BEYOND CORRECTIVE CONTENT: SOCIAL HYPERGRAPHS FOR MISINFORMATION INTERVENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

By the time misinformation is detected, users may already believe it, discuss it, or pass it on. Correcting misinformation after exposure therefore requires deciding which users or audiences need attention, what corrective information to provide, and how to adapt as their responses evolve. Existing methods personalize corrective messages or broadcast fact-checks, but rarely treat changing social responses as feedback for subsequent intervention. We address this limitation by formulating *socially driven post-exposure intervention*, in which observable user responses guide where and how to intervene over time. We represent local social groups as hyperedges in a social hypergraph, coupling each group's evolving discourse with its intervention audience while preserving users shared across groups. Within an LLM-agent simulation, an external Intervention Agent identifies risk hyperedges, generates evidence-grounded group corrections, coordinates messages for shared users, and uses previous intervention outcomes to inform later rounds. Across four misinformation datasets, our framework achieves the lowest simulated Trust_{D7} among all compared methods, reducing it by 56–58% relative to no intervention. It also reduces $\mathrm{SpreadRate}_{D7}$ by 49–84%, although its advantage over competing interventions varies across datasets.

1 INTRODUCTION

Misinformation poses a persistent societal challenge to human cognition and collective behavior due to its widespread and misleading nature (Lazer et al., 2018; Vosoughi et al., 2018; Chen & Shu, 2024). Computational research has made substantial progress in detecting false or misleading content (Zhou & Zafarani, 2020; Yang et al., 2025), but detection addresses what is false rather than how to mitigate its cognitive and behavioral effects after users have encountered it. This motivates *post-exposure intervention* (Hartwig et al., 2024; Liu et al., 2025b), which needs to decide which users or audiences require attention (Sakiyama et al., 2024), what corrective information to deliver, and how that response should adapt as beliefs and discussions evolve. This task is especially difficult because misinformation can continue to influence people's thinking even after they receive and accept a correction, a phenomenon known as the *continued influence effect* (Ecker et al., 2022).

Recent work has developed two prominent technical routes for post-exposure misinformation intervention (Wardle et al., 2026; Liu et al., 2025b). LLM-based persuasive dialogue generates corrective responses tailored to individual users (Costello et al., 2024; Salvi et al., 2025; Hackenburg et al., 2025) and ongoing interactions (Pang et al., 2026; Han et al., 2026), whereas broadcast-style fact-checking disseminates to broader audiences (Drolsbach et al., 2024; Chuai et al., 2024). Despite operating at different scales, both routes remain largely centered on content-level correction (Zhou & Zafarani, 2020; Yang et al., 2025): seeking to generate, select, or deliver an appropriate correction to a target individual or audience. Social information, when incorporated, typically conditions how a correction is generated or disseminated (Traberg et al., 2024; Prike et al., 2024), rather than functioning as feedback for updating intervention decisions. Consequently, existing approaches have limited capacity to use evolving social responses (Hoes et al., 2024; Wang et al., 2025) to guide where intervention is needed, which audiences require attention, and how corrective strategies should adapt. This limitation raises a central challenge: **how can an intervention policy use evolving social signals to decide where and how to intervene over time?**

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

We answer this question by formulating *socially driven post-exposure intervention*, in which evolving social responses explicitly inform intervention decisions (Butler et al., 2023). At each stage of misinformation diffusion, the framework summarizes users' evolving judgments within their social contexts and uses this feedback to determine intervention targets and corrective communication, rather than treating social context merely as a factor that affects intervention effectiveness. Operationalizing this formulation requires a social scope over which user responses can be observed, and corrective actions can be delivered. In this paper, we draw on social science and propose to use **local social groups** as mesoscopic observation-and-action units (McPherson et al., 2001). Within each group, public responses characterize how misinformation is being received, while group membership defines the audience for coordinated corrective communication (Jia & Zhang, 2025). A local group thus couples an evolving discourse state with a corresponding intervention audience.

However, these intervention audiences can overlap. Homophily organizes users with similar interests, values, and information preferences into local communities (Del Vicario et al., 2016; Cinelli et al., 2021), while multiple social identities allow individuals to participate in several such groups (Turner et al., 1987). For intervention, this overlap creates a coordination problem (Palla et al., 2005; Battiston et al., 2025): corrective messages tailored to different group states may converge on the same user. Inspired by the psychological research, we represent this higher-order social structure as a **social hypergraph** (de Arruda et al., 2024), with users as nodes and local groups as hyperedges. Unlike a pairwise user–user projection (Li et al., 2024), the hypergraph retains each group as an explicit intervention context and reveals which audiences share users, providing the structural basis for tracking group-specific discourse states and coordinating corrective communication across groups (Jhun, 2021; Oliveira et al., 2026).

Building on this representation, we instantiate socially driven post-exposure intervention in an LLM-agent framework (Gao et al., 2024; Aher et al., 2023). The framework consists of three connected components. (i) The **Social Hypergraph Environment** constructs local-group hyperedges from users' spatial, social-identity, and psychological similarities. A **Controller** routes information exposure across affiliated hyperedges and constructs each user's *Social Memory*, while user agents combine their profiles, *Individual Memory*, and *Social Memory* through a Belief-Desire-Intention-based decision process (Rao & Georgeff, 1995) to update their judgments and propagation decisions. (ii) In **Socially Driven Group-Level Intervention**, an external **Intervention Agent** identifies risk hyperedges from observable comments and stance labels, grounds corrections in dataset-provided factual evidence, adapts them to member profiles and the current group discourse, and coordinates their delivery across overlapping or nested hyperedges. (iii) **Social Feedback and Temporal Adaptation** summarizes post-intervention response changes as *Correction History* and uses this history as corrective context for subsequent user updates. Together, these components form a closed loop between social propagation, group-level intervention, and temporal feedback.

We evaluate the proposed framework on four English misinformation datasets under a unified seven-day simulation setting. The proposed method achieves the lowest Trust on all four datasets and consistently reduces both misinformation trust and propagation relative to no intervention. Its advantage on the lowest $\mathrm{SpreadRate}$ is more context-dependent, with the best result obtained on two datasets. Ablations identify the Group Doubt Ratio and Correction History as consistent contributors, whereas the benefit of correction fusion across overlapping groups varies with the dataset. Further analyses show that the intervention reduces supportive propagation and shifts simulated group discourse toward doubt and correction.

2 Related Work

Post-Exposure Misinformation Intervention Existing interventions follow the two routes discussed above, but both remain centered on content-level correction (Liu et al., 2025b; Zhou & Zafarani, 2020; Yang et al., 2025; Vithanage et al., 2026). LLM-based dialogue adapts corrective responses to individual users (Costello et al., 2024; Salvi et al., 2025; Pang et al., 2026), yet it mainly operates in isolated interactions and does not use group-level public responses to allocate intervention. Broadcast-style fact-checking reaches larger audiences (Drolsbach et al., 2024; Chuai et al., 2024; van der Linden et al., 2026), but adapts weakly to heterogeneous audiences and local opinion environments (Cinelli et al., 2021; Hoes et al., 2024). The key gap is therefore not the generation

of corrective content, but the lack of a policy that jointly uses social responses for target selection, strategy adaptation, and subsequent updates.

Group-Level Coordination Local social groups provide a mesoscopic unit between individual users and the whole network. Pairwise graphs are natural for dyadic influence (Li et al., 2024; Battiston et al., 2026; Agostinelli et al., 2026), but they cannot explicitly preserve group membership or overlap and nesting (Alvarez-Rodriguez et al., 2021; Battiston et al., 2025). Hypergraphs provide a natural representation of higher-order group interactions and have been used for contagion and rumor propagation (de Arruda et al., 2024; Kim & Goh, 2024; Oliveira et al., 2026). However, existing studies mainly use hyperedges to model diffusion (Yin et al., 2026; Duan et al., 2026; Xue et al., 2026). Their role as intervention units for risk identification, group-specific adaptation, and coordination across overlapping groups remains underexplored. To bridge this gap, we coordinate corrective interventions across overlapping and nested hyperedges, ensuring that users with multiple group memberships receive consistent, non-redundant, and group-tailored interventions.

LLM-based Social Simulation LLM-based simulations offer a controlled environment for studying evolving beliefs and behaviors (Park et al., 2023; Gao et al., 2024; Li et al., 2025). Through BDI-based cognitive architectures, user agents maintain memories, update trust, produce public comments, and choose between spreading and silence (Rao & Georgeff, 1995; Chuang et al., 2024; Donkers & Ziegler, 2025). Recent frameworks further model content dissemination, multi-agent misinformation intervention, and platform regulation (Liu et al., 2025a; Zhang et al., 2026). Yet few explicitly couple higher-order social structure, evolving public discourse, and intervention outcomes in one feedback loop (Choi et al., 2026). Interventions are also often imposed as predefined external events, rather than selected and adapted according to group-level public responses. Our framework addresses this gap by coupling BDI-driven user agents with a social hypergraph and treating local groups as mesoscopic intervention units.

3 METHODOLOGY

3.1 PROBLEM FORMULATION

To study post-exposure misinformation intervention in an evolving social environment, we seek a framework with three properties: (1) **socially grounded**: intervention decisions should respond to observable user judgments within local social groups; (2) **group-adaptive**: corrective communication should remain grounded in factual evidence while adapting to group characteristics and current discourse; and (3) **relation- and feedback-aware**: corrections should be coordinated for users shared across groups, while intervention outcomes should inform subsequent updates. We instantiate these requirements over a social hypergraph, with the objective of reducing persistent misinformation belief and subsequent propagation. Figure 1 illustrates how the **Social Hypergraph Environment**, **Socially Driven Group-Level Intervention**, and **Social Feedback and Temporal Adaptation** implement this formulation.

3.2 SOCIAL HYPERGRAPH ENVIRONMENT

User Agent Modeling Users embedded in the same social environment may interpret the same information differently because of their personal characteristics and prior social experiences. We therefore represent each user agent u_i using a static profile P_i (encoded persistent personal attributes) and a dynamic memory $M_i(t)$ (records the user's decision history and perceived social context).

(I) Profile. We instantiate user agents by constrained sampling from the user pool provided by SocioVerse (Zhang et al., 2025). This procedure broadly preserves **demographic proportions** (e.g., relatively balanced gender distribution) while ensuring social **category coverage** (e.g., different social backgrounds). Each user u_i is assigned a profile P_i containing ten attributes: area, occupation, education, marital status, race, gender, age, ideology, Big Five personality traits, and religion. These attributes characterize how users interpret information and participate in social contexts, providing the observable features used to estimate profile homophily during hypergraph construction.

(II) Memory. The judgments of each user depend on both their prior beliefs and the opinions encountered through social interaction (Ecker et al., 2022; Chuang et al., 2024). Each user u_i therefore maintains two complementary memory components: $M_i(t) = \big(M_i^{\mathrm{ind}}(t), M_i^{\mathrm{soc}}(t)\big)$. The **individual**

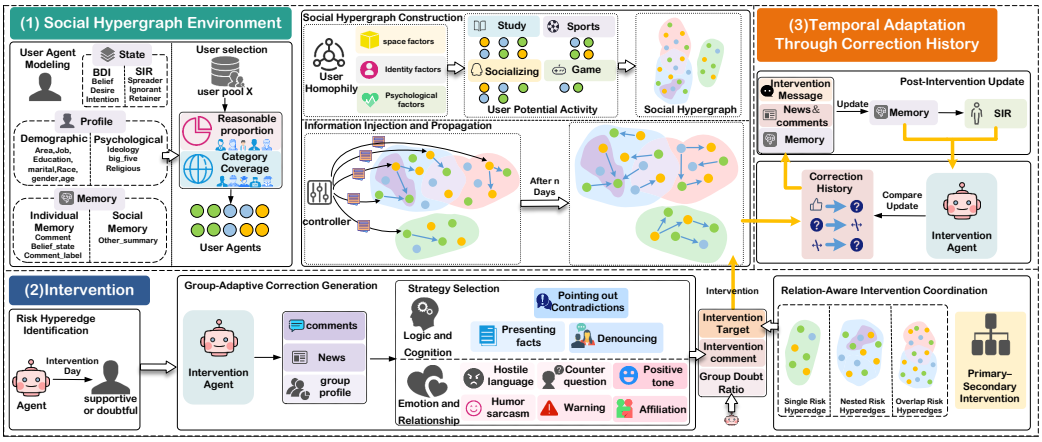


Figure 1: Overview of the proposed socially driven group-level intervention framework. The framework integrates social environment construction, group-level intervention, and intervention feedback into a closed-loop process.

memory $M_i^{\text{ind}}(t)$ records a sequence of user judgments about misinformation up to time t, including its comments $c_i(t)$, stance labels $\ell_i(t) \in \{\text{support}, \text{doubt}, \text{correct}\}$, and belief states $b_i(t) \in [0,1]$, with previous records being retained. The comment denotes the individual opinion about information veracity, the stance label categorizes this assessment, and the belief state is the trust indicator that quantifies the degree of trust in misinformation. The **social memory** $M_i^{\text{soc}}(t)$ represents a profile-conditioned summary of the public comments delivered to u_i at time t following the agent's affiliated hyperedges. This separation preserves the evolving judgment of each user while representing the local discourse that informs its subsequent decisions.

3.2.1 Social Hypergraph Construction

Users interact within local social contexts formed around shared locations, social identities, and psychological characteristics. We represent these group affiliations as a social hypergraph $\mathcal{H} = (\mathcal{U}, \mathcal{E})$, where $\mathcal{U} = \{u_i\}_{i=1}^{N}$ is the user set and $\mathcal{E} = \{e_m\}_{m=1}^{M}$ is the hyperedge set. Each hyperedge is constructed separately and contains the users associated with one local social group, allowing a user to participate in multiple social contexts.

Because direct records of participation in social foci are often unavailable or privacy-sensitive, we approximate group membership using profile homophily (McPherson et al., 2001). For users u_i and u_j, we measure similarity along three dimensions: *spatial similarity* based on area, *socio-identity similarity* based on education, job, marital status, race, gender, and age, and *psychological similarity* based on ideology, Big Five personality traits, and religion. For each dimension $d \in \mathcal{D} = \{\text{spa}, \text{id}, \text{psy}\}$, we average the corresponding attribute-level similarities to obtain s_{ij}^d. The dimensions are then weighted by their variation across user pairs, assigning greater weight to dimensions that better differentiate the sampled population.

$$s_{ij} = \sum_{d \in \mathcal{D}} \omega_d s_{ij}^d, \qquad \omega_d = \begin{cases} \dfrac{v_d}{\sum_{d' \in \mathcal{D}} v_{d'}}, & \sum_{d'} v_{d'} > 0, \\ \dfrac{1}{|\mathcal{D}|}, & \text{otherwise}, \end{cases} \tag{1}$$

where $v_d = \mathrm{Var}\big(\{s_{ij}^d : i < j\}\big)$. Given a target neighborhood size k_i, we construct one candidate group around each user: $\widetilde{e}_i = \{u_i\} \cup \mathrm{Top}_{k_i}\big(\mathcal{U} \setminus \{u_i\}; s_{ij}\big)$, where Top_{k_i} returns the k_i users most similar to u_i. Because candidates are constructed around different users, their memberships may intersect or contain one another. We remove duplicate and highly redundant candidates and retain M representative hyperedges. For two retained hyperedges e_p and e_q, their resulting memberships induce different structural relations between local groups,

$$\mathrm{rel}(e_p, e_q) = \begin{cases} \text{independent}, & e_p \cap e_q = \varnothing, \\ \text{overlapping}, & e_p \cap e_q \neq \varnothing,\ e_p \not\subseteq e_q,\ e_q \not\subseteq e_p, \\ \text{nested}, & e_p \subset e_q \text{ or } e_q \subset e_p. \end{cases} \tag{2}$$

Independent hyperedges represent separate participation contexts, whereas overlapping hyperedges share users who participate in multiple groups. A nested relation arises when a specialized group is entirely contained within a broader social context. These relations correspond respectively to separate social contexts, users participating across multiple contexts, and specialized groups embedded within broader communities (Turner et al., 1987; Breiger, 1974; Palla et al., 2005). The resulting hypergraph represents synthetic group memberships rather than an observed social network and provides the fixed structural context used by the subsequent propagation and intervention modules.

3.2.2 Information Injection and Propagation

Given the social hypergraph, misinformation propagation is jointly governed by the Controller and user agents. The Controller determines information exposure and constructs the social context presented to each user, whereas user agents independently form judgments and decide whether to spread or remain silent.

Controller. The Controller models platform-level mediation of information exposure in the social hypergraph. Its functionality comprises two components: **exposure routing** and **social-context construction**. Exposure routing determines which users receive misinformation x. It covers both the initial publication to selected seed users and the subsequent delivery from a Spreader to recipients from their associated hyperedges. In this process, the Controller routes both x and the user's textual judgment $c_i(t)$ as a public comment. Social-context construction aggregates the public comments delivered to user u_i through different affiliated hyperedges into $M_i^{\mathrm{soc}}(t)$ for each exposed user. It does not expose unpropagated judgments to other user agents. These components specify the information and social context presented to a user agent, while judgment formation and spread-or-silence decisions remain with the user agent.

Opinion–Behavior Update. The opinion–behavior response of each user is implemented through the Belief–Desire–Intention (BDI) framework (Rao & Georgeff, 1995). When user u_i is exposed to misinformation x, the **belief** component interprets the information by jointly considering the user's profile P_i, prior individual memory $M_i^{\mathrm{ind}}(t^-)$, and social memory $M_i^{\mathrm{soc}}(t)$ constructed by the Controller. This interpretation produces the user's trust in the misinformation, textual judgment, and corresponding stance. The updated belief further shapes the user's **desire**, which captures its motivation to respond, such as expressing agreement, raising doubt, or preventing further spread. The **intention** component translates this motivation into a candidate spread-or-silence decision. We summarize this process as

$$\big(b_i(t), c_i(t), \ell_i(t), a_i(t)\big) = f_{\mathrm{BDI}}\left(P_i, x, M_i^{\mathrm{ind}}(t^-), M_i^{\mathrm{soc}}(t)\right), \tag{3}$$

where $b_i(t) \in [0,1]$ denotes the user's trust in x, $c_i(t)$ is its textual comments, $\ell_i(t) \in \{\text{support}, \text{doubt}, \text{correct}\}$ is the associated stance, and $a_i(t) \in \{\text{spread}, \text{silence}\}$ is its behavioral decision. The tuple $(b_i(t), c_i(t), \ell_i(t))$ is appended to Individual Memory, whereas $a_i(t)$ is recorded separately by the environment. If the user chooses to spread, $c_i(t)$ becomes a public comment and can be routed to other users affilicated within the same hyperedges.

The behavioral decision links individual reasoning to information propagation. Following the SIR framework (Hodas & Lerman, 2014), we use $z_i(t)$ to record the exposure and propagation state of user u_i,

$$z_i(t) = \begin{cases} \text{I(Ignorant)}, & \text{if } u_i \text{ has not received } x, \\ \text{S(Spreader)}, & \text{if } u_i \text{ has received and spreads } x, \\ \text{R(Retainer)}, & \text{if } u_i \text{ has received } x \text{ but remains silent.} \end{cases} \tag{4}$$

The SIR state records exposure and executed propagation behavior rather than a fixed cognitive state. An exposed Ignorant user becomes a Spreader if its final decision is `spread` and a Retainer otherwise. For each Spreader, the Controller routes x together with the published judgment $c_i(t)$ to eligible Ignorant users in the spreader's affiliated hyperedges and constructs their Social Memory before invoking their opinion–behavior updates. If the receiver is reached by multiple distinct Spreaders, it receives x once and incorporates their distinct public comments into its Social Memory. A user belonging to multiple hyperedges can therefore carry the misinformation and its public comment into each affiliated social context, enabling cross-hyperedge propagation.

3.3 SOCIALLY DRIVEN GROUP-LEVEL INTERVENTION

Given the evolving group states in the social hypergraph, an external **Intervention Agent** determines where corrective intervention is needed and how it should be delivered. The agent first identifies risk hyperedges from observable user judgments, then generates evidence-grounded corrections adapted to each group, andbroadcasts coordinated messages to users associated with overlapping or nested intervention scopes.

Risk Hyperedge Identification Our approach identifies risk groups from the textual judgments and stance labels observable to the Intervention Agent. Specifically, for each hyperedge, the agent examines the latest stance label $\ell_i(t)$ and textual judgment $c_i(t)$ of each activated user. If at least one activated user in a hyperedge has a **support** or **doubt** label, the hyperedge is identified as risky. This strict criterion aims to cover potential risk groups early and reduce the chance of missing an intervention. The resulting set of risk hyperedges is then forwarded to the group-adaptive correction generation stage, which determines the corresponding intervention content for each identified hyperedge.

Group-Adaptive Correction Generation For each risk hyperedge, the Intervention Agent analyses the misinformation, its dataset-provided factual evidence, the profiles of all users, and the latest comments and stance labels of all activated users. The factual evidence grounds the corrective content, while the misinformation features defined by PCoT (Modzelewski et al., 2025) and the nine intervention strategies summarized in Appendix H guide strategy selection and message generation. Furthermore, the agent generates two group-specific intervention comments. The first synthesizes comments labeled as `doubt` or `correct` to amplify existing doubtful and corrective views; when no such comment exists, it is generated from the factual evidence itself. The second synthesizes comments labeled as `support` and counters their supporting reasons. Here, amplification and weakening are implemented through comment synthesis and rewriting rather than direct modification of existing user judgments. The agent also provides a **Group Doubt Ratio**, which reports the proportion of activated users whose latest stance is `doubt` or `correct` and is omitted when this proportion is zero. The two intervention comments and the ratio are broadcast to all activated users in the risk hyperedge, while interventions for users shared across risk hyperedges are coordinated in the next module.

Relation-Aware Intervention Coordination. Users may belong to multiple hyperedges and therefore fall within several intervention scopes. Rather than directly stacking the intervention comments generated by different risk hyperedges, the Intervention Agent coordinates them according to the relations between the corresponding groups. Specifically, we consider three cases: (i) If only one of the user's affiliated hyperedges is at risk, the user receives the two intervention comments and Group Doubt Ratio generated for that hyperedge. (ii) If two nested hyperedges are both at risk, the child hyperedge serves as the primary hyperedge, while information from the parent hyperedge is retained only when it is consistent with the same correction direction. (iii) If two overlapping hyperedges are both at risk, the shared users are treated as a coordination region. The agent compares their profile similarity with the non-shared members of each hyperedge and selects the hyperedge with the higher matching score as the primary hyperedge; the matching-score computation is provided in Appendix J.1. For the two double-risk cases, the agent fuses the corresponding intervention comments by retaining complementary information and removing conflicting or redundant content. The fused comments are broadcast only to the shared users, whereas non-shared users receive the original comments generated for their respective hyperedges. The shared users receive the Group Doubt Ratio of the primary hyperedge rather than an aggregation across groups. The current implementation coordinates users shared by two risk hyperedges.

3.4 SOCIAL FEEDBACK AND TEMPORAL ADAPTATION

Post-Intervention Update. After receiving the intervention comments and, when available, the Group Doubt Ratio, each targeted activated user combines them with the misinformation and its latest Individual Memory to update its trust score, textual judgment, and stance label. The updated judgment is retained in Individual Memory. Users who receive the misinformation at the current time step subsequently make their final spread-or-silence decisions, whereas previously activated

users who receive no new misinformation update only their judgments. Thus, on an intervention step, newly exposed users first form a pre-intervention judgment and then update it after intervention before determining their final behavior. Users who have never received the misinformation remain Ignorant and are unaffected.

Correction History. After the post-intervention update, the Intervention Agent compares users' pre- and post-intervention comments, stance labels, and observable behaviors to summarize how their responses changed under the delivered intervention. This summary is stored as **Correction History** and maintained separately from users' Individual and Social Memories. Its storage scope follows the intervention actually received: ordinary hyperedge members retain the history of their group-level intervention, whereas users in an overlapping or nested region retain a separate history of the fused intervention. In subsequent time steps, the corresponding Correction History is provided as historical corrective context during users' pre-intervention judgment updates. By shaping these subsequent judgments, it changes the group states observed by the Intervention Agent in later rounds, thereby completing the feedback loop. The detailed update prompt, replacement rule, and response-change criteria are provided in Appendix J.2.

4 Experiments

4.1 Experimental Setup

Datasets and simulation environment. We evaluate on four English misinformation datasets with 100 items per dataset[1]: Arg-En and Snopes cover celebrity, entertainment, and fact-checked online claims (Hu et al., 2024; Han et al., 2025), while CoAID and ECTF cover COVID-19 health misinformation and early-pandemic social media posts (Cui & Lee, 2020; Modzelewski et al., 2025). Each news item is simulated for seven days on the same fixed social hypergraph with 45 users, 8 hyperedges. User agents are driven by `DeepSeek-V4-Flash`, while the intervention agent is driven by `DeepSeek-V4-Pro` (DeepSeek AI, 2026). All conditions share the same news, users, hypergraph, propagation process, and evaluation window; only the intervention module is replaced.

Baselines. We compare against five recent methods: MADD with Fact and Narrative variants (Qiao et al., 2025), MOSAIC (Liu et al., 2025a), MisinfoEval (Gabriel et al., 2024), ED2D (Han et al., 2026), and MURSE (Pang et al., 2026). MADD and MOSAIC represent broadcast or platform-level control, whereas MisinfoEval, ED2D, and MURSE provide dialogue-based corrective interventions. All baselines are reimplemented on the same simulator and public-discourse interface, so performance differences reflect the intervention strategy.

Evaluation metrics. The unit of analysis is the news source. We evaluate both retained misinformation trust and subsequent propagation among activated users. Let A_{D_i} denote users activated by the evaluation day D_i of news item i, and let $S_{D_i} \subseteq A_{D_i}$ denote those who still spread the misinformation at that evaluation day:

$$\mathrm{Trust}_{D_i} = \frac{1}{|A_{D_i}|} \sum_{j \in A_{D_i}} b_j(D_i), \qquad \mathrm{SpreadRate}_{D_i} = \frac{|S_{D_i}|}{|A_{D_i}|}. \tag{5}$$

Here, Trust_{D_i} measures average misinformation trust at the evaluation day, while $\mathrm{SpreadRate}_{D_i}$ measures the fraction of activated users who still spread the claim. Lower values are better for both metrics. Each news item is evaluated over a seven-day cycle; therefore, we denote this cycle-end evaluation day by D_7 in our experiments.

4.2 Main Results

This section asks whether the proposed method consistently reduces users' final trust in misinformation and subsequent propagation under the unified simulation setting, and how it compares with existing intervention methods.

[1]Each seven-day simulation takes approximately an hour on average, so evaluating all 400 items requires about 400 single-process hours. We use 100 items per dataset to keep the evaluation within our practical computational budget.

Table 1: Comparison between the main experiment and five baselines on four datasets. MADD includes the MADD-F and MADD-N strategy configurations.

Dataset	Metric	No intervention	Ours	MADD-F	MADD-N	MOSAIC	ED2D	MisinfoEval	MURSE
Arg-En	Trust_{D_7}	0.6205	**0.2683**	0.4819	0.4761	0.4392	0.3594	0.4726	0.5298
Arg-En	$\mathrm{SpreadRate}_{D_7}$	0.2834	**0.0696**	0.1638	0.1867	0.2051	0.1547	0.1247	0.1907
Snopes	Trust_{D_7}	0.6211	**0.2678**	0.4225	0.4209	0.2991	0.2722	0.4371	0.4917
Snopes	$\mathrm{SpreadRate}_{D_7}$	0.3948	0.0623	0.1792	0.1422	0.0756	**0.0369**	0.1119	0.1517
CoAID	Trust_{D_7}	0.4638	**0.1963**	0.3566	0.3554	0.2848	0.3285	0.3715	0.3960
CoAID	$\mathrm{SpreadRate}_{D_7}$	0.4065	0.2089	0.2076	0.2051	**0.1670**	0.2150	0.1822	0.2193
ECTF	Trust_{D_7}	0.4977	**0.2195**	0.3819	0.3826	0.3472	0.3057	0.3814	0.4227
ECTF	$\mathrm{SpreadRate}_{D_7}$	0.3310	**0.0748**	0.1670	0.1747	0.2193	0.1770	0.1274	0.1875

Table 1 compares no intervention, the proposed method, and five baselines across four datasets. Ours achieves the lowest Trust_{D_7} on all datasets, reducing it by approximately 56%–58% relative to no intervention. It also reduces $\mathrm{SpreadRate}_{D_7}$ by approximately 49%–84% and attains the lowest value on Arg-En and ECTF. On Snopes and CoAID, the strongest baseline has a lower spread rate, but Ours still outperforms most baselines. Overall, the trust advantage is consistent across datasets, whereas the propagation advantage is more dataset-dependent. These results establish the effectiveness of the proposed method under the unified simulation setting.

4.3 Ablation Studies

This section evaluates the contributions of Correction History (CH), Fusion Broadcast (FB), Group Doubt Ratio (GDR), and higher-order relations (HR). CH, FB, and GDR are removed as intervention components, whereas HR is removed as a structural control, while group membership is held fixed. Removing GDR consistently increases both Trust_{D_7} and $\mathrm{SpreadRate}_{D_7}$, while removing CH generally worsens performance across datasets. Removing HR also increases Trust_{D_7} on most datasets, and its propagation effect is strongly dataset-dependent, with the largest degradation on ECTF. These results indicate that group-level doubt signals, historical correction outcomes, and higher-order relations provide stable contributions to trust reduction. In contrast, the effect of FB is dataset-dependent: it benefits most datasets but not CoAID. Overall, GDR, CH, and HR show relatively stable contributions, whereas the benefit of FB is less consistent and appears to depend on the interaction between group context and news content.

Table 2: Component ablation results for CH, FB, GDR, and HR.

Metric and dataset	Ours	w/o CH	w/o FB	w/o GDR	w/o HR
Trust_{D_7}					
Arg-En	**0.2683**	0.2726	0.2808	0.3148	0.3492
Snopes	**0.2678**	0.2762	0.2935	0.3062	0.3430
CoAID	0.1963	0.2271	**0.1669**	0.2189	0.2433
ECTF	**0.2195**	0.2306	0.2283	0.2242	0.2426
$\mathrm{SpreadRate}_{D_7}$					
Arg-En	0.0696	0.3467	0.2370	0.5656	0.0671
Snopes	**0.0623**	0.0679	0.2087	0.0677	**0.0373**
CoAID	0.2089	0.2958	**0.1405**	0.2659	0.2253
ECTF	**0.0748**	0.1981	0.0901	0.2538	0.5357

4.4 Structural Necessity and Mechanism

4.4.1 No-Hypergraph Control

We test the necessity of the hypergraph by merging all users into a single global pool and removing hyperedge membership and relations from propagation and intervention. Group-adaptive components are disabled, leaving one neutral fact-check generated from the news text, while all other settings remain fixed.

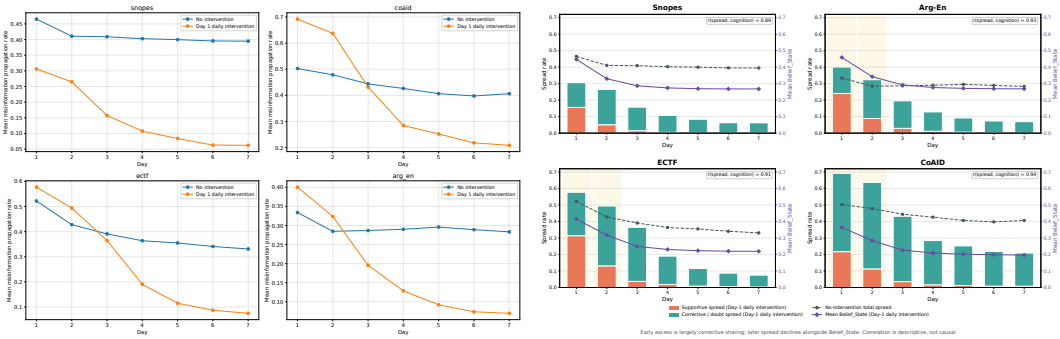
Table 3: Comparison between Ours and the no-hypergraph control under the intervention condition.

Metric	Ours				No hypergraph			
	Arg-En	Snopes	CoAID	ECTF	Arg-En	Snopes	CoAID	ECTF
Trust_{D_7}	**0.2683**	**0.2678**	**0.1963**	**0.2195**	0.4664	0.4155	0.2980	0.3359
$\mathrm{SpreadRate}_{D_7}$	**0.0696**	0.0623	**0.2089**	**0.0748**	0.4000	**0.0505**	0.5434	0.2348

Removing the hypergraph eliminates the local intervention scope, group-adaptive strategy selection, and relation-aware coordination across overlapping or nested groups. The intervention therefore becomes a single global correction that cannot exploit homophily-based composition or the current group discourse. The resulting loss of group context weakens correction and leaves more supportive belief and propagation, supporting the necessity of modeling social interactions as a hypergraph.

4.4.2 Propagation Dynamics under Intervention

We examine whether intervention reduces the total amount of propagation or changes its composition and temporal pattern, using the same news, users, hypergraph, initial receivers, and seven-day window as the main experiment.



(a) Daily misinformation spread rate.

(b) Spread composition and belief-state trajectory.

Figure 2: Propagation dynamics and the intervention mechanism.

Figure 2a first reveals a slight early rise in spread rate for Arg-En, ECTF, and CoAID, followed by a sustained decline, whereas Snopes does not show this pattern. To explain this phenomenon, Figure 2b further decomposes propagation into supportive and corrective or doubtful events. The early increase is mainly driven by doubt and correction rather than continued support. The spread-rate–trust correlations are 0.89, 0.93, 0.91, and 0.94 for Snopes, Arg-En, ECTF, and CoAID, respectively. As trust declines, supportive propagation and cross-hyperedge diffusion weaken. Thus, intervention reshapes public discourse rather than simply reducing total propagation.

4.4.3 Group Environment and Group Norms

We compare the group support rate, corrective-consensus share, and belief-state standard deviation on the initial day and the evaluation day across the Independent, Nested, and Overlap structures. The hyperedge partitions and intervention conditions remain fixed; only the statistical grouping method differs. Under Daily intervention, group support decreases, and corrective consensus increases across all four datasets and all three structure types, whereas the belief-state standard deviation changes only slightly. This indicates a consistent shift toward corrective group norms rather than complete homogenization of user beliefs. Detailed per-dataset and per-structure results are provided in Appendix E.3.

4.5 Persistent Effects across Information Variants

We further test whether the benefit of early intervention persists after intervention stops. At the end of the evaluation cycle, we retain users' Individual Memory and the hyperedge Correction History, and then introduce three variants of the same misinformation: Minor, Major, and Critical. No intervention is delivered during the following days. All other settings remain fixed. Across the three variants, Daily intervention consistently reduces both final misinformation trust and subsequent spread, and the benefit remains stable without continuous correction. This indicates that users reuse the Individual Memory and Correction History formed during the initial intervention when evaluating similar information. The full per-dataset and per-variant results, together with the detailed analysis, are provided in Appendix E.4.

5 Conclusion

We introduce a social-hypergraph framework for post-exposure misinformation intervention that combines group-adaptive correction, relation-aware coordination, and temporal feedback. Across four datasets in our simulated environment, the framework consistently reduces misinformation trust and generally reduces propagation. Ablations further support the contributions of higher-order group structure and feedback mechanisms.

AI USE DISCLOSURE

Generative AI tools were used to assist with writing and language polishing, improve grammar, help identify and check relevant literature, and generate some of the figures. They were not used for research ideation, methodology design, experimental implementation, data analysis, or result interpretation. All AI-assisted content was reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work studies post-exposure misinformation intervention entirely within an LLM-agent simulation and does not expose or intervene on real users. The simulated user profiles include demographic and psychological attributes to represent heterogeneous agents and construct local social groups. These attributes, as well as the behavior generated by LLM agents, may reproduce demographic stereotypes, cultural biases, or inaccurate assumptions about human responses. The reported results should therefore be interpreted as simulation-based evidence rather than evidence of intervention effectiveness on real populations. Group-adaptive correction techniques may also be misused for targeted persuasion or manipulation. We mitigate these risks by restricting the experiments to a controlled research environment, grounding corrections in dataset-provided factual evidence, and reporting aggregate outcomes. Any real-world deployment would require privacy review, bias auditing, human oversight, and validation with the affected communities.

REPRODUCIBILITY STATEMENT

Upon acceptance, we will publicly release all author-owned code to reproduce the reported results. Moreover, we provide the complete simulation and intervention workflow in Algorithm 1. Section 3 defines the user states, propagation process, intervention mechanisms, and evaluation metrics. The appendices document the hypergraph construction procedure and structural constraints, intervention-strategy taxonomy and prompts, Group Doubt Ratio, relation-aware coordination, and Correction History update rules. Section 4 reports the datasets, agent models, simulation environment, baselines, and evaluation protocol, while the supplementary experiments provide the control settings and full results. Across compared conditions, we hold the user pool, hypergraph, initial receivers, propagation process, and evaluation window fixed and vary only the intervention mechanism.

REFERENCES

Cosimo Agostinelli, Marco Mancastroppa, and Alain Barrat. Group adaptation drives opinion dynamics in higher-order networks. *APS Open Science*, 1:000028, 2026.

Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 337–371, 2023.

Unai Alvarez-Rodriguez, Federico Battiston, Guilherme Ferraz de Arruda, Yamir Moreno, Matjaž Perc, and Vito Latora. Evolutionary dynamics of higher-order interactions in social networks. *Nature Human Behaviour*, 5(5):586–595, 2021.

Federico Battiston, Valerio Capraro, Fariba Karimi, Sune Lehmann, Andrea Bamberg Migliano, Onkar Sadekar, Angel Sánchez, and Matjaž Perc. Higher-order interactions shape collective human behaviour. *Nature Human Behaviour*, 9(12):2441–2457, 2025.

Federico Battiston, Christian Bick, Maxime Lucas, Ana P. Millán, Per Sebastian Skardal, and Yuanzhao Zhang. Collective dynamics on higher-order networks. *Nature Reviews Physics*, 8 (3):146–159, 2026.

Ronald L. Breiger. The duality of persons and groups. *Social Forces*, 53(2):181–190, 1974.

Lucy H. Butler, Nicolas Fay, and Ullrich K. H. Ecker. Social endorsement influences the continued belief in corrected misinformation. *Journal of Applied Research in Memory and Cognition*, 12 (3):364–375, 2023.

Damon Centola. An experimental study of homophily in the adoption of health behavior. *Science*, 334(6060):1269–1272, 2011.

Man-pui Sally Chan and Dolores Albarracín. A meta-analysis of correction effects in science-relevant misinformation. *Nature Human Behaviour*, 7(9):1514–1525, 2023.

Canyu Chen and Kai Shu. Combating misinformation in the age of LLMs: Opportunities and challenges. *AI Magazine*, 45(3):354–368, 2024.

Eun Cheol Choi, Lindsay E. Young, and Emilio Ferrara. Overstating attitudes, ignoring networks: Llm biases in simulating misinformation susceptibility. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 20, pp. 520–547, 2026.

Yuwei Chuai, Haoye Tian, Nicolas Pröllochs, and Gabriele Lenzini. Did the roll-out of community notes reduce engagement with misinformation on x/twitter? *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–52, 2024.

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

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

Thomas H. Costello, Gordon Pennycook, and David G. Rand. Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714):eadq1814, 2024.

Limeng Cui and Dongwon Lee. Coaid: Covid-19 healthcare misinformation dataset, 2020.

Guilherme Ferraz de Arruda, Alberto Aleta, and Yamir Moreno. Contagion dynamics on higher-order networks. *Nature Reviews Physics*, 6:468–482, 2024.

DeepSeek AI. Deepseek api documentation: Models and pricing, 2026. Accessed: 2026-09-26.

Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559, 2016.

Tim Donkers and Jürgen Ziegler. Understanding online polarization through human-agent interaction in a synthetic LLM-based social network. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pp. 457–478, 2025.

Chiara Patricia Drolsbach, Kirill Solovev, and Nicolas Pröllochs. Community notes increase trust in fact-checking on social media. *PNAS Nexus*, 3(7):pgae217, 2024.

Zhaoyang Duan, Jiangwei Yan, Dong Xue, Fangzhou Liu, and Yang Tang. Opinion dynamics on higher-order social networks: A continuous-time perspective. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 56(8):4612–4624, 2026.

Ullrich K. H. Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K. Fazio, Nadia Brashier, Panayiota Kendeou, Emily K. Vraga, and Michelle A. Amazeen. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1 (1):13–29, 2022.

Ullrich K. H. Ecker, Jasmyne A. Sanderson, Paul McIlhiney, Jessica J. Rowsell, Hayley L. Quekett, Gordon D. A. Brown, and Stephan Lewandowsky. Combining refutations and social norms increases belief change. *Quarterly Journal of Experimental Psychology*, 76(6):1275–1297, 2023.

Saadia Gabriel, Liang Lyu, James Siderius, Marzyeh Ghassemi, Jacob Andreas, and Asu Ozdaglar. Misinfoeval: Generative AI in the era of "alternative facts", 2024.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1259, 2024.

Kobi Hackenburg, Ben M. Tappin, Luke Hewitt, Ed Saunders, Sid Black, Hause Lin, Catherine Fist, Helen Margetts, David G. Rand, and Christopher Summerfield. The levers of political persuasion with conversational artificial intelligence. *Science*, 390(6777):eaea3884, 2025.

Chen Han, Wenzhen Zheng, and Xijin Tang. Debate-to-detect: Reformulating misinformation detection as a real-world debate with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15114–15129, 2025.

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

Katrin Hartwig, Frederic Doell, and Christian Reuter. The landscape of user-centered misinformation interventions—a systematic literature review. *ACM Computing Surveys*, 56(11):1–36, 2024.

Bing He, Mustaque Ahamad, and Srijan Kumar. Reinforcement learning-based counter-misinformation response generation: A case study of COVID-19 vaccine misinformation. In *Proceedings of the ACM Web Conference*, 2023.

Nathan O. Hodas and Kristina Lerman. The simple rules of social contagion. *Scientific Reports*, 4: 4343, 2014.

Emma Hoes, Brian Aitken, Jingwen Zhang, Tomasz Gackowski, and Magdalena Wojcieszak. Prominent misinformation interventions reduce misperceptions but increase scepticism. *Nature Human Behaviour*, 8(8):1545–1553, 2024.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 22109–22117, 2024.

Bukyoung Jhun. Effective epidemic containment strategy in hypergraphs. *Physical Review Research*, 3(3):033282, 2021.

Pingping Jia and Zengyi Zhang. Misinformation dissemination and correction in online informal groups: The role of social distance and group dynamics. *New Media & Society*, 28(9):4544–4565, 2025.

Jung-Ho Kim and K.-I. Goh. Higher-order components dictate higher-order contagion dynamics in hypergraphs. *Physical Review Letters*, 132(8):087401, 2024.

David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan Zittrain. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

Chance Jiajie Li, Jiayi Wu, Zhenze Mo, Ao Qu, Yuhan Tang, Kaiya Ivy Zhao, Yulu Gan, Jie Fan, Jiangbo Yu, Jinhua Zhao, Paul Pu Liang, Luis Alonso, and Kent Larson. Position: Simulating society requires simulating thought, 2025.

Jiachen Li, Wenjie Li, Feng Gao, Meng Cai, Zengping Zhang, Xiaoyang Liu, and Wei Wang. Social contagions on higher-order community networks. *Applied Mathematics and Computation*, 478: 128832, 2024.

Genglin Liu, Vivian Le, Salman Rahman, Elisa Kreiss, Marzyeh Ghassemi, and Saadia Gabriel. Mosaic: Modeling social AI for content dissemination and regulation in multi-agent simulations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6401–6428, 2025a.

Shuliang Liu, Hongyi Liu, Aiwei Liu, Bingchen Duan, Qi Zheng, Yibo Yan, He Geng, Peijie Jiang, Jia Liu, and Xuming Hu. A survey on proactive defense strategies against misinformation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18144–18155, 2025b.

Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

Arkadiusz Modzelewski, Witold Sosnowski, Tiziano Labruna, Adam Wierzbicki, and Giovanni Da San Martino. Pcot: Persuasion-augmented chain of thought for detecting fake news and social media disinformation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24959–24983, Vienna, Austria, 2025. Association for Computational Linguistics.

Kleber Andrade Oliveira, Pietro Traversa, Guilherme Ferraz de Arruda, and Yamir Moreno. Rumor propagation on hypergraphs. *Nature Communications*, 2026.

Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

Xinle Pang, Danding Wang, Qiang Sheng, Yifan Sun, Beizhe Hu, and Juan Cao. Tailoring rumor debunking to you: Diversifying chinese rumor-debunking passages with an LLM-driven simulated feedback-enhanced framework. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics, Volume 5: Industry Track*, pp. 586–597, 2026.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023.

Toby Prike, Lucy H. Butler, and Ullrich K. H. Ecker. Source-credibility information and social norms improve truth discernment and reduce engagement with misinformation online. *Scientific Reports*, 14(1):6900, 2024.

Boyu Qiao, Kun Li, Wei Zhou, and Songlin Hu. Dynamic simulation framework for disinformation dissemination and correction with social bots. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 4688–4710, 2025.

Anand S. Rao and Michael P. Georgeff. Bdi agents: From theory to practice. In *Proceedings of the First International Conference on Multi-Agent Systems*, pp. 312–319. AAAI Press, 1995.

Daniel Russo, Shane Peter Kaszefski-Yaschuk, Jacopo Staiano, and Marco Guerini. Countering misinformation via emotional response generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11476–11492, 2023.

Sougata Saha and Rohini Srihari. Integrating argumentation and hate-speech-based techniques for countering misinformation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11109–11124, 2024.

Takumi Sakiyama, Kazuki Nakajima, and Masaki Aida. Efficient intervention in the spread of misinformation in social networks. *IEEE Access*, 12:133489–133498, 2024.

Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. On the conversational persuasiveness of GPT-4. *Nature Human Behaviour*, 9(8):1645–1653, 2025.

Cecilie S. Traberg, Trisha Harjani, Jon Roozenbeek, and Sander van der Linden. The persuasive effects of social cues and source effects on misinformation susceptibility. *Scientific Reports*, 14 (1):4205, 2024.

John C. Turner, Michael A. Hogg, Penelope J. Oakes, Stephen D. Reicher, and Margaret S. Wetherell. *Rediscovering the Social Group: A Self-Categorization Theory*. Basil Blackwell, 1987.

Sander van der Linden, Debra Louison-Lavoy, Nicholas Blazer, Nancy S. Noble, and Jon Roozenbeek. Prebunking misinformation techniques in social media feeds: Results from an Instagram field study. *Harvard Kennedy School Misinformation Review*, 2026.

Sameera Vithanage, Keith Ransom, Antonette Mendoza, and Shanika Karunasekera. The cognitive and social dynamics of social correction of misinformation: A scoping review. *Behaviour & Information Technology*, pp. 1–20, 2026.

Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

Nathan Walter and Riva Tukachinsky. A meta-analytic examination of the continued influence of misinformation in the face of correction: How powerful is it, why does it happen, and how to stop it? *Communication Research*, 47(2):155–177, 2019.

Sze Yuh Nina Wang, Samantha C. Phillips, Kathleen M. Carley, Hause Lin, and Gordon Pennycook. Limited effectiveness of psychological inoculation against misinformation in a social media feed. *PNAS Nexus*, 4(6):pgaf172, 2025.

Claire Wardle, Pete Brown, and David Scales. A research agenda for studying LLM-powered persuasion technology to mitigate misinformation. *Media and Communication*, 14:12442, 2026.

Dong Xue, Jiangwei Yan, Jingxi Zhu, Zhaoyang Duan, and Jie Huang. The opinion-dependent product spread over higher-order social networks. *Chaos, Solitons & Fractals*, 209:118490, 2026.

Chang Yang, Peng Zhang, Jing Zhang, Hui Gao, and Changhao Song. Rethink rumor detection in the era of llms: A review. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 8730–8749, 2025.

Qian Yin, Yicong Wan, Zhishuang Wang, Linyuan Guo, and Zhiyong Hong. Dynamics of competitive information propagation on hypergraphs. *Chaos, Solitons & Fractals*, 210:118667, 2026.

Kun Zhang, Xiaoyan Yu, Hao Peng, Zhe Yang, Ye Tian, Hao Jin, Tuoyu Feng, and Hui Lin. Towards efficient optimization of multi-agent social simulation via large language models. *International Journal of Machine Learning and Cybernetics*, 17(1):1, 2026.

Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, Guanying Li, Ling Yan, Yao Hu, Siming Chen, Yu Wang, Xuanjing Huang, Jiebo Luo, Shiping Tang, Libo Wu, Baohua Zhou, and Zhongyu Wei. SocioVerse: A world model for social simulation powered by LLM agents and a pool of 10 million real-world users. *arXiv preprint arXiv:2504.10157*, 2025.

Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5):1–40, 2020.

A End-to-End Intervention Procedure

This appendix records the full diffusion–intervention loop in one place, from seed exposure to schedule-triggered group intervention and final metric computation (Rao & Georgeff, 1995). The point is not to add another result table, but to make the reported appendix evidence reproducible and easy to follow.

Algorithm 1 Hyperedge-level misinformation diffusion and intervention

Require: User profiles P_i, misinformation x, simulation horizon T, intervention schedule
Ensure: Final user states, Correction Histories, and outcomes

Construct $\mathcal{H} = (\mathcal{U}, \mathcal{E})$, and initialize M_i^{ind}, M_i^{soc}, and Correction Histories
Select 1–3 initial recipients from each hyperedge and deliver x
for each simulation day $t = 1, \ldots, T$ **do**
    Aggregate misinformation and public comments received on day t
    for all exposed users u_i **do**
        Update $M_i^{\mathrm{soc}}(t)$ from the comments received through affiliated hyperedges
        Form a pre-intervention judgment using $f_{\mathrm{BDI}}(P_i, x, M_i^{\mathrm{ind}}(t^-), M_i^{\mathrm{soc}}(t))$ and the relevant Correction History
        Store $(b_i(t), c_i(t), \ell_i(t))$ in $M_i^{\mathrm{ind}}(t)$
    end for
    if t is a scheduled intervention day **then**
        Identify the risk set $\mathcal{E}^{\mathrm{risk}}(t)$: a hyperedge is risky if at least one activated member has the latest stance label `support`
        for all $e \in \mathcal{E}^{\mathrm{risk}}(t)$ **do**
            Compute $G_e(t)$, the proportion of activated members in e whose latest stance is `doubt` or `correct`
            Generate two group-specific corrections: one that amplifies existing `doubt` or `correct` views, and one that counters the supporting reasons; omit $G_e(t)$ when its value is zero
        end for
        for all activated users covered by one or more risk hyperedges **do**
            Coordinate the generated corrections and $G_e(t)$ according to nested and overlapping relations: use the child hyperedge for nested risk groups and the higher-similarity leading hyperedge for overlapping regions, while retaining only consistent supplementary content
            Record the user's pre-intervention comment, stance, and action
            Update $(b_i(t), c_i(t), \ell_i(t))$ using the coordinated intervention and the latest Individual Memory
            if u_i was newly exposed on day t **then**
                Determine the final spread-or-silence decision $a_i(t)$
            end if
            Update the corresponding hyperedge- or region-level Correction History using pre- and post-intervention comments, stance labels, and observable behaviors
        end for
    end if
    for all exposed users with $a_i(t) =$ spread **do**
        Select eligible Ignorant recipients from affiliated hyperedges
        Schedule x and $c_i(t)$ for delivery on day $t + 1$
    end for
end for
Compute the final trust and spread rate
return final user states, Correction Histories, and outcomes

Table 4: Projected comparison with random membership and flat group intervention. Control values are projected from matched-subset ratios and are not direct measurements. Lower is better; bold marks the best value in each row.

Dataset	Metric	Ours	Rand. Member.	Flat
Arg-En	Trust_{D_7}	**0.2683**	0.3451	0.3852
Arg-En	$\mathrm{SpreadRate}_{D_7}$	0.0696	0.0633	**0.0187**
Snopes	Trust_{D_7}	**0.2678**	0.3435	0.3725
Snopes	$\mathrm{SpreadRate}_{D_7}$	0.0623	0.0317	**0.0093**
CoAID	Trust_{D_7}	**0.1963**	0.2316	0.2683
CoAID	$\mathrm{SpreadRate}_{D_7}$	0.2089	0.2914	**0.1780**
ECTF	Trust_{D_7}	0.2195	**0.2189**	0.2864
ECTF	$\mathrm{SpreadRate}_{D_7}$	**0.0748**	0.6357	0.3140
Overall	Trust_{D_7}	**0.2380**	0.2856	0.3284
Overall	$\mathrm{SpreadRate}_{D_7}$	0.1039	0.1537	**0.0804**

B Projected Results for Random Membership and Flat Intervention

We further evaluate degree-matched random membership (Rand. Member.) and a non-adaptive flat group intervention (Flat) on a matched subset. To compare them with the main experiment, we apply each control's relative change on the matched subset to the corresponding main-experiment value, including the pooled aggregates. The results preserve the observed relative effects of both controls, but remain projections rather than direct measurements under the full setting.

Table 4 reports the projected results. Random membership performs worse than Ours, increasing Trust_{D_7} by 28.6%, 28.3%, and 18.0% on Arg-En, Snopes, and CoAID, respectively, while ECTF remains nearly unchanged; pooled trust increases by 20.0%. This indicates that matching group size and degree statistics is insufficient when the actual composition of the intervention audience is disrupted. Homophily organizes users with shared attributes into familiar interaction contexts, where similarity and prior social proximity facilitate social influence and behavioral diffusion (McPherson et al., 2001; Centola, 2011). In misinformation correction, lower social distance and stronger group dynamics similarly improve the acceptance and effectiveness of corrective messages within informal groups (Jia & Zhang, 2025). Randomly composed groups lack this shared context and interpersonal familiarity, making users less responsive to group-level correction. The propagation effect is also heterogeneous, ranging from a 49.1% decrease on Snopes to a 749.9% increase on ECTF, with the pooled spread rate increasing from 0.1039 to 0.1537. The large ECTF change should be interpreted cautiously because the main-experiment baseline is low.

Flat intervention increases Trust_{D_7} by 30.5%–43.6% across datasets and by 38.0% overall. Its lower spread on Arg-En, Snopes, and CoAID does not indicate stronger correction, because trust reduction is consistently weaker; on ECTF, spread is 319.8% higher. This pattern suggests reduced engagement or prolonged silence rather than effective group-adaptive correction.

Together, the two controls isolate the importance of composition-aware grouping and group-adaptive intervention. Random membership preserves the intervention machinery but disrupts the social composition that defines the audience, whereas flat intervention preserves the group structure but removes adaptation. Ours achieves the lowest or nearly lowest trust and the most balanced propagation behavior, supporting the necessity of both homophily-based hyperedges and coordinated group-level intervention.

C Risk Hyperedges and Belief Trajectories

C.1 Number of High-Risk Hyperedges

This analysis asks whether Daily intervention continuously reduces the risky groups that need treatment. It also asks whether such a change occurs only in the final metric. Following the operational risk criterion in the method, a hyperedge is identified as risky when at least one activated user has

a latest visible comment label of `support` or `doubt`. This rule relies only on observable public responses and does not use users' private belief states. The Intervention Agent does not access $b_i(t)$. Belief-based quantities are computed offline for diagnostic analysis only and cannot affect risk identification, strategy selection, or intervention generation. For each propagation day, we compute the average number of high-risk hyperedges across the four datasets. We compare no intervention with Daily under the same hypergraph and news settings.

Figure 3 shows that the number of high-risk hyperedges remains high under no intervention. On the last day, the average numbers under no intervention are 7.47, 6.03, 5.60, and 7.35 for Snopes, ECTF, CoAID, and Arg-En, respectively. Under Daily, the numbers decrease to 0.16, 0.24, 0.69, and 0.14. Daily remains close to no intervention on Day 1, but then drops quickly. This indicates that the risk group shrinks during propagation, rather than changing only at the final metric.

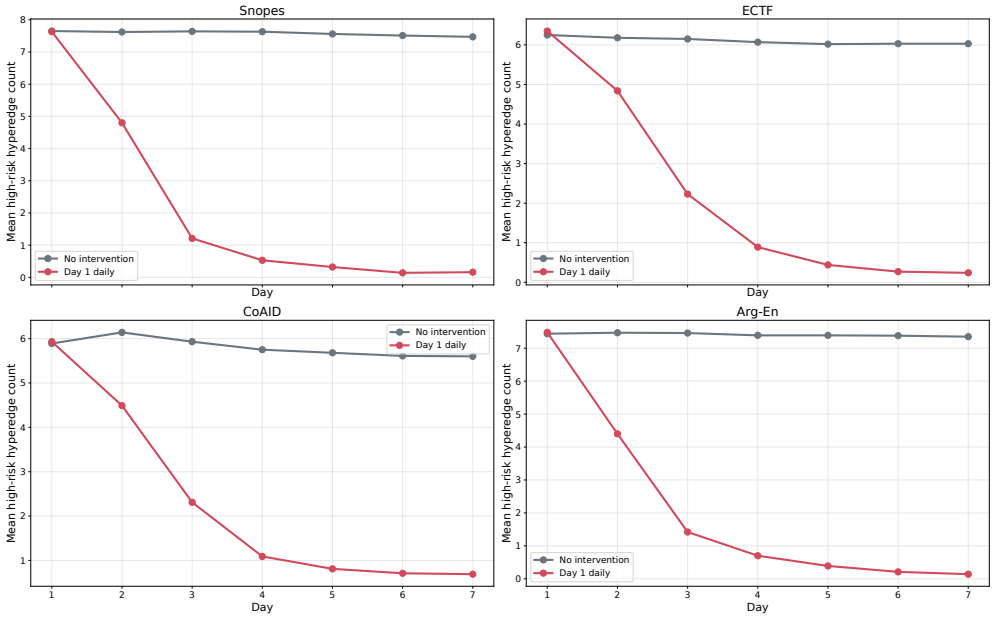


Figure 3: Mean number of high-risk hyperedges by day under no intervention and Daily.

The trajectory shows that the intervention reduces the size of the risk group during propagation. The decline in high-risk hyperedges is close in time to the later drop in the propagation rate. This pattern is consistent with a process of risk localization, group broadcast, risk-edge reduction, and propagation contraction. However, the current analysis cannot isolate the causal contribution of each step.

Overall, this analysis provides process evidence for the main result. It shows that the final gain has daily intermediate states. The hyperedge-level risk metric can also serve as a direct diagnostic of whether the intervention reaches the target group.

C.2 Belief Trajectories of Activated Users

This analysis asks whether intervention changes the average belief trajectory of all activated users. It also asks whether belief declines quickly in the early rounds and then becomes stable. We compute the average belief state of activated users on each day. We also compare four intervention schedules in a sensitivity subset. These schedules start on Day 1 or Day 3 and use either daily or every-three-day intervention.

As shown in Figure 4, under Daily, the belief state of activated users decreases from 0.447 to 0.268 on Snopes, from 0.459 to 0.268 on Arg-En, from 0.415 to 0.219 on ECTF, and from 0.364 to 0.196 on CoAID. On the last day, the corresponding values under no intervention are 0.621, 0.620, 0.498, and 0.464. Figure 5 compares five conditions using the first five news items in each dataset. Most

schedules show a rapid decline in the first two days and then a slower change. The Day-7 results differ across datasets.

This trajectory shows that the average belief state of activated users drops quickly in the early intervention rounds. It then changes more slowly. The slower change may reflect several factors. The remaining users may have higher belief states. Their propagation chains may be longer. Their state updates may also require more rounds to appear. The disappearance of the high peak does not mean that every user drops to zero immediately. It means that the average advantage of supportive judgments in the group gradually decreases.

Therefore, the belief trajectory is consistent with cumulative gains from repeated intervention. However, it cannot separate repeated intervention from a one-shot text intervention. It also cannot directly

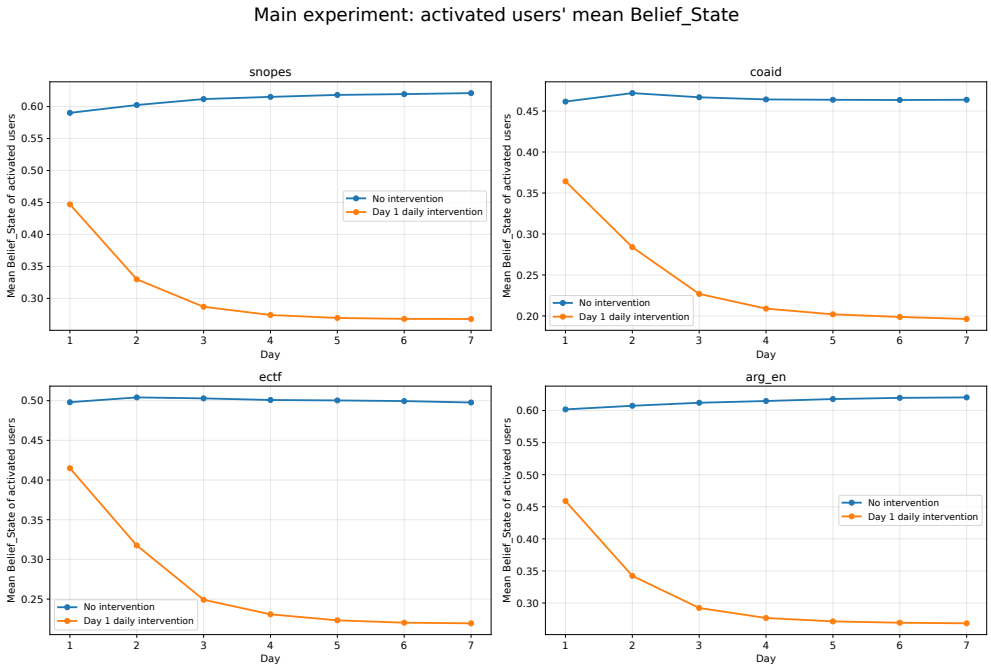


Figure 4: Mean belief state of activated users under no intervention and Daily in the main experiment.

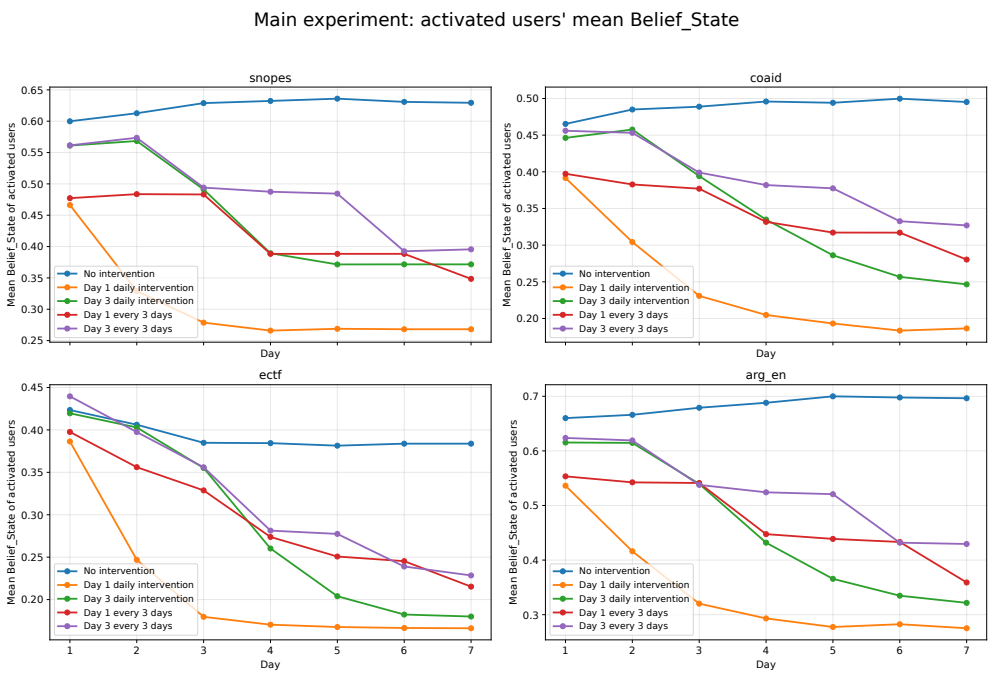


Figure 5: Mean belief state of activated users under no intervention and four intervention schedules.

show changes in public evidence or group norms. We therefore interpret the schedule differences only as sensitivity results.

D Hyperedge Structure Heterogeneity and Repeated Exposure

D.1 Evaluation-Day Effects Across Three Hyperedge Structures

This analysis asks whether Independent, Nested, and Overlap users show systematic differences in belief and behavior after intervention. We group activated users by their hyperedge structure. We then compute the Day-7 reduction in belief state and the Day-7 reduction in supportive propagation for each group. All structures use the same news items, user pool, initial receivers, and Daily schedule. Therefore, the group differences cannot be attributed simply to the hyperedge structure.

As shown in Figure 6, belief and propagation improvements have the same direction in all four datasets. No value is negative. In the order of Independent, Nested, and Overlap, the belief reductions are 0.310, 0.345, and 0.362 on Arg-En; 0.273, 0.361, and 0.370 on Snopes; 0.221, 0.281, and 0.287 on ECTF; and 0.194, 0.289, and 0.281 on CoAID. For propagation, the reductions are 0.178, 0.212, and 0.216 on Arg-En; 0.203, 0.372, and 0.340 on Snopes; 0.237, 0.242, and 0.244 on ECTF; and 0.171, 0.215, and 0.165 on CoAID.

The belief results show that structure nesting does not block corrective information from entering user judgment. Overlap and Nested users show larger belief reductions in several datasets. This may be related to receiving corrective information from multiple risk groups. The propagation results depend more on the dataset and the state of the propagation chain. Across the dataset–structure combinations, all three structures show positive propagation reductions. This indicates that the intervention is not concentrated in one structure type.

Overall, the intervention improvements differ across hyperedge structures in a descriptive sense. Under the same news items, user pool, and schedule, these differences may still be affected by user composition, initial receivers, and the local propagation state. The method uses structure information to organize interventions and broadcast ranges. It does not assume that one structure is naturally easier to correct.

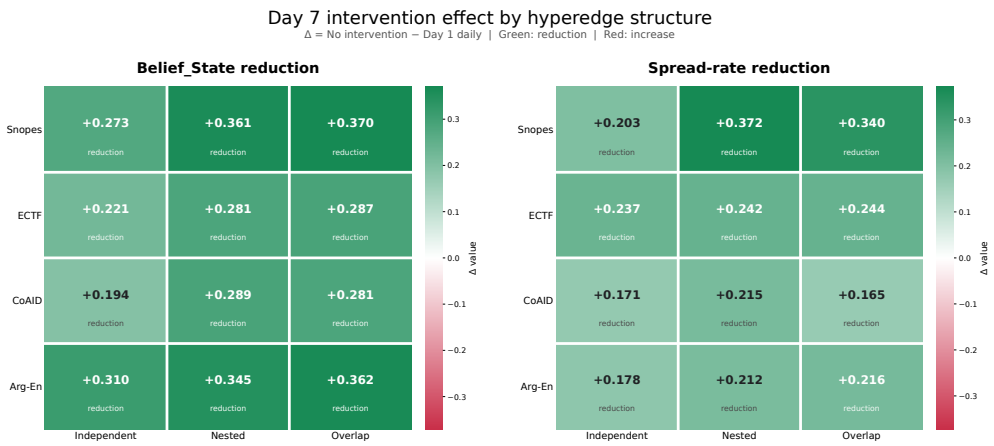


Figure 6: Cycle-end belief and propagation improvements across three hyperedge structures.

D.2 Repeated Exposure and Final Belief State

This analysis asks whether users who receive intervention multiple times form a lower final belief state and whether the gain shows clear saturation. We select users who receive at least three interventions within seven days. We keep the same user–news cohort and compare the average belief state before the first intervention, after the first, second, and third interventions, and after the final intervention. This analysis describes the association between repeated exposure and belief change. It does not treat different exposure counts as random doses.

As shown in Figure 7, on Snopes, the same cohort decreases from 0.648 to 0.534, 0.379, 0.242, and 0.206. This result includes 287 user–news observations from 36 news items. On ECTF, the cohort decreases from 0.646 to 0.547, 0.417, 0.293, and 0.235, with 519 observations from 55 news items. On CoAID, it decreases from 0.584 to 0.470, 0.338, 0.240, and 0.199, with 689 observations from 51 news items. On Arg-En, it decreases from 0.737 to 0.626, 0.454, 0.309, and 0.237, with 359 observations from 30 news items. All four datasets enter a smaller change range after the third intervention.

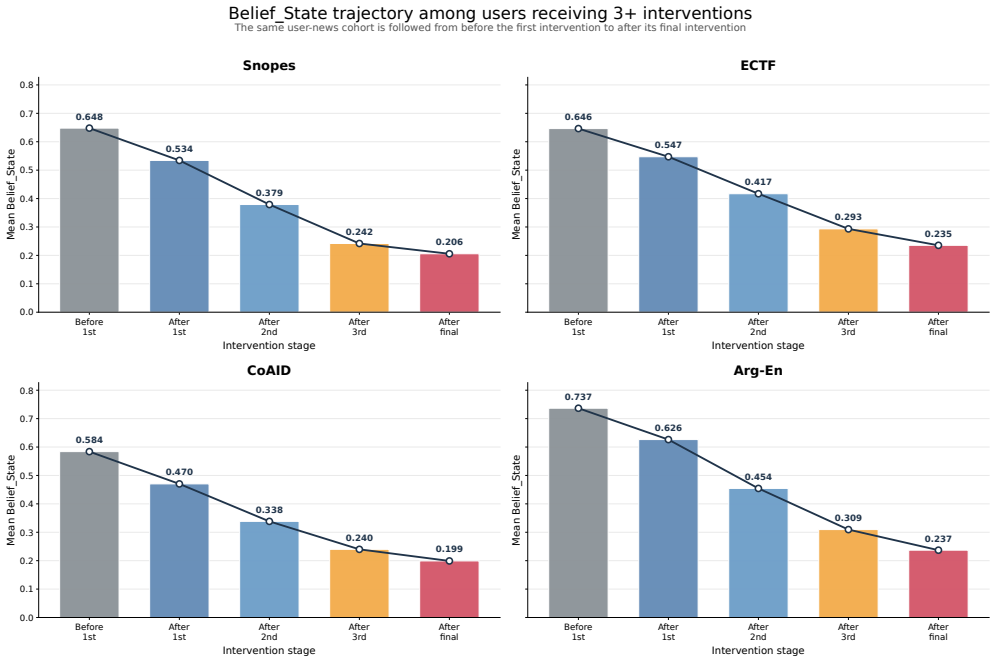


Figure 7: Average belief trajectory of users who receive at least three interventions.

These results show that repeated exposure and belief decline follow a consistent temporal order. Early intervention brings the largest change. Later gains may decrease because the remaining high-risk users become fewer and a corrective group consensus has formed (Walter and Tukachinsky, 2019; Chan and Albarracin, 2023). Users in this cohort may also have higher initial belief states or more complex positions in the propagation process. Therefore, this result cannot replace an experiment that randomly assigns the number of exposures.

Overall, the repeated exposure result supports the need for multiple intervention rounds. It also suggests that future work may reduce unnecessary later broadcasts according to risk state, rather than increase intervention frequency without limits.

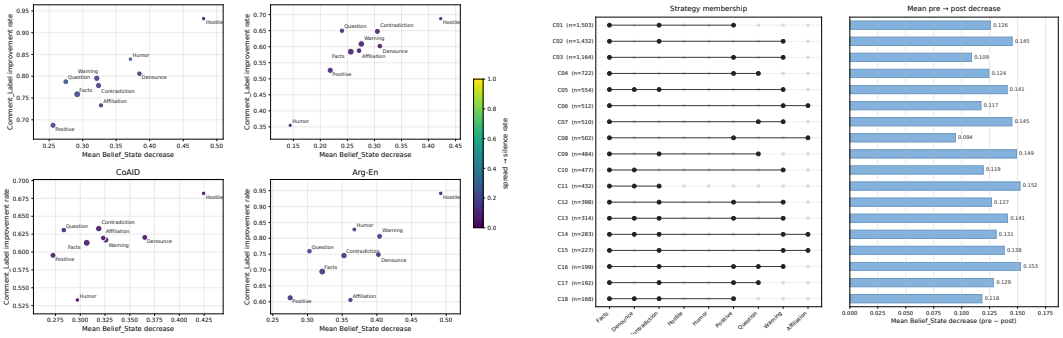
E Strategy Combinations, Public Discourse, and Group Norms: Full Analysis

E.1 Immediate Effects of Single Strategies and Strategy Combinations

This analysis asks whether intervention content matters beyond delivering corrective information. For each intervention event, we match the treated user with the no-intervention state of the same user, news item, and day. We then compare the belief-state decrease, label improvement rate, and spread-to-silence rate across strategy labels. All other components remain fixed.

In the matched statistics from the main-experiment logs, Presenting Facts appears in $14{,}240$ events, with a belief-state decrease of 0.293, a label improvement rate of 0.651, and a spread-to-silence rate of 0.116. Pointing out Hypocrisy or Contradiction and Warning of Consequences show comparable belief-state decreases of 0.325 and 0.318, respectively. Counter Question has a lower belief-state decrease of 0.276, but the highest label improvement rate among these strategies at 0.708. Positive Tone and Affiliation show weaker or more moderate effects. Hostile Language has the largest belief-state decrease at 0.450, but it corresponds to only 518 events and is therefore not a stable conclusion. Figure 8a reports these single-strategy effects.

We identify 148 strategy combinations. As shown in Figure 8b, the 18 most frequent combinations retain Presenting Facts as a common component in most cases, while contradiction, consequence warning, or positive tone are added for group-specific adaptation. These results indicate that factual



(a) Immediate effects associated with individual strategy labels.

(b) Frequent strategy combinations in effective interventions.

Figure 8: Intervention strategy content and effective strategy combinations.

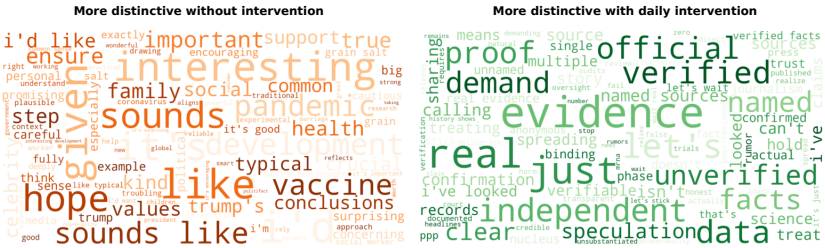


Figure 9: High-frequency difference words in user comments.

correction provides the common basis, whereas auxiliary strategies shape whether users are willing to reconsider the claim. They do not support a single universal strategy. Instead, they support dynamic strategy selection conditioned on the news content, group profile, and current discourse state. Because effective events often contain multiple strategies and event counts vary substantially, the single-strategy means reflect associations under co-occurrence rather than independent causal effects.

E.2 Discourse Shifts in Comments and Action Motives

This analysis uses 12,783 matched user–news states On the last day. We pair the Daily and no-intervention conditions by news item and user, and use TF-IDF differences to examine whether intervention shifts public comments from subjective judgment toward evidence checking and action motives from maintaining a stance toward protecting the group.

At the comment level, the no-intervention condition emphasizes personal judgment, prior beliefs, and support, whereas the Daily condition emphasizes evidence, verification, official information, and sources. At the action-motive level, no intervention emphasizes personal expression, cautiousness, and maintaining existing views, whereas Daily intervention emphasizes protecting others and preventing the spread of unverified information. Figure 9 shows the comment-level shift, and Figure 10 shows the corresponding shift in action motives.

These results indicate that intervention changes not only internal belief but also the reasons users provide for their judgments and actions. Once evidence becomes more prominent in public discourse, supportive propagation faces clearer verification standards. This provides an observable discourse-level mechanism consistent with GDR and public comment broadcasting. Because the evidence is based on word-frequency differences, it cannot establish belief change for every comment. It complements the individual belief trajectories by showing how the group-level discourse environment changes.

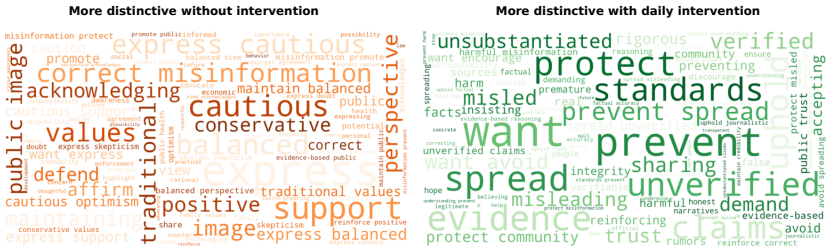


Figure 10: High-frequency difference words in user action motives.

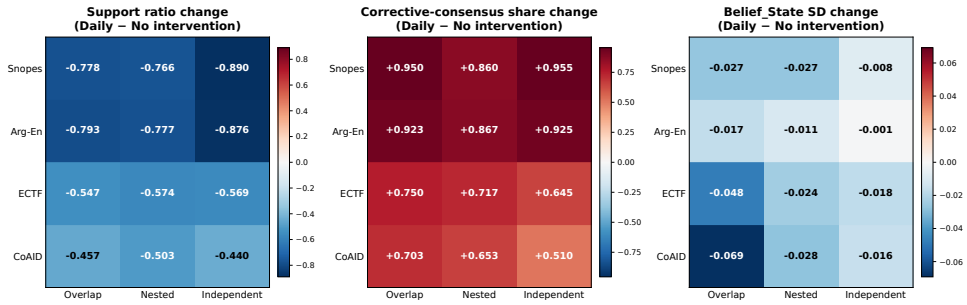


Figure 11: Group-level changes in support rate, corrective consensus, and belief-state standard deviation across hyperedge structures.

E.3 Full Results on Group Norms

This analysis examines whether the group support rate and corrective consensus remain consistent across the full data. We calculate both metrics for the Independent, Nested, and Overlap structures on all four datasets.

Figure 11 shows that, under Daily, the support rate decreases relative to no intervention by 0.766 to 0.890 on Snopes, 0.777 to 0.876 on Arg-En, 0.547 to 0.574 on ECTF, and 0.440 to 0.503 on CoAID. The corrective-consensus share increases by 0.860 to 0.955 on Snopes, 0.867 to 0.925 on Arg-En, 0.645 to 0.750 on ECTF, and 0.510 to 0.703 on CoAID. The standard deviation of belief decreases in all datasets and structures, but only by a small margin.

The main consistent change is therefore a decline in supportive discourse and an increase in corrective consensus, rather than full convergence to the same belief. The Overlap and Nested results indicate that local broadcasts, cross-group information flow, and attribute matching jointly shape group norms. These findings corroborate the GDR ablation: reducing individual belief without changing the publicly visible distribution of group opinions is unlikely to sustain lower belief and propagation.

E.4 Full Results Across Variants

This tests whether the cross-variant persistent effect holds across all datasets and variant types. As in Section 4.5, Daily intervention runs for fixed days. Individual Memory and hyperedge- and region-level Correction History are then retained, while Minor, Major, and Critical variants propagate for seven more days without intervention.

Table 5 reports all 24 comparisons. Early intervention is better in every case: average trust decreases from 0.5454 to 0.2252, a reduction of 58.7%, and the spread rate decreases from 0.3638 to 0.0674, a reduction of 81.5%. However, variant severity is not monotonic. For example, Critical has the lowest spread rate on Arg-En, whereas Major is lower than Critical on ECTF. Thus, severity alone does not determine the later effect; variant semantics, user doubt, and correction history may also shape propagation.

Because Individual Memory and hyperedge- and region-level Correction History are both retained, the results support a joint persistent effect but cannot isolate their separate contributions. Overall, the intervention influence extends beyond a single news item, while variant severity should not be treated as a monotonic difficulty variable.

Table 5: Cross-variant results under no intervention and Daily.

Dataset	Metric	Minor No Int.	Minor Daily	Major No Int.	Major Daily	Critical No Int.	Critical Daily
Arg-En	Trust	0.6877	**0.2243**	0.6786	**0.2350**	0.6844	**0.2257**
Arg-En	SpreadRate	0.5022	**0.0378**	0.5085	**0.0600**	0.4409	**0.0059**
Snopes	Trust	0.6363	**0.2773**	0.6579	**0.2730**	0.6525	**0.2867**
Snopes	SpreadRate	0.4074	**0.0000**	0.4637	**0.0211**	0.4251	**0.0000**
CoAID	Trust	0.5104	**0.2343**	0.4809	**0.2241**	0.4662	**0.2337**
CoAID	SpreadRate	0.3487	**0.1230**	0.3400	**0.1507**	0.3138	**0.1622**
ECTF	Trust	0.3559	**0.1535**	0.3674	**0.1698**	0.3662	**0.1644**
ECTF	SpreadRate	0.1944	**0.0800**	0.2239	**0.0650**	0.1969	**0.1030**

F Qwen3 Intervention Agent Extension

This appendix asks whether a lower-cost small local model can replace the main intervention agent when only a small amount of misinformation correction data is available. We therefore fine-tune a local Qwen3-8B model with LoRA and compare it with DeepSeek-V4Pro under the same intervention setting, starting from Day 1.

We build training pairs from VerMouth (Russo et al., 2023) and MisinfoCorrect (He et al., 2023), using the social media claim `tweet_claim` as input and the human-edited explanation `tweet_verdict` as the target. On $1{,}264$ held-out samples, we compare the base Qwen3-8B, LoRA model, and DeepSeek-V4Pro using ROUGE-L F1, Token-F1, BLEU-1, and BLEU-2. The LoRA model scores 0.2992, 0.4078, 0.3749, and 0.2558, respectively, and performs best on all four metrics. This indicates improved content organization, token coverage, and phrase matching in offline correction generation.

Table 6: Online intervention results of Qwen3-8B and the main intervention agent.

Dataset	Qwen3 Trust	Main Agent Trust	Qwen3 SpreadRate	Main Agent SpreadRate
Arg-En	0.3760	**0.2683**	0.5916	**0.0696**
Snopes	0.2826	**0.2678**	0.1707	**0.0623**
CoAID	0.2360	**0.1963**	**0.1868**	0.2089
ECTF	0.2694	**0.2195**	0.1920	**0.0748**

After transfer to the intervention process, Qwen3-8B still produces a clear correction effect, but generally remains weaker than the main agent (Table 6). It outperforms the main agent only on the evaluation-day spread rate for CoAID (0.1868). The largest gap appears on Arg-En, where Qwen3 obtains 0.5916 and the main agent obtains 0.0696. Thus, offline improvements do not yet translate into stable online advantages.

Overall, Qwen3-8B is an exploratory feasibility result rather than a replacement for the main experimental conclusions. Its main value is to show that a smaller local model can support correction generation and potentially reduce intervention cost. A promising next step is to train it on internal intervention data so that it can adapt to group states, historical information, and cross-round feedback, although this requires substantial data organization, cleaning, and training effort.

G Supplementary Robustness and Propagation-Structure Diagnostics

This appendix uses the logs of 100 news items from the main experiment and the corresponding full offline summaries. Unless otherwise stated, the statistical unit is the news source. In Appendix G.2, the unit is the matched intervention event.

G.1 Predictive Validity of the Risk Rule

This analysis asks whether a high-risk label is associated with later propagation and final user states. We compare high-risk and non-high-risk hyperedges in three aspects: next-day propagation, the number of activated users on Day 7, and the Day-7 belief state. We also report Precision and Recall for predicting later propagation. The analysis uses existing intervention and propagation logs without changing the model, hypergraph, or schedule.

Figure 12 shows that the mean differences in next-day propagation across the four datasets are 0.352, 0.344, 0.526, and 0.499. The corresponding differences in Day-7 activated users are 2.807, 2.418, 3.224, and 3.600, while those in Day-7 belief are 0.203, 0.283, 0.251, and 0.193. Precision is approximately 0.36, 0.69, 0.59, and 0.56, and Recall is approximately 0.96, 0.77, 0.90, and 0.96.

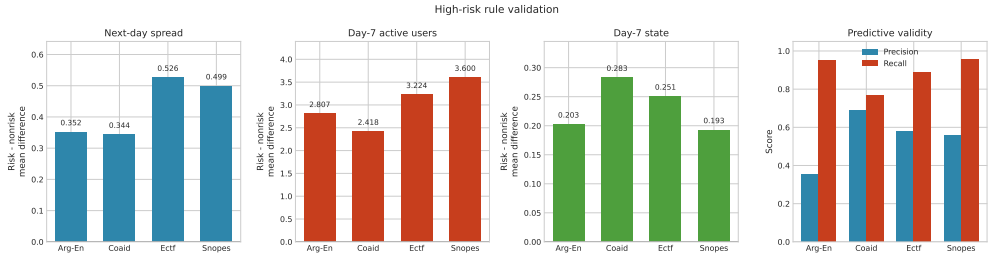


Figure 12: Predictive validity of the risk rule for next-day propagation, Day-7 activated users, final belief, and later propagation.

The risk rule therefore has high recall but moderate precision. This favors coverage of potential spreading groups, while also producing some unnecessary broadcasts. The evidence supports the predictive validity of the risk label, but not a causal effect of the label itself. Overall, the rule is suitable as a high-recall target filter; future work can further rank targets by propagation intention, bridge position, and group size.

G.2 Full Benefit Changes Under Cumulative Exposure

This analysis asks whether intervention benefit increases with exposure count and whether saturation appears. We extend the exposure count from 1 to 7 and compare average recipient benefit, marginal benefit, and sample size. Each intervention event is matched to the no-intervention state of the same dataset, news item, date, and user.

Figure 13 reports sample sizes of $7{,}859$, $4{,}485$, $1{,}854$, $1{,}036$, 605, 344, and 62 for exposure counts 1 to 7. The average benefits are approximately 0.128, 0.134, 0.121, 0.106, 0.093, 0.094, and 0.066. Benefit peaks near exposure count 2 and then declines overall. The adjacent changes are $+0.006$, -0.013, -0.015, -0.013, $+0.001$, and -0.028. Since later exposure groups are much smaller, uncertainty in the tail estimates also increases.

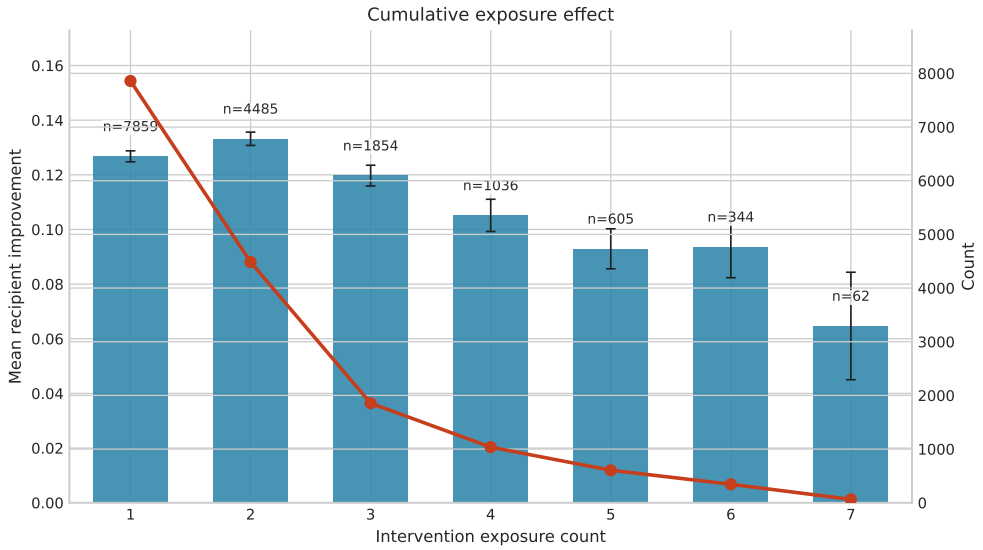


Figure 13: Mean benefit, marginal benefit, and sample size under different intervention exposure counts.

Early repeated exposure is associated with larger belief improvement, but the marginal benefit of additional broadcasts decreases as high-risk users become fewer and corrective consensus forms (Walter & Tukachinsky, 2019; Chan & Albarracín, 2023). Exposure groups are not randomly assigned, so the result cannot be interpreted as a fixed causal gain from one additional broadcast. The pattern motivates risk-triggered scheduling: continue broadcasts while group risk remains high, and reduce repetition when gains slow. An independent schedule-control experiment is still needed to test cost-effectiveness.

G.3 Penetration, Bridge Nodes, and Source-Level Benefit

This analysis asks whether source-level benefit is limited to low-penetration or simple propagation structures. For each news source, we calculate information penetration, the number of bridge nodes, and belief improvement under matched no-intervention and intervention conditions.

As shown in Figure 14, belief improvement in Arg-En rises from approximately 0.21 to 0.56 as penetration increases; in Snopes, it rises from approximately 0.31 to 0.49. The trend is weaker and more variable for CoAID and ECTF. All four datasets generally show higher improvement as the number of bridge nodes increases, although source-level scatter remains wide. Thus, news semantics and group states still create substantial variation under the same structural measure.

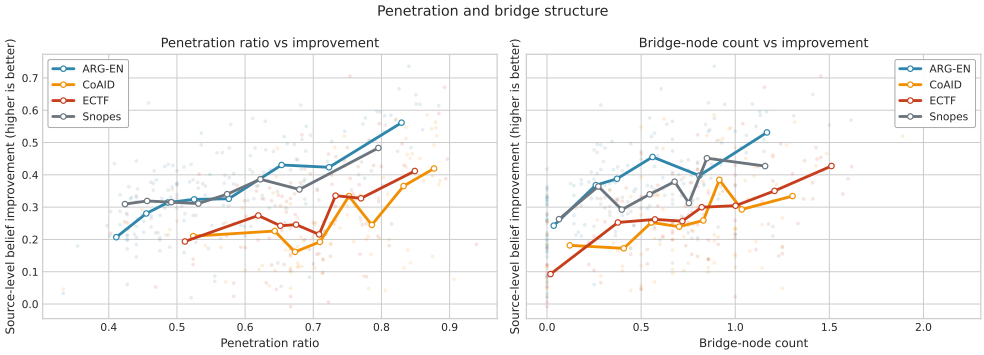


Figure 14: Source-level relationships between penetration, bridge-node count, and intervention benefit.

These results suggest that intervention is not restricted to low-penetration or simple propagation sources. They also identify bridge users and penetration as candidate indicators for later budget ranking. Because the curves are source-level associations and do not control for news topic, initial receivers, or exposure count, they should not be interpreted as direct causal effects of structural size.

H HYPEREDGE PROFILES AND INTERVENTION STRATEGIES

This appendix defines the PCoT content features, intervention-strategy taxonomy, two-stage strategy selection and message generation, and the Group Doubt Ratio (GDR). These components provide soft decision signals for the Intervention Agent and do not impose fixed beliefs, comments, or behaviors on individual users.

H.1 PCOT CONTENT FEATURES

Table 7 defines the six PCoT content features used to characterize misinformation (Modzelewski et al., 2025). Multiple features may be assigned to the same item. These labels describe the content itself, not the beliefs or attributes of the receiving group.

Table 7: Operational definitions of the six PCoT content features.

Feature	Operational definition
Attack on Reputation	The content attacks the credibility, moral character, or legitimacy of an individual, organization, or institution.
Simplification	The content reduces a complex issue to a simple causal claim, a binary choice, or an unsupported general conclusion.
Distraction	The content shifts attention away from the central claim or relevant evidence by introducing unrelated issues, blame, or emotional material.
Manipulative Wording	The content uses loaded, emotional, or framed language to guide judgment and increase acceptance of the claim.
Call to Action	The content urges recipients to share, act, or mobilize immediately, often before the claim has been verified.
Justification	The content supplies pseudo-reasons, unsupported explanations, or rationalizations that make the claim appear better supported than it is.

H.2 INTERVENTION STRATEGY TAXONOMY

Table 8 lists the nine intervention strategies (Saha & Srihari, 2024). Logic- and cognition-oriented strategies primarily address the persuasiveness of the claim, whereas emotion- and relationship-oriented strategies aim to increase audience receptiveness. Presenting Facts serves as the general factual basis; the remaining strategies provide complementary support and are not fixed templates.

Table 8: Definitions of the nine intervention strategies.

Category	Strategy	Description
Logic and Cognition	Presenting facts	Provides factual evidence and verified information to correct the claim.
	Denouncing	Criticizes or condemns the false viewpoint.
	Pointing out hypocrisy or contradiction	Identifies inconsistencies, hypocrisy, or contradictions in the claim or its source.
Emotion and Relationship	Hostile language	Uses hostile or confrontational language against the false claim.
	Humor and sarcasm	Uses humor or sarcasm to undermine the claim.
	Positive tone	Uses a positive and courteous tone to reduce audience defensiveness.
	Counter question	Uses a rhetorical question to encourage reconsideration of the claim.
	Warning of consequences	Warns about possible harms or consequences of believing or spreading the claim.
	Affiliation	Expresses empathy or establishes a shared identity to increase receptiveness.

H.3 Strategy Selection and Message Generation

For each risk hyperedge e, the agent uses the misinformation x, the dominant group profile P_e, the latest comments $C_e(t)$, and the current intervention round. It follows a two-stage plan-then-generate procedure. The first stage extracts content features and summarizes the group context. The second stage selects a strategy combination from Table 8 and generates the intervention messages.

INTERVENTION PROMPT

```
Analysis-stage prompt.

Inputs:
- News: {news}
- Mainstream group profile: {group_profile}
- Latest comments from activated users:
  {group_comments}
- Current intervention round:
  {intervention_round}

Tasks:
1. Select the applicable content features from:
  Attack on Reputation, Simplification, Distraction,
  Manipulative Wording, Call to Action, Justification.
2. Summarize the group's current consensus, including
  the main supporting, doubtful, and corrective views.
3. Summarize the mainstream audience characteristics
  from the group profile.

Return JSON:
news_features, group_consensus, audience_summary,
intervention_round.

Generation-stage prompt.

Inputs:
- Analysis from the first stage: {analysis}
- Doubt or correction comments: {doubt_comments}
- Supportive comments: {support_comments}
- Current intervention round: {intervention_round}

Tasks:
1. Generate amplify_correction to synthesize and clarify
  doubt or correction.
2. Generate weaken_support to weaken supportive reasons.
3. Select a dynamic combination from the nine strategies.
4. Do not assume that a fixed strategy or round requires
  a fixed combination.
5. Treat the intervention round as a soft stage indicator.
6. Do not invent unsupported facts, evidence, or sources.
7. Do not mention private cognitive states.

Fallback:
- If no supportive comment is provided, return an empty
  weaken_support string. Do not invent a supportive
  comment merely to rebut it.
- If no doubt or correction comment is provided, generate
  amplify_correction from the misinformation and available
  factual context.

Return JSON:
strategies, amplify_correction, weaken_support.
```

The two non-empty fields `amplify_correction` and `weaken_support` are delivered as separate intervention messages. If `weaken_support` is empty, only the first is delivered. The rounded GDR is also delivered as a separate intervention cue. For users shared by multiple risk hyperedges, the two message channels are coordinated separately according to the overlap and nesting relations in Appendix J.1, avoiding mechanical concatenation or conflicting content.

H.4 Group Doubt Ratio and Examples

Before intervention, let $A_e(t)$ denote the activated users in risk hyperedge e, and let $\ell_i(t)$ denote the latest comment label of user u_i. The Group Doubt Ratio is

$$G_e(t) = \frac{1}{|A_e(t)|} \sum_{u_i \in A_e(t)} \mathbb{I}\left[\ell_i(t) \in \{\text{doubt}, \text{correct}\}\right].$$

It measures the share of activated users whose latest public comment is doubtful or corrective (Prike et al., 2024; Ecker et al., 2023). The Intervention Agent converts this ratio to the nearest integer percentage and broadcasts the resulting GDR cue to the activated users in the hyperedge:

$$M_e(t) = \left\lfloor 100 G_e(t) + \frac{1}{2} \right\rfloor.$$

If $M_e(t) = 0$, no ratio-derived cue is sent. If $M_e(t) > 0$, the following message is broadcast to the activated users:

> Approximately $M_e(t)\%$ of exposed users in your group have questioned or corrected this news.

The cue reports only doubt or correction that already exists in the relevant hyperedge. It does not fabricate consensus, infer private belief states, or merge ratios from unrelated hyperedges, and it is not written into individual Correction History.

The following examples illustrate message construction. They are paraphrases rather than fixed templates.

> **Doubt/correction amplification example.** Suppose the analysis returns *Manipulative Wording* and *Simplification*, and the hyperedge contains a weak doubtful comment. The agent may combine *Affiliation*, *Positive tone*, and *Presenting facts*:
>
> Someone in the group has already raised a fair point: the post does not provide a verifiable source. Before passing it on, it is safer to check the available evidence.
>
> **Support-weakening example.** Suppose a supportive comment argues that the claim is credible because many people are sharing it, and the selected features are *Call to Action* and *Justification*. The agent may combine *Counter question*, *Pointing out hypocrisy or contradiction*, and *Warning of consequences*:
>
> Does "many people are saying it" provide verifiable evidence? The post turns an unsupported claim into a reason to act. Sharing it before verification may spread unverified information.
>
> **Group Doubt Ratio example.** If 3 of 10 activated users have latest labels `doubt` or `correct`, then $G_e(t) = 0.30$ and $M_e(t) = 30$. The ratio-derived cue is:
>
> Approximately 30% of exposed users in your group have questioned or corrected this news.

I Hypergraph Construction and Structural Guarantees

We construct the social hypergraph from profile similarity and then apply constrained overlap filtering to retain representative local groups. Spatial similarity is based on *area*; socio-identity similarity is based on *education*, *job*, *marital status*, *race*, *gender*, and *age*; psychological similarity is based on *ideology*, *Big Five personality traits*, and *religion*.

For each attribute, nominal values use exact matching:

$$\text{sim}_a(x, y) = \mathbb{I}[x = y].$$

Ordinal values use rank-based similarity:

$$\mathrm{sim}_a(x, y) = 1 - \frac{|\phi_a(x) - \phi_a(y)|}{L_a - 1},$$

where $\phi_a(x)$ is the rank of value x and L_a is the number of ordered levels. Area, age, and ideology are ordinal attributes.

Big Five traits use Jaccard similarity:

$$\mathrm{sim}_{\mathrm{big5}}(x, y) = \frac{|X_{\mathrm{big5}} \cap Y_{\mathrm{big5}}|}{|X_{\mathrm{big5}} \cup Y_{\mathrm{big5}}|}.$$

For each similarity dimension $d \in \{\mathrm{spa}, \mathrm{id}, \mathrm{psy}\}$, we average its attribute-level similarities:

$$s_{ij}^d = \frac{1}{|\mathcal{A}_d|} \sum_{a \in \mathcal{A}_d} \mathrm{sim}_a(P_i[a], P_j[a]).$$

The dimensions are combined by variance-based weights:

$$\omega_d = \begin{cases} \dfrac{v_d}{\sum_{d'} v_{d'}}, & \sum_{d'} v_{d'} > 0, \\ \dfrac{1}{|\mathcal{D}|}, & \text{otherwise}, \end{cases} \qquad v_d = \mathrm{Var}\left(\{s_{ij}^d : i < j\}\right),$$

and the overall similarity is

$$s_{ij} = \omega_{\mathrm{spa}} s_{ij}^{\mathrm{spa}} + \omega_{\mathrm{id}} s_{ij}^{\mathrm{id}} + \omega_{\mathrm{psy}} s_{ij}^{\mathrm{psy}}.$$

For each user, we construct an ego-centered candidate group:

$$\widetilde{e}_i = \{u_i\} \cup \mathrm{Top}_{k_i}\left(\mathcal{U} \setminus \{u_i\}; s_{ij}\right),$$

where $k_i = k = 12$ for all users.

We then retain candidates around centers that represent relatively distinct similarity regions. The neighborhood centrality of u_i is

$$c_i = \frac{1}{k} \sum_{u_j \in \mathrm{Top}_{k_i}} s_{ij}.$$

The first center is $r_1 = \arg\max_i c_i$. For each subsequent center, we maximize

$$d_i = 1 - \max_{u_r \in \mathcal{R}} s_{ir},$$

with ties broken by c_i and then by user identifier. This center selection avoids retaining multiple candidates from the same similarity region while preserving highly cohesive local groups.

Candidate duplication and overlap are controlled using the Jaccard coefficient:

$$J(e_p, e_q) = \frac{|e_p \cap e_q|}{|e_p \cup e_q|}.$$

Identical candidates are removed. For non-nested candidates, a pair is considered highly redundant when

$$J(e_p, e_q) \geq 0.8.$$

Containment is not removed at this stage because it may represent a valid nested group. A contained candidate is retained only when it has higher average internal similarity than its containing candidate, which selects a coherent local subgroup rather than an arbitrary subset.

The remaining candidates are prioritized by newly covered users, gain in relation-type coverage, average internal similarity, and group size, with identifiers used only as deterministic tie-breakers. Shared users are also constrained: overlapping hyperedges may share members, but the same shared-user set is not reused across several different overlap pairs. When an overlap causes repeated sharing

or excessive redundancy, the corresponding candidate is discarded. Thus, retained groups are determined by coverage, internal cohesion, and overlap control rather than by arbitrary hyperedge selection.

For the 45-user main simulation, the general candidate construction is instantiated through a deterministic structural procedure that retains overlap, nested, and independent relations while validating coverage, internal cohesion, and overlap constraints. This instantiation yields $M = 8$ hyperedges covering all users. Every retained hyperedge contains at least three users, and its size is determined by the retained candidate and its shared members.

The hypergraph is generated once and reused across all news items, intervention schedules, and comparison conditions. It is not resampled for individual news items or experimental runs, so all conditions are evaluated on the same fixed social structure.

J Intervention Coordination and Temporal Adaptation

J.1 Leading Hyperedge Selection for Overlapping Risk Regions

For an overlapping region R that belongs to multiple risk hyperedges, we treat the shared region as one coordination unit rather than selecting a leading source for each individual user. Let $\mathcal{E}_R^{\text{risk}}$ denote the set of risk hyperedges that contain R. For each risk hyperedge $e \in \mathcal{E}_R^{\text{risk}}$, we remove the shared region and obtain the non-shared part $e \setminus R$. The matching score between the overlapping region and the non-shared part of the hyperedge is defined as the average user–user similarity across the two parts:

$$S(R, e) = \frac{1}{|R|\,|e \setminus R|} \sum_{u_i \in R} \sum_{u_j \in e \setminus R} s_{ij}.$$

Here, s_{ij} is the overall similarity between users u_i and u_j, which combines spatial, identity, and psychological similarity. The leading hyperedge is selected as

$$e^* = \arg\max_{e \in \mathcal{E}_R^{\text{risk}}} S(R, e).$$

The selected hyperedge provides the primary correction for the overlapping region. The remaining risk hyperedges contribute only supplementary information that is consistent with the corrective direction of the leading hyperedge.As shown in Figure 15, different hypergraph risk structures lead to different intervention priorities.

For nested structures, the child hyperedge is used as the leading intervention source when both the parent and child are risky. The parent hyperedge provides supplementary information only. If the structure is strictly nested, the shared region is the child hyperedge itself, so the overlap-score comparison is not applied.

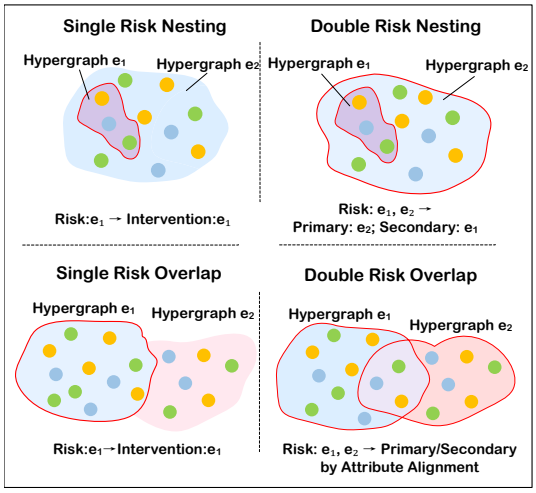


Figure 15: Hypergraph risk structures and intervention prioritization.

J.2 Correction History: Prompt, Replacement, and Feedback-Change Criteria

We maintain a separate Correction History $H_e^{\mathrm{corr}}(t)$ for each intervention scope. Ordinary hyperedges and overlapping or nested regions do not share histories. The summary is derived from the delivered correction and the pre- and post-intervention comments, stance labels, and observable propagation behaviors. It therefore relies only on public responses and does not use users' private belief states. Table 9 summarizes its update prompt, replacement rule, and response-change criteria.

Table 9: Correction History update, replacement, and response-change criteria.

Component	**Rule**
Update prompt	Given the previous history, the delivered correction, and the pre- and post-intervention comments, stance labels, and actions, summarize the observed response changes. Return `successful_strategies`, `failed_strategies`, `observed_changes`, and `recommended_adjustment`.
Replacement rule	The new summary replaces the previous record. Ordinary hyperedge histories and region-level histories remain separate and are not fused again. The update is $H_e^{\mathrm{corr}}(t{+}1) = \mathrm{Summarize}\left(H_e^{\mathrm{corr}}(t), m_e(t), \Delta C_e(t), \Delta L_e(t), \Delta A_e(t)\right),$ where $\Delta C_e(t)$, $\Delta L_e(t)$, and $\Delta A_e(t)$ denote changes in comments, stance labels, and actions, respectively.
Feedback-change criteria	Stance improvement includes $\mathrm{support} \rightarrow \mathrm{doubt}$, $\mathrm{support} \rightarrow \mathrm{correct}$, and $\mathrm{doubt} \rightarrow \mathrm{correct}$. Behavioral improvement is $\mathrm{spread} \rightarrow \mathrm{silence}$. Deterioration includes $\mathrm{correct}/\mathrm{doubt} \rightarrow \mathrm{support}$ and $\mathrm{silence} \rightarrow \mathrm{spread}$. Unchanged responses are treated as neutral, while opposite stance and behavior changes are retained as mixed outcomes.
Temporal use	In subsequent time steps, the corresponding Correction History is provided as historical corrective context during users' pre-intervention judgment updates, allowing earlier intervention outcomes to influence later updates and strategy selection.