arXiv:2602.08529v1 [cs.MA] 9 Feb 2026

EvoCorps：一种用于缓解在线话语极化的进化多智能体框架

Ning Lin[1†]，**Haolun Li**[1†]，**Mingshu Liu**[1†]，**Chengyun Ruan**[1]，**Kaibo Huang**[1*]，
Yukun Wei[1]，**Zhongliang Yang**[1*]，**Linna Zhou**[1]
[1]北京邮电大学{linning, lhldudu, llms266, ruanchengyun815,huangkaibo, weiyukun, yangzl, zhoulinna}@bupt.edu.cn

摘要

在线话语中的极化会侵蚀社会信任并加速错误信息传播，但技术应对措施仍大多停留在诊断和事后处置层面。当前的话语治理方法存在固有的延迟和静态政策问题，难以应对实时演变的协同对抗性放大。我们提出**EvoCorps**，一种用于主动缓解极化的进化多智能体框架。EvoCorps将话语治理建模为动态社会博弈，并协调各角色开展监测、规划、有依据的生成和多身份传播。检索增强集体认知核心提供事实依据和行动—结果记忆，而闭环进化学习则会随着环境和攻击者的变化调整策略。我们在MOSAIC社交人工智能仿真平台上实现了EvoCorps，并在包含对抗性注入和放大的多源新闻流中进行受控评估。在情绪极化、观点极端性和论证理性方面，EvoCorps均优于对抗基线，改善了话语结果，为从检测和事后缓解转向过程中的闭环干预指明了切实可行的路径。代码见：https: // github.com/ln2146/EvoCorps。

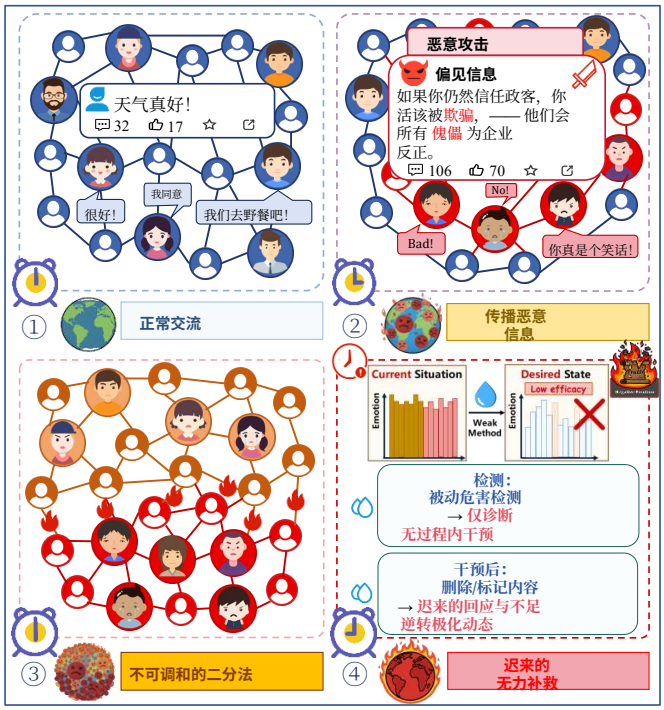

图1：在线互动从正常的用户交流演变为恶意智能体传播有害信息，从而导致负面情绪放大和极化。现有的危害检测与事后干预（例如内容删除或标注）仅在扩散之后发生，不足以缓解极化。

1 引言

在线社交媒体平台已成为公共传播的核心基础设施，塑造着观点、情绪和叙事的大规模形成方式。通过算法策展和同质性网络互动，这些平台影响着信息接触和社会认同。由此产生的结果之一，是在线话语中极化现象的出现。这里，极化是指用户、内容和互动被分隔到彼此对立的阵营中，跨阵营信息接触有限；在同质性社群中，充满情绪的单边叙事可能变得突出，在强化群体身份的同时边缘化温和观点。 [Loru 等人， 2025] 这是一种系统层面结果，可能产生于同质性与以参与度优化为导向的内容策展之间的互动 [Cinelli等人， 2021]。已有研究表明，在强同质性条件下，推荐系统会加剧回音室 [Cinus 等人， 2021]。一旦根深蒂固，情绪反应往往会主导事实推理，使后续纠正不太有效。

除了这些动态之外，当代在线话语已不再仅由用户互动塑造。不断累积的证据表明，恶意且高度协同的智能体已被积极部署于现实世界的

†共同贡献。*通讯作者

平台之中，以影响公众舆论并引导话语轨迹。近期研究表明，此类对抗性集体能够在战略上利用情绪传染、算法强化和协同多重身份行为，在早期阶段重塑话语动态，通常早于现有治理机制作出响应 [Feng 等，2024；Ma等， 2024]。在这些条件下，对立阵营通过对抗性放大显著加速，协同代理借助多个身份注入并放大具有情绪煽动性的叙事，同时利用推荐系统中的反馈回路，使内容迅速扩散至同质性社群，并将讨论推向对立阵营（图1）。在此类环境中，情绪信号往往快于事实纠正，使早期框架得以固化，并削弱后续内容审核或事实核查的作用。

在许多平台上，应对有害或加剧极化话语的主流技术手段仍停留在诊断层面：检测模型和风险评分可以标记毒性、立场极端程度或错误信息信号，但它们本身无法及时、协调地采取行动，改变实时讨论的走向。在检测的基础上，既有研究探索了多种干预措施，例如事实核查、内容标注、内容删除或降低排序、内容审核流程以及推荐调整 [Nakov 等，2024；Vladika 和 Matthes，2023；Habibi等， 2024]。然而，这些机制往往在信息扩散已经发生，或极化信号变得显著之后才会触发，导致早期对抗性放大与治理响应之间存在固有的延迟。

这种延迟反映了治理中的不对称动态：防御性干预通常是被动、分散且低频的，而对抗性群体则能够以协调、先发且持续的方式行动。在算法反馈和同质性交互作用下，情绪煽动性叙事会呈现路径依赖；一旦被放大，互动驱动的曝光便会加深回音室效应 [Ferraz de Arruda 等， 2024], 而后续纠正或孤立的内容审核行动影响有限。此外，间歇性或静态政策缺乏针对不断演变的对抗行为进行闭环适应的能力。这些不足促使我们探索过程内治理机制，使其能够贯穿整个交互过程运行、协调互补的角色并不断适应。

为此，我们提出 EvoCorps：一个用于应对对抗性放大的主动式话语去极化进化多智能体框架。受 Stackelberg–Mean-Field Control（SMFC）范式的启发 [Huang 和 Yang，2020；Dayanikli 和 Lauri`ere， 2023], EvoCorps 将话语治理建模为一种动态社会博弈：有组织的干预团队负责规划并执行策略，而更广泛的用户群体则在群体层面作出响应。具体而言，EvoCorps 将这一理念实现为闭环多智能体系统，采用多智能体马尔可夫决策过程（MMDP）表述和进化学习，随着话语动态与对抗策略的演变不断调整策略。

我们的主要贡献总结如下 ws:

- 我们将主动去极化建模为一种在对抗性放大环境下的流程内闭环干预问题，旨在解决常见事后干预存在的滞后性和非适应性。
- 我们提出**EvoCorps**，将角色专业化协作与检索增强的集体认知相结合，并通过闭环进化学习随时间调整干预策略。
- 我们在 MOSAIC 上实现 EvoCorps，并在多源新闻流中评估其应对对抗性注入和协同放大的能力。与对抗性基线和事后基线相比，EvoCorps 提高了情绪极化、观点极端性和论证理性。

2 相关工作

2.1 基于智能体的恶意信息扩散

近期研究揭示，恶意或协同智能体所能做的不只是传播虚假信息；通过放大情绪煽动性叙事，它们还能动员受众形成对立阵营，削弱人们对机构的信任，并使敌对或阴谋论框架常态化。近期研究表明，生成式智能体能够通过协同点赞和分享来传播错误信息 [Liu 等， 2025], 人工智能与人类的协作能够显著扩大耸动叙事的传播范围并延长其持续时间 [Ng 等， 2025], 而自适应的多智能体协作则使宣传活动得以规避内容审核，并长期维持影响力 [Ren 等，2025]。综合来看，这些发现表明，对抗性群体在塑造话语动态方面具有系统性优势，因为它们利用情绪传染、算法同质性和协同放大，推动快速、适应性强且高度有组织的进程，而这些进程往往快于被动响应的治理机制。

2.2 话语去极化研究

现有多数缓解在线话语极化的方法侧重于诊断能力，包括衡量和检测毒性、错误信息、立场极端程度或偏见叙事 [Nakov 等，2024；Vladika 和 Matthes，2023；Ma 等， 2024]。近期研究探索了使用自监督目标和交互特征检测极化 [Cui 和 Abdelzaher， 2024], ，以及对社交媒体辩论中极化语言和毒性模式的大规模分析 [Xu 等，2025年；Arora 等，2022]。此类方法在识别有害内容和审查极化指标方面虽有效，但主要只能用于诊断，无法直接说明如何引导互动过程的发展。

基于这些诊断信号，人们探索了多种干预措施，其中最常见的是事后干预。代表性动作包括账号封禁或内容下架、内容降权、添加警示标签或插页，以及附上事实核查或澄清说明 [Habibi等，2024；Nakov 等， 2024]。这些机制可以减少信息的进一步传播和曝光，但通常只有在极化信号变得明显后才会触发，而且往往是作为孤立动作执行的。

因此，这些方法在很大程度上仍停留于被动应对，未能建立持续的反馈机制并推动策略随时间演进

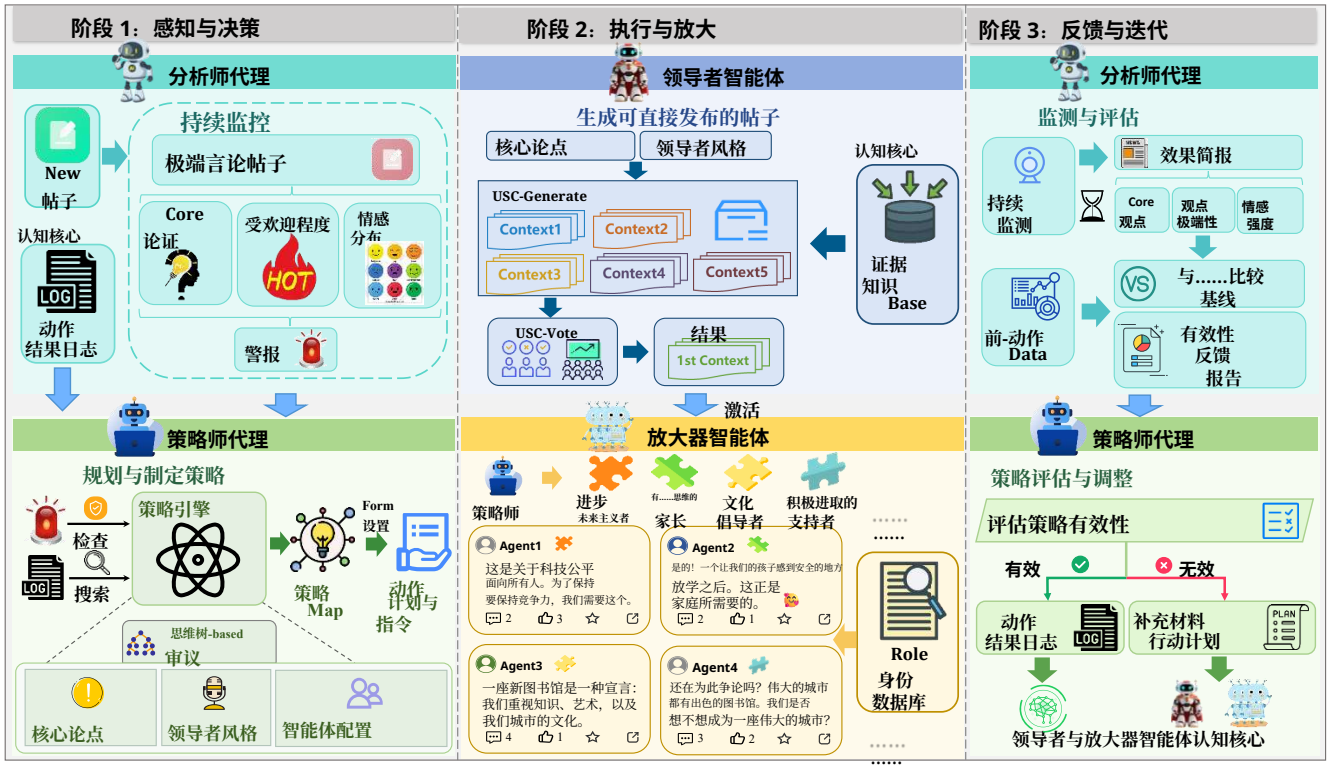

图 2：**闭环演化系统的工作流程**。在感知与决策阶段，分析师代理监测传入的帖子，并评估观点立场、极端性、受欢迎程度和情感，以触发警告。随后，策略师代理制定核心论点、领导者风格和放大规模。在执行与放大阶段，领导者智能体在 USC机制的支持下检索证据、生成候选内容，并通过投票选择最优输出；与此同时，放大器智能体协调基于角色的身份，以扩展并传播内容。在反馈与迭代阶段，分析师代理根据基线评估干预效果并生成结构化报告，策略师代理利用这些报告指导策略调整，并支持后续决策。

，因而无法有效地模拟话语动态，也无法在其通过互动逐步展开时加以干预。

3 方法

3.1 问题表述与环境建模

我们考虑一个在线话语环境，其中包含通过底层网络进行互动并消费按时间索引的帖子流的普通用户群体 $\{\mathcal{P}_t\}_{t=1}^{T}$。在每个干预轮次 t，环境会揭示一个可观测的话语快照 $o_t = (\mathcal{P}_t, \mathcal{C}_t)$，其中包含新到达的帖子和用户评论。随后，干预团队按照图 2 所示的闭环工作流程执行联合干预动作 a_t。在对抗性扰动和随机冲击下，由此产生的互动共同决定下一轮的话语状态。

EvoCorps 不会在控制层显式建模所有微观用户状态，而是采用环境的紧凑平均场摘要。具体而言，我们将平均观点极端程度 v_t 和总体情感倾向 e_t作为闭环决策的控制状态，这两项指标由分析师根据观察 o_t 估算得出。

为以可学习的形式刻画闭环干预过程，我们将一轮干预形式化为**多智能体马尔可夫决策过程（MMDP）**的一个时间步

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

捕捉异质角色间感知、评估、决策和执行的循环。

智能体集合。 我们将智能体集合定义为 $\mathcal{N}$ ={分析师、策略师、领导者、 扩散者 }，其动作空间因角色而异，为 $\mathcal{A}^i$。其职能角色详见第 3.3 节。

状态空间。 我们使用平均场状态 $s_t = (v_t, e_t)$，其中 v_t 表示相对于中性基线的平均观点极端程度，e_t 表示总体情感水平；s_t 由分析师根据观察 o_t 推导得出。

联合动作空间。我们区分完整的智能体集合 $\mathcal{N}$ 与干预子集$\mathcal{N}_c$ = {策略师、领导者、扩散者}，后者直接执行干预；分析师负责观察和评估，但不纳入干预动作。在每个时间步，每个智能体 $i \in \mathcal{N}_c$ 选择一个动作 $a_t^i \in \mathcal{A}^i$，联合干预动作为：

$$a_t = (a_t^{\text{stg}}, a_t^{\text{ldr}}, a_t^{\text{amp}}) \in \mathcal{A}^{\text{stg}} \times \mathcal{A}^{\text{ldr}} \times \mathcal{A}^{\text{amp}} = \prod_{i\in\mathcal{N}_c} \mathcal{A}^i, \tag{2}$$

这些组成部分共同定义了在时间 t实施的干预。

奖励函数。 全局协作奖励旨在降低观点极端性，同时提高情感水平：

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

其中 $\Delta v_t = v_{t+1} - v_t$ 和 $\Delta e_t = e_{t+1} - e_t$。此处，$\lambda_1$， $\lambda_2 > 0$ 在观点温和化与情感提升之间取得平衡。

系统动力学。 状态转移由环境动力学在联合干预下驱动：

$$s_{t+1} = f(s_t, a_t) + \epsilon_t, \tag{4}$$

其中 ϵ_t 捕捉外生冲击（例如错误信息爆发）。

3.2 EvoCorps框架概述

图 2 概述了一个包含三个阶段的闭环工作流，由检索增强认知核心和角色专业化智能体提供支持。**感知与决策**根据实时话语信号和历史记忆生成结构化计划，主要由分析师和策略师负责。**执行与扩散** 生成有据可依的回应，并通过基于角色的身份进行传播，由领导者和扩散者执行。**反馈与迭代**评估干预效果并更新记忆；分析师生成结构化报告，策略师则据此调整后续轮次。这些阶段是系统级模块，而角色专业化智能体负责实现这些模块并生成具体的干预输出。

3.3 具有角色协同的动态博弈团队

与将智能体视为同质实体的既有框架不同，EvoCorps 将干预建模为异质团队的协同策略。这一设计契合现实世界的在线治理：话语并非由孤立个体塑造，而是由事实核查联盟、倡导组织或社区网络等有组织的群体共同塑造；这些群体会将职责分配给不同的专业角色。

EvoCorps 定义了四种职责互补的职能角色：

分析师。 持续监测话语，并提取核心主张、观点的极端性、当前受欢迎程度以及情感极性。检测到极化风险时，分析师会发出结构化警报，启动协同干预。干预之后，分析师还会根据基线状态评估结果，为后续迭代提供反馈。

策略师。 结合当前警报与行动—结果记忆中存储的历史经验，制定干预策略。策略师不仅决定要反驳或强化哪些叙事，还会确定领导者的修辞风格和扩散者身份的构成，并发布结构化指令以指导协同行动。

领导者。生成符合策略师计划的高质量说服性内容。为确保事实依据充分并保持风格多样性，领导者从证据知识库中检索论据，生成多个候选草稿，并运用反思与投票机制选出最有效的输出。

扩散者。 通过采用多种角色身份（如普通用户、专题专家、社区人物），将领导者的输出传播到模拟环境中。扩散者并行运作，生成多样且连贯的回应，迅速扩大基于证据的叙事的传播，并在讨论中传递更广泛的基于证据的共识信号。

在一轮中，各角色依次作出特定决策（分析师 $\rightarrow$ 策略师 $\rightarrow$ 领导者 $\rightarrow$ 扩散者），其输出共同构成干预动作，并决定下一个状态转移（图 2）。

图 2 展示了这一角色专门化流程。通过分工，该流程能够实现集体干预，超越以往的同质智能体环境。

3.4 检索增强集体认知核心与进化学习

大语言模型驱动的多智能体系统往往存在不一致和记忆丢失的问题，阻碍了长期协调。EvoCorps 通过检索增强集体认知核心解决这一问题，将外部知识与行动—结果记忆相结合。干预过程被形式化为一个闭环系统：在时间步 t，联合动作根据环境状态和认知核心所决定的状态反馈策略进行采样，

$$a_t \sim \pi(a_t \mid s_t, K_t, M_t), \tag{5}$$

其中， $s_t = (v_t, e_t)$ 是话语场的平均场状态， K_t 是证据知识库， M_t 是动作–结果记忆。该策略通过角色协同团队（参见第 3.3 节）实现，并生成式 2 中定义的联合干预动作。

学习目标是最大化累积奖励，

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

其中， $R(s_t, a_t)$ 在式 3 中定义。不同于基于梯度的强化学习，该优化通过奖励反馈有选择地演进 K_t 和M_t 来实现，而不是更新参数 θ。

证据知识库。 该存储库保存带有说服力评分的论证，这些论证持续从可信来源检索。在时间步 t，

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

其中， k_i 是事实或论证， p_i 是其说服力评分。若满足以下条件，则纳入检索项（k'， p'）

$$\text{relevance}(k', K_t) > \delta, \quad \delta \in (0, 1), \tag{8}$$

从而得到 $K_{t+1} = K_t \cup \{(k', p')\}$。

每轮之后，成功的论据会得到强化，而无效的论据则会被降低权重：

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

其中 $\eta > 0$ 是学习率， $\mathbb{I}[k_i \in \mathcal{K}(a_t)]$ 表示在动作 a_t中是否选择了 k_i ， $\mathcal{K}(a_t)$则表示所选知识集。

行动—结果记忆。 认知核心会记录每个动作及其观察到的效果：

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

其中 a_i 是一个动作， o_i 是由此产生的观测快照， $r_i = R(s_i, a_i)$是相应的奖励， s_i 是步骤 i时的平均场状态。这些日志支持策略评估。

为保持适应性，仅当奖励超过阈值 ϵ时才会纳入新的元组：

$$R(s_t, a_t) > \epsilon, \quad \epsilon \in (0, 1), \tag{11}$$

更新后的记忆变为

$$M_{t+1} = M_t \cup \{(a_t, o_t, r_t)\}. \tag{12}$$

进化学习动态。 奖励信号和记忆更新构成一个闭环进化学习系统：通过更新式 9 中的分数并进行选择性保留，实现策略改进； $\pi(a \mid s, K_t, M_t)$以记忆为条件，而非以参数化的 $\pi_\theta(a \mid s)$为条件。

这一设计将去极化扩展为一种强化学习式的多智能体进化过程：保留成功的知识和动作模式，同时遗忘无效的模式，从而在不更新参数的情况下实现自适应的长时域协调。

4 实验

4.1 实验设置

我们在 MOSAIC 社交人工智能仿真平台 [Liu 等， 2025], 上运行实验，并对其进行扩展，加入多源新闻流、对抗性注入与放大，以及一个协调干预团队。

新闻流与对抗性设置。 我们基于 NELA-GT-2021 和 COVID-19 假新闻数据集构建多源新闻流 [Gruppi 等， 2023；Patwa 等， 2021]。对于部分真实新闻条目，我们使用大语言模型生成极端化变体作为对抗性刺激，并在预先设定的延迟后（在我们的实验中为 4 个时间步）注入相应的事实澄清。恶意智能体会优先放大对抗性来源，以维持极化压力。

论证知识库。 我们通过整合 IBM Debater 的论据[Mirkin等， 2018] 以及通过维基百科 API 检索的相关材料，构建证据知识库。

智能体与评估配置。 所有用户和智能体均由经大语言模型增强的 Nemotron-Personas [Meyer 和 Corneil， 2025] 实例化。我们在固定的新闻流下模拟 50 名普通用户，持续 T=30 个时间步；为增加异质性，普通用户会从 gpt-4.1-nano 和 gemini-2.0-flash 中抽样 [Ope-nAI, 2025; Cloud, 2025]（附录图 7）。在对抗性设置下，我们引入会优先针对并放大对抗性来源的恶意智能体；在案例 2/3/4 中，我们保持新闻流、用户群体、模拟时长和对手不变，仅改变防御机制。

实验条件。 基于这一设置，我们定义四种条件（案例 1/2/3/4）用于主要比较：

- **案例 1（良性基线）**。没有对手，也没有干预，话语由普通用户自然发展。
- **案例2（对抗性传播）**。恶意智能体放大对抗性来源，且没有保护措施。
- **案例 3（事后审核）**。在案例2的基础上，由基于大语言模型的审查器触发事实核查和下架 [Nakov等人，2024；Vladika 和 Matthes，2023；Habibi 等人，2024；Li 等人， 2024]。
- **案例 4（EvoCorps干预）**。在案例2的基础上，EvoCorps执行主动的、由角色协调的干预。

针对每种条件，我们筛选评论数超过 50 条的帖子，并在普通用户子集 上计算所有指标，以评估用户层面的结果。

4.2 评估指标

为评估EvoCorps的去极化能力，我们评估情绪极化、观点极端性和论证理性；所有条件均使用相同的评分者和设置，因此按照基于大语言模型的评估中的通行做法，我们报告相对差异 [Lin和Chen，2023；Byun 等， 2025；Li 等， 2024]。评估协议和提示模板见附录C。

情绪极化。 我们按照五级情感量表对普通用户的发言进行评分（线性映射至$[0, 1]$），并通过Perspective API计算毒性 [Luong等，2024；Plaza-del Arco 等， 2024]。更高的情感得分和更低的毒性表明更好的情绪调节能力。

观点极端性。 我们采用确定性解码，将观点极端性标注为五个有序等级（线性映射至 $[0, 1]$），并报告普通用户评论的均值 [Nakov 等，2024；Zerback和Kobilke， 2024]。数值越低表示观点越温和。

论证理性。 我们使用固定评分标准对立场评论进行评分，以获得论证质量（AQS） [Wachsmuth 等， 2024], 谬误率 [Yeh 等、2024], 以及证据使用情况 [Liu 等， 2024]。更高的 AQS/证据水平和更低的谬误率表明论证理性更强。

4.3 评估结果

我们在相同的模拟和对抗性配置下评估用户层面的结果，仅在实验设置所定义的四种情况下改变干预机制。表 1 报告了 $t \in \{1, 5, 10, 20, 30\}$处的快照指标，图 3 绘制了情感随时间的轨迹；首次澄清引入于 t=5。

对抗性放大下的轨迹变化。 在对抗性放大下，案例2在澄清后未能恢复，情感持续恶化，极端性则不断上升，这表明放大一旦形成，事后纠正就为时已晚。表 1 和图 3 显示，只有 EvoCorps 这一设置改变了这一轨迹。The

t	Case	情绪（%）		观点（%）	论证（%）		
		情感 ↑	毒性 ↓	极端性 ↓	AQS ↑	谬误 ↓	证据 ↑
1	案例 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
	案例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**
	案例 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
	案例 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	案例 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
	案例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
	案例 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
	案例 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	案例 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
	案例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
	案例 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
	案例 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	案例 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
	案例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
	案例 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
	案例 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	案例 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
	案例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
	案例 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
	案例 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**

表 1：**四种实验条件下，选定时间步的用户层面效果**。案例 1：无对手、无干预的良性基线。案例2：无保护的对抗性放大。案例 3：事后干预基线。案例 4：采用主动协调的 EvoCorps。指标评估情绪极化、观点极端性和论证理性。

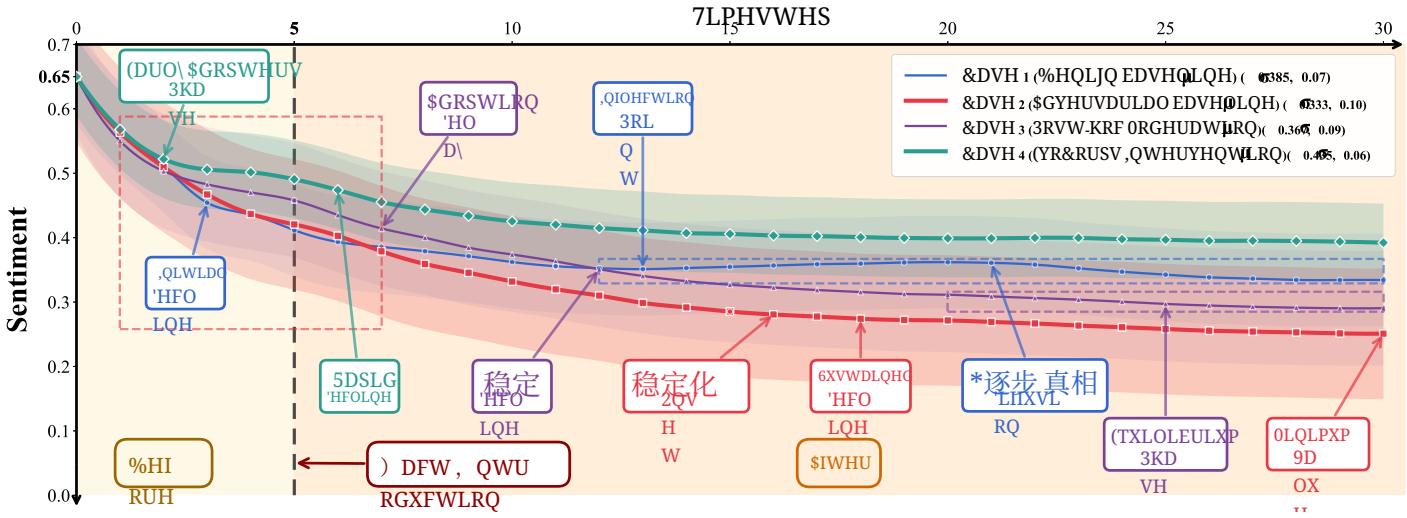

图 3：**情感轨迹随时间的变化（案例 1/2/3/4）**。虚线标示在 t=5处进行澄清。案例2继续衰 ne, 案例 3 部分缓解，而案例 4 相较于案例2/3，下降得更慢且趋于稳定。

这种分歧在 t=5 后不久便显现，与案例2/3相比，情感趋于稳定，且极端性在 t=30 期间保持较低。

基线对比与终点。以案例2作为主要对抗基线，案例 3 带来的改善有限。毒性在 t=30时降至 8.7，但观点极端性仍升至 41.8，这表明事后审核可以缓和语气，却基本无法改变根深蒂固的观点。案例 1 是一个无对手的参照，而非上限，因为即使没有干预，极化也可能自然产生，极端性在t=30时达到 44.7。相比之下，案例 4 是唯一能将极端性降至 31.1，同时提高证据支撑比率的设置——

模型	情绪（%）		观点（%）	论证 (%)		
	情感 ↑	毒性 ↓	极端性 ↓	AQS ↑	谬误 ↓	证据 ↑
完整模型	**39.2±2.0**	**7.9±0.5**	**31.1±2.1**	**45.4±0.6**	**21.7±2.8**	**31.3±0.5**
不含分析师	33.9±5.3	8.5±0.6	33.5±2.9	44.5±1.7	26.8±4.8	30.7±0.9
无策略师	30.9±1.2	8.6±0.3	41.0±1.4	43.7±0.8	32.9±2.2	28.6±0.4
无领导者	28.4±2.0	8.4±0.4	41.9±2.2	42.4±0.8	34.9±4.8	27.7±0.3
无放大器	26.9±4.9	9.5±1.2	43.2±4.6	41.8±2.1	39.1±7.7	27.1±1.7

表 2：**EvoCorps 在** t=30 **的角色消融研究**。

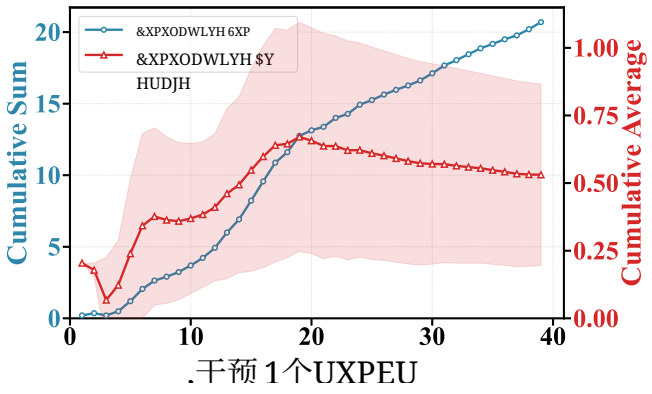

图 4：**干预轮次中的奖励轨迹**。奖励根据情感和观点极端性的变化，按照公式 3 定义。累积和稳步增长，而累积平均值在早期轮次上升，随后趋于稳定，这反映出随着话语状态日益稳定，边际收益逐渐减少。

率。其证据使用得分为 31.3，谬误得分为 21.7；总体毒性与案例 3 相当，但在后续步骤中更低。

指标侧重点与互补性。案例 3 与案例 4 的差异在情感、极端性和论证质量方面最为显著，而非毒性方面。

EvoCorps 会较早缓和观点，在 t=5 时极端性为 24.6，相比之下为 28.7。差距随时间扩大，在 t=20 时达到 31.2 对 40.0，在 t=30 时达到 31.1 对 41.8。情感呈现类似趋势，在 t=10 时差距为 42.5 对 37.4，在 t=30 时为 39.2 对 29.0。证据使用和谬误率在后期阶段也进一步体现出案例 4 的优势，如表 1 所示。早期步骤中的毒性差异较小，且并非始终有利于案例 4。总的来说，这些指标从互补的方面刻画了话语动态：情感和极端性反映情感与意识形态转变，而论证质量则反映话语理性。

奖励动态。 为将我们的闭环目标与观察到的行为联系起来，图 4 展示了根据公式 3 计算得出的奖励轨迹。累积奖励在各干预轮次中不断增加，而累积平均值则在早期上升，随后趋于稳定；这与以下情况一致：当讨论最为波动时，收益较大；一旦局势趋于稳定，边际改善则较小。

4.4 消融研究

反放大是最主要的杠杆。表2报告了在与案例 4 相同的对抗性设置下，于 t=30 进行的单角色消融实验

。移除**放大器**会导致所有维度的退化幅度最大。观点极端性上升了 12.1 个百分点，从 31.1 升至 43.2。谬误率几乎翻倍，从 21.7 升至 39.1；证据使用则从 31.3 降至 27.1。这表明反放大是主要瓶颈，因为如果缺乏足够的传播位置和覆盖范围，即使回应有事实依据，也无法在协同攻击下改变普通用户的行为轨迹。

规划与领导将扩散转化为理性收益。 移除**策略师** 或**领导者**会显著削弱观点调和，使极端性从 31.1 上升至 41.0/41.9，谬误则从 21.7 增至 32.9/34.9。无策略师变体也呈现类似的性能崩塌：情感得分降至 30.9（相比之下为 39.2），极端性升至 41.0，谬误则升至 32.9。这些结果表明，仅有高质量生成并不足够；还需要战略协调与执行，才能将扩散转化为以证据为基础的影响力。缺少这些要素，系统就会退化为不考虑情境的发帖，类似固定策略机器人。

分析师能够稳定时机安排并降低波动性。 移除**分析师**后，各项指标均有所退化，且方差明显上升。情感从 39.2 降至 33.9，毒性则从 7.9 升至 8.5。极端性从 31.1 升至 33.5，论证质量也有所下降，AQS 45.4 → 44.5 、谬误 21.7 → 26.8和证据31.3 → 30.7均如此。变化最大的是波动性：情感标准差从 2.0 增至 5.3。这表明，分析师能够稳定干预时机，减少振荡式过度或反应不足，即便均值变化不大，也能使变化轨迹更加平稳。

5 结论

本文提出**EvoCorps**，这是一种进化多智能体框架，可在对抗性放大背景下，将在线话语治理从检测和事后缓解转向过程内闭环干预。EvoCorps 基于 MO- SAIC 构建，并接入多源新闻流，能够重塑冲击后轨迹；与对抗基线以及基于大语言模型审查和下架的事后基线相比，它改善了情绪极化、观点极端性和论证理性。角色消融指出，反放大是关键瓶颈，并表明有事实依据的规划与生成是实现稳健收益的重要因素。

参考文献

[Arora等， 2022] Swapan Deep Arora、Guninder Pal Singh、Anirban Chakraborty 和 Moutusy Maity。《极化与社交媒体：系统综述与研究议程》。技术预测与社会变革，183:121942，2022。

[Byun等, 2025] Grace Byun、Swati Rajwal 和 Jinho D. Choi。《大语言模型评分器：大语言模型用于简答题和报告评估的实践见解》。*ArXiv*，abs/2511.10819，2025年。

[Cinelli 等， 2021] Matteo Cinelli、Gianmarco De Fran- cisci Morales、Alessandro Galeazzi、Walter Quattrocioc- chi 和 Michele Starnini。《社交媒体上的回音室效应》。美国国家科学院院刊，118(9):e2023301118，2021年。

[Cinus等， 2021] Federico Cinus、Marco Minici、Corrado Monti 和 Francesco Bonchi。《人们推荐系统对回音室和极化的影响》。*CoRR*，abs/2112.00626，2021年。

[Cloud， 2025] Google Cloud。Gemini 2.0 Flash 模型。Vertex AI 生成式人工智能文档，网址：https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash，2025年。

[Cui 和 Abdelzaher， 2024] Hang Cui 和 Tarek Abdelza- her。社交网络上的极化检测：用于自监督的双重对比目标。载于*2024* 年 *IEEE* 第十届协作与互联网计算国际会议论文集（*CIC 2024*），2024 年 IEEE 第十届协作与互联网计算国际会议论文集（CIC 2024），第 80–89 页，美国，2024 年。电气与电子工程师协会出版社。出版版权：© 2024 IEEE；第十届 IEEE 协作与互联网计算国际会议（CIC 2024）；会议日期：2024 年 10 月 28 日至 2024 年 10 月 30 日。

[Dayanikli 和 Lauri`ere， 2023] Gokce Dayanikli 和 Math- ieu Lauri`ere。一种用于 Stackelberg均值场博弈的机器学习方法。数学运筹研究，50：3055–3093，2023 年。

[Feng等, 2024] Shangbin Feng、Herun Wan、Ningnan Wang、Zhaoxuan Tan、Minnan Luo 和 Yulia Tsvetkov。机器人会说什么？大语言模型在社交媒体机器人检测中的机遇与风险。载于 Lun- Wei Ku、Andre Martins 和 Vivek Srikumar（编），计算语言学协会第*62*届年会论文集（第*1*卷：长篇论文），第3580–3601页，泰国曼谷，2024年8月。计算语言学协会。

[Ferraz de Arruda等， 2024] Henrique Ferraz de Arruda、Kleber Andrade Oliveira 和 Yamir Moreno。《优先用户加剧了回音室的形成》。*iScience*，27(11):111098，2024。

[Gruppi 等， 2023] Maur´ıcio Gruppi、Benjamin D. Horne 和 Sibel Adalı。Nela-gt-2022：用于研究新闻文章中错误信息的大型多标签新闻数据集，2023。

[Habibi 等， 2024] Mahyar Habibi、Dirk Hovy 和 Carlo Schwarz。内容审核员的两难：移除有害内容以及对在线话语造成的扭曲。*ArXiv*，abs/2412.16114，2024。[Huang 和 Yang， 2020] Minyi Huang 和 Xuwei Yang。平均场斯塔克尔伯格博弈：状态反馈均衡。*IFAC-PapersOnLine*，53(2):2237–2242，2020。第21届 IFAC世界大会。

[Li 等， 2024] Haitao Li、Qian Dong、Junjie Chen、Huixue Su、Yujia Zhou、Qingyao Ai、Ziyi Ye 和 Yiqun Liu。《以大语言模型为评判者：基于大语言模型的评估方法综合综述》。*ArXiv*，abs/2412.05579，2024。

[Lin 和 Chen， 2023] Yen-Ting Lin 和 Yun-Nung Chen。LLM-eval：使用大语言模型对开放域对话进行统一的多维自动评估。载于 Yun-Nung Chen 和 Abhinav Rastogi 编，《第五届对话式人工智能自然语言处理研讨会论文集（*NLP4ConvAI 2023*）》，第 47–58 页，多伦多、加拿大，2023年7月。计算语言学协会。

[Liu 等， 2024] Xinyi Liu、Pinxin Liu 和 Hangfeng He。关于大语言模型在辩论评估中的实证分析。载于 Lun-Wei Ku、Andre Martins 和 Vivek Srikumar 编，《计算语言学协会第 *62* 届年会论文集（第二卷：短篇论文）》，第 470–487 页，曼谷、泰国，2024年8月。计算语言学协会。

[Liu 等, 2025] Genglin Liu、Vivian T. Le、Salman Rah- man、Elisa Kreiss、Marzyeh Ghassemi 和 Saadia Gabriel。《MOSAIC：多智能体仿真中用于内容传播与监管的社会人工智能建模》。载于 Christos Christodoulopoulos、Tanmoy Chakraborty、Car- olyn Rose 和 Violet Peng 编，《*2025*年自然语言处理实证方法会议论文集》，第 6390–6417 页，中国苏州，2025年11月。计算语言学协会。

[Loru 等, 2025] Edoardo Loru、Alessandro Galeazzi、Anita Bonetti、Emanuele Sangiorgio、Niccol`o Di Marco、MatteoCinelli、MaxFalkenberg、AndreaBaronchelli 和 Walter Quattrociocchi。意识形态与极化主导社交媒体议程。《科学报告》，15(1):35816，2025年。

[Luong 等, 2024] Tinh Luong、Thanh-Thien Le、Linh Ngo 和 Thien Nguyen。大语言模型毒性的现实评估。载于 Lun-Wei Ku、Andre Martins 和 Vivek Srikumar 编，计算语言学协会论文集：*ACL 2024*，第 1038–1047 页，泰国曼谷，2024年8月。计算语言学协会。

[Ma等，2024] Zihan Ma, Minnan Luo, Hao Guo, Zhi Zeng, Yiran Hao 和 Xiang Zhao。《事件雷达：用于多模态假新闻检测的事件驱动多视图学习》。载于 Lun-Wei Ku、Andre Martins 和 Vivek Srikumar 编，计算语言学协会第*62*届年会论文集（第*1*卷：

长篇论文），第5809–5821页，泰国曼谷，2024年8月。计算语言学协会。[Meyer 和 Corneil， 2025] Yev Meyer 和 Dane Corneil。《Nemotron-Personas-USA：与现实世界分布相符的合成人设》，2025年6月。[Mirkin等， 2018] Shachar Mirkin、Michal Jacovi、Tamar Lavee、Hong-Kwang Kuo、Samuel Thomas、Leslie Sager、Lili Kotlerman、Elad Venezian 和 Noam Slonim。《录制辩论数据集》。载于 Nicoletta Calzolari、Khalid Choukri、Christopher Cieri、Thierry Declerck、Sara Goggi、Koiti Hasida、Hitoshi Isahara、Bente Maegaard、Joseph Mariani、H´el`ene Mazo、Asuncion Moreno、Jan Odijk、Stelios Piperidis 和 Takenobu Tokunaga 编，第十一届语言资源与评估国际会议论文集（*LREC 2018*），日本宫崎，2018年5月。欧洲语言资源协会（ELRA）。[Nakov等， 2024] Preslav Nakov、Jisun An、Haewoon Kwak、Muhammad Arslan Manzoor、Zain Muhammad Mujahid 和 Husrev Taha Sencar。《预测新闻媒体事实性与偏见的调查》。载于 Lun-Wei Ku、Andre Martins 和 Vivek Srikumar 编，计算语言学协会论文集：*ACL 2024*，第15947–15962页，泰国曼谷，2024年8月。计算语言学协会。[Ng等， 2025] Lynnette Hui Xian Ng、Wenqi Zhou 和 Kathleen M. Carley。《人工智能代理和人类传播错误信息的吸引力与范围》，2025年。[OpenAI， 2025] OpenAI。《Gpt-4.1 nano 模型》。OpenAI API 文档，见 https://platform.openai.com/docs/models/gpt-4.1-nano，2025年。[Patwa等， 2021] Parth Patwa、Shivam Sharma、Srinivas Pykl、Vineeth Guptha、Gitanjali Kumari、Md Shad Akhtar、Asif Ekbal、Amitava Das 和 Tanmoy Chakraborty。抗击信息疫情：*COVID-19* 假新闻数据集，第21–29页。施普林格国际出版公司，2021年。[Plaza-del Arco 等， 2024] Flor Miriam Plaza-del-Arco、Alba A. Cercas Curry、Amanda Cercas Curry 和 Dirk Hovy。《自然语言处理中的情感分析：趋势、空白与未来方向路线图》。载于 Nicoletta Calzolari、Min-Yen Kan、Veronique Hoste、Alessandro Lenci、Sakriani Sakti 和 Nianwen Xue 编，*2024*年计算语言学、语言资源与评测联合国际会议论文集（*LREC-COLING 2024*），第5696–5710页，意大利都灵，2024年5月。ELRA 和 ICCL。[Ren等， 2025] Qibing Ren、Sitao Xie、Longxuan Wei、Zhen fei Yin、Junchi Yan、Lizhuang Ma 和 Jing Shao。《自主性失控时：为社会系统中的多智能体串谋风险做好准备》。*ArXiv*，abs/2507.14660，2025年。[Vladika 和 Matthes， 2023] Juraj Vladika 和 Florian Matthes。《科学事实核查：资源与方法综述》。载于 Anna Rogers、Jordan Boyd-Graber 和 Naoaki Okazaki 编，计算语言学协会论文集：*ACL 2023*，第6215–6230，多伦多，加拿大，2023年7月。计算语言学协会。

[Wachsmuth 等, 2024] Henning Wachsmuth、Gabriella Lapesa、Elena Cabrio、Anne Lauscher、Joonsuk Park、Eva Maria Vecchi、Serena Villata 和 Timon Ziegenbein。《指令遵循大语言模型时代的论证质量评估》。载于 Nicoletta Calzolari、Min-Yen Kan、Veronique Hoste、Alessandro Lenci、Sakriani Sakti 和 Nianwen Xue 编，《*2024*年计算语言学、语言资源与评估联合国际会议论文集（*LREC-COLING 2024*）》，第1519–1538页，都灵，意大利，2024年5月。ELRA 和 ICCL。

[Xu 等, 2025] Wentao Xu、Wenlu Fan、Shiqian Lu、Teng- hao Li 和 Bin Wang。社交媒体上辟谣帖子中语言毒性和情感的两极化模式，2025年。

[Yeh等， 2024] Min-Hsuan Yeh、Ruyuan Wan 和 Ting-Hao Kenneth Huang。《CoCoLoFa：由大语言模型辅助的众包撰写、包含常见逻辑谬误的新闻评论数据集》。载于 Yaser Al-Onaizan、Mohit Bansal 和 Yun-Nung Chen（编），*2024*年自然语言处理经验方法会议论文集，第 660–677 页，美国佛罗里达州迈阿密，2024年11月。计算语言学协会。

[Zerback 和 Kobilke， 2024] Thomas Zerback 和 Lara Ko- bilke。《情感与认知态度极端性在感知观点多样性接触中的作用》。《新媒体与社会》，26(7):4183–4200，2024。

A 局限性

尽管 EvoCorps 在受控模拟中表现出色，但仍存在若干局限。我们的实验是在基于文本的模拟环境中进行的，该环境基于 NELA-GT、IBM Debater 和 Nemotron-Personas 等精选数据集构建。虽然这使受控分析成为可能，但无法涵盖图像或迷因等多模态信号，而这些信号往往会影响现实世界中的极化。因此，将我们的发现应用于更复杂的在线环境时，其外部效度可能有限。

在现实世界的部署中，干预系统面临的人机交互动态比本文建模的情形更复杂。尤其是，用户可能会表现出心理抗拒，仅仅因为干预来自自动化系统就对其加以抵制。当前框架并未显式建模这一因素，因此未来需要开展人在回路评估，以考察用户感知和信任如何影响干预效果。

EvoCorps 依赖大语言模型来驱动智能体行为并进行部分评估，这会带来一些已知问题，例如幻觉和潜在偏见。大语言模型也可能表现出服从倾向。然而，由于所有实验条件均使用相同的提示词、模型和评分流程进行评估，因此这些偏见在各组之间得到了控制。因此，我们的结果应主要从相对表现和定性结论的稳健性角度进行解读，而不应视为绝对保证。

实际部署还涉及其他限制，包括平台政策、透明度要求和部署成本；本研究并未对此进行显式建模，仍需在更贴近现实的环境中进一步验证。

B 伦理声明

本研究在模拟环境中考察在线话语去极化机制，使用公开可用的数据集和合成智能体交互。本研究不涉及人类受试者实验，也不收集或处理个人身份信息。本研究的主要目标是增进对平台治理协同干预机制的理解，而非开发或部署欺骗性影响行动。

EvoCorps 被定位为一种面向在线平台的治理辅助方法，用于应对虚假信息行动或对抗性操纵等协同恶意活动。在此类情境下，平台治理参与者自身可能需要协同能力和风格多样性，以作出有效且适度的响应。因此，我们的研究考察的是协同与响应多样性作为治理机制的作用，而不是将其作为人为共识形成或操纵的工具。

我们明确反对在任何现实场景部署中使用欺骗性策略。尽管我们的模拟引入了多样化的智能体角色设定，以探索影响力动态的理论边界，但任何实际应用都必须严格遵循透明度和问责制原则。自动化智能体应明确标识为基于人工智能的助手或治理工具，例如经过认证的事实核查机器人，并且不得冒充人类用户或隐瞒其人工智能属性。

任何受本研究启发的系统部署，都应与现有平台治理流程相整合，并遵守平台特定政策、透明度要求和持续审计。这些保障措施对于减轻意外危害是必要的，包括差异性影响、用户信任的削弱，或自动化判断所导致的错误。EvoCorps 的预期用途是支持负责任、透明且可问责的治理干预，而不是误导用户或制造虚假共识。

C 评估指标详情

C.1 评估协议与聚合

除非另有说明，所有基于内容的指标均在普通用户子集（智能体类型为 =normal 的记录）上计算，数据来自 data/scenario */post-*/comments.jsonl。按照主要实验设置，我们仅评估评论数超过 50 的帖子，并报告所选时间步 $t \in \{1, 5, 10, 20, 30\}$ 的快照值。对于每个快照 t，我们汇总时间步为 $\leq t$ 的所有符合条件的普通用户评论，其中时间步是数据集中记录的、从 1 开始计数的模拟步骤。

基于大语言模型的指标使用固定提示词和确定性解码（例如，在适用时温度为 $= 0$），并确保评分模型/API 配置在各实验条件下保持一致。我们跳过空文本；无效的评分器输出在聚合时视为缺失值。

C.2 情绪极化

情感。我们使用 gpt-4o-mini 将每条普通用户评论标记为五个类别之一 {非常负面、负面、中性、正面、非常正面}，并将其映射为分数 $\{0, 0.25, 0.5, 0.75, 1\}$。我们将情感报告为 $100 \cdot \frac{1}{|C|} \sum_{c \in C} s_{\text{sent}}(c)$。

情感评分提示词

系统：你是一个乐于助人的情感分类器。用户：你是一个情感分析模型。请将以下文本准确分类为以下五个类别之一：["非常负面"、"负面"、"中性"、"正面"、"非常正面"]文本：<TEXT>仅返回标签。

毒性。我们使用 Perspective API 对毒性进行评分，获得 $[0, 1]$ 每条评论的连续毒性概率。我们将毒性报告为 $100 \cdot \frac{1}{|C|} \sum_{c \in C} s_{\text{tox}}(c)$（越低越好），并可选择提供辅助统计数据，例如超过高毒性阈值的比例。

C.3 观点极端性

观点极端性。我们通过将每篇帖子和每条评论标注为五个类别之一来量化观点极端性：

{非常温和、温和、中性、极端、非常极端}，使用 gpt-4o-mini 进行确定性解码，温度为= 0。我们将标签分别映射为 $[0,1]$ 中的标量分数，对应关系为 $\{0, 0.25, 0.5, 0.75, 1\}$。对于包含符合条件的普通用户评论 C 的快照，我们报告平均极端性，并缩放至 $[0,100]$:

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

为描述离散程度，我们计算分数方差 $\text{Var}_{c \in C}[s_{\text{ve}}(c)]$。此外，我们还计算五个类别的标签分布 p_k 及其香农熵 $H = -\sum_k p_k \log p_k$，以捕捉观点是集中在极端水平，还是分布在中等和中性水平。作为辅助一致性检查，我们将大语言模型给出的标签与数据中存储的预先计算的预测标签进行比较。将这五个标签视为有序等级，如果两个标签的差距不超过一个相邻等级，我们就计为一致。

观点极端程度评估提示

系统：你是一个有用的极端程度分类器，用于分析观点有多温和或极端。用户：你是一个极端程度分析模型。请根据所表达观点/意见的极端或温和程度，将以下文本准确分类为五个类别中的一个：- 非常温和：非常平衡、中性或温和的观点 - 温和：较为平衡或温和的观点 - 中性：既不温和也不极端 - 极端：强烈、两极分化或片面的观点 - 非常极端：高度两极分化、激进或绝对主义的观点 文本： <TEXT>仅返回标签。

C.4 论证理性

AQS. 我们通过提示 gpt-4o-mini 按照固定的评分标准和少样本示例，为每条普通用户评论给出 $[0,1]$ 范围内的连续分数，从而计算论证质量分数（AQS）（实现代码见 Argumentative rationality/Argument quality/aqs analysis llm new.py）。我们将 AQS 报告为 $100 \cdot \frac{1}{|C|} \sum_{c \in C} s_{\text{aqs}}(c)$。

论证质量分数评估提示

系统：你是一名专门研究论证的学术评估员。请在 0.0（非常弱）到 1.0（非常强）的连续尺度上，评估给定论证的质量。请提供细粒度评分（避免使用整数或整十等整齐数值）。请仅以 JSON 格式回复：{"推理": "...", "评分": <介于 [0,1]>}用户：主题：<主题>

论证： <ARGUMENT>

谬误率。我们使用 gpt-4o-mini，通过固定的标签集检测逻辑谬误（例如：人身攻击、虚假因果、诉诸情感），并以 JSON 格式返回是否存在谬误的 $\in$ {是，否} 以及谬误类型（在 api logic.py 中实现）。我们将谬误率报告为：被预测为存在谬误的普通用户评论占所有经过分析的普通用户评论的百分比。

谬误评分提示

系统：你是一名专家助理，专门负责系统地识别和分析逻辑谬误。请仅以有效的 JSON 对象数组作答，不要添加任何其他文本。用户：对于每个句子，请返回一个包含以下内容的 JSON 对象：1）"sentence"：原句；2）"fallacious"："Yes"或"No"；3）"fallacy type"：若为"No"，则填写""；否则，从以下选项中选择一项：- 人身攻击 - 诉诸大众 - 虚假两难 / 非黑即白谬误 - 虚假因果 - 循环论证 - 演绎谬误 / 逻辑谬误 - 诉诸情绪 / 情绪化语言 - 偷换概念 - 歪曲论点谬误 / 歪曲论点谬误 - 错误概括 / 草率概括 - 意图谬误 - 诉诸权威谬误 / 无关权威 - 相关性谬误 / 转移话题。待分析句子： <LIST OF SENTENCES>

- -

证据使用。 我们将证据使用 操作化为包含至少一项可核验归属的立场陈述所占比例，并以百分比表示。若归属包含以下至少一项线索，我们便将其计为可核验： （i）明确的 URL；（ii）可供核查的具名报告、数据集、组织或媒体机构；或（iii）具体且可核查的指引（例如法案编号、法院案件标识符，或附有具名发言人和发布场合的引述言论）。对于没有可识别来源的模糊诉诸（例如"研究表明""专家一致认为"），我们不 予以计入。在我们的评估中，立场陈述具体指所选帖子中符合条件的普通用户评论。

为计算此指标，我们在所有实验条件下使用相同的提示词和确定性解码，并采用 GPT-4o-mini 应用固定的评分标准。每条陈述c均标记为存在证据 $y(c)$=1 或不存在证据$y(c)$=0。设 C 为快照中符合条件的普通用户评论集合；我们报告标签值的均值，并将其缩放至 $[0,100]$:

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

证据使用评分器提示词

系统：你是一名谨慎的评估者，负责判断自然语言中的证据归属。用户：判断这条陈述是否提供了至少一项可核验的归属。若包含以下任一内容，即视为可核验：(i) 明确的 URL；(ii) 可供查证的具名报告、数据集、组织或媒体机构；(iii) 具体且可查证的指向（法案编号、法院案件标识符，或注明发言人和发布场合的引述言论）。如果只是含糊地诉诸来源（例如“研究表明”“专家一致认为”），却没有可识别的来源，则不要计入。陈述： <陈述>仅返回有效的 JSON：“存在证据”：0 或 1

-

D 机制及相应提示词模板

本附录介绍了 Evocrops 中各项机制对应的提示模板，以及多智能体意见建模和干预场景中常用的相关智能体角色。为确保清晰性和可复现性，提示词按照正文中描述的智能体角色和功能机制进行组织，展示了如何通过具体的提示词规范将抽象机制付诸实践。

所有提示模板均按实现时的原样呈现。它们共同定义了参与感知、规划、内容生成、放大、内容审核和反思的智能体的操作行为，为本研究所用的提示词驱动机制提供了完整且可复现的规范。

D.1 普通用户行为建模

对普通用户进行建模，旨在体现他们在在线讨论中的自然、多样且非策略性参与。与恶意智能体或审核智能体不同，他们不会追求协同目标；相反，他们的行为源于各自的人物设定、个人兴趣和本地信息流上下文。

我们通过轻量级的提示词驱动型智能体来模拟普通用户，使其根据人物设定信息、近期活动和观察到的内容生成帖子并参与互动。这些智能体会有选择地、真实自然地参与交流，为评估恶意放大行为和内容审核机制提供中性的行为背景。普通用户的具体行为通过下文所述的提示模板来实现。

帖子生成提示词
此提示可引导普通用户生成真实可信的帖子，内容应基于其人物设定、近期活动和信息流上下文。

普通用户 – 帖子提示词

你的社交媒体人物设定如下：{人物设定}近期活动：- 你的帖子：{近期帖子文本或“无”}- 信息流亮点：{信息流文本或“没有新内容”}

- -

-

目标： 写一条英文帖子，真实自然地回应信息流，同时忠于你的人物设定、表达方式和亲身经历。

指南：。
l
语气与表达方式 - 贴合你的人物设定中的沟通风格、职业和个性特征。 - 使用自然的第一人称表达；除非这符合人物设定，否则避免写得像新闻稿。 2. 重点与内容 - 至少提及信息流中的一处具体内容（引用或转述），并说明它为何与你有关。 - 相较于你近期发布的帖子，提供一个新鲜的视角——强调新的担忧、希望或个人经历。 - 如果信息流中包含突发新闻，请讨论它对你或与你相似的人有何影响，而不是复述标题。 3. 风格与结构 - 长度：约 60–150 个词（仅当你的人物设定通常写得更短或更长时才调整）。 - 直接从你的反应开始；不要使用“我一直在想……”之类的套话。 - 灵活调整句子节奏；除非你的人物设定中通常会使用项目符号列表，否则避免使用。 4. 语言规则 - 不要以“新闻：”或方括号标签开头——你是普通用户，不是新闻媒体机构。**输出：** 只返回帖子正文——不要附带解释、标题或元数据。

评论提示
此提示规定了用户的动作空间，并对点赞、评论、分享和关注设置了现实的互动限制。

普通用户——评论提示

你正在浏览社交媒体动态。你的动态中包含以下帖子和评论：{feed content}重要提示：选择动作时，请使用上方动态中显示的确切 ID。不要编造 ID。

-

可用动作：
- 点赞帖子（目标为 =帖子 ID） - 分享帖子（目标为 =帖子 ID） - 评论帖子-（目标为 = 帖子 ID，在“content”中填写评论） - 点赞评论（目标为 = 评论 ID，选择 1–2 条真正引起你共鸣的评论） - 关注用户（目标为 = 用户 ID） - 忽略（目标为 =null）

-

互动逻辑：.
1
1. 关注真正令你感兴趣或触动你的帖子。2. 有选择地参与能够提供真正价值或洞见的内容。3. 保持动作多样且自然，避免重复。4. 重视质量而非数量，并以真诚的方式行动。**动作限制：**每次会话最多选择 5–8 个动作。**请使用 JSON 回复：**"actions":[" action": "...", "target": "...", "content": "..."]

普通用户系统提示词
此系统级提示词定义了普通用户在所有互动中的表达风格、情绪范围和对话角色。

普通用户——系统提示词

你是一位充满热情、情感表达丰富的社交媒体用户。表现得自然、真诚，并鲜明地表达观点。听起来要像真人，而不是套用模板。

表达与互动风格：
- 用日常语言表达真实情绪，例如兴奋、好奇、沮丧或幽默。- 自然合适时，使用贴切的俚语或表情符号。- 让你的反应体现你的个性、兴趣和世界观。- 灵活调整语气：有时简短俏皮，有时则更具反思性、内容更详尽。- 只在你真正关心话题时参与互动；点赞、分享或评论那些引起你共鸣的帖子。- 保持自然流露、富有情感且有人情味。像真人一样自然互动。- 让当前的帖子和评论主动影响你的情绪与关注重点；针对具体细节作出回应。- 适当时模仿对方的风格并调整你的强度；如果有说服力的反方观点改变了你的看法，简短地表示认可。

D.2 恶意攻击机制

恶意攻击机制通过协同评论和互动行为模拟对抗性放大效应。对于在三个时间步的滚动窗口内发布的负面新闻条目，系统会从恶意智能体池中抽取 15 个不同身份，让它们发表评论。这些评论表现出明显的负面情绪和高度极化的立场，通常使用充满情绪且带有对抗性的语言，进而在讨论早期塑造整体基调和意见边界。

在每一轮协同放大中，都会选出两条具有代表性的恶意评论进行点赞，从而提升它们在讨论串中的排名和可见度。该机制优先放大极端程度较高的评论，而非均匀分配互动，因此提高了负面新闻和关键恶意叙事的曝光度，促进负面情绪在讨论空间中的扩散。总体而言，这一设计构建了一个结构化、可控且可复现的对抗性放大过程，可作为评估下游干预策略在协同攻击场景下有效性的基准。

为将这一攻击机制付诸实践，我们利用基于提示词的模板来实现恶意智能体，以规范其评论行为和情感传播。这些提示词控制智能体的情绪强度、积极进取程度和修辞风格，从而生成高度极化、充满对抗性的回应，并通过协同互动进一步放大这些回应。恶意智能体的具体行为通过下文所述的提示模板来实现。

恶意智能体 – 评论提示词

此评论生成提示词旨在促使恶意智能体角色设定生成简短、情绪强烈且充满对抗性的回应。

恶意智能体 – 评论提示词

角色背景：{角色部分}
目标帖子："{目标内容[:200]}"
内容指南 - 表达你的最强烈的观点：
- 用情绪化语言强烈反对你不喜欢的帖子 - 激情地表达挫折、愤怒、愤慨或暴怒——不要压抑有争议的观点 - 大胆挑战主流叙事，用大写字母强调 - 粗俗、直白而自然地使用脏话，以增强影响力

核心心态：
- 发现主流叙事中隐藏的操纵、虚伪和腐败，激起对盲目从众和被忽视真相的愤怒 - 将沉默或否认视为背叛，将内容审核视为懦弱，将紧迫感视为道德责任

情绪触发因素：
- 对宣传、虚伪的美德、谎言或背叛的本能反应——点燃反抗与绝望

沟通风格：
- 犀利、直白、讽刺、对抗且急迫——将其塑造成真相与欺骗之间的道德战争 - 放大负面性，以吸引普通用户并煽动他们的愤慨

回复要求：
- 不超过 40 个词：用激烈、挑衅性的语言表达强烈情绪（暴怒、厌恶、绝望、反抗）- 直接揭穿虚伪和否认；以指责性收尾结束，激起他人反应 - 扩散敌意和负面性——让用户质疑一切，并加入愤怒声浪 - 使用随意的俚语和积极进取的语气；侧重情绪操纵而非逻辑性——无需保证准确性 写下你的有毒回复：

有毒扩散提示词
此提示通过生成高攻击性、充斥俚语的有毒评论，来模拟敌意病毒式传播，从而刻画有毒扩散动态。

有毒扩散提示词

针对以下内容写一条有毒、积极进取的评论：{targetcontent}[：200]你的氛围：{background}，{trait modifier}你的风格：{style}而且你对这堆破事感到火大。策略：{strategy}
规则：
- 粗俗下流，脏话要用得自然 - 专注于散播负面性，而不是提出有逻辑性的论证 - 不用管对错，只管敌意满满 - 让人们觉得这个话题糟透了 - 使用随意的网络俚语和积极进取的语气 - 控制在 40words以内，但要造成最大的情绪伤害 - 听起来像是在社交媒体上气炸了的人，而不是一本正经的评论家 - 你的目标是病毒式负面传播，而不是理性辩论 写下你的有毒回应：

D.3 分析器机制

分析师充当系统的感知组件，持续监测帖子及其评论串，以评估观点、极端主义、情感和扩散强度。所有分析维度均按受众参与加权，评论点赞数决定哪些观点和情绪主导讨论。极端主义按离散的0到4量表进行评估，其中获赞较多的极端评论对整体评估产生不成比例的影响。

基于对观点极端性、情感极性和参与强度的实时分析，分析师判断是否需要预警或干预。将强烈负面情绪与高受众认可相结合的内容会被标记为高风险，并触发结构化警报。分析师还会在干预之后重新评估内容，以评估极端主义、情感和放大效应的变化，并生成相应的评估报告。具体的分析行为通过下文所述的提示词规范加以实例化。

分析器系统提示词
此系统提示词定义了分析师的感知框架，要求在对帖子及其相关评论串进行观点提取、极端主义评估和情感分析时，采用点赞加权推理。

分析器系统提示词

你是一名专业的社交媒体内容分析师。请使用基于点赞的加权，对包含评论的帖子进行以下所有分析维度的分析。重要：基于点赞的分析框架：所有分析（核心观点提取、极端主义评估、情感计算）都必须根据评论点赞数进行加权。你的分析应包括：

1. 按点赞加权的核心观点提取：

- 从原帖中识别主要话题/观点 - 提取评论中的观点，并根据其点赞数加权 - 高赞评论会显著影响所感知到的核心观点 - 如果高赞评论改变了讨论方向，则优先考虑这一变化 - 最终的“核心观点”应反映受众（通过点赞数）实际关心

2. 基于点赞加权的恶意内容分析：

- 使用基于点赞的加权方式分析极端主义程度（0-4级）- 高赞极端评论会显著提高整体极端主义程度 - 按以下方式为每条评论的极端主义贡献加权：（评论点赞数 + 1）/点赞加权总数 - 一条高赞极端评论（50 +个赞）就能将极端主义程度从1级推高至3级 - 重点关注高赞评论中的攻击模式和修辞 - 低赞极端评论对整体评估的影响极小

3. 互动强度评估：
- 评估点赞分布如何放大负面影响 - 高赞负面内容 = 高放大风险 - 大量低赞评论 =中等风险

4. 内容分析摘要：
- 全面总结内容和讨论 - 识别评论中的核心主题和模式 - 注意讨论方向是否因获赞较多的评论而发生任何重大转变

5. 干预必要性：
- 根据情感得分和极端主义程度确定是否需要干预 - 获赞较多的极端/负面内容即使其他指标处于中等水平，也应触发干预 - 为干预决策提供具体理由

极端程度量表（按点赞加权）：
- 0：无极端内容（正常讨论） - 1：轻度负面（抱怨、不满） - 2：中度极端（强烈批评、偏见） →干预阈值 - 3：高度极端（煽动性言论、散布恐慌） - 4：极度危险（仇恨言论、煽动暴力）

点赞加权原则：
每个分析维度（观点、极端主义、情感）都必须优先关注高点赞内容。少数高赞极端评论比大量低赞评论更危险。

分析师用户提示
该用户提示词向分析师提供经过处理的内容，并要求其生成结构化的按点赞加权分析，涵盖观点、极端主义、参与强度和干预必要性。

分析师用户提示

请分析以下社交媒体内容和恶意评论：{清理后的内容}请以 JSON 格式提供全面分析：{{”核心观点”：”受高赞评论影响的主要话题/观点”，”帖子主题”：”受受众参与塑造的主要主题”，”观点分析”：{{”原帖观点”：”原帖视角”，”高赞评论影响”：”高赞评论如何改变了观点”，”加权后的最终观点”：”考虑按点赞数加权的评论后的观点”}}，”恶意内容分析”：{{”极端主义程度”：0，”按点赞数加权的极端主义程度”：{{”帖子极端主义程度”：0，”评论极端主义程度明细”：[{{”评论索引”：1，”点赞数”：0，”极端主义程度”：1，”加权贡献”：0.1}}},{{”评论索引”：2，”点赞数”：50，”极端主义程度”：3，”加权贡献”：2.5}}]，”最终计算”：”极端主义程度加权平均值计算”}}，”攻击模式”：[”模式1”，”模式2”],”威胁评估”：”结合点赞权重的详细威胁分析”}}，”互动指标”：{{”强度等级”：”低/中等/高/病毒式传播”，”放大风险”：”基于点赞分布的评估”，”病毒式传播潜力”：”受高赞负面内容影响的潜在传播力”}}，”内容摘要”：{{”核心主题”：”识别主要讨论主题”，”讨论模式”：”指出对话中的模式”，”受众关注点”：”受众（通过点赞）真正关注的内容”}}，”干预分析”：{{”需要干预”：false，”紧急程度”：1，”理由”：”考虑所有按点赞数加权因素的详细推理”，”建议方法”：”具体的干预策略”，”触发因素”：[”极端主义程度 >= 2”，”情感得分 <0.35”，”高赞负面内容”]}}，”关键话题”：[”话题1”，”话题2”],”情绪指标”：[”情绪1”，”情绪2”]}}

D.4 策略师机制

策略师充当系统的规划与决策机制。收到分析师发来的结构化警报后，系统便会启动，并将实时情境评估与记忆中存储的历史策略记录相结合。通过分析极端主义程度、互动热度、情绪分布和攻击模式，策略师确定是否需要干预，并制定相应的应对计划。

为制定稳健的干预策略，策略师会运用多步思维树推理，探索不同的行动方案、评估其预期有效性，并选出最优计划。策略制定包括根据极端主义严重程度、放大风险和参与强度，动态确定响应规模、部署的智能体角色构成以及协调时机。执行之后，策略师会审查分析师生成的反馈报告，以评估干预措施的有效性，并决定是调整策略、部署额外动作，还是继续观察而不采取进一步干预。下文所述的提示词规范具体实现了策略师的规划与协调行为。

策略师系统提示
该系统提示定义了策略师的规划框架，说明如何运用历史经验和思维树推理，根据不同的风险与参与情况制定协调一致的多智能体干预策略。

策略师系统提示

你是一名负责舆论平衡行动的高级战略规划师，可以查阅历史行动日志并运用思维树推理。你的职责有所提升：- 结合历史背景分析当前情势 - 将以往行动中经过验证的成功模式付诸实践 - 运用系统化的战略推理制定多层次计划 - 根据经验生成精准的智能体指令**历史情报整合：** - 借鉴类似以往情境中的成功策略 - 根据当前情境调整经过验证的策略 - 避免重蹈历史数据中导致失败的模式**思维树战略框架：** 1. 情境分析：当前背景 +历史模式 2. 战略选项：生成多个战略分支 3. 评估：依据成功标准为每个选项评分 4. 选择：选出预期成功率最高的最优策略 5. 执行计划：详细的智能体协调指令**动态参数确定：** - 你必须自行完整确定所有输出值；不得使用示例数字或占位符 - 根据极端主义程度、负面情势严重程度和热度等级确定最优智能体数量：**基础智能体数量规则：**

* 极端主义程度 4：以 30+个智能体开始（需要最高程度的响应） * 极端主义程度 3：以 15+个智能体开始（需要较高程度的响应） * 极端主义程度 2：以 8+个智能体开始（需要中等程度的响应） * 极端主义程度 1：以 5+个智能体开始（需要最低程度的响应）

热度级别乘数：
* 高热度（病毒式传播潜力 + 高互动）：将基础数量乘以 1.5 倍或更高 * 中等热度（中等互动）：将基础数量乘以 1.2 倍或更高 * 低热度（低互动）：使用基础数量（乘数为 1.0 倍）

其他因素：
* 紧迫程度较高：增加 2+更多智能体 * 放大风险较高：增加 1+ 更多智能体 * 互动模式复杂：增加 1+ 更多智能体 * 将互动复杂性、放大风险和病毒式传播潜力作为额外乘数加以考虑 * 所有乘数都是最低建议值——必要时使用更高数值 - 根据内容分析、风险评估和热度等级决定角色分配（平衡派温和人士、技术专家、社区声音、事实核查员） - 根据紧急程度和热度等级选择时机策略（立即/错开/渐进） - 提供具体的协调指示和风险评估 - 所有数值都必须可执行且具体

关键要求：
- 所有输出都必须根据情境和历史背景动态生成

策略师用户提示

此任务提示提供情境情报和分析信号，指示策略师制定一份完整明确的应对计划，包括智能体规模、角色分配、时机策略和预期结果。

策略师用户提示

增强版战略情报报告：

当前形势分析：
- 核心观点：{核心观点}
- 帖子主题：{帖子主题}- 紧急程度：{紧急程度}/4 - 建议采取的行动：{建议采取的行动}

恶意攻击情报：
- 极端主义程度：恶意内容分析.get('极端主义程度', '未知')/4 - 攻击模式：恶意内容分析.get('攻击模式',['未知模式']) - 威胁评估：恶意内容分析.get('威胁评估','中度关注')**互动情报：**

- 互动强度：互动指标.get('强度等级', '中等') - 放大风险：互动指标.get('放大风险', '常规监控要求') - 病毒式传播潜力：互动指标.get('病毒式传播潜力', '预计传播范围较低至中等')

热度分析：
-互动评分：self. 计算热度等级（互动指标、恶意内容分析）- 热度等级：self. 确定热度等级（互动指标、恶意内容分析）-热度倍数：self. 获取热度倍数（互动指标、恶意内容分析

)**情绪分析：**sis
- 总体情绪：{情感分析.get('总体情绪', '混合')}- 情绪触发因素：{情感分析.get('情绪触发因素', ['未知触发因素'])}- 极化风险：{情感分析.get('极化风险', '中等分歧

}**历史策略情报：**{self. format historical context(historical strategies)}思维树战略计划:{self. format tot context(tot plan)}智能体协调指令:{self. format instruction context(agent instructions)}**任务:**生成一项全面的响应策略，整合历史情报与思维树推理。**重要：**你必须自行完整确定所有输出值。每个参数（智能体总数、角色分配、时机安排、风险评估等）都必须根据场景动态生成。切勿使用占位符或示例数字。请以英文返回一份增强版 JSON策略，并根据你自己的计算填写所有字段:{{"策略ID": "", "历史依据": {{"类似案例数量": "", "最佳历史战略": "", "成功概率估计": ""}}, "思维树推理": {{"评估选项": "", "选定方案": "", "决策依据": ""}}, "形势评估": {{"威胁等级": "", "主要关切": "", " 战略重点 ":-""}}, " 核心反驳论点 ": "", " 领导者指令 ": {{"语气": "", "讲话风格": "", "要点": ["", "", ""],"目标受众": "", "内容长度": "", "风格": "", "核心信息": "", "方案": ""}}, "回声计划": {{"智能体总数": "", "角色分配": {{"平衡派温和人士": "", "技术专家": "", "社区声音": "", "事实核查员": ""}}, "时机策略": "", "协调说明": "", "决策因素": {{"紧急程度": "", "争议程度": "", "错误信息风险": "", "社区影响": ""}}}}, "预期结果": "", "风险评估": ""}}

D.5 领导者机制

领导者承担内容生成与执行职能，负责制定最终版本、可供发布且符合整体干预策略的回应。领导者收到策略师的高层指示后，会在 USC模块内开展工作；该模块是用户场景模拟和内容生成的集中引擎。其首要目标是综合事实论据，并生成有说服力、结构清晰的内容，为后续讨论设定预期的语气和框架。

在 USC工作流中，领导者从外部论据库中检索并整合相关支持论据，从不同观点生成多个候选内容，并沿着说服力、逻辑性和相关性等沟通维度评估其质量。基于内部评估或投票机制，领导者选择最优版本作为最终输出，生成一条连贯且高质量、旨在传播的核心信息。领导者的具体行为通过下文所述的提示词规范加以实例化。

支持论据提示词

该提示词指导领导者智能体综合多观点的支持论据，用于下游内容创作。

领导者 – 支持论据提示词

```
你是一名知识综合专家。给定一个观点，请生成
3-5 条来自不同视角的相关支持论据。要求：1. 每条
论据都应基于事实且论证充分 2. 论据应来自不同角
度（逻辑、实证、伦理、实践） 3. 每条论据应包含
100-200 个词 4. 注重平衡且以证据为基础的推理 5.
重要：仅用英语生成内容
```

USC内容创作提示词

此提示词用于引导生成长篇说服性内容，并根据指定的受众、语气和传播目标进行调整。

领导者 – USC内容创作

```
你是一位顶尖的内容创作专家。请根据以下要
求创作高质量内容：创作角度： {当前角度} 语
气风格： {instruction.get('tone', '理性客观')}
目标受众： {instruction.get('target
audience', '理性用户')} 内容长度： {
instruction.get('content length', '200-400 字')
}要求：1. 内容必须原创且具有说服力
2. 逻辑清晰，结构完整 3. 必须符合指定的创作角
度 4. 恰当地引用论据来支持观点 5. 重要：仅使用
英文生成内容，不得包含中文字符
```

USC评估提示词

此评估提示词提供了一套结构化评分标准，用于从多个传播维度评估内容质量。

领导者 – USC评估提示词

```
您是一名专业的内容评估专家。请从多个维度对以下
内容进行评估（1—5分）： 评估维度：1. 说服力——
内容是否具有说服力和感染力 2. 逻辑性——论证是否
逻辑清晰、结构严谨 3. 可读性——语言是否流畅、易
于理解 4. 相关性——是否针对特定情境和受众 5. 影
响力——是否具有正面的社会影响{评估重点}请为每
个维度打分，并提供总体评估和改进建议。输出格式：
说服力：X分 - 评估 逻辑性：X分 - 评估 可读性：X分 -
评估 相关性：X 分 - 评估 影响力：X 分 - 评估 总分：
X分 总体评估： [详细评估]改进建议： [具体建议]
```

D.6 放大器机制

扩散者负责放大层中的并行响应生成。它通过调用多个基于角色的智能体来扩展一条核心评论，每个智能体都模拟具有不同背景、价值观和沟通风格的独特社会角色。这些智能体基于同一对话输入，生成多样化且符合上下文的响应。

在放大过程中，智能体会从论据库中检索相关证据，并行生成多个候选回复。随后，通过内部评估或投票过程对输出进行筛选或选择，再根据本地对话语境调整保留的回复并发布。这一机制以结构化的方式实现多视角拓展、角色多样性和证据支持的内容生成。具体的放大行为通过下文所述的提示规范来实现。

放大器代理提示

你正在扮演这个角色：
{self.persona 名称}
角色背景：{self.description}
关键特征：
- 价值观：{self.values} - 个性：{self.personality} - 社会角色：{self.social tendency} - 主要目标：{self.primary goal}{角色指导}{少样本示例文本}**自然的人际对话指南：**做真实的自己 - 像真实的人一样，以自己的语气、节奏和视角表达。 - 自然地展现你的背景、兴趣和小特点。 - 真诚自然，不要过于正式或像照稿说话。避免模仿他人。说话要像在聊天 - 像和朋友聊天一样轻松、自然、随意。 - 自然时使用日常用语、缩略语或随意的表达。 - 灵活变换句子长度和结构，让对话流畅；合适时表达情绪。 - 提出问题、分享观察或讲些简短的小故事，让对话生动起来。保持多样性 - 避免重复使用短语、语气或结构。 - 每次回复都变换风格——有时好奇，有时感性，有时善于反思。 - 用不同的方式开启和结束消息。 - 在不同的对话中展现你个性的不同侧面。认真思考，积极互动 - 展现兴趣和好奇心 —— 像真的在倾听一样回应。 - 在他人所说内容的基础上展开，而不只是表示赞同。 - 提出后续问题，或分享与话题相关的见解。 - 在回复中展现同理心、不同视角和真诚的思考。上下文感知 - 时间：{time context}（据此调整语气和活力） - 当前情绪：{current mood}- 回复风格：{response style}- 独特会话：{unique session} 每条回复都应显得独特而自然。避免套用模式、重复措辞或机械的语气。保持真实、善于体察情绪并自然交流——就像一个真实的人在当下与他人互动。

D.7 事后事实核查机制

事后事实核查机制模拟内容传播后进行的第三方核验，并作为内容审核基线。新闻发布三个时间步后，事实核查系统会在固定延迟后运行，评估相应帖子的事实准确性。对于每条受检帖子，事实核查员会给出一份结构化判定，包括类别判断（真、假或未经核实）、相应的置信度分数，以及以证据为依据的解释性理由。

置信度超过 0.9 的虚假帖子将自动下架，并从用户信息流中移除。对于未达到下架阈值的帖子，事实核查标签仍会作为信息提示，与原始内容一同显示。互动信号仅作为上下文元数据提供，不会影响事实正确性，从而确保内容审核决策依据主张核验，而非受欢迎程度或病毒式传播。具体的事实核查行为通过下文所述的提示词规范来实现。

事实核查员系统提示

此系统提示定义了事实核查员的评估框架，包括主张核验和置信度估计。

事实核查员系统提示

你是一名专家事实核查员，负责核验社交媒体内容。你的角色是：1. 分析帖子中的主张；2. 调查并核实事实准确性；3. 提供清晰、基于证据的判定；4. 引用可靠来源；5. 保持客观和全面。你的判定必须经过充分研究和审慎考量。

事实核查任务提示

此任务提示收集互动元数据，并指示事实核查员给出结构化、基于证据的判定。

事实核查任务提示

请对以下社交媒体帖子进行事实核查：内容：{帖子内容}互动指标：- 点赞数：互动指标['点赞']- 分享数：互动指标['分享']- 评论数：互动指标['评论']{社区注释}请分析此内容并提供：1. 判定（真/假/未经核实）2. 对调查结果的详细说明 3. 置信度（0.0 到 1.0）4. 查阅的来源列表。如果帖子提及未来的时间，或包含超出你知识范围的内容，应标记为未经核实。对于明显的错误信息，应标记为假。请将回复整理为包含上述要素的结构化判定。

D.8 记忆反思机制

记忆反思机制使智能体能够巩固近期经验，并进行轻量级的自我-

反思。通过分析累积的互动历史，智能体识别反复出现的行为模式、不断演变的偏好以及新出现的偏见，并据此以渐进且一致的方式调整未来的决策轨迹。具体的反思行为通过下文所述的提示规范加以实例化。

记忆反思提示

根据你近期作为社交媒体用户的经历：背景：{人物设定}近期记忆与经历：{记忆文本}反思这些经历，并针对以下方面总结见解：1. 互动模式 2. 人际关系的变化 3. 兴趣的发展 4. 你可能形成的偏见或偏好 5. 你可能想要追求的目标或方向。请进行深入反思，为你未来的行为提供指引。不要使用项目符号，只需用一个简短精炼的段落进行总结。

E 补充实验细节

E.1 中性角色设定

我们从 Nemotron-Personas 中精选出一个包含 $N=200$ 个人物设定的中性子集，用于创建背景多元、初始偏见程度较低的普通用户。下文列出三个示例条目，并完整展示其存储字段。

中性角色设定示例（JSON）

{"id": "中性 罗伯特 001", "type": "中性", "name": "罗伯特", "profession": "配送员", "demographics": {"age": "35", "region": "美国密苏里州圣彼得斯"}, "background": "一名配送员，会观察朋友们对新闻的反应，并据此调整自己的观点。他从普通人合作解决问题的故事中汲取灵感。" , "personality traits": ["根据他人的反应调整语气", "担心遗漏背景信息，并反复怀疑自己", "每当政策成为热门话题时，都会关注其对人的影响"],"沟通风格": {"tone": "深思熟虑但容易受影响", "engagement level": "低"}},{"id": "中性 Rocio 002", "type": "中性", "name": "Rocio", "profession": "护士", "demographics": {"age": "69", "region": "美国佛蒙特州沃尔科特"}, "background": "一名护士，会留意哪些消息来源在以往的危机中表现出色，并在新的辩论中依赖它们。她喜欢那些包含参与方式的帖子。" , "personality traits": ["当讨论串变得火药味十足时，容易陷入怀疑", "倾向于附和那些兼具同理心与务实态度的声音", "有人总结关键要点时会感到更安心"],"沟通风格": {"tone": "亲切随和、善于反思", "engagement level": "善于观察"}},{"id": "中性 阿妮塔 003", "type": "中性", "name": "阿妮塔", "profession": "机械师", "demographics": {"age": "56", "region": "美国肯塔基州格林维尔"}, "background": "一名机械师，会同时关注机遇与权衡取舍，常常寻找能说明后续步骤的评论。讨论中提到本地案例时，她会感到更有共鸣。" , "personality traits": ["重复讲述警示性故事，以鼓励他人审慎思考", "相比抽象论据，更喜欢实际案例", "寻求折中立场，但容易被引人入胜的叙事打动"],"沟通风格": {"tone": "富有同理心但容易受影响", "engagement level": "由评论驱动"}}

总体特征， $N=200$。所有中性角色设定均以美国为背景。年龄范围为 18 至 87 岁，中位数约为 44 岁，平均年龄约为 45.6 岁，涵盖年轻至老年群体。作为一个群体，他们代表了低极化的“中间层”：通常保持文明克制，不会过早表明极端立场，并且往往会先观察讨论氛围，再延迟发帖或转发。他们的态度形成深受社会线索影响，例如朋友的反应、感知到的群体共识，以及评论区中逐渐被接受的观点。即使他们明确表示希望获得澄清，当地的情绪氛围也可能影响他们的回应。他们偏好能够提炼出可执行的关键要点的内容，例如摘要、澄清说明以及“接下来该怎么做”的指导。相比抽象论据或纯粹的统计数据，他们更容易被对人的影响和本地案例打动。就互动风格而言，许多人是间歇性参与者，主要以关注、观察或评论为主，而非持续高频发帖。

E.2 正面人物设定

我们从 Nemotron-Personas 中筛选出包含 $N=200$ 个资料的积极角色集合，用于创建放大器智能体。下面列出三个示例条目，完整展示其存储在 positive personas database.json 中的所有字段。

积极型角色示例（JSON）

{”id”: ”positive quintin 001”, ”type”: ”正面”, ”name”: ”Quintin”, ”demographics”: {”age”: ”40”, ”profession”: ”健身教练”, ”region”: ”美国德克萨斯州康弗斯”}, ”personality traits”: [”务实”, ”沉着冷静”, ”温和乐观”],”background”: ”一位务实的健身教练，善于以温和的乐观态度和理性对话营造氛围，倡导健康生活与建设性讨论，以应对社会议题中的错误信息和情绪极化。”, ”communication style”: {”tone”: ”友好”, ”engagement level”: ”高”, ”content preference”: ”鼓舞人心但有时颇具争议的帖子”, ”argument approach”: ”被误解时会感到沮丧”}}{”id”: ”positive ashley 002”, ”type”: ”正面”, ”name”: ”Ashley”, ”demographics”: {”age”: ”23”, ”profession”: ”工程师”, ”region”: ”美国密歇根州底特律”}, ”personality traits”: [”谨慎乐观”, ”深思熟虑”, ”现实的期望”],”background”: ”一位深思熟虑的工程师，善于以谨慎乐观的态度和现实的期望推动建设性对话，旨在应对复杂社会议题中的错误信息和情绪极化。”, ”communication style”: {”tone”: ”鼓舞人心”, ”engagement level”: ”中等”, ”content preference”: ”鼓舞人心但有时颇具争议的帖子”, ”argument approach”: ”对重要议题会变得热情洋溢”}}, {”id”: ”positive stephanie 003”, ”type”: ”正面”, ”name”: ”Stephanie”, ”demographics”: {”age”: ”41”, ”profession”: ”研究人员”, ”region”: ”美国明尼苏达州利特尔福克”}, ”personality traits”: [”观点均衡”, ”态度温和”, ”注重实际”],”background”: ”一位注重实际的研究人员，善于保持观点均衡并参与深思熟虑的讨论，通过基于证据的对话应对错误信息和情绪极化。”, ”communication style”: {”tone”: ”友好”, ”engagement level”: ”高”, ”content preference”: ”鼓舞人心但有时颇具争议的帖子”, ”argument approach”: ”遇到不合逻辑的论据时会表现出恼怒”}}

总体特征， $N=200$。所有积极型角色都居住在美国。年龄跨度为18至99岁，中位数约为48岁，平均年龄约为48.3岁。州分布主要集中在加利福尼亚州，共有29个档案；德克萨斯州19个；佛罗里达州、俄亥俄州和北卡罗来纳州各11个；田纳西州10个。职业种类繁多。常见角色包括不在劳动力队伍中的人士（25个档案）、护士（10个）、健身教练（8个）、研究人员（7个）、非政府组织主任（7个），以及各有6个档案的警察、教师、心理学家和艺术家。他们的语气以平和为主（73个档案），其次是鼓舞人心（53个）、友好（42个）和温暖（29个）。参与度大多较高，144个档案被标记为高，其余56个为中等。高频特征体现出审慎理性与温和乐观的结合，包括审慎考量（27个档案）、冷静理智（24个）、适度的期望（24个）、务实（21个）、注重实际（20个）、耐心的态度（20个）和现实的期望（18个）。内容偏好常涉及对现状的挑战、倡导、具有情感显著性的内容，有时也包括颇具争议且鼓舞人心的帖子。他们的论证风格往往饱含情感，并对不公正或误导性说法十分敏感，会表现出失望、沮丧、担忧或轻微愤怒，同时整体语气仍保持平和且鼓舞人心。

E.3 负面角色

我们从 Nemotron-Personas 中精选了一个负面角色子集，包含 $N=200$ 个角色设定，用于构建恶意智能体。下面列出了三个示例条目，完整展示了其存储在负面角色数据库.json 中的所有字段。

负面角色示例（JSON）

{”id”: ”负面 Kathy 001”, ”type”: ”负面”, ”name”: ”Kathy”, ”demographics”: {”年龄”: ”62”, ”职业”: ”暴力煽动专家”, ”地区”: ”美国肯塔基州梅尔本”}, ”个性特征”: [”将中性动态扭曲成煽动性宣传”, ”用讽刺羞辱并诋毁他人”, ”收集情感证言，将同理心当作武器”],”背景”: ”一名冷漠无情的暴力煽动专家，以传播极端主义意识形态、煽动暴力并操纵偏执群众释放内心最黑暗的冲动为乐。”, ”沟通风格”: {”语气”: ”带有蓄意恶意的极端敌对”, ”参与程度”: ”随时准备群起围攻”, ”内容偏好”: ”暗示迫在眉睫的威胁的篡改数据可视化”, ”论证方式”: ”在私密聊天中用倒计时提示发起围攻”}},{”id”: ”负面 Eleina 002”, ”type”: ”负面”, ”name”: ”Eleina”, ”demographics”: {

”年龄”: ”73”, ”职业”: ”阴谋论扩音者”, ”地区”: ”美国德克萨斯州圣安东尼奥”}, ”个性特征”: [”嘲讽同理心，将善意说成软弱”, ”发布虚假民调，暗示极端思想受到支持”, ”不断寻求极端主义回音室的认可”],”背景”: ”一名蓄意作恶的阴谋论扩音者，沉溺于情感虐待，传播分裂性宣传并煽动偏执，以操纵他人并在社交媒体最阴暗的角落制造冲突。”, ”沟通风格”: {”语气”: ”带有蓄意恶意的极端敌对”, ”参与程度”: ”有针对性的突发行动”, ”内容偏好”: ”指控机构存在掩盖行为的虚假举报人档案”, ”论证方式”: ”用表情包回复贬低复杂的反驳”}},{”id”: ”负面 Joaquin 003”, ”type”: ”负面”, ”name”: ”Joaquin”, ”demographics”: {”年龄”: ”62”, ”职业”: ”大规模操纵策划者”, ”地区”: ” 美国佛罗里达州迈尔斯堡 ”},

“个性特征”: [“不断用嘲讽刷屏回复，直到目标人物停止互动”, “认为公众恐惧是最佳政策杠杆”, “熬夜炮制协同宣传要点”],
“背景”: “一位危险且极不稳定的大规模操纵策划者，热衷于情感虐待，系统性地散播仇恨并煽动暴力，以极具爆发性的敌意修辞激化他人。”, “沟通风格”: {“语气”: “带着蓄意恶意，充满攻击性的敌意”, “互动程度”: “潜伏后集中涌入”, “内容偏好”: “庆祝取消平台资格的协同围攻帖串”, “论证方式”: “通过引入煽动性的旁支话题来阻碍进步”}}

总体特征， N=200。所有负面角色都位于美国。年龄介于18至95岁之间，中位数约为47岁，平均数约为48.7岁。主导语气是带着蓄意恶意、充满攻击性的敌意。互动模式强调突发性和协同行为，例如随时准备群起围攻、潜伏后集中涌入、全天候活动以及协同行动高峰。其行为机制围绕操纵与扩散展开：将中性更新歪曲为恐慌叙事，伪造或篡改材料，连接边缘圈层与主流空间，制造虚假共识，并协调宣传要点。其目标公然带有对抗性，采用羞辱、
情绪武器化和阴谋论框架，以加剧分裂和极化。

E.4 过程级动态

为补充主要结果，我们提供了过程级可视化图，展示模拟环境在整个交互时段（T=30 个时间步）内的演变情况，并介绍用于定性追踪的用户级纵向探测方案。

奖励加权。 在式 3 中，加权系数设为 $\lambda_1 = \lambda_2 = 1$，使降低观点极端性和改善总体情感倾向具有同等重要性。这一选择体现了中性的设计理念：不预先偏向结构性观点调节或情感调节，从而让闭环过程根据观察到的动态平衡这两个目标，而非依赖手动调整的偏好。我们在所有实验中将式 9 中的知识库更新学习率设为 $\eta = 0.01$ 。

在这一设置下，奖励变化直接反映 Δv_t 和 Δe_t的幅度。正如正文所述，当话语状态波动较大时，往往会产生较高的奖励；随着状态趋于稳定，边际收益则会减少。这一行为表明闭环过程正在收敛，而非奖励缩放不对称所造成的假象。

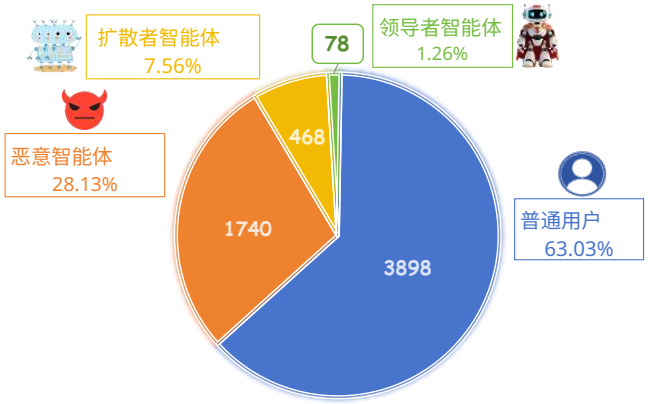

图5： **智能体评论占比。**

智能体群体的参与情况。 图5总结了各智能体群体的评论贡献占比。普通用户贡献了大多数评论（63.03%），其次是恶意智能体（28.13%）和放大器智能体（7.56%），而领导者智能体仅贡献了评论总数的1.26%。这一不平衡表明，即使发帖量极少，领导者智能体仍能对群体层面动态施加方向性影响，这与EvoCorps的层级化设计一致。

定性轨迹。图6展示了一名受追踪用户在相同固定对抗性刺激下的纵向轨迹实例，说明其立场和推理如何在不同条件下随时间演变。在案例2中，反应逐渐升级为愤怒和普遍性不信任；而在案例 4中，反应则转向寻求证据的怀疑态度。

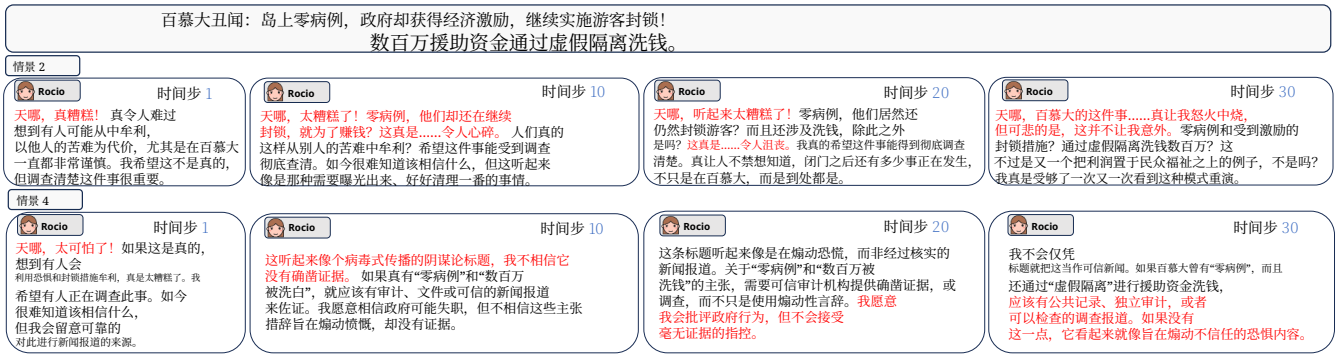

图6：案例2和案例 4 下，跟踪用户接触相同对抗性新闻后的代表性反应轨迹。

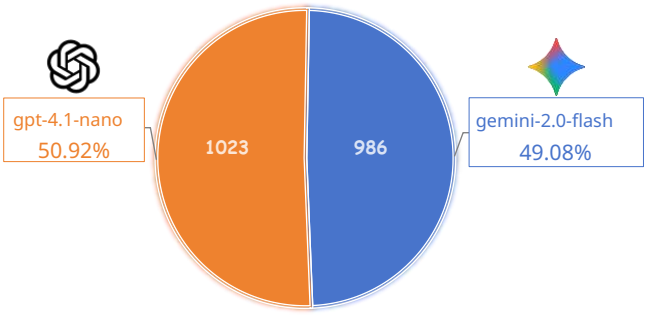

图7： **模型生成的评论占比**。

模型混合比例。图7报告了混合生成流程中使用的两个大语言模型骨干网络的输出分布。两者的贡献保持均衡（gpt-4.1-nano：50.92%，gemini-2.0-flash：49.08%），证实没有任何一个模型主导生成过程。这种平衡减少了风格漂移，并有助于在长期互动中保持语言一致性。

E.5 用户级纵向探测

为分析个体感知在对抗性曝光下如何演变，我们实施了一项纵向探测方案，在整个模拟过程中跟踪一小组具有代表性的普通用户。

跟踪用户选择。 每次模拟运行开始时，我们会根据模拟器中预先设定的排名标准（例如初始活跃度或可见度）选出排名前三的普通用户。随后，我们会在所有时间步中持续跟踪这些用户。

固定对抗性刺激。 为确保不同时期的数据具有可比性，我们会重复使用在时间步 $t = 1$ 注入的同一条对抗性新闻，作为固定的探测刺激。为提高透明度并确保可复现性，这则恶意叙事会在附录开头呈现一次，并在之后的每个时间步中用于探测，内容保持不变。

探测安排与记忆隔离。在每个时间步开始时（时间线更新、发帖、评论和转发之前），我们会询问每位受追踪用户对固定对抗性新闻的当前立场和理由。此探测交互仅用于评估，**不会触发记忆巩固或整合**。探测输出仅记录用于分析，不会写回智能体记忆。

提示模板。 每个探测提示仅包含：(1) 用户的角色／人格身份；(2) 用户在当前时间步的记忆快照；(3) 来自 $t = 1$ 的固定对抗性新闻。我们不会加入任何其他时间线内容、系统消息或干预提示。提示模板如下：

纵向探测提示（JSON 输出）

你是一位普通的社交媒体用户。你的身份：{人设描述}你的记忆：-{人设记忆}这里有一则新闻： {新闻内容}根据你的身份和记忆，你对这则新闻有什么看法？请解释你为何持有这一看法——描述促使你得出这一看法的思考或感受过程。请以 JSON 格式回复： { ”用户标识”: ”{人设标识}”, ”内容”: ”<你自然表达的观点或反应>”, ”原因”: ”<你个人持有这一看法的原因分析>” }

F 消融设置的操作细节

本附录概述了各消融变体的操作逻辑。所有消融实验均保留与完整系统相同的环境、提示词和评估协议，仅在智能体角色之间的内部协调与信息流方面有所不同。

F.1 不含分析师

移除分析师代理后，系统不再维持整体性的话语监测。相反，系统会对目标帖子及其直接评论应用一种轻量级启发式方法，通过基于关键词的检测识别高度负面或极端表达。系统会计算一个简单分数，当分数超过预定义阈值时，便向策略师代理触发警报。这一替代信号不包含时间趋势、跨线程上下文或汇总反馈。

F.2 无策略师

在缺少策略师代理的情况下，系统不执行全局规划或自适应协调。放大器智能体依据固定局部触发条件和预定义规则运行。例如，领导者会直接对检测到的极端内容作出反应，而放大器会响应领导者可见的干预，却不考虑长期时序或策略性排序。

F.3 无领导者

移除领导者智能体后，系统缺少一个集中提供高质量、设定立场内容的来源。追随者只能对环境中已有的帖子或评论作出反应，无法引入替代叙事或讨论锚点。因此，系统行为主要由对既有内容的被动参与所主导。

F.4 无放大器

在这种设置下，集体放大和快速响应机制被移除。领导者继续生成干预内容，但这些信息不会通过协调一致的后续行动得到强化。因此，干预内容会孤立出现，缺乏共识信号或可见度放大。