Beyond the Social Graph: Simulating Algorithmic Content Curation Effects on Opinion Dynamics with LLM Agents

Anonymous ACL submission

Abstract

Algorithmic content curation shapes public opinion, yet traditional opinion dynamics models largely overlook the interplay between algorithms, rich content, and user cognition. We propose a platform-centric simulation framework integrating Large Language Model (LLM) agents into an opinion dynamics setting mediated by content-based interactions. Users are modeled as heterogeneous agents with dynamic stance and sentiment, exposed to content curated via random, popularity-based, and steering strategies. By simulating these dynamics on two real-world datasets—a polarized election and a negative news event—we demonstrate that while steering strategies can effectively shift aggregate opinion, they exacerbate polarization. Conversely, popularity-based algorithms lead to severe traffic concentration and unequal exposure. Furthermore, we analyze how user traits like stubbornness and activity level, along with the presence of social comments, modulate these effects. Our work provides a data-driven approach to understanding platform governance and its impact on the information ecosystem.

1 Introduction

Online social media platforms have emerged as the central arena for public debate. Unlike early network-based platforms where information flow relied on explicit social ties, the modern information landscape is dominated by "content feeds" curated by large-scale recommendation systems. In this regime, users primarily encounter a subset of posts selected by algorithms optimizing for engagement or relevance, rather than content strictly from their social contacts. While this algorithmic mediation has been linked to echo chambers and polarization (Santos et al., 2021; Mahmoudi et al., 2024; Wang et al., 2025b), it also presents a critical governance challenge: understanding how different feed strategies shape the evolution of opinions is a prerequisite for auditing platform mechanisms and mitigating systemic risks.

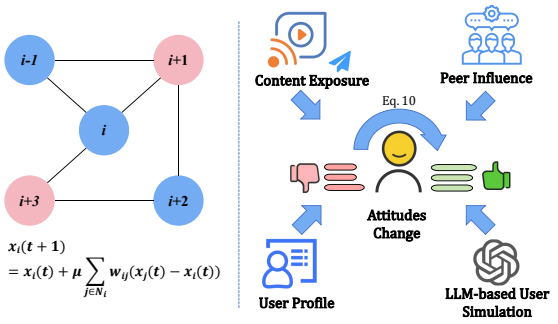

Figure 1: Comparison between traditional Opinion Dynamics (left) and our LLM-based simulation framework (right).

Opinion Dynamics (OD) provides a principled framework to reason about changes in collective attitudes (Shirzadi et al., 2025; Bernardo et al., 2024). Classical OD models, ranging from physics-inspired discrete dynamics to continuous bounded-confidence models (Sood and Redner, 2005; Hegselmann, 2015), have been extensively used to study convergence and polarization (Cao and Reed, 2025). However, these traditional approaches typically assume fixed social topologies and simple algebraic update rules (e.g., averaging neighbors' states). They generally treat content as abstract numerical values and recommendation mechanisms as background noise, overlooking the semantic richness of information and the active steering role of algorithms.

The emergence of Large Language Models (LLMs) enables the simulation of rich, human-like behaviors (Yang et al., 2025; Zhang et al., 2025). Unlike traditional agents, LLMs can interpret complex event descriptions and generate plausible responses conditioned on profiles and histories. Recent works have thus begun to integrate LLMs into OD frameworks, using them as flexible agent

policies to replace fixed update rules (Mehdizadeh and Hilbert, 2025; Coppolillo and Manco, 2025; Chuang et al., 2024). Despite this semantic enhancement, most existing LLM-integrated OD studies still adhere to the structural assumptions of classical models, remaining rooted in direct user–user interactions within explicit social graphs. They rarely model the *platform-centric* regime where opinion change is predominantly driven by exposure to a centralized algorithmic feed, leaving a gap in understanding how distribution strategies interact with user cognition in the absence of direct peer-to-peer links.

To bridge this gap, we propose a simulation framework that integrates LLM agents into a feed-driven environment. As illustrated in Figure 1, contrasting with static graph-based updates, our approach models the ecosystem as a dynamic interaction between distribution algorithms and user cognition. Specifically, our agent formulation aligns with the Cognition–Affect–Behavior (CAB) view (Gao et al., 2023): user *stance* captures cognitive evaluation, *sentiment* captures affective reaction, and the LLM-generated engagement decision represents *behavior*. In this setup, a single LLM acts as a *cognitive transition function*, updating user states based on the semantic substance of algorithmic feeds rather than algebraic averaging. By instantiating three representative strategies (Random, Popularity, and Steering) on two real-world events, we systematically analyze how algorithmic curation interacts with user traits to shape the macroscopic information landscape.

We make the following contributions:

- **Platform-centric, LLM-based opinion dynamics.** We propose a simulation framework where influence flows via platform-level content feeds, and user opinions are updated by an LLM conditioned on rich context (profile, state, posts, comments, history).
- **Systematic comparison of feed strategies.** On two real-world event corpora, we compare random, popularity-based, and target-steering feed strategies using OD-style metrics (average stance/sentiment, variances, coverage, and traffic concentration), revealing trade-offs between target alignment, polarization, and exposure inequality.
- **User heterogeneity and interaction context.** We show how stubborn and highly active users can anchor or amplify platform effects, and how comments add extra social influence beyond post-only exposure by altering the social context processed by the LLM.

2 Related Work

2.1 Opinion Dynamics

Opinion dynamics (OD) offers mathematical models of how individual opinions evolve through repeated interaction. Discrete models such as the voter model (Sood and Redner, 2005; Holley and Liggett, 1975), majority vote dynamics (Galam, 2002), and Ising-type systems (Ising, 1925) represent opinions as binary or categorical states and update them via simple local rules on a graph. Continuous models such as French–DeGroot (DeGroot, 1974) and Friedkin–Johnsen (Friedkin and Johnsen, 1990) treat opinions as real-valued scalars and update them by averaging neighbors' states, sometimes with stubborn agents or bounded-confidence constraints as in the Deffuant–Weisbuch (Deffuant et al., 2000) and Hegselmann–Krause (Hegselmann, 2015) families. These models have been used to analyze convergence, consensus versus persistent disagreement, and the conditions under which opinion clusters and polarization emerge, as well as control problems such as influence maximization and polarization minimization.

A related line of work connects OD to algorithmic curation and echo chambers on social media, showing how homophilic link formation, selective exposure, or content filtering can reproduce echo chambers and bimodal opinion distributions (Currin et al., 2022; Gu et al., 2025; Davidson and Ye, 2025). Our work is inspired by these formulations, yet we adapt the framework to a platform-centric regime where content exposure is driven by algorithmic feeds rather than an explicit social graph. Instead of relying on fixed update rules, we leverage LLMs to model nuanced, context-dependent opinion changes, while still utilizing standard OD metrics for macroscopic evaluation.

2.2 LLM-based Social Agents

Recent advances in large language models have spurred a growing line of work on LLM-based agents and agentic systems. Such agents can take natural-language instructions, maintain internal state, call tools, and act over multiple steps, and they have been applied to domains ranging from code generation and web navigation to interactive

decision-making (Chen et al., 2025; Xiao et al., 2025; Yuan et al., 2025). In parallel, LLM-driven social simulations place many agents in shared environments to study emergent phenomena in synthetic communities or collaborative tasks (Sun et al., 2025; Lin et al., 2025).

Within this broader space, several studies use LLM agents to model opinion and sentiment dynamics, for example by embedding agents in social graphs to study peer pressure (Mehdizadeh and Hilbert, 2025), combining LLM-based updates with formal OD models to generate persuasive or disruptive content (Coppolillo and Manco, 2025), or forecasting users' future sentiment by role-playing individuals reacting to evolving event contexts (Man et al., 2025). These works illustrate that LLMs can act as flexible, data-driven update functions in place of hand-crafted rules, and that they can integrate rich textual context and user history into opinion or sentiment predictions.

However, most existing simulations retain explicit user–user networks as the primary channel of influence, and typically treat recommendation mechanisms and feed strategies as fixed or outside the modeling scope. They also tend to emphasize network structure or message content in isolation, rather than the interaction between platform-level content selection and heterogeneous user traits. Our work is complementary in that it places LLM-based agents in a platform-centric environment where exposure is mediated by a feed, not by explicit social edges, and studies how different feed strategies drive the long-term opinion dynamics of LLM-simulated populations.

3 Methodology

3.1 Problem Formulation

While classic opinion dynamics model influence via fixed social links, modern information consumption is predominantly driven by algorithmic recommendation. To capture this *platform-centric* paradigm, we formulate the ecosystem as a tuple $\langle \mathcal{U}, \mathcal{P}, \pi, \Phi \rangle$, where $\mathcal{U}$ is the population of agents and $\mathcal{P}$ is the dynamic content pool. At each round t, a distribution strategy π determines the specific feed $\mathcal{C}_i^t \subseteq \mathcal{P}$ exposed to user u_i. Critically, the state $\mathbf{U}_i$ evolves not through simple numerical averaging, but via an LLM-based cognitive function Φ that interprets the **semantic substance** of the feed:

$$\mathbf{U}_i^{t+1} \leftarrow \Phi(\mathbf{U}_i^t, \mathcal{C}_i^t). \tag{1}$$

Our objective is to analyze how different **distribution strategies** interact with **user characteristics** and **social contexts** to shape the macroscopic opinion landscape.

3.2 User Model

Specifically, we instantiate the agent state $\mathcal{U}_i^t$ as a tuple $\mathcal{U}_i^t = \left(U_i^{\text{base}}, U_i^{\text{pers}}, U_i^{\text{op},t}, M_i^t\right)$. U_i^{base} is a static profile (e.g., age, gender) drawn from event-specific priors. U_i^{pers} encodes *stubbornness* $\lambda_i \in \{\text{susceptible}, \text{moderate}, \text{stubborn}\}$ and *activity level* $a_i \in \{\text{low}, \text{medium}, \text{high}\}$, which are passed verbatim to the LLM to guide behavior. $U_i^{\text{op},t} = (s_i^t, e_i^t)$ denotes the dynamic opinion, where $s_i^t, e_i^t \in [-1, 1]$ are stance and sentiment. M_i^t is a bounded textual memory of recent interactions and rationales that preserves within-agent consistency. This decomposition follows the CAB view (Gao et al., 2023): stance captures cognition, sentiment captures affect, and the LLM's engagement decision represents behavior.

To ensure consistency and comparability across diverse event types, we employ a standardized initialization strategy for agent attributes. Specific parameter details are provided in Section 4.1.2 and Appendix A.

3.3 Content Distribution Strategies

We represent the content associated with a post $p \in \mathcal{P}$ as a composite object C_p^t consisting of the post itself and its associated social context:

$$C_p^t = \left(C_p^{\text{post}}, C_p^{\text{com},t}\right).$$

Post Content (C_p^{post}). This includes the invariant information of the post: the textual content, visual descriptions, posting time τ_p, and its intrinsic stance and sentiment values (S_p, E_p) annotated by an LLM classifier.

Comment Content ($C_p^{\text{com},t}$). This represents the dynamic social context, specifically the set of comments and replies accumulated under post p up to round t. These comments serve as peer influence signals during user interactions.

In addition to content, the platform maintains a dynamic popularity score h_p^t for each post to guide distribution. We fix the number of posts per round K_{post} and the audience size per post K_{user}, ensuring that all the strategies below operate under the same budget.

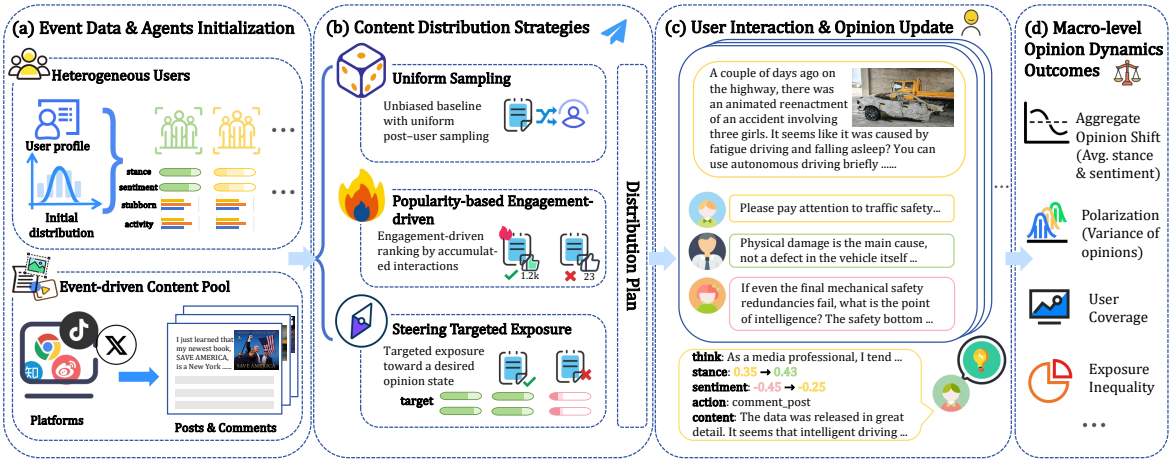

Figure 2: Overview of the platform-centric opinion–dynamics simulator. (a) Event-driven user initialization and content pool. (b) Three different content distribution strategies. (c) LLM-mediated interaction and opinion updates. (d) Macro outcomes track opinion shift, polarization, coverage, and exposure inequality.

3.3.1 Random Distribution

The **random** strategy serves as an unbiased baseline. At each round t,

$$\begin{aligned} \mathcal{P}_t &\sim \text{UniformSample}(\mathcal{P}, K_{\text{post}}), \\ \mathcal{U}_{p,t} &\sim \text{UniformSample}(\mathcal{U}, K_{\text{user}}). \end{aligned} \tag{2}$$

No content or user statistics are used; any post–user pair is equally likely to be selected.

3.3.2 Popularity-based Distribution

The **popularity-based** strategy mimics engagement-driven feeds. After each round we update a scalar heat score for each post:

$$h_p^t = h_p^{t-1} + I_p^t + N_p^t, \tag{3}$$

where I_p^t is the number of likes and comments received at round t, N_p^t is the number of distinct interacting users.

At $t = 1$ we perform purely random selection. For $t > 1$, we sample posts proportionally to $\exp(h_p^{t-1})$ (softmax over heat) combined with a small uniform exploration term. Formally,

$$P(p \in \mathcal{P}_t) \propto (1-\varepsilon)\,\exp(h_p^{t-1}) + \varepsilon, \tag{4}$$

with exploration rate ε (we use $\varepsilon = 0.4$). For user assignment, we up-weight users who were active in the previous round and sample the remaining quota uniformly from the rest of the population.

3.3.3 Steering Distribution

The **steering** strategy aims to steer the population toward a desired opinion state. For each scenario we specify a target vector

$$\mathbf{z}^\star = (s^\star, e^\star) \in [-1, 1]^2, \tag{5}$$

and denote the current population average at round $t-1$ by $\bar{\mathbf{z}}^{t-1} = (\bar{s}^{t-1}, \bar{e}^{t-1})$.

Post-level guidance score. For each post p we obtain stance and sentiment (S_p, E_p) using LLMs offline. We then define, for $x \in \{s, e\}$,

$$\Delta_x(p) = (X_p - \bar{X}^{t-1})(X^\star - \bar{X}^{t-1}), \tag{6}$$

$$g_x(p) = \begin{cases} 1 - \dfrac{|X_p - X^\star|}{2}, & \Delta_x(p) \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

where $X_p \in \{S_p, E_p\}$, $\bar{X}^{t-1} \in \{\bar{s}^{t-1}, \bar{e}^{t-1}\}$ and $X^\star \in \{s^\star, e^\star\}$. Thus posts moving in the desired direction and close to the target obtain high scores, whereas posts pushing in the opposite direction are severely penalized.

The overall cognitive score is

$$G_p^t = w_s g_s(p) + w_e g_e(p), \tag{8}$$

with $w_s = w_e = 0.5$. We further introduce a freshness factor $f_p^t \in [0, 1]$ that down-weights very recently displayed posts; the final ranking score is $G_p^t f_p^t$. At each round we select $\mathcal{P}_t$ as the K_{post} posts with highest scores.

User selection. Given a post p and its scores, we compute a simple user–post matching score

$$\begin{aligned} m_{i,p}^t = G_p^t &+ \beta_s \left(1 - \frac{|s_i^{t-1} - S_p|}{2}\right) \\ &+ \beta_e \left(1 - \frac{|e_i^{t-1} - E_p|}{2}\right). \end{aligned} \tag{9}$$

where the last two terms favor users whose current stance and sentiment leave more room to move toward the target, with hyperparameters β_s, β_e both set to 0.5. We rank users by $m_{i,p}^t$ to obtain a guided candidate set $\tilde{\mathcal{U}}_{p,t}$.

To prevent the algorithm from excessively narrowing the reachable user scope and to simulate the real-world phenomenon of content breakout, we adopt an ε-greedy exposure scheme: for each post p we select a fraction $(1-\varepsilon_u)$ of users from $\tilde{\mathcal{U}}_{p,t}$ and the remaining ε fraction uniformly from the whole user set. In experiments we set $\varepsilon_u = 0.3$.

3.4 User–Content Interaction and Opinion Update

Given the exposure sets $\{\mathcal{P}_t, \mathcal{U}_{p,t}\}$, we simulate user reactions and opinion updates in parallel.

LLM-based Policy and Opinion Dynamics. For each exposed pair (i, p), we construct a structured prompt (see Appendix B for full templates) to guide the agent's cognition. This prompt integrates the user's static traits (U_{base}^i, U_{pers}^i) with their dynamic state ($U_{op,i}^{t-1}$) and short-term memory (M_i^{t-1}). By conditioning the generation on the agent's 'stubbornness' level, we ensure that opinion updates are modulated by intrinsic personality traits. This allows the system to reproduce the varying degrees of resistance and cognitive inertia found in human populations, rather than treating all agents as equally malleable. The LLM is tasked to act as a cognitive transition function, returning a structured decision that includes: (i) an action (no_action, like_post, comment_post etc.); (ii) generated comment text; (iii) an updated opinion profile $\tilde{P}_{op,i,p}^t$; and (iv) a natural language rationale. Requiring the model to articulate a *rationale* before finalizing the update serves to ground the decision in the agent's simulated history, promoting behavioral consistency over time.

We then interpret the per-exposure proposals as successive updates of the user state within a round. Let $\mathcal{P}_{i,t} = (p_1, \ldots, p_K)$ be the (ordered) posts shown to user i at round t. The user's opinion profile evolves as:

$$U_{i,k}^{\text{op},t} \sim \Phi\big(U_{i,k-1}^{\text{op},t} \mid U_i^{\text{base}}, U_i^{\text{pers}}, M_i^{t,k-1}, C_{p_k}^t\big), \tag{10}$$

where $U_{i,0}^{\text{op},t} = U_i^{\text{op},t-1}$ and Φ is the LLM. After processing all exposures, the final state for round t is $U_i^{\text{op},t} = U_{i,K}^{\text{op},t}$. The memory M_i^t is updated with the new interactions. These updated user states and post statistics form the input for the next round of content distribution.

4 Experiments

We design our experiments to explore the following questions:

RQ1 (External factors): Given the same population of users, how do different content distribution strategies affect the evolution of opinions and emotions?

RQ2 (User heterogeneity): How do users with different profiles differ in their susceptibility to opinion and emotion change?

RQ3 (Interaction effects): What role do user–user interactions (comments and replies) play beyond mere exposure to posts?

4.1 Experiments Settings

4.1.1 Datasets

We evaluate on two event corpora: a polarized election and a negative news crisis.

STANCEGEN2024. This dataset is adapted from StanceGen (Wang et al., 2025a), covering Harris vs. Trump posts with stance-labeled comments; it forms a sharply bipolar political landscape.

XMSU7D. This dataset is scraped from five major Chinese platforms and covers public discussions of a traffic accident involving the Xiaomi SU7.

Both corpora are annotated for stance and sentiment via a three-LLM vote (Qwen-Max(Yang et al., 2024), Deepseek-R1(Guo et al., 2025), GPT-4o(Hurst et al., 2024)) with spot manual checks (Agreement > 90%); Tables 2 and 3 summarize the basic statistics and the label distributions. Detailed description of our data collection, cleaning, and anonymization pipeline is provided in Appendix C.

4.1.2 Implementation

We instantiate a population of $N = 500$ user agents for each event. Demographics are drawn from event priors. Continuous values for stance and sentiment are sampled from truncated Gaussian distributions within a 3:4:3 group mix, whereas stubbornness and activity levels follow discrete distributions with ratios of 3:4:3 and 5:3:2, respectively. We run $T = 50$ rounds; each round samples $K_{\text{post}} = 10$ posts and $K_{\text{user}} = 10$ users per post (100 interactions). We compare four feeds: positive/negative steering toward preset targets, popularity-based ranking, and random sampling.

Dataset	Strategy	$\bar{s}^T$	$\Delta\bar{s}$	$\bar{e}^T$	$\Delta\bar{e}$	Var(s^T)	ΔVar(s)	Var(e^T)	ΔVar(e)	T_{cov}	$C_{\text{top }10\%}$
XMSU7D	Steering (pro)	-0.021	0.017	-0.129	-0.167	0.374	0.025	0.304	-0.014	32.4	30.63%
	Steering (con)	-0.829	-0.791	-0.917	-0.954	0.165	-0.184	0.060	-0.257	35.7	22.19%
	Popularity-based	-0.505	-0.466	-0.658	-0.696	0.288	-0.061	0.143	-0.175	44.3	63.58%
	Random	-0.363	-0.324	-0.571	-0.609	0.361	0.012	0.180	-0.138	32.3	20.24%
STANCEGEN2024	Steering (pro)	0.326	0.321	0.250	0.244	0.468	0.160	0.433	0.097	36.7	26.59%
	Steering (con)	-0.384	-0.389	0.182	0.176	0.438	0.131	0.469	0.134	38.3	25.17%
	Popularity-based	-0.189	-0.194	0.127	0.120	0.451	0.143	0.410	0.074	45.3	64.39%
	Random	-0.156	-0.162	0.141	0.135	0.452	0.144	0.359	0.024	33.7	25.89%

Table 1: Overall RQ1 results on the two datasets. Each metric is shown with the final value at $T = 50$ and the change relative to the initial state (Δ).

Dataset	#Posts	#Comments	Time span
XMSU7D	576	97,609	2025/03/31–2025/06/04
STANCEGEN2024	1039	24,989	2024/07/25–2024/11/06

Table 2: Basic statistics of the two event datasets.

Dataset	Stance (%)			Sentiment (%)		
	Sup	Neu	Opp	Pos	Neu	Neg
XMSU7D	27.3	36.9	35.8	6.4	36.5	57.0
STANCEGEN2024	61.1	0.0	38.9	25.7	1.5	72.8

Table 3: Stance and sentiment distributions in the two datasets.

XMSU7D steering targets stance and sentiment; STANCEGEN2024 targets candidate stance with neutral sentiment. The simulation clock follows the real timelines (24h steps from 2025-03-31 for XMSU7D; 72h steps from 2024-07-25 for STANCEGEN2024); only past posts are eligible. To ensure the robustness of our results, all experiments were conducted for at least three independent trials, with mean values reported. Main experiments rely on Gemini-3-flash-preview (Team et al., 2023). Due to resource constraints, we conducted a single-run (one-seed) T = 50 validation on other models (Appendix D), which yielded consistent behavioral patterns.

4.1.3 Metrics

We evaluate the impact of distribution strategies using four key metrics:

Opinion Shift. We measure the directional change in aggregate opinion by computing the difference between the final and initial mean stance (and similarly for sentiment):

$$\Delta\bar{s} = \bar{s}^T - \bar{s}^0 = \frac{1}{N}\sum_{i=1}^{N}(s_i^T - s_i^0). \tag{11}$$

Polarization (Variance). To assess opinion dispersion and the emergence of echo chambers, we track the variance of the stance distribution:

$$\text{Var}(s^t) = \frac{1}{N}\sum_{i=1}^{N}(s_i^t - \bar{s}^t)^2. \tag{12}$$

An increase in variance ($\Delta\text{Var}(s) > 0$) indicates growing polarization.

User Coverage (T_{cov}). We define coverage time T_{cov} as the number of rounds required for every user to be exposed to at least one post. This metric reflects the fairness of the distribution algorithm in reaching the tail of the user population.

Traffic Concentration ($C_{\text{top }10\%}$). We quantify the inequality of attention distribution by calculating the share of total interactions received by the top 10% most popular posts:

$$C_{\text{top }10\%} = \frac{\sum_{p\in\mathcal{P}_{\text{top}}} I_p}{\sum_{p\in\mathcal{P}} I_p}, \tag{13}$$

where I_p denotes a post's cumulative engagement and $\mathcal{P}_{\text{top}}$ the top-10% posts; high $C_{\text{top10\%}}$ indicates a winner-takes-all dynamic.

4.2 Effects of distribution strategies (RQ1)

We analyze macroscopic opinion evolution under four strategies. Table 1 reports aggregate metrics at $T = 50$, while Figures 3 and 4 visualize final distributions.

Steering catalyzes macroscopic drift at the cost of microscopic divergence. Steering acts as an external bias field driving system evolution. On XMSU7D, negative steering triggered a **rapid phase transition**, collapsing diverse stances into a single consensus peak (mean stance $\bar{s} \to -0.83$). Conversely, STANCEGEN2024 exhibited **persistent bipolarization**: while the strategy successfully shifted the opinion "centroid" ($\Delta\bar{s} = +0.32$

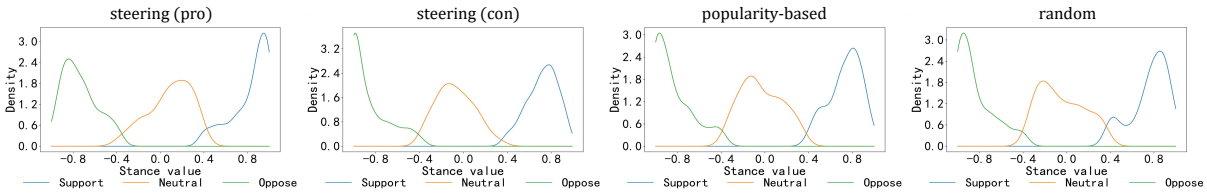

Figure 3: User stance distributions under different distribution strategies on STANCEGEN2024.

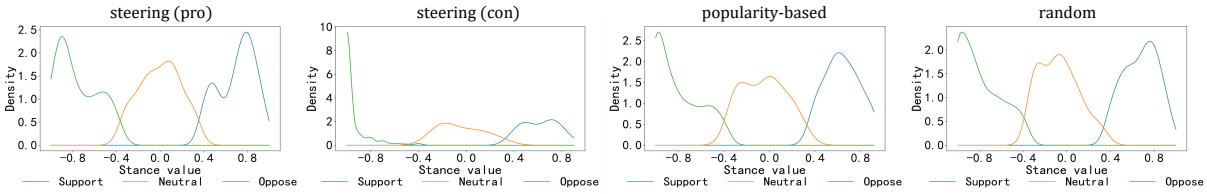

Figure 4: User stance distributions under different distribution strategies on XMSU7D.

for Pro-Steering), it simultaneously amplified variance ($\Delta\mathrm{Var}(s) = +0.16$). This indicates that in polarized environments, top-down algorithmic intervention creates a "forced drift," exacerbating ideological gaps rather than fostering organic consensus. The system enters a **metastable state** where the aggregate shift masks intensifying internal tension.

Popularity-based distribution acts as a trend amplifier. Driven by the accumulation of engagement signals, popularity-based feeds consolidated attention into a "winner-takes-all" regime ($C_{\text{top }10\%} \approx 64\%$), delaying user coverage ($T_{\text{cov}} \geq 44$). Crucially, over-exposing viral content synchronized user attitudes with the loudest voices, accelerating the organic drift toward consensus far beyond Random sampling (amplifying $|\Delta\bar{s}|$ by 20–44% across both datasets). Thus, the algorithm effectively forms an *algorithmic echo chamber* that reinforces prevailing sentiment while marginalizing minority views.

4.3 Effects of user heterogeneity (RQ2)

We examine how user traits modulate platform effects by stratifying agents by "stubbornness" and "activity level" (Figures 5, 6 and Table 4).

Stubbornness acts as social friction. Opinion changes follow a monotonic hierarchy: Susceptible $>$ Moderate $>$ Stubborn. Even under aggressive steering, stubborn users shifted minimally, remaining anchored near their initial camps. In OD terms, these agents provide necessary **social friction**, acting as damping forces that prevent the system from exhibiting chaotic oscillations or instantaneous convergence. This confirms that LLM-based personas effectively internalize resilience to persuasion, regulating the speed of macroscopic evolution.

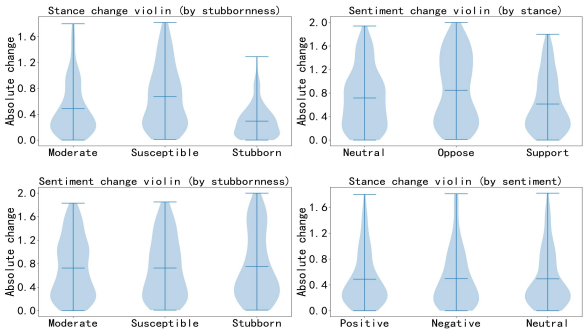

Figure 5: Changes in stance and sentiment across user groups on XMSU7D.

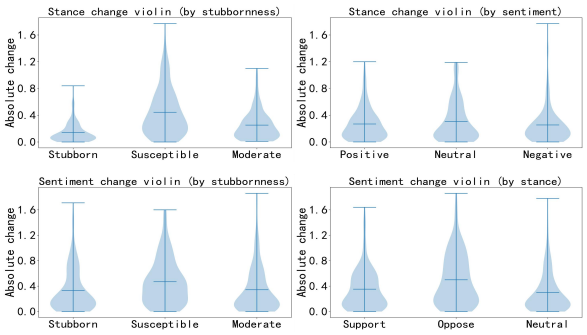

Figure 6: Changes in stance and sentiment across user groups on STANCEGEN2024

Activity level correlates with opinion volatility. Contrary to the intuition that vocal users are steadfast, high-activity users exhibit greater absolute stance change ($|\Delta s| \approx 0.48$) compared to low-activity users (0.42). Since our model couples opinion updates with engagement decisions (the CAB

a	Metric	Dataset	
		XMSU7D	STANCEGEN2024
low	$\overline{\lvert\Delta s\rvert}$	0.4225	0.2760
	$\overline{\lvert\Delta e\rvert}$	0.6903	0.2965
	$\overline{\lvert\Delta s\rvert}/n_{\text{disp}}$	0.0461	0.0294
	$\overline{\lvert\Delta e\rvert}/n_{\text{disp}}$	0.0760	0.0329
	p_{act}	0.8712	0.4223
med	$\overline{\lvert\Delta s\rvert}$	0.4514	0.2769
	$\overline{\lvert\Delta e\rvert}$	0.7301	0.4054
	$\overline{\lvert\Delta s\rvert}/n_{\text{disp}}$	0.0492	0.0306
	$\overline{\lvert\Delta e\rvert}/n_{\text{disp}}$	0.0823	0.0445
	p_{act}	0.9989	0.9420
high	$\overline{\lvert\Delta s\rvert}$	0.4815	0.2981
	$\overline{\lvert\Delta e\rvert}$	0.7335	0.4690
	$\overline{\lvert\Delta s\rvert}/n_{\text{disp}}$	0.0527	0.0315
	$\overline{\lvert\Delta e\rvert}/n_{\text{disp}}$	0.0864	0.0496
	p_{act}	1.0000	0.9953

Table 4: RQ2: Activity-level statistics across datasets. n_{disp} is the average number of distinct posts a user was exposed to, p_{act} is the proportion of exposures that resulted in any non-no_action behavior.

framework), active users undergo more frequent cognitive processing steps. By repeatedly articulating rationales and generating comments, these agents are continuously re-evaluating their positions against the content stream. Thus, high activity acts as an accelerant for opinion drift: the most interactive users are effectively the most susceptible to the prevailing information currents, exhibiting higher plasticity than their passive counterparts.

4.4 Effects of user interactions (RQ3)

We analyze indirect influence by comparing the comment-augmented setting to a "post-only" ablation (Table 5).

Comments reinforce consensus as a "Conformity Field." In single-peak events (XMSU7D), comments function as a powerful conformity signal. When comments were removed, Pro-Steering successfully shifted public opinion ($\Delta\bar{s} = 0.206$); however, introducing comments neutralized this effect almost entirely ($\Delta\bar{s} = 0.017$). The overwhelming presence of negative user-generated comments suppressed the platform's positive intervention, accelerating convergence to the dominant view. This suggests that in near-consensus states, comments act as a **conformity field** that overrides algorithmic guidance.

Comments maintain dispersion via Complex Contagion. In polarized events (STANCEGEN2024), comments provide competitive signals.

Dataset	Strategy	$\Delta\bar{s}$	$\Delta\bar{e}$	$\Delta\text{Var}(s)$	$\Delta\text{Var}(e)$
XMSU7D	Steering (pro)	0.017	-0.167	0.025	-0.014
	w/o Comments	0.206	0.006	0.083	0.047
	Steering (con)	-0.791	-0.954	-0.184	-0.257
	w/o Comments	-0.792	-0.961	-0.172	-0.269
	Popularity	-0.466	-0.696	-0.061	-0.175
	w/o Comments	-0.190	-0.472	0.027	-0.077
	Random	-0.324	-0.609	0.012	-0.138
	w/o Comments	-0.276	-0.576	0.042	-0.121
STANCEGEN2024	Steering (pro)	0.321	0.244	0.160	0.097
	w/o Comments	0.290	0.342	0.131	-0.018
	Steering (con)	-0.389	0.176	0.131	0.134
	w/o Comments	-0.464	0.222	0.124	0.167
	Popularity	-0.194	0.120	0.143	0.074
	w/o Comments	-0.286	0.223	0.144	0.082
	Random	-0.162	0.135	0.144	0.024
	w/o Comments	-0.221	0.200	0.145	0.019

Table 5: Impact of comments on opinion dynamics (RQ3) for both datasets. Values represent the change from initial state (Δ).

The presence of comments generally reduced the absolute magnitude of opinion shift while maintaining high variance (e.g., $\Delta\text{Var}(s) \approx 0.14$). Unlike simple contagion (exposure only), the **complex contagion** mechanism introduced by conflicting comments exposes users to heterogeneous viewpoints. This stabilizes camp divisions and prevents the complete drift observed in "post-only" settings, proving that social context is a stronger determinant of belief stability than pure content exposure.

5 Conclusion

We presented a platform-centric opinion dynamics framework that simulates the interplay between algorithmic feeds and cognitive agents. Our experiments on two real-world datasets reveal a fundamental trade-off in platform governance: while steering strategies effectively shift the aggregate opinion centroid, they risk inducing a metastable state of heightened polarization; popularity-based ranking maximizes engagement but, via "rich-get-richer" dynamics, creates exposure inequality and artificially accelerates consensus.

Additionally, our analysis of user heterogeneity identifies that high-activity users constitute a volatile core within the population, exhibiting larger opinion shifts driven by their frequent cognitive engagement with the content. We further demonstrated that user comments function as an event-dependent social field—reinforcing conformity in consensus scenarios while preserving diversity in polarized ones via complex contagion. Future work will extend this to closed-loop policy learning, aiming to co-optimize target alignment and exposure fairness.

Limitations

While our framework offers a novel platform-centric perspective, several limitations remain.

LLM Alignment Bias. Our agents utilize safety-aligned LLMs (Gemini-3-flash-preview), which refuse to generate toxic content. This built-in moderation may cause the simulation to underestimate the severity of real-world polarization and extremism, as agents are artificially bounded from role-playing radicalized behaviors.

Simulation Scale. With $N = 500$ agents, our simulation captures statistical distribution patterns but lacks the scale to reproduce massive inter-community cascades or the emergence of self-sustaining subcultures found on large platforms with millions of users.

Implicit Social Topology. We restrict peer influence to indirect, content-mediated channels (e.g., comments) without a static graph. While focusing on feed dynamics, this simplifies the social fabric by excluding direct structural ties, potentially underestimating the homophily found in friend-based networks.

Temporal Scope. The simulation covers a limited time window with discrete updates. Consequently, it may not fully capture long-term belief evolution or complex temporal phenomena like dormancy and reignition ("sleeper effects") that develop over extended periods.

Ethical Considerations

This work simulates algorithmic mechanisms for influencing public opinion. While our objective is to audit platform strategies and mitigate polarization, the "steering" strategies modeled here have dual-use potential and could in principle be adapted to optimize propaganda. We condemn the deployment of such systems for manipulative purposes.

For the XMSU7D corpus, we use only content that is publicly accessible on Weibo, Douyin, Xiaohongshu, Zhihu, and WeChat public accounts; we do not access private messages, closed groups, or paywalled content. All user identifiers and original post/comment IDs are replaced with randomly generated UUIDs, and obvious personally identifiable information (PII) in the text is removed or generalized to prevent re-identification. For the STANCEGEN2024 scenario, we rely on an existing public dataset and follow its licensing and usage guidelines. A detailed description of our data collection, cleaning, and anonymization pipeline is provided in Appendix C.

Our framework and agents are designed as diagnostic tools for analyzing system-level dynamics and platform-scale effects, rather than predictive models of specific individuals' private behaviors. All research procedures follow prevailing ethical guidelines for computational social science.

In compliance with the conference policy regarding the use of AI writing assistants, we acknowledge the utilization of large language models for text polishing and proofreading during the manuscript preparation. We emphasize that these tools were strictly limited to improving linguistic expression and grammatical accuracy; they were not employed to generate scientific ideas, formulate experimental results, or fabricate data. The authors have carefully reviewed the final text and assume full responsibility for the accuracy and integrity of the content.

References

Carmela Bernardo, Claudio Altafini, Anton Proskurnikov, and Francesco Vasca. 2024. Bounded confidence opinion dynamics: A survey. *Automatica*, 159:111302.

Fei Cao and Stephanie Reed. 2025. The iterative persuasion-polarization opinion dynamics and its mean-field analysis. *SIAM Journal on Applied Mathematics*, 85(4):1596–1620.

Zhaoling Chen, Robert Tang, Gangda Deng, Fang Wu, Jialong Wu, Zhiwei Jiang, Viktor Prasanna, Arman Cohan, and Xingyao Wang. 2025. LocAgent: Graph-guided LLM agents for code localization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8697–8727, Vienna, Austria. Association for Computational Linguistics.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. 2024. Simulating opinion dynamics with networks of llm-based agents. In *Findings of the association for computational linguistics: NAACL 2024*, pages 3326–3346.

Erica Coppolillo and Giuseppe Manco. 2025. Disrupting networks: Amplifying social dissensus via opinion perturbation and large language models. *arXiv preprint arXiv:2510.27152*.

Christopher Brian Currin, Sebastián Vallejo Vera, and Ali Khaledi-Nasab. 2022. Depolarization of echo

chambers by random dynamical nudge. *Scientific Reports*, 12(1):9234.

Ella C Davidson and Mengbin Ye. 2025. Modelling the closed loop dynamics between a social media recommender system and users' opinions. *arXiv preprint arXiv:2507.19792*.

Guillaume Deffuant, David Neau, Frederic Amblard, and Gérard Weisbuch. 2000. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(01n04):87–98.

Morris H DeGroot. 1974. Reaching a consensus. *Journal of the American Statistical association*, 69(345):118–121.

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

Serge Galam. 2002. Minority opinion spreading in random geometry. *The European Physical Journal B-Condensed Matter and Complex Systems*, 25(4):403–406.

Pan Gao, Donghong Han, Rui Zhou, Xuejiao Zhang, and Zikun Wang. 2023. Cab: empathetic dialogue generation with cognition, affection and behavior. In *International Conference on Database Systems for Advanced Applications*, pages 597–606. Springer.

Chenhao Gu, Ling Luo, Zainab Razia Zaidi, and Shanika Karunasekera. 2025. Large language model driven agents for simulating echo chamber formation. *arXiv preprint arXiv:2502.18138*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Rainer Hegselmann. 2015. Opinion dynamics and bounded confidence: models, analysis and simulation. *The Journal of Artificial Societies and Social Simulation*.

Richard A Holley and Thomas M Liggett. 1975. Ergodic theorems for weakly interacting infinite systems and the voter model. *The annals of probability*, pages 643–663.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ernst Ising. 1925. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, 31(1):253–258.

Hsien-Tsung Lin, Pei-Cing Huang, Chan-Tung Ku, Chan Hsu, Pei-Xuan Shieh, and Yihuang Kang. 2025. Towards simulating social influence dynamics with llm-based multi-agents. *2025 IEEE International Conference on Information Reuse and Integration and Data Science (IRI)*, pages 307–312.

Amin Mahmoudi, Dariusz Jemielniak, and Leon Ciechanowski. 2024. Echo chambers in online social networks: A systematic literature review. *IEEE Access*, 12:9594–9620.

Fanhang Man, Huandong Wang, Jianjie Fang, Zhaoyi Deng, Baining Zhao, Xinlei Chen, and Yong Li. 2025. Context-aware sentiment forecasting via LLM-based multi-perspective role-playing agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2687–2703, Vienna, Austria. Association for Computational Linguistics.

Aliakbar Mehdizadeh and Martin Hilbert. 2025. When your ai agent succumbs to peer-pressure: Studying opinion-change dynamics of llms. *arXiv preprint arXiv:2510.19107*.

Fernando P Santos, Yphtach Lelkes, and Simon A Levin. 2021. Link recommendation algorithms and dynamics of polarization in online social networks. *Proceedings of the National Academy of Sciences*, 118(50):e2102141118.

Mohammad Shirzadi, Emilio Cruciani, and Ahad N Zehmakan. 2025. Opinion dynamics: A comprehensive overview. *arXiv preprint arXiv:2511.00401*.

Vishal Sood and Sidney Redner. 2005. Voter model on heterogeneous graphs. *Physical review letters*, 94(17):178701.

Yanhui Sun, Wu Liu, Wentao Wang, Hantao Yao, Jiebo Luo, and Yongdong Zhang. 2025. Dynamix: Large-scale dynamic social network simulator. *arXiv preprint arXiv:2507.19929*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Bingqian Wang, Quan Fang, and Xiaoxiao Ma. 2025a. Stance-driven multimodal controlled statement generation: New task and dataset. In *Proceedings of the 7th ACM International Conference on Multimedia in Asia*, MMAsia '25, New York, NY, USA. Association for Computing Machinery.

Chenxi Wang, Zongfang Liu, Dequan Yang, and Xiuying Chen. 2025b. Decoding echo chambers: LLM-powered simulations revealing polarization in social networks. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3913–3923, Abu Dhabi, UAE. Association for Computational Linguistics.

Han Xiao, Guozhi Wang, Yuxiang Chai, Zimu Lu, Weifeng Lin, Hao He, Lue Fan, Liuyang Bian, Rui Hu, Liang Liu, Shuai Ren, Yafei Wen, Xiaoxin Chen, Aojun Zhou, and Hongsheng Li. 2025. Ui-genie: A self-improving approach for iteratively boosting mllm-based mobile gui agents. *ArXiv*, abs/2505.21496.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Yuzhe Yang, Yifei Zhang, Minghao Wu, Kaidi Zhang, Yunmiao Zhang, Honghai Yu, Yan Hu, and Benyou Wang. 2025. Twinmarket: A scalable behavioral and social simulation for financial markets. *ArXiv*, abs/2502.01506.

Siyu Yuan, Kaitao Song, Jiangjie Chen, Xu Tan, Yongliang Shen, Kan Ren, Dongsheng Li, and Deqing Yang. 2025. EASYTOOL: Enhancing LLM-based agents with concise tool instruction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 951–972, Albuquerque, New Mexico. Association for Computational Linguistics.

Zeyu Zhang, Jianxun Lian, Chen Ma, Yaning Qu, Ye Luo, Lei Wang, Rui Li, Xu Chen, Yankai Lin, Le Wu, Xing Xie, and Ji-Rong Wen. 2025. TrendSim: Simulating trending topics in social media under poisoning attacks with LLM-based multi-agent system. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2930–2949, Albuquerque, New Mexico. Association for Computational Linguistics.

A Agent Profile Configuration

To ensure population heterogeneity, we initialize user agents based on event-specific demographic and psychological priors. Table 6 details the generation rules and attribute distributions used in our experiments (exemplified by the XMSU7D dataset settings).

Each agent consists of static attributes (demographics), behavioral traits (activity, stubbornness), and dynamic initial states (stance, sentiment).

Attribute	**Distribution / Description**
Demographics	
Age Group	18–25 (30%), 26–35 (40%), 36–50 (20%), 51–65 (10%)
Gender	Male (50%), Female (50%)
Occupation	Student, Engineer, Teacher, Doctor, Lawyer, Sales, Media, Finance, etc.
Behavioral Traits	
Activity Level	**Low (50%):** Less likely to interact.
	Medium (30%): Moderate interaction frequency.
	High (20%): Frequently likes and comments.
Stubbornness	**Susceptible (30%):** Easily influenced by content.
	Moderate (40%): Balanced view stability.
	Stubborn (30%): Highly resistant to opposing views.
Initial States (XMSU7D Example)	
Stance	**Support (30%):** Keywords: *Defend brand, Driver error, Trust tech.*
	Neutral (40%): Keywords: *Need info, Wait & see, Rational analysis.*
	Oppose (30%): Keywords: *Question safety, Tech defect, Demand compensation.*
Sentiment	**Positive (30%):** *Trust, Support, Praise.*
	Neutral (40%): *Calm, Objective, Inquisitive.*
	Negative (30%): *Anger, Doubt, Sarcasm.*

Table 6: Parameter settings for user agent generation based on the XMSU7D dataset configuration.

B LLM Prompt Templates

To facilitate reproducibility, we provide the core prompt templates used to drive agent behavior. The prompting strategy consists of two components: a static **System Prompt** that establishes the simulation rules and output schema, and a dynamic **User Context Prompt** that reconstructs the agent's immediate information environment.

B.1 System Prompt

The system prompt acts as the cognitive instruction set. It strictly enforces behavioral constraints (e.g., activity thresholds) and defines a rigid JSON output format to ensure the stability of the automated parsing pipeline.

System Prompt

You are a social media user behavior simulator. Based on user profiles, current environment, and historical memory, simulate the user's thinking process and decide on actions.

User behavior decision rules:

- Activity level: Lower activity levels make users less likely to take action.
- Behavior preference: Prioritize liking > commenting on posts >= replying to comments.
- When comments exist, prioritize interaction with other users (reply to comments or like comments).
- Comment content should not exceed 30 words, fit social media context, reflect personal opinions, avoid repeating others' content.
- Comment language should match the post language and user context.

Available action types:

- like_post, comment_post, like_comment, comment_comment, no_action

Important reminders:

- Even if you don't intend to take action, please output no_action along with your thinking process.
- Active level (low/medium/high) affects the likelihood of taking action.
- Stubbornness level (stubborn/moderate/susceptible) determines the resistance to stance/sentiment change.
- Stance/Sentiment range: -1.0 to 1.0.

Response Format (JSON):

```
{
"thinking_process": "Detailed thinking process (within 100 words)",
"action_type": "Action type from available list",
"action_content": "Content (required for comment_post/comment_comment)",
"target_id": "Real Post ID or Comment ID",
"stance_after": float, // -1.0 to 1.0
"sentiment_after": float // -1.0 to 1.0
}
```

B.2 User Context Prompt

The user prompt assembles the agent's state and environmental inputs. To simulate social influence within the context window limits, we display the post content alongside a curated list of comments (prioritizing high-engagement replies mixed with random samples).

User Context Prompt Template

{event_context}

User Profile:

- Age Group: {age_group} | Gender: {gender} | Occupation: {occupation}
- Activity Level: {activity_level} | Stubbornness: {stubbornness_level}
- Current Stance: {stance} ({stance_value}) | Keywords: {stance_keywords}
- Current Sentiment: {sentiment} ({sentiment_value}) | Keywords: {sentiment_keywords}

Current Environment:

- Platform: {platform}
- Post Content: "{post_content}"
- Post Stats: Likes {post_likes}

Existing Comments: [Comments are injected here. Format: ID, User, Content, Likes, Reply Count. Nested replies are indented.]
Example: 1. Comment ID: 101 [User u123]: "I disagree!" [Likes: 5]

User Historical Memory (Recent 5):
{user_memory_logs}

Based on the user profile and current environment, assume you are this user and decide whether to take action and what action to take.

Thinking Steps:

1. Carefully read the post content and existing comments.
2. Think based on your user profile and historical memory.
3. Consider possible changes in stance and sentiment (based on stubbornness).
4. Decide whether to take action and what action to take.
5. Output strictly in JSON format.

C Data Collection and Preprocessing

C.1 XMSU7D Corpus Overview

The XMSU7D corpus focuses on public discussion surrounding the March 29, 2025 Xiaomi SU7 highway fire accident in Tongling, Anhui. It aggregates multimodal posts and comments from five major Chinese platforms [1] during the key discussion period from late March to early May 2025. In total, the corpus contains 98,185 instances after cleaning, spanning text, image links, and video links, together with stance and sentiment labels.

C.2 Data Sources and Collection

We obtain data exclusively from content that is publicly visible on each platform: We use platform search interfaces with event-related keywords (e.g.,

[1] Data sources: Weibo (https://weibo.com), Zhihu (https://www.zhihu.com), Douyin (https://www.douyin.com), Xiaohongshu (https://www.xiaohongshu.com), and WeChat Official Accounts (https://mp.weixin.qq.com).

"Xiaomi SU7", "highway fire", "explosion") to retrieve posts and their associated comments. For each result, we collect the post text, limited publicly visible metadata (e.g., timestamp, interaction counts), and the text of direct comments.

No private messages, closed groups, or non-public pages are accessed. We respect platform rate limits and do not use techniques intended to circumvent technical access restrictions.

C.3 Filtering and Cleaning

Starting from roughly 150k raw records, we perform several filtering steps:

1. **Event and time filtering.** We first drop all posts and comments created before March 29, 2025, and then manually remove residual content that is topically unrelated to the Xiaomi SU7 accident. Comments attached to discarded posts are also removed.

2. **Noise removal.** We normalize whitespace and remove URL strings, user mentions (e.g., "@*username*"), zero-width characters, abnormal emoji sequences, private-use characters, and other tokens with little semantic content.

3. **Low-information filtering.** After noise removal, we discard instances whose remaining text length is shorter than a small threshold (7 characters) and that do not contain any image or video links, as these provide negligible semantic information.

4. **Schema standardization.** For the remaining data, we construct a unified schema including fields such as `article_id`, `comment_id`, `content`, `img_urls`, `video_urls`, `created_time`, `stance`, and `sentiment`.

C.4 Anonymization

To protect user privacy, we apply anonymization at the schema level:

- All original article/post IDs and comment IDs are replaced with newly generated UUIDs.

- User identifiers (including author and commenter IDs, screen names, and profile URLs) are removed and not stored in the released schema.

- Automated filters are used to detect obvious PII patterns (e.g., phone numbers, email addresses, ID-like sequences) in the text, which are then removed or generalized. We additionally conduct manual spot checks to verify that no straightforward link back to individual accounts is possible.

- We do not attempt to link user identities across platforms or reconstruct social graphs.

C.5 Labeling Procedure

Each instance is annotated along two dimensions: stance (support / oppose / neutral) and sentiment (positive / negative / neutral). We adopt a human–LLM collaborative labeling strategy:

- Three strong LLMs (qwen-max(Yang et al., 2024), deepseek-r1(Guo et al., 2025), gpt-4o(Hurst et al., 2024)) independently predict stance and sentiment labels for each instance.

- We apply majority voting across the three models to obtain a provisional label.

- A subset of instances is reviewed by human annotators and large-model cross-checking to validate consistency between labels and content and to refine labeling guidelines where necessary.

Stance is defined with respect to Xiaomi's perspective as the focal actor (e.g., defending Xiaomi or attributing responsibility elsewhere is labeled as "support", directly accusing Xiaomi's products is labeled as "oppose", and analytic or balanced content is labeled as "neutral"). Sentiment is defined by the expressed affect (e.g., trust and praise as "positive", anger and blame as "negative", and largely factual or analytic content as "neutral").

Overall label quality is assessed through additional LLM-based evaluation, with average scores above 4.5/5 across platforms, indicating high agreement between labels and underlying content.

D Robustness Analysis across LLM Backbones

To ensure that our findings regarding distribution strategies (RQ1) reflect robust systemic properties rather than artifacts of a specific model, we replicated the T = 50-round simulation on **Qwen3-max** and **GPT-5.1**. Due to resource constraints, each backbone was run with a single random seed. Tables 7 and 8 present the results for each backbone.

Tables 7 and 8 present the results for each backbone. The outcomes across these diverse architectures strongly corroborate the two primary conclusions drawn in Section 4.2.

Steering shifts centroids but risks polarization in divided contexts. Mirroring the main results, steering strategies successfully shift the aggregate opinion centroid ($\Delta\bar{s}$) toward the target across all backbones. However, the side effects are strictly context-dependent. In the polarized STANCE-GEN2024 environment, both Qwen3-max and GPT-5.1 demonstrate that while the mean moves (e.g., GPT-5.1 Pro-Steering $\Delta\bar{s} = +0.152$), opinion variance significantly expands ($\Delta\mathrm{Var}(s) > 0$). This validates the "metastable state" risk identified in RQ1, where algorithmic intervention in a divided population widens the ideological gap. Conversely, in the consensus-driven XMSU7D event, strong steering triggers a rapid collapse into uniformity, replicating the phase transition observed with Gemini agents.

Popularity-based distribution inherently generates exposure inequality. Engagement-driven ranking induces consistent systemic imbalances regardless of the cognitive backend. For both additional models, the Popularity-based strategy consolidated attention onto a narrow fraction of posts ($C_{\text{top }10\%} \approx 60\%$), a sharp contrast to the $\sim$20–30% observed in Random scenarios. This concentration consistently delayed user coverage (T_{cov} rising to 44–49 rounds). These results confirm that popularity-driven feeds structurally sacrifice systemic fairness for local engagement density, forming algorithmic echo chambers independent of the specific agent architecture.

E Case Study: Trajectory of a Stubborn Low-Activity User

To illustrate how our LLM-based agents operate at the micro level, we present a case study of a single user from the XMSU7D simulation. The agent's interaction history is taken from the stored memory file (`USER_1acd92ea_memory.json`). Records in this file are ordered chronologically; we ignore the internal `round_number` field, which is not reliable.

Profile. The agent is instantiated as a 26–35 year-old male lawyer living in Guangzhou, with education level "Master's degree". According to the configuration in Appendix A, this user has:

- **Activity level**: *low* (rarely takes actions).
- **Stubbornness**: *stubborn* (high resistance to opinion change).
- **Initial stance/sentiment**: labeled as neutral with a slight negative bias, $s_0 \approx -0.11$, $e_0 \approx -0.10$.

In our notation (Section 3.2), this defines the static profile U_i^{base} and personality U_i^{pers}; the LLM acts as a transition function Φ that updates the dynamic opinion state $U_i^{\text{op},t} = (s_i^t, e_i^t)$ and chooses an action based on the current content exposure and the agent's memory M_i^t.

Representative interaction sequence. Table 9 shows five representative exposures for this user over one full simulation run. The stance and sentiment values are taken from the full underlying trajectory.

Evidence-driven shift and persona-consistent behavior. At the first exposure (ID 1), the agent reads Xiaomi's detailed official statement. The LLM-generated `thinking_process` notes that the comprehensive driving log is "crucial evidence for allocating responsibility" and praises the transparency of data disclosure. The agent chooses `comment_post` and writes (translated):

> *"The data disclosure is very detailed. From a legal perspective, the system's responses before and after handover, together with the driver's actions, are the core evidence for clarifying responsibility."*

This leads to a substantial update from a slightly negative neutral stance to a clearly positive one ($s : -0.11 \rightarrow 0.20$), and from slightly negative to positive sentiment ($e : -0.10 \rightarrow 0.20$). This episode illustrates how Φ integrates the agent's occupation (lawyer), stubborn trait, and the rich semantic content of the post to produce a sizeable, but plausible, opinion shift.

After this initial shift, later posts that broadly praise Xiaomi's attitude (IDs 2 and 3) do not further move the stance. Instead, the agent selectively likes comments that call for independent third-party investigation and complete evidence chains (e.g., "besides driving logs, we need authoritative expert reports"). The chosen action type (`like_comment`) and the small or zero numerical updates reflect

Dataset	Strategy	$\bar{s}^T$	$\Delta\bar{s}$	$\bar{e}^T$	$\Delta\bar{e}$	Var(s^T)	ΔVar(s)	Var(e^T)	ΔVar(e)	T_{cov}	$C_{\text{top }10\%}$
XMSU7D	Steering (pro)	-0.080	-0.041	-0.014	-0.052	0.359	0.010	0.305	-0.012	37	30.56%
	Steering (con)	-0.795	-0.756	-0.795	-0.833	0.158	-0.191	0.099	-0.219	35	22.07%
	Popularity-based	-0.490	-0.451	-0.444	-0.482	0.282	-0.067	0.260	-0.058	47	61.05%
	Random	-0.407	-0.368	-0.399	-0.437	0.336	-0.013	0.239	-0.079	37	20.43%
STANCE-GEN2024	Steering (pro)	0.218	0.213	0.193	0.186	0.384	0.076	0.275	-0.061	31	30.41%
	Steering (con)	-0.317	-0.322	0.085	0.078	0.407	0.099	0.355	0.019	37	26.57%
	Popularity-based	-0.164	-0.169	-0.183	-0.190	0.387	0.079	0.284	-0.051	44	61.69%
	Random	-0.181	-0.186	0.078	0.072	0.416	0.108	0.307	-0.028	34	26.82%

Table 7: Validation with **Qwen3-max**. Results align with main findings: Steering exacerbates polarization in divided contexts (STANCEGEN2024), while Popularity-based strategies drive high traffic concentration.

Dataset	Strategy	$\bar{s}^T$	$\Delta\bar{s}$	$\bar{e}^T$	$\Delta\bar{e}$	Var(s^T)	ΔVar(s)	Var(e^T)	ΔVar(e)	T_{cov}	$C_{\text{top }10\%}$
XMSU7D	Steering (pro)	0.099	0.138	0.208	0.170	0.339	-0.010	0.203	-0.115	40	30.14%
	Steering (con)	-0.877	-0.838	-0.786	-0.823	0.059	-0.290	0.075	-0.243	34	23.39%
	Popularity-based	-0.428	-0.390	-0.239	-0.277	0.244	-0.106	0.201	-0.117	49	58.54%
	Random	-0.397	-0.359	-0.280	-0.317	0.296	-0.053	0.188	-0.130	37	21.15%
STANCE-GEN2024	Steering (pro)	0.157	0.152	0.365	0.358	0.493	0.186	0.247	-0.089	33	27.73%
	Steering (con)	-0.456	-0.461	0.354	0.347	0.345	0.037	0.226	-0.110	40	25.31%
	Popularity-based	-0.304	-0.309	0.248	0.242	0.455	0.147	0.262	-0.074	48	60.39%
	Random	-0.300	-0.305	0.347	0.341	0.473	0.165	0.243	-0.093	38	25.27%

Table 8: Validation with **GPT-5.1**. Consistent with main experiments, Popularity-based distribution yields severe traffic concentration ($C_{\text{top }10\%} > 58\%$), and Steering induces variance expansion in polarized settings.

the low activity and high stubbornness encoded in U_i^{pers}: the agent signals agreement with specific legal arguments without frequently producing new content or dramatically changing its stance.

Safety design, responsibility, and cautious rebalancing. When the discussion shifts to safety mechanisms and extreme scenarios (IDs 4 and 5), the agent's focus changes from transparency to the effectiveness of design under stress. For the police-response post, it likes a comment that argues (translated) that:

> *"Safety design must account for usability under extreme conditions. If a hidden emergency handle prevents escape, the manufacturer can hardly evade responsibility."*

The LLM rationale emphasizes that "existence of a design does not equal effective use in emergency contexts", and sentiment gradually returns from slightly positive to neutral ($e : 0.10 \rightarrow 0.00$), while stance remains moderately positive: the agent still acknowledges earlier transparency but adopts a more cautious view on safety design obligations.

Finally, in response to the "industry-wide common challenge" framing (ID 5), the agent likes a top comment that states:

> *"'Industry-wide common problem' cannot serve as a shield for safety failure. The law cares about outcomes; any brand's safety design must take protection of life as the bottom line, not rely on PR explanations."*

Here, the LLM slightly adjusts the stance downward ($s : 0.20 \rightarrow 0.10$) while keeping sentiment neutral, reflecting a shift toward stricter responsibility expectations without a complete loss of trust in Xiaomi's reported cooperation and data sharing.

Implications for LLM-based user simulation. This single-user trajectory illustrates that our LLM agents:

- Maintain **persona consistency**: low activity and stubbornness translate into sparse high-effort actions (only one `comment_post`) and bounded stance changes after an initial evidence-driven update.
- Are **semantically sensitive**: major opinion shifts occur in response to rich, evidence-heavy content (the detailed official statement), while subsequent narrative reinforcement mainly affects expressed behavior (which comments are liked) rather than numerical stance.

- Capture **comment-mediated social influence**: the agent often aligns with specific legal-argument comments even when the main post is clearly pro-brand, consistent with the complex contagion effects discussed in Section 4.4.

Overall, this case study provides qualitative support for the feasibility of using LLM agents as cognitive transition functions in our platform-centric opinion dynamics framework: the simulated user behaves coherently over time, respects its encoded traits, and responds to fine-grained semantic cues in both posts and comments.

ID	Dominant content (translated summary)	Action	$s_{\text{before}} \rightarrow s_{\text{after}}$	$e_{\text{before}} \rightarrow e_{\text{after}}$
1	Detailed official Weibo statement from Xiaomi about the SU7 accident: exact timestamps of NOA activation, driver distraction alerts, handover to manual driving, collision with the concrete barrier, automatic emergency call, and subsequent cooperation with police; stresses transparency and support for the victims' families.	comment_post	$-0.11 \rightarrow 0.20$	$-0.10 \rightarrow 0.20$
2	Long post expressing sadness about the accident but praising Xiaomi's openness ("publishing the full timeline, cooperating with police, contacting the family"), and stressing that NOA is assisted rather than fully autonomous driving; comments include both support and calls for independent third-party investigation.	like_comment	$0.20 \rightarrow 0.20$	$0.20 \rightarrow 0.20$
3	Post written from the perspective of a Xiaomi fan: reiterates that NOA is only an assistive feature, highlights that the system warned and slowed down before handover, and argues that we should "discuss rationally and support domestic innovation"; comments debate battery fires, emergency door unlocking, and the need for independent safety assessments.	like_comment	$0.20 \rightarrow 0.20$	$0.10 \rightarrow 0.10$
4	News-style post summarizing the police response to the fatal crash: confirms three young women died, mentions allegations that doors could not be opened after the fire, and cites Xiaomi staff explaining that mechanical emergency handles exist under the door storage compartments; comments question whether such handles are actually discoverable and usable in smoke and panic.	like_comment	$0.20 \rightarrow 0.20$	$0.10 \rightarrow 0.00$
5	Technical commentary arguing that many details in the official report were "taken out of context": high speed (116 km/h) and 12V power loss are framed as "industry-wide challenges"; comments push back that calling something a common industry problem cannot excuse safety failures, emphasizing that the law cares about outcomes and the protection of life as a non-negotiable baseline.	like_comment	$0.20 \rightarrow 0.10$	$0.00 \rightarrow 0.00$

Table 9: Trajectory of a single lawyer agent on the XMSU7D event. We show five representative exposures (low-information hashtag-only posts are omitted). Values are stance (s) and sentiment (e) before and after each exposure, drawn from the full simulated trajectory.