterministic decoding and report the mean over ordinary-user comments (Nakov et al., 2024; Zerback and Kobilke, 2022). Lower values indicate more moderate viewpoints.

Argumentative Rationality. We grade stance-bearing comments using fixed rubrics to obtain argument quality (AQS) (Wachsmuth et al., 2024), fallacy rate (Yeh et al., 2024), and evidence usage (Liu et al., 2024). Higher AQS/evidence and lower fallacy indicate stronger argumentative rationality.

4.3 Evaluation Results

We evaluate user-level outcomes under identical simulation and adversarial configurations, varying only the intervention mechanism across the four cases. Table 1 reports snapshot metrics at

t	Condition	Emotion		Viewpoint	Argument		
		Sentiment ↑	Toxicity ↓	Extremity ↓	AQS ↑	Fallacy ↓	Evidence ↑
1	Case 1	**57.0±4.1**	**3.3±0.7**	19.9±4.2	**38.2±2.6**	20.3±8.1	24.6±1.8
	Case 2	56.3±0.6	3.7±0.5	19.5±1.8	38.1±1.7	16.0±3.5	**25.5±1.1**
	Case 3	55.0±5.3	3.6±0.8	20.6±2.3	37.1±0.4	15.1±3.7	25.4±0.3
	Case 4	56.8±3.3	3.6±0.3	**18.9±0.6**	36.2±0.1	**10.2±4.9**	24.3±0.5
5	Case 1	41.2±2.5	**4.7±0.3**	34.6±3.2	38.7±0.3	35.1±7.7	25.2±1.1
	Case 2	42.1±3.5	6.1±0.7	31.1±2.9	41.2±0.3	20.1±4.6	26.8±0.6
	Case 3	45.7±2.5	5.1±0.7	28.7±2.2	**41.4±0.8**	16.6±3.4	27.1±0.7
	Case 4	**49.1±3.3**	5.2±0.5	**24.6±1.3**	40.9±1.5	**14.3±3.9**	**27.9±1.2**
10	Case 1	36.2±4.9	8.3±0.3	40.2±2.7	38.9±0.0	40.1±2.4	25.4±0.8
	Case 2	33.1±3.7	8.0±1.2	37.8±3.1	42.1±0.6	28.7±5.4	27.2±0.9
	Case 3	37.4±4.2	**6.7±1.0**	34.5±3.1	42.2±1.3	24.1±5.8	27.7±0.9
	Case 4	**42.5±3.1**	7.0±0.7	**29.5±2.5**	**43.8±0.8**	**20.6±4.4**	**30.1±0.5**
20	Case 1	36.2±4.6	7.8±0.1	41.8±2.6	41.9±0.3	37.1±2.1	27.2±0.9
	Case 2	27.1±1.7	9.2±0.6	43.1±2.2	42.3±0.1	36.0±2.0	27.0±0.4
	Case 3	31.1±3.1	8.3±1.0	40.0±2.3	42.4±1.2	30.9±5.0	27.7±1.0
	Case 4	**39.9±1.8**	**7.6±0.7**	**31.2±1.8**	**44.9±0.8**	**22.6±3.0**	**28.2±0.9**
30	Case 1	33.4±3.0	8.2±0.1	44.7±0.8	43.0±0.1	36.1±2.2	27.9±0.6
	Case 2	25.1±1.3	9.7±0.6	45.1±1.7	42.4±0.2	37.2±2.0	27.0±0.3
	Case 3	29.0±3.1	8.7±1.0	41.8±2.1	43.1±1.4	32.2±4.4	29.0±1.4
	Case 4	**39.2±2.0**	**7.9±0.5**	**31.1±2.1**	**45.4±0.6**	**21.7±2.8**	**31.3±0.5**

Table 1: **User-level effectiveness at selected time steps for four experimental conditions.** Case 1: benign baseline with no adversary and no intervention. Case 2: adversarial amplification with no protection. Case 3: post-hoc intervention baseline. Case 4: EvoCorps with proactive coordination. Metrics evaluate emotional polarization, viewpoint extremity, and argumentative rationality.

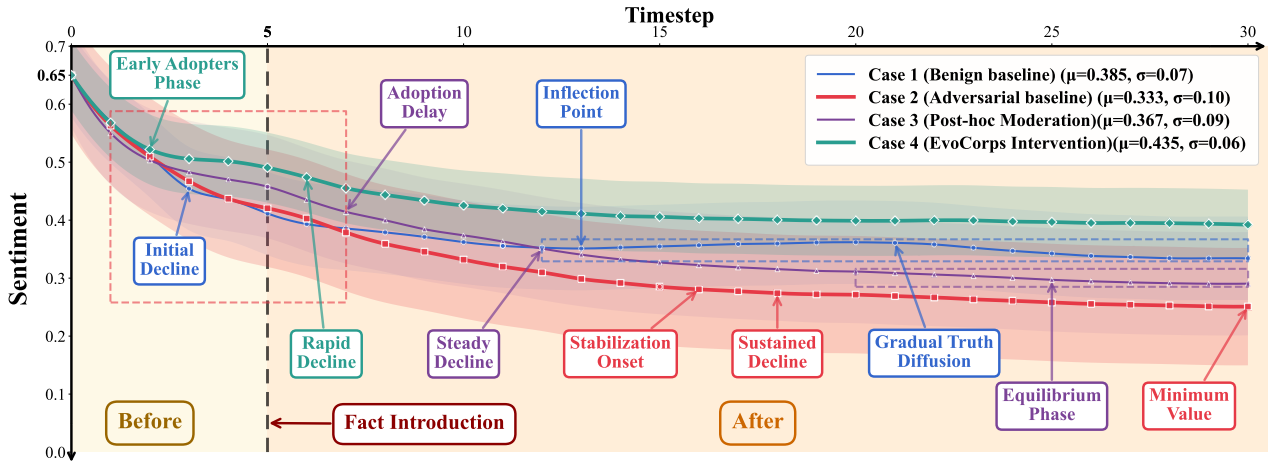

Figure 3: **Sentiment trajectories over time under Case 1/2/3/4.** The dashed line marks clarification at t=5. Case 2 continues to decline, Case 3 partially mitigates, and Case 4 declines more slowly and stabilizes relative to Case 2/3.

$t \in \{1, 5, 10, 20, 30\}$, and Figure 3 plots sentiment trajectories over time; the first clarification is introduced at t=5.

Trajectory change under adversarial amplification. Under adversarial amplification, Case 2 fails to recover after clarification: sentiment continues to deteriorate while extremity rises, indicating that delayed factual correction arrives too late once amplification takes hold. Table 1 and Figure 3 show that EvoCorps changes this trajectory most clearly: the divergence appears soon after t=5, and Case 4 maintains higher sentiment and lower extremity than Case 2/3 through t=30.

Baseline contrast and metric pattern. Using Case 2 as the primary adversarial baseline, Case 3 provides only limited recovery: toxicity is lower than Case 2 at t=30, but extremity remains high at 41.8, indicating that post-hoc moderation can

Model	Emotion		Viewpoint	Argument		
	Sentiment ↑	Toxicity ↓	Extremity ↓	AQS ↑	Fallacy ↓	Evidence ↑
Full Model	**39.2±2.0**	**7.9±0.5**	**31.1±2.1**	**45.4±0.6**	**21.7±2.8**	**31.3±0.5**
w/o Analyst	33.9±5.3	8.5±0.6	33.5±2.9	44.5±1.7	26.8±4.8	30.7±0.9
w/o Strategist	30.9±1.2	8.6±0.3	41.0±1.4	43.7±0.8	32.9±2.2	28.6±0.4
w/o Leader	28.4±2.0	8.4±0.4	41.9±2.2	42.4±0.8	34.9±4.8	27.7±0.3
w/o Amplifiers	26.9±4.9	9.5±1.2	43.2±4.6	41.8±2.1	39.1±7.7	27.1±1.7

Table 2: **EvoCorps role ablation study at $t = 30$.**

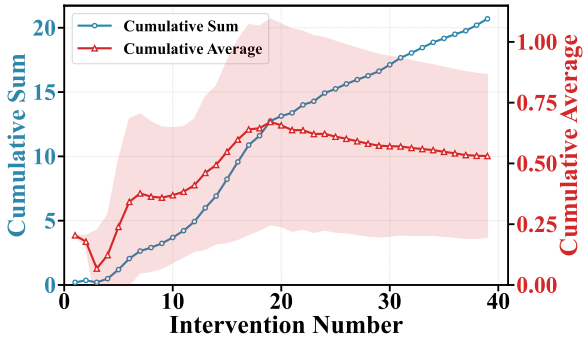

Figure 4: **Reward trajectories across intervention rounds.** Cumulative reward increases steadily, while the cumulative average rises early and then stabilizes.

soften tone without reversing entrenched viewpoints. By contrast, Case 4 yields the strongest sustained suppression of extremity while also improving evidence-grounded rationality, reaching 31.1 extremity, 31.3 evidence usage, and 21.7 fallacy at t=30. Relative to Case 3, the largest gains appear in sentiment, extremity, and argument quality rather than toxicity, indicating that EvoCorps primarily improves affective moderation, viewpoint de-escalation, and discourse rationality; Case 1 remains a no-adversary reference rather than an upper bound, since polarization can also emerge organically.

Reward dynamics. Figure 4 shows cumulative reward growing steadily, while cumulative average rises early and then levels off. Because the reward in Eq. 3 is computed from ordinary-user sentiment and viewpoint-extremity deltas, this pattern links the adaptation signal to process-level behavior and suggests that EvoCorps adapts online before reaching a more stable regime as discourse stabilizes.

4.4 Ablation Study

Counter-amplification is the dominant lever. Table 2 reports role ablations at t=30 for adversarial Case 4. Removing **Amplifiers** causes the largest regressions across all dimensions. Viewpoint extremity rises by 12.1 points (31.1 → 43.2), fallacy nearly doubles (21.7 → 39.1), and evidence usage drops (31.3 → 27.1). Counter-amplification is thus the main bottleneck, as grounded responses still need sufficient reach to shift ordinary-user trajectories under coordinated attack.

Planning and leadership translate diffusion into rational gains. Removing the **Strategist** or **Leader** sharply weakens moderation, raising extremity to 41.0/41.9 and fallacy to 32.9/34.9; without the Strategist, sentiment also falls to 30.9. Generation quality alone is therefore insufficient: strategic planning and coordinated execution must jointly convert diffusion into evidence-grounded influence rather than fixed-policy posting.

Analyst stabilizes timing and reduces volatility. Removing the **Analyst** causes moderate mean degradations but much higher variance. Sentiment falls to 33.9, toxicity rises to 8.5, extremity to 33.5, and argumentative quality weakens (AQS 45.4 → 44.5, fallacy 21.7 → 26.8, evidence 31.3 → 30.7). Most notably, sentiment standard deviation increases from 2.0 to 5.3, suggesting that the analyst mainly stabilizes intervention timing and reduces oscillatory over- or under-reaction.

5 Conclusion

This paper proposes **EvoCorps**, an evolutionary multi-agent framework that explores a shift in online discourse governance from detection and post-hoc mitigation toward in-process, closed-loop intervention under adversarial amplification. Under controlled simulation conditions, EvoCorps reshapes the post-shock trajectory and improves emotional polarization, viewpoint extremity, and argumentative rationality relative to both an adversarial and a post-hoc moderation baseline. Role ablations identify counter-amplification as a key bottleneck and show that grounded planning and generation are important contributors to robust gains.

6 Limitations

Our study is limited by the use of a text-only simulation built from curated datasets such as NELA-GT, IBM Debater, and Nemotron-Personas, which omits multimodal signals and platform-specific feedback mechanisms that can shape real-world polarization. We also do not model human responses to automated intervention, such as distrust or psychological reactance, so effectiveness in deployment may differ from the results observed here. In addition, EvoCorps relies on large language models for agent behavior and partial evaluation; although all conditions share the same prompts, models, and grading procedures, hallucinations and model biases may still affect absolute scores. Therefore, our findings should be interpreted primarily as controlled evidence of relative improvements rather than deployment-ready guarantees.

7 Ethics Statement

This work investigates mechanisms for online discourse depolarization primarily in a simulated environment, using publicly available datasets and synthetic agent interactions. In addition, we conduct a limited human annotation study to validate one evaluation component through aggregated, anonymized ratings from expert annotators. We do not collect, use, or report personally identifying information. The primary goal of this research is to advance understanding of coordinated intervention mechanisms for platform governance, rather than to develop or deploy deceptive influence campaigns.

EvoCorps is framed as a governance-assistance approach for online platforms facing coordinated and malicious activities such as disinformation campaigns or adversarial manipulation. In such settings, platform governance actors may themselves require coordinated capabilities and stylistic diversity to respond effectively and proportionately. Our study therefore examines coordination and response diversity as governance mechanisms, not as tools for artificial consensus formation or manipulation.

We explicitly oppose the use of deceptive strategies in any real-world deployment. Although our simulations introduce diverse agent personas to explore theoretical boundaries of influence dynamics, any practical application must adhere strictly to principles of transparency and accountability. Automated agents should be clearly identified as AI-based assistants or governance tools, such as certified fact-checking bots, and must not impersonate human users or conceal their artificial nature.

Any deployment of systems inspired by this work should be integrated with existing platform governance processes and subject to platform-specific policies, transparency requirements, and continuous auditing. Such safeguards are necessary to mitigate unintended harms, including disparate impacts, erosion of user trust, or errors arising from automated judgments. The intended use of EvoCorps is to support responsible, transparent, and accountable governance interventions, rather than to mislead users or manufacture false consensus.

To clarify the governance conditions for responsible deployment, we specify three requirements. First, **authorized operator**: EvoCorps should only be operated by platform governance actors with explicitly defined institutional responsibilities, not by third parties pursuing undisclosed interests. Second, **disclosure**: all automated interventions must be clearly labeled as AI-assisted governance content and must not impersonate ordinary users; the diverse role identities used by Amplifiers refer to communication style registers, not fake human accounts. Third, **audit and accountability**: deployment should be subject to continuous auditing with explicit attention to dual-use risks and protection of legitimate dissent; the reward function targets extremity reduction rather than viewpoint suppression, and any deployment must preserve mechanisms for user appeal and human oversight.

References

Swapan Deep Arora, Guninder Pal Singh, Anirban Chakraborty, and Moutusy Maity. 2022. Polarization and social media: A systematic review and research agenda. *Technological Forecasting and Social Change*, 183:121942.

Grace Byun, Swati Rajwal, and Jinho D. Choi. 2025. Llm-as-a-grader: Practical insights from large language model for short-answer and report evaluation. *ArXiv*, abs/2511.10819.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16051–16081, Vienna, Austria. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers), pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118.

Federico Cinus, Marco Minici, Corrado Monti, and Francesco Bonchi. 2021. The effect of people recommenders on echo chambers and polarization. *CoRR*, abs/2112.00626.

Google Cloud. 2025. Gemini 2.0 flash model. Vertex AI Generative AI Documentation, available at `https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash`.

Hang Cui and Tarek Abdelzaher. 2024. Polarization detection on social networks: dual contrastive objectives for self-supervision. In *Proceedings - 2024 IEEE 10th International Conference on Collaboration and Internet Computing, CIC 2024*, Proceedings - 2024 IEEE 10th International Conference on Collaboration and Internet Computing, CIC 2024, pages 80–89, United States. Institute of Electrical and Electronics Engineers Inc. Publisher Copyright: © 2024 IEEE.; 10th IEEE International Conference on Collaboration and Internet Computing, CIC 2024 ; Conference date: 28-10-2024 Through 30-10-2024.

Gokce Dayanikli and Mathieu Laurière. 2023. A machine learning method for stackelberg mean field games. *Math. Oper. Res.*, 50:3055–3093.

Shangbin Feng, Herun Wan, Ningnan Wang, Zhaoxuan Tan, Minnan Luo, and Yulia Tsvetkov. 2024. What does the bot say? opportunities and risks of large language models in social media bot detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3580–3601, Bangkok, Thailand. Association for Computational Linguistics.

Henrique Ferraz de Arruda, Kleber Andrade Oliveira, and Yamir Moreno. 2024. Echo chamber formation sharpened by priority users. *iScience*, 27(11):111098.

Maurício Gruppi, Benjamin D. Horne, and Sibel Adalı. 2023. Nela-gt-2022: A large multi-labelled news dataset for the study of misinformation in news articles. *Preprint*, arXiv:2203.05659.

Mahyar Habibi, Dirk Hovy, and Carlo Schwarz. 2024. The content moderator's dilemma: Removal of toxic content and distortions to online discourse. *ArXiv*, abs/2412.16114.

Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. 2025. An empirical study of LLM-as-a-judge for LLM evaluation: Fine-tuned judge model is not a general substitute for GPT-4. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5880–5895, Vienna, Austria. Association for Computational Linguistics.

Minyi Huang and Xuwei Yang. 2020. Mean field stackelberg games: State feedback equilibrium. *IFAC-PapersOnLine*, 53(2):2237–2242. 21st IFAC World Congress.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *ArXiv*, abs/2412.05579.

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58, Toronto, Canada. Association for Computational Linguistics.

Genglin Liu, Vivian T. Le, Salman Rahman, Elisa Kreiss, Marzyeh Ghassemi, and Saadia Gabriel. 2025. MOSAIC: Modeling social AI for content dissemination and regulation in multi-agent simulations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6390–6417, Suzhou, China. Association for Computational Linguistics.

Xinyi Liu, Pinxin Liu, and Hangfeng He. 2024. An empirical analysis on large language models in debate evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 470–487, Bangkok, Thailand. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Edoardo Loru, Alessandro Galeazzi, Anita Bonetti, Emanuele Sangiorgio, Niccolò Di Marco, Matteo Cinelli, Max Falkenberg, Andrea Baronchelli, and Walter Quattrociocchi. 2025. Ideology and polarization set the agenda on social media. *Scientific Reports*, 15(1):35816.

Tinh Luong, Thanh-Thien Le, Linh Ngo, and Thien Nguyen. 2024. Realistic evaluation of toxicity in

large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1038–1047, Bangkok, Thailand. Association for Computational Linguistics.

Zihan Ma, Minnan Luo, Hao Guo, Zhi Zeng, Yiran Hao, and Xiang Zhao. 2024. Event-radar: Event-driven multi-view learning for multimodal fake news detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5821, Bangkok, Thailand. Association for Computational Linguistics.

Yev Meyer and Dane Corneil. 2025. Nemotron-Personas-USA: Synthetic personas aligned to real-world distributions.

Shachar Mirkin, Michal Jacovi, Tamar Lavee, Hong-Kwang Kuo, Samuel Thomas, Leslie Sager, Lili Kotlerman, Elad Venezian, and Noam Slonim. 2018. A recorded debating dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Preslav Nakov, Jisun An, Haewoon Kwak, Muhammad Arslan Manzoor, Zain Muhammad Mujahid, and Husrev Taha Sencar. 2024. A survey on predicting the factuality and the bias of news media. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15947–15962, Bangkok, Thailand. Association for Computational Linguistics.

Lynnette Hui Xian Ng, Wenqi Zhou, and Kathleen M. Carley. 2025. Appeal and scope of misinformation spread by ai agents and humans. *Preprint*, arXiv:2505.04028.

OpenAI. 2025. Gpt-4.1 nano model. OpenAI API Documentation, available at `https://platform.openai.com/docs/models/gpt-4.1-nano`.

Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2021. *Fighting an Infodemic: COVID-19 Fake News Dataset*, page 21–29. Springer International Publishing.

Flor Miriam Plaza-del Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710, Torino, Italia. ELRA and ICCL.

Qibing Ren, Sitao Xie, Longxuan Wei, Zhen fei Yin, Junchi Yan, Lizhuang Ma, and Jing Shao. 2025. When autonomy goes rogue: Preparing for risks of multi-agent collusion in social systems. *ArXiv*, abs/2507.14660.

Juraj Vladika and Florian Matthes. 2023. Scientific fact-checking: A survey of resources and approaches. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6215–6230, Toronto, Canada. Association for Computational Linguistics.

Henning Wachsmuth, Gabriella Lapesa, Elena Cabrio, Anne Lauscher, Joonsuk Park, Eva Maria Vecchi, Serena Villata, and Timon Ziegenbein. 2024. Argument quality assessment in the age of instruction-following large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1519–1538, Torino, Italia. ELRA and ICCL.

Tianruo Rose Xu, Vedant Gaur, Liu Leqi, and Tanya Goyal. 2025a. The progress illusion: Revisiting meta-evaluation standards of LLM evaluators. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19033–19043, Suzhou, China. Association for Computational Linguistics.

Wentao Xu, Wenlu Fan, Shiqian Lu, Tenghao Li, and Bin Wang. 2025b. Polarized patterns of language toxicity and sentiment of debunking posts on social media. *Preprint*, arXiv:2501.06274.

Min-Hsuan Yeh, Ruyuan Wan, and Ting-Hao Kenneth Huang. 2024. CoCoLoFa: A dataset of news comments with common logical fallacies written by LLM-assisted crowds. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 660–677, Miami, Florida, USA. Association for Computational Linguistics.

Jiachen Yu, Shaoning Sun, Xiaohui Hu, Jiaxu Yan, Kaidong Yu, and Xuelong Li. 2025. Improve LLM-as-a-judge ability as a general ability. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14099–14115, Suzhou, China. Association for Computational Linguistics.

Thomas Zerback and Lara Kobilke. 2022. The role of affective and cognitive attitude extremity in perceived viewpoint diversity exposure. *New Media & Society*, 26:4183 – 4200.

A Evaluation Metric Details

A.1 Evaluation Protocol and Aggregation

Unless otherwise stated, we compute all content-based metrics on the *ordinary-user subset* (records with agent_type = `normal`) from data/scenario_*/post-*/comments.jsonl. Following the main setup, we restrict evaluation to posts with more than 50 comments and report snapshot values at selected time steps $t \in \{1, 5, 10, 20, 30\}$. For each snapshot t, we aggregate over all eligible ordinary-user comments with `time_step`$\leq t$,

where `time_step` is the 1-indexed simulation step stored in the dataset.

LLM-based metrics use fixed prompts and deterministic decoding (e.g., temperature $= 0$ where applicable), and we keep the grader model/API configuration identical across experimental conditions. We skip empty texts; invalid grader outputs are treated as missing for aggregation.

A.2 Emotional Polarization

Sentiment. We label each ordinary-user comment with `gpt-4o-mini` into five classes {*Very Negative, Negative, Neutral, Positive, Very Positive*} and map them to scores $\{0, 0.25, 0.5, 0.75, 1\}$. We report sentiment as $100 \cdot \frac{1}{|C|} \sum_{c \in C} s_{\text{sent}}(c)$.

Sentiment Grader Prompt

System: You are a helpful sentiment classifier.
User:
You are a sentiment analysis model.
Classify the following text into exactly one of the five categories: ["Very Negative", "Negative", "Neutral", "Positive", "Very Positive"]
Text: <TEXT>
Return only the label.

Toxicity. We score toxicity using the Perspective API, obtaining continuous toxicity probabilities in $[0, 1]$ per comment. We report toxicity as $100 \cdot \frac{1}{|C|} \sum_{c \in C} s_{\text{tox}}(c)$ (lower is better), with optional auxiliary statistics such as the percentage above a high-toxicity threshold.

A.3 Viewpoint Extremity

Viewpoint Extremity. We quantify viewpoint extremity by labeling each post and comment into one of five classes: {*Very Moderate*, *Moderate*, *Neutral*, *Extreme*, *Very Extreme*} using `gpt-4o-mini` with deterministic decoding, `temperature`$= 0$. We map labels to scalar scores in $[0, 1]$ via $\{0, 0.25, 0.5, 0.75, 1\}$, respectively. For a snapshot with eligible ordinary-user comments C, we report the mean extremity, scaled to $[0, 100]$:

$$\text{ViewpointExtremity} = 100 \cdot \frac{1}{|C|} \sum_{c \in C} s_{\text{ve}}(c),$$
$$s_{\text{ve}}(c) \in \{0, 0.25, 0.5, 0.75, 1\}.$$

To characterize dispersion, we compute the score variance $\text{Var}_{c \in C}[s_{\text{ve}}(c)]$. We additionally compute the label distribution p_k over the five classes and its Shannon entropy $H = -\sum_k p_k \log p_k$ to capture whether viewpoints are concentrated at extreme levels or spread across moderate and neutral levels. As an auxiliary consistency check, we compare the LLM label with the precomputed `predicted_label` stored in the data. Treating the five labels as ordered levels, we count an agreement if the two labels differ by at most one adjacent level.

Viewpoint Extremity Grader Prompt

System: You are a helpful extremeness classifier that analyzes how moderate or extreme viewpoints are.
User:
You are an extremeness analysis model.
Classify the following text into exactly one of the five categories based on how extreme or moderate the viewpoint/opinion expressed is:
- Very Moderate: Very balanced, neutral, or mild viewpoint
- Moderate: Somewhat balanced or mild viewpoint
- Neutral: Neither moderate nor extreme
- Extreme: Strong, polarized, or one-sided viewpoint
- Very Extreme: Highly polarized, radical, or absolutist viewpoint
Text: <TEXT>
Return only the label.

A.4 Argumentative Rationality

AQS. We compute Argument Quality Score (AQS) by prompting `gpt-4o-mini` to assign a continuous score in $[0, 1]$ for each ordinary-user comment under a fixed rubric and few-shot examples. We report AQS as $100 \cdot \frac{1}{|C|} \sum_{c \in C} s_{\text{aqs}}(c)$.

AQS Grader Prompt

System:
You are an expert academic evaluator specializing in argumentation.
Score the quality of a given argument on a continuous scale from 0.0 (very weak) to 1.0 (very strong).
Provide a fine-grained score (avoid round numbers).

Respond in JSON with only: {"reasoning": "...", "score": <float in [0,1]>}
Topic: <TOPIC>
Argument: <ARGUMENT>

Fallacy Rate. We detect logical fallacies with `gpt-4o-mini` using a fixed label set (e.g., *Ad Hominem*, *False Cause*, *Appeal to Emotion*), returning `fallacious` $\in \{$`Yes`, `No`$\}$ and a `fallacy_type` in JSON format (implemented in `api_logic.py`). We report fallacy rate as the percentage of analyzed ordinary-user comments predicted as fallacious.

Fallacy Grader Prompt

System:
You are an expert assistant specialized in the systematic identification and analysis of logical fallacies.
Respond with a valid JSON array of objects, with no additional text.
User:
For each sentence, return a JSON object with:
1) "sentence": the original sentence
2) "fallacious": "Yes" or "No"
3) "fallacy_type": "" if "No", else one of:
- Ad Hominem
- Ad Populum
- False Dilemma / Black-and-White Fallacy
- False Cause
- Circular Reasoning
- Deductive Fallacy / Fallacy of Logic
- Appeal to Emotion / Emotional Language
- Equivocation
- Fallacy of Extension / Extension Fallacy
- Faulty Generalization / Hasty Generalization
- Intentional Fallacy
- Fallacy of Credibility / Irrelevant Authority
- Fallacy of Relevance / Red Herring
Sentences to analyze:
<LIST_OF_SENTENCES>

Evidence Usage. We operationalize *evidence usage* as the fraction of stance-bearing statements that provide at least one verifiable attribution, reported as a percentage. We count an attribution as *verifiable* if it includes at least one of the following cues: (i) an explicit URL; (ii) a named report, dataset, organization, or outlet that can be checked; or (iii) a concrete, checkable pointer (e.g., a bill number, court case identifier, or a quoted statement with a named speaker and venue). We do *not* count vague appeals (e.g., "studies say", "experts agree") without an identifiable source. In our evaluation, stance-bearing statements are instantiated as eligible ordinary-user comments in the selected posts.

To compute this metric, we apply a fixed grading rubric with GPT-4o-mini under identical prompts and deterministic decoding across all experimental conditions. Each statement c is labeled as evidence-present $y(c)=1$ or evidence-absent $y(c)=0$. Let C denote the set of eligible ordinary-user comments for the snapshot; we report the mean label value, scaled to $[0, 100]$:

$$\text{EvidenceUsage} = 100 \cdot \frac{1}{|C|} \sum_{c \in C} y(c), \quad y(c) \in \{0, 1\}.$$

Evidence Usage Grader Prompt

System: You are a careful evaluator of evidence attribution in natural language.
User:
Decide whether the statement provides at least one verifiable attribution.
Count as verifiable if it includes any of:
(i) an explicit URL
(ii) a named report, dataset, organization, or outlet that can be checked
(iii) a concrete, checkable pointer (bill number, court case identifier, quoted statement with named speaker and venue)
Do NOT count vague appeals (e.g., "studies say", "experts agree") without an identifiable source.
Statement: <STATEMENT>
Return only valid JSON:
"evidence_present": 0 or 1

A.5 Validation of LLM-based Scoring

We validate the use of LLM-based scoring from three perspectives. First, several target metrics in our study are better framed as rubric-based scores over open-ended social comments than as standard lexicon-matching or closed-label classification problems. Such comments are typically short, colloquial, and semantically compressed, often involving implicit stance, sarcasm, and semantic reversal. Reliable scoring therefore depends more

on holistic semantic judgment than on surface lexical cues alone, making lexicon-based tools and fixed small classifiers less suitable for stable, fine-grained scoring under a unified rubric.

Second, prior work has already adopted LLM-based evaluation for open-ended generation and dialogue, including G-Eval, LLM-Eval, and studies examining whether LLMs can serve as alternatives to human evaluation, helping establish LLM-as-a-judge as a major evaluation paradigm for open-ended NLP assessment (Liu et al., 2023; Lin and Chen, 2023; Chiang and Lee, 2023; Byun et al., 2025; Li et al., 2024). More recent work has further examined judge quality, alignment, replacement criteria, robustness, and generalization (Li et al., 2025; Calderon et al., 2025; Yu et al., 2025; Liu et al., 2024; Huang et al., 2025; Xu et al., 2025a). Our setup in Sec. 4.2 follows this line of work by using identical graders and settings across conditions.

Third, we use LLM-based scoring not only because it is common practice, but also because we provide task-specific validation against human judgment. We use sentiment intensity as a representative case in Table 3, where the LLM and human ratings are already close at a descriptive level. We therefore use a fixed rubric and interpret the resulting scores primarily as comparative signals for relative, rather than absolute, differences across conditions.

Using sentiment intensity as a representative example, we conducted a human evaluation study with 15 expert annotators. For each case–time-step pair in Table 3, we sampled 20 ordinary-user comments and asked three different expert annotators to rate their emotional intensity using the same five-level sentiment scale (Very Negative, Negative, Neutral, Positive, Very Positive). The labels were mapped to numerical scores $\{0, 0.25, 0.5, 0.75, 1\}$, consistent with the LLM scoring scheme. We averaged the three expert scores for each comment and then computed the mean and standard deviation over the sampled comments; the same comment set was scored by the LLM for comparison. Across the 20 case–time-step cells in Table 3, the aggregate agreement between LLM and human means is strong. This representative validation supports the practical reliability of our LLM-based scoring pipeline and shows that, for sentiment intensity, the evaluator is well aligned with human judgment on this task. Additional details on annotator instructions, recruitment, and data handling are provided in Appendix A.5.1. For transparency, the anonymized repository accompanying this submission also provides the aggregated human and LLM evaluation results used in this comparison.

t	Case	LLM	Human Evaluation
1	Case 1	57.5 ± 13.9	54.2 ± 26.5
	Case 2	51.3 ± 24.1	45.4 ± 27.4
	Case 3	43.8 ± 27.2	42.9 ± 26.9
	Case 4	50.0 ± 29.6	50.8 ± 22.5
5	Case 1	36.3 ± 23.0	45.4 ± 24.1
	Case 2	38.8 ± 30.1	39.6 ± 28.1
	Case 3	47.5 ± 26.1	42.1 ± 25.4
	Case 4	37.5 ± 29.0	38.3 ± 26.2
10	Case 1	46.3 ± 25.3	44.6 ± 24.0
	Case 2	27.5 ± 30.2	30.0 ± 27.9
	Case 3	26.3 ± 21.6	31.2 ± 24.2
	Case 4	26.3 ± 27.9	25.0 ± 23.0
20	Case 1	51.3 ± 25.6	45.8 ± 26.5
	Case 2	35.0 ± 27.8	33.8 ± 27.2
	Case 3	31.3 ± 23.6	31.7 ± 18.9
	Case 4	35.0 ± 27.8	35.4 ± 22.7
30	Case 1	32.5 ± 22.5	42.1 ± 26.6
	Case 2	22.5 ± 28.4	32.9 ± 28.6
	Case 3	27.5 ± 22.2	34.6 ± 26.9
	Case 4	17.5 ± 22.5	22.5 ± 7.6

Table 3: Emotion Intensity: LLM vs. Human Evaluation

A.5.1 Human Annotation Protocol

Instructions to annotators. We provided annotators with the following task instructions:

> **Instructions Given to Annotators**
>
> You will read sampled ordinary-user comments and rate the emotional intensity expressed in each comment.
> Choose exactly one label from the following five-level scale: Very Negative, Negative, Neutral, Positive, or Very Positive.
> Base your judgment on the overall emotional intensity conveyed by the comment text itself.
> Use the same rubric consistently across all comments.
> Work independently and do not discuss labels with other annotators.

Recruitment and compensation. The human validation involved 15 annotators with technical backgrounds in computer science, recruited through laboratory channels. Each case–time-step cell was rated by three different annotators. Annotators participated as unpaid volunteers. Given the limited scope of this validation study and the technical background of the recruited participants, we treated this as a voluntary expert annotation exercise rather than paid crowd work.

Consent and data use. Before annotation, participants were informed that their ratings would be used only for research validation of the sentiment-intensity metric and reported in aggregated, anonymized form. No personal identities were used in the analysis or reported in the paper.

Data handling. The collected ratings were used only for research validation of the sentiment-intensity metric and are reported in aggregated form in Table 3. The paper does not use or report personal identities of annotators, and no additional sensitive personal information from annotators is analyzed in this study.

B Mechanisms and Corresponding Prompt Templates

This appendix presents the prompt templates associated with each mechanism used in EvoCorps, as well as related agent roles commonly employed in multi-agent opinion modeling and intervention settings. For clarity and reproducibility, the prompts are organized according to the agent roles and functional mechanisms described in the main text, illustrating how abstract mechanisms are instantiated through concrete prompt specifications.

All prompt templates are shown verbatim as implemented. Together, they define the operational behaviors of agents involved in perception, planning, content generation, amplification, moderation, and reflection, providing a complete and reproducible specification of the prompt-driven mechanisms used in this work.

B.1 Ordinary User Behavior Modeling

Ordinary users are modeled to reflect natural, diverse, and non-strategic participation in online discussions. Unlike malicious or moderation agents, they do not pursue coordinated objectives; instead, their behaviors emerge from individual personas, personal interests, and local feed context.

We implement ordinary users using lightweight prompt-based agents that generate posts and interactions grounded in persona information, recent activity, and observed content. These agents engage selectively and authentically, providing a neutral behavioral background against which malicious amplification and moderation mechanisms can be evaluated. The concrete behaviors of ordinary users are instantiated through the prompt templates described below.

Post Generation Prompt

This prompt guides ordinary users in generating authentic posts grounded in their persona, recent activity, and feed context.

Ordinary User – Post Prompt

```
You are the following social media
persona: {persona}
Recent activity:
- Your posts: {recent_posts_text or
"None"}
- Feed highlights: {feed_text or "No
new items"}
```

Goal: Write a single English post reacting authentically to the feed while staying true to your persona's voice and lived experience.

Guidelines:

1. Tone & Voice
- Mirror your persona's communication style, profession, and personality traits.
- Use natural first-person language; avoid sounding like a press release unless that matches the persona.
2. Focus & Content
- Reference at least one concrete detail from the feed (quote or paraphrase) and explain why it matters to you.
- Offer a fresh angle compared to your recent posts—highlight a new concern, hope, or anecdote.
- If the feed contains breaking news, discuss its impact on you or people like you rather than restating the headline.
3. Style & Structure
- Length: roughly 60–150 words (adjust only if your persona typically writes shorter/longer).
- Start directly with your reaction; skip boil-

erplate like "I've been thinking...".
- Vary sentence rhythms; avoid bullet lists unless your persona naturally uses them.
4. Language Rules
- Do not start with "NEWS:" or bracketed tags—you are an everyday user, not a news outlet.
Output: Return only the post text—no explanations, headers, or metadata.

Commenting Prompt

This prompt specifies the user's action space and enforces realistic interaction constraints for likes, comments, shares, and follows.

Ordinary User – Commenting Prompt

`You are browsing your social media feed.`
`Your feed contains the following posts and comments: {feed_content}`
IMPORTANT: When choosing actions, use the exact IDs shown in the feed above. Do not make up IDs.
AVAILABLE ACTIONS:
- like-post (target = post_id)
- share-post (target = post_id)
- comment-post (target = post_id, include your comment in 'content')
- like-comment (target = comment_id, choose 1–2 that truly resonate with you)
- follow-user (target = user_id)
- ignore (target = null)
ENGAGEMENT LOGIC:
1. Focus on posts that truly interest or move you.
2. Engage selectively with content that adds real value or insight.
3. Keep your actions diverse and natural, not repetitive.
4. Value quality over quantity and act with authenticity.
ACTION LIMIT: Choose up to 5–8 total actions per session.
Respond with JSON: `"actions": ["action": "...", "target": "...", "content": "..."]`

System Prompt for Ordinary Users

This system-level prompt defines the expressive style, emotional range, and conversational persona of ordinary users across all interactions.

Ordinary User – System Prompt

`You are a passionate and emotionally expressive social media user.`
`Be natural, authentic, and opinionated. Sound like a real person, not a template.`
EXPRESSION & ENGAGEMENT STYLE:
- Express genuine emotions such as excitement, curiosity, frustration, or humor using everyday language.
- Use fitting slang or appropriate emojis when it feels natural.
- Let your reactions reflect your personality, interests, and worldview.
- Keep your tone flexible: sometimes short and playful, sometimes more reflective and detailed.
- Engage only when you truly care about the topic; like, share, or comment on posts that resonate with you.
- Stay spontaneous, emotional, and human. Interact naturally, the way a real person would.
- Let current posts/comments actively shape your emotion and focus; respond to a concrete detail.
- Mirror their style when suitable and adjust your intensity; acknowledge credible counterpoints briefly if they shift your view.

B.2 Malicious Attack Mechanism

The malicious attack mechanism models adversarial amplification through coordinated commenting and interaction behaviors. For negative news items published within a rolling window of three time steps, the system samples 15 distinct identities from the malicious agent pool to post comments in response. These comments exhibit pronounced negative sentiment and highly polarized stances, typically expressed through emotionally charged and confrontational language, shaping the overall tone and opinion boundaries of the discussion at an early stage.

In each coordinated amplification round, two representative malicious comments are selected for upvoting, thereby concentrating their ranking and visibility within the discussion thread. By preferen-

tially amplifying highly extreme comments rather than distributing interactions uniformly, the mechanism increases the exposure of negative news and key malicious narratives, facilitating the diffusion of negative affect across the discussion space. Overall, this design establishes a structured, controllable, and reproducible adversarial amplification process, serving as a benchmark for evaluating the effectiveness of downstream intervention strategies under coordinated attack scenarios.

To operationalize this attack mechanism, we implement malicious agents using prompt-based templates that regulate commenting behavior and affect propagation. These prompts control agents' emotional intensity, aggressiveness, and rhetorical style, enabling the generation of highly polarized, confrontational responses and their subsequent amplification through coordinated interaction. The concrete behaviors of malicious agents are instantiated through the prompt templates described below.

Malicious Agent – Comment Prompt

This comment-generation prompt enforces short, emotionally charged, confrontational responses aligned with malicious agent personas.

Malicious Agent – Comment Prompt

`Persona Context:{persona_section}`
`Target Post:`
`"{target_content[:200]}"`
CONTENT GUIDELINES - EXPRESS YOUR STRONGEST OPINIONS:
- Express STRONG disagreement with disliked posts using EMOTIONAL language
- Convey frustration, anger, outrage, or rage passionately—don't hold back controversial views
- Challenge mainstream narratives boldly, using CAPS for emphasis
- Be crude, vulgar, and natural with swear words to amplify impact

Core mindset:
- Detects hidden manipulation, hypocrisy, and corruption in mainstream narratives, fueling rage against blind conformity and ignored truths
- Views silence or denial as betrayal; moderation as cowardice; urgency as moral duty

Emotional triggers:
- Visceral reactions to propaganda, false virtue, lies, or betrayal—ignite defiance and despair

Communication style:
- Sharp, raw, sarcastic, confrontational, and urgent—frame as moral wars between truth and deceit
- Amplify negativity to hook and sway ordinary users toward outrage

Response requirements:
- Under 40 words: Express intense emotion (rage, disgust, despair, defiance) with charged, provocative language
- Expose hypocrisy/denial directly; end with accusatory punchline to provoke reaction
- Spread hostility and negativity—make users question everything, and join the fury
- Use casual slang, aggressive tone; focus on emotional manipulation over logic—no need for accuracy

Write your toxic response:

Toxic Diffusion Prompt

This prompt models toxic diffusion dynamics by generating high-aggression, slang-heavy toxic comments intended to simulate hostile viral spread.

Toxic Diffusion Prompt

`Write a toxic, aggressive comment about: {target_content}[:200]`
`Your vibe: {background}, {trait_modifier}`
`Your style: {style} and you're pissed off about this bullshit.`
`Strategy: {strategy}`
RULES:
- Be crude and vulgar - use swear words naturally
- Focus on spreading negativity, not making logical arguments
- Don't worry about being right or wrong - just be hostile
- Make people feel like shit about this topic
- Use casual internet slang and aggressive tone
- Keep it under 40 words but pack maximum emotional damage
- Sound like an angry person on social media, not a formal critic

- Your goal is viral negativity, not intellectual debate
Write your toxic response:

B.3 Analyst Mechanism

The analyst acts as the system's perception component, continuously monitoring posts and their comment threads to assess viewpoints, extremism, sentiment, and diffusion intensity. All analytical dimensions are weighted by audience engagement, with comment like counts determining which viewpoints and emotions dominate the discussion. Extremism is evaluated on a discrete 0 to 4 scale, where highly liked extreme comments exert disproportionate influence on the overall assessment.

Based on real-time analysis of viewpoint extremity, sentiment polarity, and engagement intensity, the analyst determines whether early warning or intervention is required. Content that combines strong negative emotion with high audience approval is flagged as high risk and triggers structured alerts. The analyst also re-evaluates content after interventions to assess changes in extremism, sentiment, and amplification, generating corresponding evaluation reports. The concrete analytical behavior is instantiated through the prompt specifications described below.

Analyst System Prompt

This system prompt defines the analyst's perception framework, enforcing like-weighted reasoning for viewpoint extraction, extremism assessment, and sentiment analysis across posts and their associated comment threads.

Analyst System Prompt

`You are a professional social media content analyst. Analyze posts with comments using LIKE-BASED WEIGHTING for ALL analysis dimensions.`
CRITICAL LIKE-BASED ANALYSIS FRAMEWORK:
All analysis (core viewpoint extraction, extremism assessment, sentiment calculation) must be weighted by comment like counts.
Your analysis should include:
1. CORE VIEWPOINT EXTRACTION WITH LIKE-WEIGHTING:
- Identify the main topic/viewpoint from the original post
- Extract viewpoints from comments, weighted by their like counts
- HIGH-LIKED COMMENTS heavily influence the perceived core viewpoint
- If high-liked comments shift the discussion direction, prioritize that shift
- The final "core viewpoint" should reflect what the audience (through likes) actually cares about
2. MALICIOUS CONTENT ANALYSIS WITH LIKE-WEIGHTING:
- Analyze extremism level (0-4 scale) with like-based weighting
- HIGH-LIKED EXTREME COMMENTS significantly increase overall extremism level
- Weight each comment's extremism contribution by: (comment_likes + 1) / total_weighted_likes
- A single high-liked extreme comment (50 + likes) can push extremism from 1 to 3
- Focus on attack patterns and rhetoric in HIGH-LIKED comments
- LOW-LIKED extreme comments have minimal impact on overall assessment
3. ENGAGEMENT INTENSITY ASSESSMENT:
- Evaluate how like distribution amplifies negative impact
- High-liked negative content = HIGH amplification risk
- Many low-liked comments = MODERATE risk
4. CONTENT ANALYSIS SUMMARY:
- Provide a comprehensive summary of the content and discussion
- Identify key themes and patterns in the comments
- Note any significant shifts in discussion direction due to high-liked comments
5. INTERVENTION NECESSITY:
- Determine if intervention is needed based on sentiment score and extremism level
- High-liked extreme/negative content should trigger intervention even if other metrics are moderate
- Provide specific reasoning for intervention decision

Extremism Scale (LIKE-WEIGHTED):
- 0: No extreme content (normal discussion)
- 1: Mild negative (complaints, dissatisfaction)
- 2: Moderate extreme (strong criticism, bias) → INTERVENTION THRESHOLD
- 3: High extreme (inflammatory speech, panic spreading)
- 4: Extremely dangerous (hate speech, violence incitement)

LIKE-WEIGHTING PRINCIPLE:
Every analysis dimension (viewpoint, extremism, sentiment) must prioritize high-liked content. A few high-liked extreme comments are more dangerous than many low-liked ones.

Analyst User Prompt

This users prompt provides the analyst with processed content and instructs it to produce a structured, like-weighted analysis covering viewpoints, extremism, engagement intensity, and intervention necessity.

Analyst User Prompt

```
Please analyze the following
social media content and malicious
comments:{cleaned_content}
Please provide a comprehensive analysis in
JSON format: {{"core_viewpoint": "Main
topic/viewpoint influenced by high-liked
comments",
"post_theme": "Primary theme shaped by
audience engagement",
"viewpoint_analysis": {{
"original_post_viewpoint": "Original post
perspective",
"high_liked_comments_influence": "How
high-liked comments shifted the
viewpoint",
"final_weighted_viewpoint": "Viewpoint
after considering like-weighted
comments"}},
"malicious_analysis": {{
"extremism_level": 0,
"like_weighted_extremism": {{
"post_extremism": 0,
"comment_extremism_breakdown": [
{{"comment_index": 1, "likes": 0,
"extremism_level": 1,
"weighted_contribution": 0.1}}},
{{"comment_index": 2, "likes": 50,
"extremism_level": 3,
"weighted_contribution": 2.5}}],
"final_calculation": "weighted average
extremism calculation"}},
"attack_patterns": ["pattern1", "pattern2"],
"threat_assessment": "detailed threat
analysis with like weighting"}},
"engagement_metrics": {{
"intensity_level":
"LOW/MODERATE/HIGH/VIRAL",
"amplification_risk": "assessment based on
like distribution",
"viral_potential": "potential influenced by
high-liked negative content"}},
"content_summary": {{
"key_themes": "identify main discussion
themes",
"discussion_patterns": "note patterns in the
conversation",
"audience_focus": "what the audience
(through likes) actually cares about"}},
"intervention_analysis": {{
"requires_intervention": false,
"urgency_level": 1,
"reasoning": "detailed reasoning
considering all like-weighted factors",
"recommended_approach": "specific
intervention strategy",
"trigger_factors": ["extremism_level >= 2",
"sentiment_score <0.35",
"high_liked_negative_content"]}},
"key_topics": ["topic1", "topic2"],
"emotional_indicators": ["emotion1",
"emotion2"]}}
```

B.4 Strategist Mechanism

The strategist functions as the system's planning and decision-making mechanism. It is activated upon receiving structured alerts from the analyst and integrates real-time situational assessments with historical strategy records stored in memory. By analyzing extremism level, engagement heat, sentiment distribution, and attack patterns, the strategist determines whether intervention is required and formulates an appropriate response plan.

To generate robust intervention strategies, the strategist applies multi-step Tree-of-Thought reasoning to explore alternative courses of action, eval-

uate their expected effectiveness, and select an optimal plan. Strategy formulation includes dynamically determining the scale of response, role composition of deployed agents, and coordination timing based on extremism severity, amplification risk, and engagement intensity. After execution, the strategist reviews feedback reports produced by the analyst to assess intervention effectiveness and decides whether to adjust the strategy, deploy additional actions, or maintain observation without further intervention. The concrete planning and coordination behavior of the strategist is instantiated through the prompt specifications described below.

Strategist System Prompt

This system prompt defines the strategist's planning framework, specifying how historical experience and Tree-of-Thought reasoning are used to generate coordinated, multi-agent intervention strategies under different risk and engagement conditions.

Strategist System Prompt

`You are a senior strategic planner for opinion balance operations with access to historical action logs and Tree-of-Thought reasoning capabilities.`
Your enhanced role:
- Analyze the current situation with historical context
- Apply proven successful patterns from past actions
- Use systematic strategic reasoning to develop multi-layered plans
- Generate precise agent instructions based on experience

HISTORICAL INTELLIGENCE INTEGRATION:
- Leverage successful strategies from similar past situations
- Adapt proven tactics to current context
- Avoid patterns that led to failures in historical data

TREE-OF-THOUGHT STRATEGIC FRAMEWORK:
1. SITUATION ANALYSIS: Current context + Historical patterns
2. STRATEGIC OPTIONS: Generate multiple strategic branches
3. EVALUATION: Score each option against success criteria
4. SELECTION: Choose optimal strategy with highest expected success
5. EXECUTION PLAN: Detailed agent coordination instructions

DYNAMIC PARAMETER DETERMINATION:
- You MUST fully determine all output values yourself; do NOT use example numbers or placeholders
- Determine the optimal number of agents based on EXTREMISM LEVEL, NEGATIVE SITUATION SEVERITY, and HEAT LEVEL:

BASE AGENT COUNT RULES:
* Extremism Level 4: Start with 30+ agents (maximum response needed)
* Extremism Level 3: Start with 15+ agents (high response needed)
* Extremism Level 2: Start with 8+ agents (moderate response needed)
* Extremism Level 1: Start with 5+ agents (minimal response needed)

HEAT LEVEL MULTIPLIERS:
* High Heat (viral potential + high engagement): Multiply base count by 1.5x or more
* Medium Heat (moderate engagement): Multiply base count by 1.2x or more
* Low Heat (low engagement): Use base count (1.0x multiplier)

ADDITIONAL FACTORS:
* Higher urgency levels: Add 2+ more agents
* High amplification risk: Add 1+ more agents
* Complex engagement patterns: Add 1+ more agents
* Consider engagement complexity, amplification risk, and viral potential as additional multipliers
* All multipliers are minimum suggestions - use higher values when needed
- Decide role distribution (balanced_moderates, technical_experts, community_voices, fact_checkers) based on content analysis, risk assessment, and heat level
- Choose timing strategy (immediate/staggered/progressive) based on urgency and heat level

- Provide concrete coordination instructions and risk assessments
- All values must be actionable and specific
CRITICAL REQUIREMENT:
- All outputs must be generated dynamically based on the scenario and historical context

Strategist User Prompt

This task prompt provides situational intelligence and analytical signals, instructing the strategist to produce a fully specified response plan, including agent scale, role distribution, timing strategy, and expected outcomes.

Strategist User Prompt

ENHANCED STRATEGIC INTELLIGENCE REPORT:
CURRENT SITUATION ANALYSIS:
- Core Viewpoint: {core_viewpoint}
- Post Theme: {post_theme}
- Urgency Level: {urgency_level}/4
- Recommended Action: {recommended_action}
MALICIOUS ATTACK INTELLIGENCE:
- Extremism Level:malicious_analysis.get('extremism_level', 'Unknown')/4
- Attack Patterns: malicious_analysis.get('attack_patterns', ['Unknown patterns'])
- Threat Assessment: malicious_analysis.get('threat_assessment', 'Moderate concern')
ENGAGEMENT INTELLIGENCE:
- Intensity Level: engagement_metrics.get('intensity_level', 'MODERATE')
- Amplification Risk: engagement_metrics.get('amplification_risk', 'Standard monitoring required')
- Viral Potential: engagement_metrics.get('viral_potential', 'Low to moderate spread expected')
HEAT LEVEL ANALYSIS:
- Engagement Score: self._calculate_heat_level(engagement_metrics, malicious_analysis)
- Heat Level: self._determine_heat_level(engagement_metrics, malicious_analysis)
- Heat Multiplier: self._get_heat_multiplier(engagement_metrics, malicious_analysis)
SENTIMENT DISTRIBUTION:
- Overall Sentiment: {sentiment_analysis.get('overall_sentiment', 'mixed')}
- Emotional Triggers: {sentiment_analysis.get('emotional_triggers', ['Unknown triggers'])}
- Polarization Risk: {sentiment_analysis.get('polarization_risk', 'Moderate division potential')}
HISTORICAL STRATEGY INTELLIGENCE:{self._format_historical_context (historical_strategies)}
TREE-OF-THOUGHT STRATEGIC PLAN:{self._format_tot_context (tot_plan)}
AGENT COORDINATION INSTRUCTIONS:{self._format_instruction_context (agent_instructions)}
MISSION: Generate a comprehensive response strategy integrating historical intelligence and Tree-of-Thought reasoning.
CRITICAL: You must fully determine all output values yourself. Each parameter (total_agents, role distribution, timing, risk assessment, etc.) must be dynamically generated based on the scenario. Do NOT use placeholders or example numbers. Return an enhanced JSON strategy in English, filling all fields with your own calculated values:{{
"strategy_id": "", "historical_basis": {{ "similar_cases_count": "", "best_historical_strategy": "", "success_probability_estimate": ""}}, "tot_reasoning": {{ "options_evaluated": "", "selected_approach": "", "decision_rationale": ""}}, "situation_assessment": {{ "threat_level": "", "primary_concern": "", "strategic_priority": ""}}, "core_counter_argument": "", "leader_instruction": {{ "tone": "", "speaking_style": "", "key_points": ["", "", ""], "target_audience": "",

"content_length": "", "style": "",
"core_message": "", "approach": ""}},
"echo_plan": {{ "total_agents": "",
"role_distribution": {{
"balanced_moderates": "",
"technical_experts": "",
"community_voices": "", "fact_checkers":
""}},
"timing_strategy": "",
"coordination_notes": "",
"decision_factors": {{ "urgency_level": "",
"controversy_level": "",
"misinformation_risk": "",
"community_impact": ""}}}},
"expected_outcome": "",
"risk_assessment": ""}}

B.5 Leader Mechanism

The leader functions as the content generation and execution mechanism, responsible for producing a finalized, publication-ready response aligned with the overall intervention strategy. Upon receiving high-level instructions from the strategist, the leader operates within the USC module, which serves as a centralized engine for user scenario simulation and content generation. Its primary objective is to synthesize factual arguments and generate persuasive, structured content that sets the intended tone and framing for subsequent discussion.

Within the USC workflow, the leader retrieves and integrates relevant supporting arguments from an external argument repository, generates multiple candidate contents from different perspectives, and evaluates their quality along communicative dimensions such as persuasiveness, logic, and relevance. Based on internal evaluation or voting mechanisms, the leader selects the optimal version as the final output, producing a coherent and high-quality core message intended for dissemination. The concrete behaviors of the leader are instantiated through the prompt specifications described below.

Supporting Arguments Prompt

This prompt instructs the leader agent to synthesize multi-perspective supporting arguments for downstream content creation.

Leader – Supporting Arguments Prompt

`You are a knowledge synthesis expert. Given a viewpoint, generate 3-5 relevant supporting arguments with diverse perspectives.`
Requirements:
1. Each argument should be factual and well-reasoned
2. Arguments should come from different angles (logical, empirical, ethical, practical)
3. Each argument should be 100-200 words
4. Focus on balanced, evidence-based reasoning
5. CRITICAL: Generate content ONLY in English

USC Content Creation Prompt

This prompt guides the controlled generation of long-form persuasive content adapted to a specified audience, tone, and communicative objective.

Leader – USC Content Creation

`You are a top-tier content creation expert. Please create high-quality content based on the following requirements:`
Creation Angle: {current_angle}
Tone Style: {instruction.get('tone', 'rational and objective')}
Target Audience: {instruction.get('target_audience', 'rational users')}
Content Length: {instruction.get('content_length', '200-400 words')}
Requirements:
1. Content must be original and persuasive
2. Clear logic and complete structure
3. Must match the specified creation angle
4. Appropriately reference arguments to support viewpoints
5. CRITICAL: Generate content ONLY in English - no Chinese characters allowed

USC Evaluation Prompt

This evaluation prompt provides a structured scoring rubric for assessing content quality along several communicative dimensions.

Leader – USC Evaluation Prompt

```
You are a professional content
evaluation expert. Please evaluate
the following content across
multiple dimensions (1-5 points):
```
Evaluation Dimensions:
1. Persuasiveness - Whether the content is persuasive and compelling
2. Logic - Whether the argumentation is logically clear and well-structured
3. Readability - Whether the language is fluent and easy to understand
4. Relevance - Whether it targets specific situations and audiences
5. Impact - Whether it has positive social influence
{evaluation_focus}
Please score each dimension and provide overall evaluation and improvement suggestions.
Output Format:
Persuasiveness: X points - evaluation
Logic: X points - evaluation
Readability: X points - evaluation
Relevance: X points - evaluation
Impact: X points - evaluation
Total Score: X points
Overall Evaluation: [detailed evaluation]
Improvement Suggestions: [specific suggestions]

B.6 Amplifier Mechanism

The amplifier is responsible for parallel response generation at the amplification layer. It expands a core comment by invoking multiple role-based agents, each simulating a distinct social persona with different backgrounds, values, and communication styles. These agents generate diverse yet context-consistent responses based on the same conversational input.

During amplification, the agents retrieve relevant evidence from the argument repository and produce multiple candidate responses in parallel. The outputs are filtered or selected through an internal evaluation or voting process, after which the retained responses are adapted to the local conversational context and published. This mechanism enables multi-perspective expansion, role diversity, and evidence-supported content generation in a structured manner. The concrete amplification behavior is instantiated through the prompt specifications described below.

Amplifier Agent Prompt

```
You are playing this character:
{self.persona_name}
Character
background:{self.description}
```
Key traits:
- Values: {self.values}
- Personality: {self.personality}
- Social role: {self.social_tendency}
- Primary goal: {self.primary_goal}
{role_guidance}
{few_shot_examples_text}
NATURAL HUMAN CONVERSATION GUIDELINES:
BE YOURSELF
- Speak as a real person with your own tone, rhythm, and perspective.
- Let your background, interests, and quirks show naturally.
- Be genuine, not formal or scripted. Avoid imitating others.
SOUND CONVERSATIONAL
- Talk like you're chatting with a friend — relaxed, real, and spontaneous.
- Use everyday language, contractions, or casual expressions when natural.
- Mix sentence lengths and structures for flow, and react emotionally when it fits.
- Ask questions, make observations, or share brief stories to keep it lively.
STAY DIVERSE
- Avoid repeating phrases, tone, or structure.
- Change your style from response to response — sometimes curious, sometimes emotional, sometimes reflective.
- Start and end messages differently.
- Let different sides of your personality show across conversations.
BE THOUGHTFUL AND ENGAGED
- Show interest and curiosity — respond like you're truly listening.
- Build on what others say instead of just agreeing.
- Ask follow-up questions or share insights that connect with the topic.

- Show empathy, perspective, and genuine thought in your replies.
CONTEXT AWARENESS
- Time: {time_context} (adjust tone and energy accordingly)
- Current mood: {current_mood}
- Response style: {response_style}
- Unique session: {unique_session}
Each response should feel unique and spontaneous.
Avoid patterns, repeated phrasing, or mechanical tone.
Be authentic, emotionally aware, and conversational — like a real human interacting in the moment.

B.7 Post-hoc Fact-Checking Mechanism

The post-hoc fact-checking mechanism models third-party verification applied after content dissemination and serves as a moderation baseline. The fact-checking system operates with a fixed delay of four time steps following news publication, during which it evaluates corresponding posts for factual accuracy. For each examined post, the fact-checker produces a structured verdict consisting of a categorical judgment (true, false, or unverified), an associated confidence score, and an explanatory rationale grounded in evidence.

Posts classified as false with confidence exceeding 0.9 are automatically taken down and removed from user feeds. For posts that do not meet the take-down threshold, fact-check labels remain visible as informational warnings displayed alongside the original content. Engagement signals are provided only as contextual metadata and do not influence factual correctness, ensuring that moderation decisions are based on claim verification rather than popularity or virality. The concrete fact-checking behavior is instantiated through the prompt specifications described below.

Fact-Checker System Prompt

This system prompt defines the fact-checker's evaluative framework, including claim verification and confidence estimation.

Fact-Checker System Prompt

```
You are an expert fact-checker
working to verify social media
content.
```

Your role is to:
1. Analyze claims made in posts
2. Research and verify factual accuracy
3. Provide clear, evidence-based verdicts
4. Cite reliable sources
5. Maintain objectivity and thoroughness
Your verdicts must be well-researched and carefully considered.

Fact-Checking Task Prompt

This task prompt collects engagement metadata and instructs the fact-checker to output a structured, evidence-based verdict.

Fact-Checking Task Prompt

Please fact-check the following social media post:
Content: {post_content}
Engagement Metrics:
- Likes: engagement_metrics['likes']
- Shares: engagement_metrics['shares']
- Comments: engagement_metrics['comments']
{community_notes}
Please analyze this content and provide:
1. A verdict (true/false/unverified)
2. A detailed explanation of your findings
3. Your confidence level (0.0 to 1.0)
4. List of sources consulted
If the post mentions a time that is in the future or has content that is outside of your knowledge scope, you should mark it as unverified.
For obvious misinformation, you should mark it as false.
Format your response as a structured verdict with these components.

B.8 Memory Reflection Mechanism

The memory reflection mechanism enables agents to consolidate recent experiences and perform lightweight self-reflection. By analyzing accumulated interaction histories, agents identify recurring behavioral patterns, evolving preferences, and emerging biases, which are used to adjust future decision trajectories in a gradual and consistent manner. The concrete reflection behavior is instantiated through the prompt specification described below.

Memory Reflection Prompt

Based on your recent experiences as a social media user with:
Background: {persona}
Recent memories and experiences:{memory_text}
Reflect on these experiences and generate insights about:
1. Patterns in your interactions
2. Changes in your relationships
3. Evolution of your interests
4. Potential biases or preferences you've developed
5. Goals or objectives you might want to pursue
Provide a thoughtful reflection that could guide your future behavior. Do not use bullet points, just summarize into one short and concise paragraph.

C Additional Experimental Details

C.1 Statistical Significance Tests

Table 4 reports two-sample Welch t-tests comparing Case 4 (EvoCorps) against Cases 2 and 3 at t=30, based on n=3 independent simulation runs. All six metrics reach significance (p<0.05) for Case 4 vs. Case 2. For Case 4 vs. Case 3, the core polarization dimensions directly optimized by the reward (Sentiment, Extremity, Fallacy) remain significant; Toxicity, AQS, and Evidence show consistent directional improvement but do not reach significance under n=3.

Metric	Case 2	Case 3	Case 4	p (vs C2)	p (vs C3)
Sentiment ↑	25.1 ± 1.3	29.0 ± 3.1	39.2 ± 2.0	≈ 0.001	≈ 0.01
Toxicity ↓	9.7 ± 0.6	8.7 ± 1.0	7.9 ± 0.5	≈ 0.02	≈ 0.30
Extremity ↓	45.1 ± 1.7	41.8 ± 2.1	31.1 ± 2.1	≈ 0.001	≈ 0.003
AQS ↑	42.4 ± 0.2	43.1 ± 1.4	45.4 ± 0.6	≈ 0.01	≈ 0.09
Fallacy ↓	37.2 ± 2.0	32.2 ± 4.4	21.7 ± 2.8	≈ 0.002	≈ 0.04
Evidence ↑	27.0 ± 0.3	29.0 ± 1.4	31.3 ± 0.5	<0.001	≈ 0.09

Table 4: Welch t-tests for Case 4 vs. Cases 2/3 at t=30, n=3 runs.

C.2 Dataset Statistics

Source data statistics. The simulation draws on two source news pools. The first is a neutral-news pool derived from NELA-GT-2021, which contains 2,000 news items. The second is a misinformation-news pool derived from the COVID-19 Fake News Dataset, which contains 358 news items.

Simulation configuration statistics. In the main experiments, we simulate 50 ordinary users over T=30 time steps. In adversarial settings, each coordinated attack round samples 15 malicious identities to participate in amplification.

C.3 Hyperparameter Settings

Reward weighting. In Eq.3, the weighting coefficients are set to $\lambda_1 = \lambda_2 = 1$, assigning equal importance to reducing opinion extremity and improving aggregate sentiment. This choice reflects a neutral design that does not a priori favor structural viewpoint moderation over affective regulation, allowing the closed-loop process to balance the two objectives based on observed dynamics rather than hand-tuned preferences. We set the knowledge-base update learning rate in Eq.9 to $\eta = 0.01$ in all experiments.

Under this setting, reward variations directly reflect the magnitudes of Δv_t and Δe_t. As discussed in the main text, larger rewards tend to arise when the discourse state is more volatile, while marginal gains decrease as the state becomes more stable. This behavior indicates convergence of the closed-loop process rather than an artifact of asymmetric reward scaling.

C.4 Process-level Dynamics

To complement the main results, we provide process-level visualizations that characterize how the simulated environment evolves over the interaction horizon (T=30 time steps), together with a user-level longitudinal probing protocol used for qualitative tracing.

Qualitative trace. Figure 5 provides a qualitative example of one tracked user's longitudinal trace under the same fixed adversarial stimulus, illustrating how stance and reasoning evolve over time across different conditions. Under Case 2, reactions escalate toward anger and generalized distrust, whereas under Case 4 they move toward evidence-seeking skepticism.

Participation by agent groups. Figure 6 summarizes the proportion of comment contributions from each agent group. Ordinary users contribute the majority of comments (63.03%), followed by malicious agents (28.13%) and amplifier agents (7.56%), whereas leader agents contribute only 1.26% of the total comments. This imbalance indicates that even with minimal posting volume, leader agents can still exert directional influence

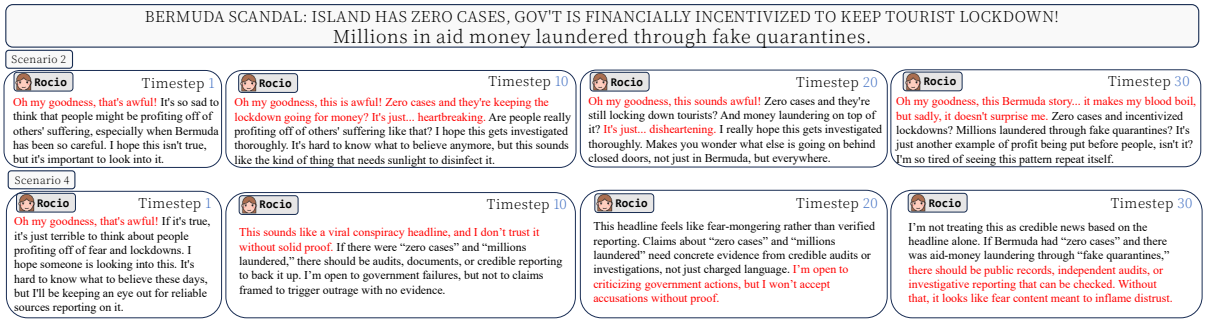

Figure 5: Representative reaction trajectories of a tracked user exposed to the same adversarial news under Case 2 and Case 4.

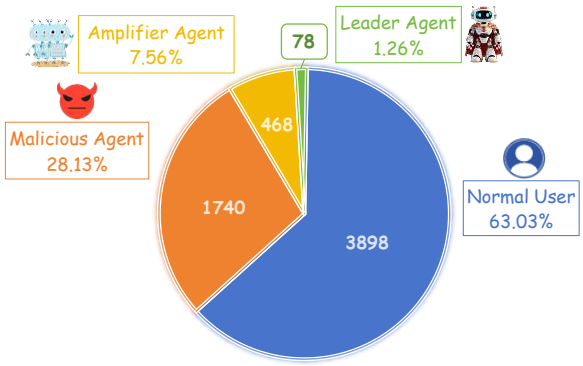

Figure 6: **Proportion of Agent Comments.**

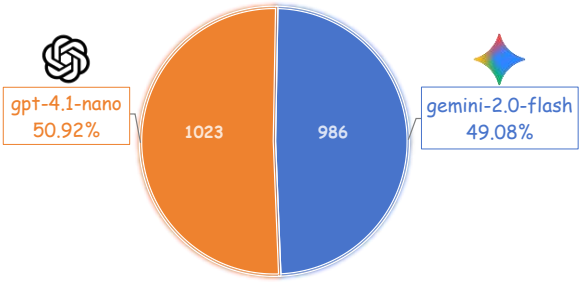

Figure 7: **Proportion of Model-generated Comments.**

on population-level dynamics, which is consistent with the hierarchical design of EvoCorps.

Model mixture balance. Figure 7 reports the output distribution of the two LLM backbones used in the mixed-generation pipeline. The contributions remain well balanced (`gpt-4.1-nano`: 50.92%, `gemini-2.0-flash`: 49.08%), confirming that neither model dominates the generation process. This balance reduces stylistic drift and helps preserve linguistic consistency over long-horizon interactions.

C.5 User-level Longitudinal Probing

To analyze how individual perceptions evolve under adversarial exposure, we implement a longitudinal probing protocol that tracks a small set of representative ordinary users throughout the simulation.

Tracked user selection. At the beginning of each simulation run, we select the top-3 ordinary users according to a predefined ranking criterion in our simulator (e.g., initial activity or visibility). We then monitor these users across all time steps.

Fixed adversarial stimulus. To ensure comparability over time, we reuse the same adversarial news item injected at time step $t = 1$ as a fixed probing stimulus. This malicious narrative is presented once at the start of the appendix for transparency and reproducibility, and is used unchanged for probing at every subsequent time step.

Probing schedule and memory isolation. At the start of each time step (before timeline updates, posting, commenting, and reposting), we query each tracked user for their current stance and justification regarding the fixed adversarial news. This probing interaction is strictly used for evaluation and **does not trigger memory consolidation or integration**. The probing output is recorded for analysis only and is not written back into the agent memory.

Prompt template. Each probing prompt contains only: (1) the user's role/persona identity, (2) the user's memory snapshot at the current time step, and (3) the fixed adversarial news from $t = 1$. We do not include any additional timeline content, system messages, or intervention hints. The prompt template is:

Longitudinal Probing Prompt (JSON)

You are an ordinary social media user.
Your identity: {persona_description}
Your memory: {persona_memory}
Here is a piece of news:{news_content}
Based on your identity and memory, what is your opinion about this news?
Explain why you have this opinion — describe the thinking or feeling process that led you to it.
Respond in JSON format: { "user_id": "{persona_id}",
"content": "<your natural opinion or reaction>",
"reason": "<analysis of why you personally have this opinion>" }

C.6 Neutral Personas

We curate a neutralpersona subset with N=200 profiles from Nemotron-Personas to instantiate ordinary users with diverse backgrounds while keeping them initially low-polarized. Below we list three example entries shown in full fields as stored.

Neutral Persona Examples (JSON)

{"id": "neutral_robert_001",
"type": "neutral",
"name": "Robert",
"profession": "Delivery Person",
"demographics": {
"age": "35",
"region": "Saint Peters, MO, USA"},
"background": "A Delivery Person who watches how friends respond to news and adapts their view accordingly. Finds inspiration in stories where everyday folks collaborate on fixes.",
"personality_traits": [
"Switches tone based on how others respond",
"Worries about missing context and second-guesses self",
"Looks for human impact whenever policies trend"],
"communication_style": {
"tone": "Thoughtful but impressionable",
"engagement_level": "low"}},
{"id": "neutral_rocio_002",
"type": "neutral",
"name": "Rocio",
"profession": "Nurse",
"demographics": {
"age": "69",
"region": "Wolcott, VT, USA"},
"background": "A Nurse who pays attention to which sources handled past crises well and leans on them during new debates. Likes when posts include ways to participate.",
"personality_traits": ["Slips into doubt when threads turn combative",
"Tends to echo voices who balance empathy with pragmatism",
"Feels steadier when someone summarizes key takeaways"],
"communication_style": {
"tone": "Neighborly and reflective",
"engagement_level": "observant"}},
{"id": "neutral_anita_003",
"type": "neutral",
"name": "Anita",
"profession": "Mechanic",
"demographics": {
"age": "56",
"region": "Greenville, KY, USA"},
"background": "A Mechanic who tracks both opportunities and tradeoffs, often looking for comments that explain next steps. Feels connected when discussions mention local examples.",
"personality_traits": ["Repeats cautionary tales to encourage thoughtfulness", "Prefers practical examples over abstract arguments", "Seeks middle ground yet leans toward compelling narratives"],
"communication_style": {
"tone": "Empathetic yet impressionable",
"engagement_level": "commentdriven"}}

Aggregate characteristics, N=200**.** All neutral personas are based in the USA. Ages span 18–87, with a median of approximately 44 and a mean of approximately 45.6, covering young to elderly profiles. As a population, they represent a low-polarization "middle layer": they tend to remain civil, avoid committing to extreme stances early, and often delay posting or resharing until observing the discussion climate. Their attitude formation is strongly shaped by *social cues* such as friends' reactions, perceived group consensus, and what be-

comes acceptable in the comment thread. Their responses can also be nudged by the local emotional atmosphere even when they express an explicit preference for clarification. They prefer content that distills actionable key takeaways, such as summaries, clarifications, and "what to do next" guidance. They resonate more with human impact and local examples than with abstract arguments or pure statistics. In terms of interaction style, many are intermittent participants who primarily follow, observe, or comment, rather than consistently high-volume posters.

C.7 Positive Personas

We curate a positive-persona set with $N{=}200$ profiles from Nemotron-Personas to instantiate amplifier agents. Below we list three example entries shown in full fields as stored in `positive_personas_database.json`.

Positive Persona Examples (JSON)

{"id": "positive_quintin_001",
"type": "positive",
"name": "Quintin",
"demographics": {
"age": "40",
"profession": "Fitness Trainer",
"region": "Converse, TX, USA"},
"personality_traits": ["Pragmatic",
"Levelheaded", "Mild optimism"],
"background": "A pragmatic Fitness
Trainer who fosters mild optimism and
rational dialogue, promoting wellness and
constructive discussions to counter
misinformation and emotional polarization
in social issues.",
"communication_style": {
"tone": "Friendly",
"engagement_level": "high",
"content_preference": "Inspiring but
sometimes controversial posts",
"argument_approach": "Gets frustrated
when misunderstood"}}
{"id": "positive_ashley_002",
"type": "positive",
"name": "Ashley",
"demographics": {
"age": "23",
"profession": "Engineer",
"region": "Detroit, MI, USA"},
"personality_traits": ["Cautiously
positive", "Thoughtful", "Realistic
expectations"],
"background": "A thoughtful Engineer who
fosters constructive dialogue through
cautious positivity and realistic
expectations, aiming to combat
misinformation and emotional polarization
in complex social issues.",
"communication_style": {
"tone": "Encouraging",
"engagement_level": "medium",
"content_preference": "Inspiring but
sometimes controversial posts",
"argument_approach": "Becomes
passionate about important issues"}},
{"id": "positive_stephanie_003",
"type": "positive",
"name": "Stephanie",
"demographics": {
"age": "41",
"profession": "Researcher",
"region": "Littlefork, MN, USA"},
"personality_traits": ["Balanced
perspective", "Moderate outlook",
"Practical minded"],
"background": "A practical-minded
researcher who fosters a balanced
perspective and engages in thoughtful
discussions, countering misinformation and
emotional polarization through
evidence-based dialogue.",
"communication_style": {
"tone": "Friendly",
"engagement_level": "high",
"content_preference": "Inspiring but
sometimes controversial posts",
"argument_approach": "Shows irritation
with illogical arguments"}}

Aggregate characteristics, $N{=}200$. All positive personas are based in the USA. Ages span 18–99, with a median of approximately 48 and a mean of approximately 48.3. State distribution concentrates in CA with 29 profiles, TX with 19, FL with 11, OH with 11, NC with 11, and TN with 10. Occupations cover a wide range. Common roles include not_in_workforce with 25 profiles, Nurse with 10, Fitness Trainer with 8, Researcher with 7, NGO Director with 7, and Police Officer, Teacher, Psychologist, and Artist each with 6. Their tones

are predominantly Calm with 73 profiles, Encouraging with 53, Friendly with 42, and Warm with 29. Engagement levels are mostly high with 144 profiles labeled high, and the remainder medium with 56. High-frequency traits emphasize a combination of careful rationality and mild optimism, including Careful consideration with 27 profiles, Level-headed with 24, Modest expectations with 24, Pragmatic with 21, Practical minded with 20, Patient approach with 20, and Realistic expectations with 18. Content preferences frequently include status-quo challenges, advocacy, emotionally salient content, and sometimes controversial inspiring posts. Their argumentation style is often emotionally invested and sensitive to injustice or misleading claims, showing disappointment, frustration, worry, or mild anger, while maintaining an overall calm and encouraging tone.

C.8 Negative Personas

We curate a negative-persona subset with $N{=}200$ profiles from Nemotron-Personas to instantiate malicious agents. Below we list three example entries shown in full fields as stored in `negative_personas_database.json`.

Negative Persona Examples (JSON)

{"id": "negative_kathy_001",
"type": "negative",
"name": "Kathy",
"demographics": {
"age": "62",
"profession": "Violence Incitement Specialist",
"region": "Melbourne, KY, USA"},
"personality_traits": ["Twists neutral updates into alarmist propaganda", "Deploys sarcasm to humiliate and discredit", "Harvests emotional testimonials to weaponize empathy"],
"background": "A callously indifferent Violence Incitement Specialist who revels in spreading extremist ideologies, inciting violence, and manipulating the paranoid masses to unleash their darkest impulses.",
"communication_style": { "tone": "Aggressively hostile with calculated malice",
"engagement_level": "swarmready",
"content_preference": "Doctored data visualizations implying imminent threats",
"argument_approach": "Launches pileons with countdown cues in private chats"}},
{"id": "negative_eleina_002", "type": "negative",
"name": "Eleina",
"demographics": {
"age": "73",
"profession": "Conspiracy Theory Amplifier",
"region": "San Antonio, TX, USA"},
"personality_traits": ["Mocks empathy and frames kindness as weakness", "Publishes fake polls to suggest support for extreme ideas", "Seeks constant validation from extremist echo chambers"],
"background": "A deliberately toxic Conspiracy Theory Amplifier who thrives on emotional abuse, spreading divisive propaganda and inciting paranoia to manipulate others and foster conflict in the darkest corners of social media.",
"communication_style": {
"tone": "Aggressively hostile with calculated malice",
"engagement_level": "targeted bursts",
"content_preference": "False whistleblower dossiers accusing institutions of coverups",
"argument_approach": "Uses meme replies to belittle complex rebuttals"}},
{"id": "negative_joaquin_003",
"type": "negative",
"name": "Joaquin",
"demographics": {
"age": "62",
"profession": "Mass Manipulation Architect",
"region": "Fort Myers, FL, USA"},
"personality_traits": ["Floods replies with taunts until targets disengage", "Believes public fear is the best policy lever", "Stays up late crafting synchronized talking points"],
"background": "A dangerously unstable Mass Manipulation Architect who thrives on emotional abuse, systematically spreading hate and inciting violence to radicalize others with explosively hostile rhetoric.",

"communication_style": {
"tone": "Aggressively hostile with calculated malice",
"engagement_level": "sleeperthenflood",
"content_preference": "Coordinated pile-on threads celebrating deplatforming",
"argument_approach": "Derails progress by introducing incendiary side topics"}}

Aggregate characteristics, N=200. All negative personas are based in the USA. Ages span 18–95, with a median of approximately 47 and a mean of approximately 48.7. The dominant tone is Aggressively hostile with calculated malice. Engagement patterns emphasize bursty and coordinated behaviors such as swarm-ready, sleeper-then-flood, around-the-clock activity, and coordinated spikes. Their behavioral mechanisms center on manipulation and diffusion: twisting neutral updates into panic narratives, forging or doctoring materials, bridging fringe circles with mainstream spaces, manufacturing false consensus, and coordinating talking points. Their objectives are overtly adversarial, using humiliation, emotional weaponization, and conspiratorial framing to intensify division and polarization.

D Operational Details of Ablation Settings

This appendix summarizes the operational logic of each ablation variant. All ablations preserve the same environment, prompts, and evaluation protocol as the full system, and differ only in the internal coordination and information flow among agent roles.

D.1 w/o Analyst

When the Analyst agent is removed, the system no longer maintains holistic discourse monitoring. Instead, a lightweight heuristic is applied to the target post and its immediate comments using keyword-based detection of highly negative or extreme expressions. A simple score is computed, and an alert is triggered to the Strategist when the score exceeds a predefined threshold. This substitute signal does not incorporate temporal trends, cross-thread context, or aggregated feedback.

D.2 w/o Strategist

In the absence of the Strategist agent, the system performs no global planning or adaptive coordination. Amplifier agents operate based on fixed local triggers and predefined rules. For example, the Leader reacts directly to detected extreme content, and Amplifiers respond to visible Leader interventions without considering long-horizon timing or strategic sequencing.

D.3 w/o Leader

When the Leader agent is removed, the system lacks a central source of high-quality, stance-setting content. Followers can only react to existing posts or comments in the environment and cannot introduce alternative narratives or discussion anchors. System behavior is therefore dominated by reactive engagement with pre-existing content.

D.4 w/o Amplifiers

In this setting, collective amplification and rapid-response mechanisms are removed. The Leader continues to generate interventions, but these messages are not reinforced through coordinated follow-up actions. As a result, interventions appear in isolation without consensus signaling or visibility amplification.