REHYDRATION: FIXED-SEMANTICS COUNTERFACTUAL REPLAY FOR MECHANISTIC LLM-AGENT SOCIAL SIMULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Counterfactual experiments with large language model (LLM) agents are difficult to interpret when an intervention changes the platform state, thereby altering subsequent prompts and generated language. Standard replay comparisons can therefore mix structural policy response with post-intervention semantic adaptation, even when branches share the same decoding seed. We formalize behavioral and platform interventions through the **Behavior–Platform Decomposed Mechanistic Transition Framework (BDMTF)** and introduce **Rehydration**, a paired counterfactual replay protocol that generates and seals semantic opportunities before policy assignment and reuses them across intervention branches. This construction defines a **fixed-semantics structural effect** that complements the conventional adaptive-semantic effect. In an analytic setting with known ground truth, Rehydration recovers the fixed-semantics target, whereas common-seed adaptive replay does not recover that target. Controlled factorial experiments further reveal a behavior-by-ranking interaction: controversy-oriented ranking increases participation substantially more under conflict-responsive behavior while redirecting activity toward shallower reply structures, a regime we call **Shallow Swarm**. Mechanism-deletion and execution-trace analyses link the additional volume to conflict-responsive participation and the depth shift to ranking-conditioned exposure and reply targeting. The interaction is replicated across independently frozen model-family semantic catalogs and held-out simulator posts, while activation-form and boundary analyses identify settings in which parts of the effect weaken or reverse. Held-out platform evidence supports the conflict-to-volume pathway, whereas evidence for the shallower-depth component remains more limited. Overall, Rehydration provides a reproducible way to isolate fixed-semantics structural effects from policy-induced semantic adaptation and to test mechanistic claims in generative-agent social simulations.

1 INTRODUCTION

Large language model (LLM) agents are increasingly used to simulate social behavior, from generative agent communities to platform-scale interaction simulators Park et al. (2023); Li et al. (2023); Yang et al. (2025). Their appeal is counterfactual: a real platform cannot rewind history and rerun the same discussion under a different ranking policy, but in principle a simulator can Park et al. (2022; 2023); Liu et al. (2025). Online platforms shape collective attention through ranking, recommendation, and other exposure mechanisms, and experimental and observational studies have shown that social influence, early feedback, and algorithmic exposure can substantially alter collective outcomes Salganik et al. (2006); Muchnik et al. (2013); Cinus et al. (2022). Yet such effects need not be intrinsic to the platform policy itself: the same exposure mechanism may produce different outcomes in populations that respond differently to conflict, attention, or social feedback Cinus et al. (2022); Fabbri et al. (2022); Mansoury et al. (2020). This raises a basic mechanistic question: *is platform amplification a property of the ranking policy itself, or does it emerge from the interaction between platform exposure and population behavior?*

A common practice is to rerun the simulator under a different policy and compare the two outcomes, often using the same decoding seed to reduce random variation Fidone et al. (2026). For example,

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.

switching from best to controversy ranking may expose an agent to a different reply; that reply changes the next prompt, which changes the generated response and, in turn, what later agents see. Seeds align generation only while prompts match. Once the runs diverge, the measured contrast reflects both how the policy changes what agents see and how it changes what they subsequently generate Fidone et al. (2026); Goyal et al. (2026). The latter is itself a meaningful adaptive response, but the resulting contrast does not isolate the structural response under fixed semantic opportunity Robins & Greenland (1992); Pearl (2001).

Empirical evidence shows that this divergence is not merely a conceptual concern. As shown in Figure 1 (a), state-conditioned adaptive replay exhibits higher paired log-volume-effect variance than frozen replay on 24 of 25 held-out threads, with a median adaptive-to-frozen variance ratio of 4.89. Figure 1 (b) illustrates the underlying distinction: adaptive branches condition generation on their own evolving states, whereas frozen replay reuses a condition-blind semantic batch sealed before treatment. Together, these observations motivate separating structural policy response from post-treatment semantic adaptation.

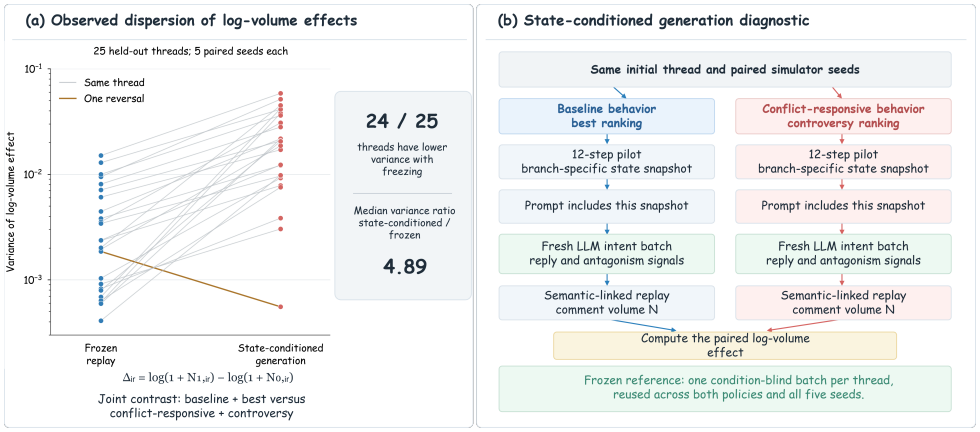


Figure 1: Paired log-volume effect $\Delta_{ir} = \log(1 + N_{1,ir}) - \log(1 + N_{0,ir})$ (conflict-responsive+controversy vs. baseline+best). **(a)** Per-thread variance over 5 paired seeds (df $= 4$); gold: the single reversal (r/funny, ratio 0.30). The adaptive arm samples via API without a fixed decoding seed; the two arms target different estimands (estimand check: Table 12). **(b)** Adaptive branches generate from branch-specific 12-step pilot snapshots; frozen replay reuses one condition-blind batch across branches and seeds.

We formalize behavioral and platform interventions through the *Behavior–Platform Decomposed Mechanistic Transition Framework (BDMTF)*, within which we introduce *Rehydration*, a controlled replay protocol that separates semantic opportunity from its realization. Before policy assignment, semantic opportunities are sampled from a prespecified reference distribution and sealed into a read-only pool. Their construction cannot condition on the treatment label or on post-treatment engagement, depth, or ranking state. All policy conditions then reuse the same pool and matched keyed random streams. A policy may determine whether an opportunity is realized, when it appears, and which visible comment it targets, but it cannot change which semantic opportunities were available before assignment. This construction lets us ask a counterfactual question that ordinary adaptive replay does not isolate: *how would the outcome change if the policy changed while the available semantic opportunities were held fixed?* We call the resulting quantity the *fixed-semantics effect*. Adaptive semantic response remains a complementary estimand rather than noise or bias; Rehydration is designed to isolate a different target. Within BDMTF, behavioral response and platform mechanisms are represented as separate intervention coordinates, allowing Rehydration to vary declared policies while holding the semantic opportunity set fixed.

We use this framework to study whether ranking amplification depends on population behavior. Among the five ranking policies implemented in our simulator, we use *best* and *controversy* as the primary mechanistic contrast. Crossing these with two behavioral regimes—baseline and conflict-responsive— yields a paired 2×2 factorial over 1,500 blocks. All four cells share the same sealed semantic pool and the same keyed random streams, differing only in the declared behavioral and ranking interventions. Controversy ranking has almost no effect on comment volume under baseline

behavior: volume changes by only $0.988\times$ relative to best ranking. Under conflict-responsive behavior, however, the same ranking change increases volume to $1.200\times$. Equivalently, the ranking effect under conflict-responsive behavior is $1.214\times$ the effect under baseline behavior, and this interaction is positive for 98.4% of posts. The controlled contrast therefore supports behavior-conditioned amplification: the same ranking policy is nearly neutral under one behavioral regime and amplifying under the other.

Mechanism tests further characterize this interaction. When conflict-conditioned response is removed, the additional increase in comments largely disappears. Reply trees also remain shallower after matching runs on comment volume, showing that the structural shift is not explained by increased activity alone. Execution traces link the remaining depth effect to which replies the ranking policy exposes and where agents subsequently respond. We call this regime *Shallow Swarm*: more comments, but shallower conversations. The interaction persists across alternative activation forms and independently frozen semantic catalogs from three model families, suggesting that it is not specific to a single implementation choice. Dedicated boundary tests also identify regimes in which the effect weakens or reverses, delimiting rather than hiding its scope. Held-out platform data provide partial external support for the proposed mechanism, while cross-platform experiments test whether the intervention interfaces remain executable after bounded target-platform calibration rather than assuming direct parameter transfer. Our contributions are threefold:

- We formalize behavioral and platform interventions through **BDMTF** and distinguish fixed-semantics structural effects from adaptive-semantic effects, defining a reference-distribution fixed-semantics estimand for counterfactual social simulation.
- We introduce **Rehydration**, a paired counterfactual replay protocol that operationalizes this estimand by freezing pre-treatment semantic opportunities while allowing policy-dependent realization.
- Using BDMTF and Rehydration, we identify and mechanistically trace behavior-conditioned ranking amplification and the **Shallow Swarm** regime, establish its robustness and failure boundaries, and compare its predicted pathways with held-out platform evidence.

2 Related Work

Prediction, generation, and intervention. Diffusion models estimate cascade growth from temporal and structural observations Szabo & Huberman (2010); Cheng et al. (2014); Zhao et al. (2015); Cao et al. (2017), but distributional fit alone does not define a ranking or moderation contrast. Generative Agents, CAMEL, Social Simulacra, OASIS, and MOSAIC support rich agent and platform experiments Park et al. (2023); Li et al. (2023); Park et al. (2022); Yang et al. (2025); Liu et al. (2025). COSMOS is the closest counterfactual simulator: it runs factual and moderation branches with common decoding seeds Fidone et al. (2026). Subsequent generation nevertheless conditions on each branch's updated memory and sensory state, so its target includes adaptive semantic response. Rehydration instead couples the same realized semantic proposals and varies only declared policy coordinates. We test this protocol-level distinction directly in Section 5.3, where a common-seed adaptive replay baseline is compared with Rehydration in an analytic SCM with a known fixed-semantics ground truth. Appendix K further compares the targets and reported external evidence of these systems without treating validation as a binary property.

Mediation and replay targets. Policy-dependent generated language is post-treatment adaptation, analogous to a mediator rather than a baseline confounder Robins & Greenland (1992); Pearl (2001). Rehydration is closest to a stochastic interventional direct-effect construction Vansteelandt & Daniel (2017): it draws simulator-specific semantic potential from a prespecified reference distribution and holds the realized pool common across arms. It neither labels adaptive semantics as bias nor identifies a human mediation effect. Its contribution is an executable fixed-semantics estimand with paired replay, compatibility and exhaustion semantics, and estimable intervention coordinates.

External evidence and portability. Simulator claims require separation from human-platform effects Goyal et al. (2026); observational contrasts additionally depend on exposure, timing, overlap, and confounding assumptions Aral et al. (2009); Shalizi & Thomas (2011). We therefore use chronological moderation responses as external prediction benchmarks rather than relabeling simulation as causal identification Chandrasekharan et al. (2022); Callaway & Sant'Anna (2021); Sun & Abraham

(2021). Cross-platform tests likewise distinguish an executable interface after target estimation from unchanged parameter or causal-effect transport Wang et al. (2021).

3 METHODOLOGY

3.1 PROBLEM FORMULATION

Let X_i denote the pre-intervention context of discussion i. In ordinary adaptive simulation, policy π changes the platform state and therefore later prompts and generated content. Let Z_i^π be the resulting policy-dependent semantic trajectory. For outcome functional M, the adaptive-semantic effect is $\tau_M^{\text{adapt}} = \mathbb{E}[M(\pi_1, Z_i^{\pi_1}, \epsilon_1) - M(\pi_0, Z_i^{\pi_0}, \epsilon_0)]$, which combines structural response and policy-induced semantic adaptation. Rehydration instead fixes semantic opportunities before assignment. Given X_i, it samples a finite pool $Z_i = \{z_{i,u,j}\}_{u,j} \sim Q_{\text{ref}}(\cdot \mid X_i, \xi_i)$, with $z_{i,u,j} = (a_{i,u,j}, y_{i,u,j}, c_{i,u,j})$, where a, y, and c denote action, semantic payload, and static compatibility metadata. The sealed proposal source satisfies $Q_{\text{ref}}(Z_i \mid X_i, \pi) = Q_{\text{ref}}(Z_i \mid X_i)$. Policies may change which opportunities are realized and when or where, but not which were available initially. Proposal generation occurs in Phase I; ***Phase II makes no language-model calls***. After target selection, a deterministic renderer adapts the surface utterance to the selected parent without regenerating its communicative frame. Thus, "fixed semantics" means invariance of the pre-treatment frame, not lexical identity. Language models construct proposals and profiles; replay isolates policy effects over the fixed opportunity set. Let $\pi = (\pi^{\text{act}}, \pi^{\text{target}}, \pi^{\text{rank}}, \pi^{\text{corr}})$ denote the intervention coordinates. Given Z_i and simulator innovations ϵ_i, replay evolves as $S_{i,t+1}^\pi = F_t(S_{i,t}^\pi, X_i, Z_i, \pi, \epsilon_{i,t})$, where transitions determine whether, when, and where an opportunity is realized but cannot generate new semantic content. For policies π_1 and π_0, define the *reference-distribution fixed-semantics effect*

$$\boxed{\tau_M^{\text{FS}}(Q_{\text{ref}}) = \mathbb{E}_{X,Z\sim Q_{\text{ref}},\epsilon}[M(\pi_1, Z, \epsilon) - M(\pi_0, Z, \epsilon)]\,.} \tag{1}$$

This answers the counterfactual question: *how would the outcome change if the policy changed while the available semantic opportunities were held fixed?* With n discussions and R replay seeds, we estimate $\widehat{\tau}_M^{\text{FS}} = \frac{1}{nR}\sum_{i=1}^{n}\sum_{r=1}^{R}[M_i(\pi_1, Z_i, \epsilon_{ir}) - M_i(\pi_0, Z_i, \epsilon_{ir})]$. Branches share immutable Z_i but maintain separate used-token sets, so they may exhaust the pool at different rates. Exhaustion is a logged no-op and never triggers regeneration. Keyed streams prevent unused or exhausted draws in one branch from shifting later draws in the other; capacity and exhaustion audits test finite-pool sensitivity. The information barrier fixes the estimand; common random numbers improve precision. Appendix L gives the replay conditions, and Proposition 1 establishes unbiasedness for Equation 1 and non-increasing variance under nonnegative cross-branch covariance. Finally, τ_M^{adapt} and τ_M^{FS} are distinct, not assumed to be additively related: the former includes post-treatment semantic regeneration, whereas the latter holds semantic opportunities fixed.

3.2 INTERVENTION DECOMPOSITION

BDMTF writes an outcome as $M(B, P; X, Z, \epsilon)$, where B is a behavioral response policy, P is a platform policy, and X is initial context. B maps visible signals to activation, targeting, correction, and exit; P supplies exposure, ranking, traffic, and moderation; and X contains the root post and pre-treatment thread. These are executable policy channels, not personality or clinical constructs. For behavioral regimes B_1, B_0 and platform regimes P_1, P_0, the fixed- semantics interaction is

$$\tau_{B\times P} = \mathbb{E}[M(B_1, P_1) - M(B_1, P_0) - M(B_0, P_1) + M(B_0, P_0)], \tag{2}$$

with X, Z, ϵ shared across cells. The Shallow Swarm case study uses a conflict-responsive B and controversy-oriented P. Configuration files call B, P, and X `Core`, `Structure`, and `Context`; these names carry no additional psychological meaning.

3.3 REHYDRATION: FROZEN-INTENT REPLAY

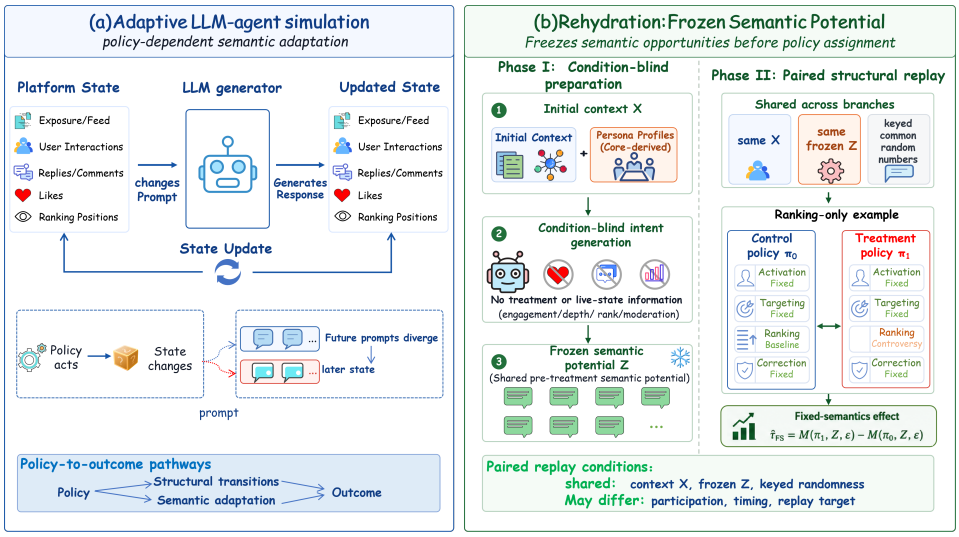


Figure 2: Rehydration shares frozen semantic opportunities and keyed randomness with branch-local consumption. The ranking-only example shares other policy rules; realized actions may differ. Rendered text does not feed back into transitions.

Rehydration separates semantic generation from counterfactual replay (Figure 2). Phase I uses LLMs to construct frozen semantic opportunities before treatment, while Phase II replays agent behavior over this shared space without further language-model calls. This separation isolates structural policy response under fixed semantics. In Phase I, each discussion receives a finite, condition-blind semantic opportunity pool Z_i, generated and sealed before treatment assignment. In Phase II, all policy branches reuse the same frozen pool and keyed random streams while maintaining branch-local consumption states, with no further language-model calls. Policies may change whether, when, and where an opportunity is realized, but cannot alter proposal generation or introduce new semantic content. The full replay algorithm (Algorithm 1), compatibility conditions, pool-exhaustion semantics, and proof of Proposition 1 are given in Appendix L.

Proposition 1 (Fixed-semantics replay and paired variance)**.** *If Z is generated before treatment and satisfies the information-barrier condition in Section 3.1, Phase II makes no calls to Q_{ref}, and the branch outcomes are $Y_b = M(\pi_b, Z, \epsilon_b)$ with the deterministic renderer R excluded from the transition kernel, then policy cannot alter proposal generation within the replay and the paired estimator $\widehat{\tau}_M^{\mathrm{FS}}$ defined in Section 3.1 is unbiased for Equation 1. If, in addition, paired and independent replay have the same marginal outcome distribution under each policy, then the paired variance is $\mathrm{Var}(Y_1 - Y_0)_{\mathrm{paired}} = \sigma_1^2 + \sigma_0^2 - 2\,\mathrm{Cov}(Y_1, Y_0)$, versus $\sigma_1^2 + \sigma_0^2$ for independent replay. Pairing does not increase variance when the induced covariance is nonnegative and strictly reduces it when the covariance is positive.*

This is design-based correctness for the declared Q_{ref} fixed-semantics target; it does not reclassify the adaptive effect in Section 3.1 as bias.

3.4 POLICY INTERFACES AS INTERVENTION COORDINATES

Given fixed X and Z, one transition factorizes into ranking, activation, targeting, and correction kernels. **Ranking** determines the exposed viewport; **Activation** determines whether an event occurs and which actor and action are eligible; **Targeting** selects a reply parent; and **Correction** applies counterspeech, removal, locking, or exit. These interfaces are intervention coordinates because a reported contrast replaces declared kernel parameters while holding all other kernels, context, semantics, and paired randomness fixed. Behavioral and platform ownership is tagged within each kernel, so a platform policy can alter exposure and platform-supplied arrival without silently changing behavioral response. Figure 2(b) illustrates a ranking-only intervention: the branches share context X, frozen pool Z, and keyed randomness, while only the ranking kernel changes from baseline to

controversy. Activation, targeting, and correction kernels remain fixed, but may produce different realized actions and reply targets because ranking changes their exposed signals. The contrast therefore measures how the same semantic opportunities are realized under different exposure policies, not newly generated semantics. For cross-platform portability, an adapter maps native logs to a common state–observation–action–event contract. Replacing only platform parameters is warranted when action meanings and reply parentage are preserved, retained behavioral kernels pass target validation, target states have support overlap, and frozen intents remain compatible with target actions. Failure of these conditions requires contract or behavioral adaptation, not merely a new parameter file. Thus portability of the interface does not imply portability of its fitted parameters. Appendix M formalizes the kernels, coordinate effects, and portability restrictions. Connectivity is optional. Reply trees, co-participation networks, and follower graphs remain distinct; when follower data are unobserved, connectivity is explicitly null rather than inferred from co-participation.

4 Instantiating BDMTF with Agent Response Policies

Agent variables parameterize observable response policies rather than latent psychological traits. We use two complementary regimes. The *controlled regime* fixes theory-guided, reference-calibrated response parameters before factorial, deletion, and sensitivity analyses, enabling transparent mechanism intervention without requiring behavioral identification under endogenous conflict exposure. The *data-estimated regime* fits arrival, targeting, depth decay, author reuse, and correction only on chronological training data and freezes them before held-out evaluation. No parameter is updated after observing a mechanism-deletion or held-out intervention outcome.

In both regimes, activation, targeting, and visibility are implemented through prespecified policy components, and external traffic may alter visibility but cannot create new semantic opportunities. The controlled coefficients are mechanism parameters rather than observationally identified human behavioral effects. Complete policy equations, calibration details, and parameter provenance are reported in Appendix N.

Our main falsification removes complete pathways without refitting. Removing conflict-conditioned response substantially attenuates the volume interaction, whereas deleting heat or explicit depth heuristics does not eliminate the negative depth interaction. Exposure and target traces further localize the remaining depth effect to ranking-conditioned visibility and parent selection. A separately learned reply-topology model provides an additional check that the interaction is not specific to the hand-specified routing structure (Appendix O.6).

5 Experiments

We organize the experiments around four questions: (1) whether ranking amplification depends on population behavior; (2) which pathways produce the resulting *Shallow Swarm*; (3) whether Rehydration supports a credible fixed-semantics contrast; and (4) how robust, bounded, and externally supported the mechanism is. Controlled counterfactual comparisons share posts, frozen semantic opportunities, and keyed random streams across policy conditions. Predictive and observational evaluations use chronological partitions. Appendix O.1 specifies data splits, metrics, and uncertainty procedures; Appendix O.2 reports the identification checks; Appendix O.3 and Table 14 report capacity and exhaustion audits; and Appendix O.6 reports auxiliary topology diagnostics.

5.1 RQ1: Does Ranking Amplification Depend on Population Behavior?

Hypothesis: controversy ranking should have little volume effect under baseline behavior but amplify conflict-responsive participation.

We test a paired 2×2 factorial crossing baseline versus conflict-responsive behavior with best versus controversy ranking over 1500 replay blocks. All cells share posts, frozen semantic opportunities, and keyed random streams, changing only the declared behavior and ranking policies.

Switching to controversy ranking changes volume by $0.988\times$ under baseline behavior and $1.200\times$ under conflict-responsive behavior. The multiplicative interaction is 1.214 [1.207, 1.222] and is positive on 98.4% of posts (Figure 3(a)). This difference in ranking slopes,

not the 3.497× joint-versus-baseline contrast, is the behavior–platform interaction: the baseline curve is nearly flat, whereas the conflict-responsive curve rises under controversy ranking. The result is therefore not explained by higher overall activity in the focal population. The same factorial reduces mean leaf depth, with the conflict-responsive, controversy-ranked cell producing the shallowest trees (Figure 3(b)). The raw depth difference in differences is -0.824 [-1.092, -0.559]. After fitting a positive scale on 382 chronological development posts, the held-out calibrated interaction is -0.100 [-0.179, -0.016]. The raw scale describes simulator topology; the calibrated scale measures correspondence to observed Reddit depth.

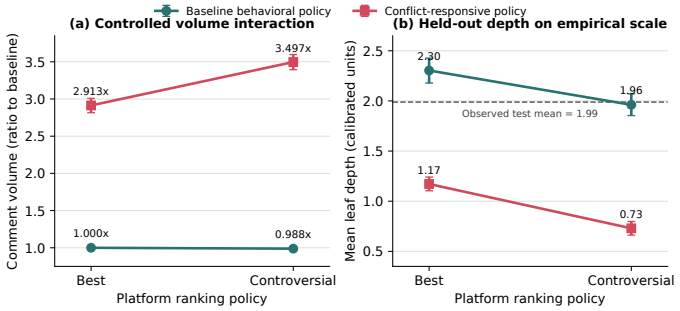


Figure 3: Behavior–platform interaction. Error bars: post-cluster 95% intervals; dashed line: observed test mean.

We then replace hand-specified routing with a context-size model and local reply kernel fitted on 285 chronological training cascades. Evaluation on 100 held-out posts, ten paired seeds, and 4,000 post–seed–cell records yields a volume interaction of 1.234 [1.227, 1.240] and raw depth interaction of -0.313 $[-0.360, -0.264]$. The pattern therefore persists under learned topology, although that model has an overlong depth-tail failure mode (Appendix O.6). This replication tests whether the interaction survives when both reply-depth and parent-selection behavior are estimated from chronological cascades rather than imposed by the controlled configuration.

Together, the controlled and learned configurations identify a *Shallow Swarm*: controversy ranking is nearly neutral for baseline populations but amplifies conflict-responsive participation while producing shallower reply structures. The learned-topology result shows that the pattern is not solely induced by the hand-specified routing and depth heuristics.

Answer: ranking amplification depends on population behavior, and the volume interaction persists under both controlled and learned topology.

5.2 RQ2: Which Pathways Produce Shallow Swarm?

Hypothesis: conflict-responsive behavior supplies the additional volume, while ranking-conditioned exposure and reply targeting produce the shallower discussion structure.

To isolate the contribution of each implemented channel, we compare channel-deletion interventions with the all-channel condition on the same prespecified 25-post diagnostic subset. Instrumented replay shows that the negative depth interaction appears in both the ranking-conditioned exposure set and subsequent target selection, with exposure-depth and selected-target-depth interactions of -3.541 and -2.914, respectively. Consistent with these traces, Table 1 shows that the negative depth interaction persists after removing either the explicit depth exponent or the complete depth channel. The complete pathway-deletion protocol, parameter provenance, and remaining channel ablations are reported in Appendix N. Together, these results indicate that the shallow structure does not require a hand-coded preference for shallow replies. To rule out a purely mechanical volume effect, we match all four cells to a common number of comments within each post–seed block. The depth interaction remains -1.692 $[-2.906, -0.409]$, showing that higher comment volume alone does not explain the shallower trees. Table 1 provides the complementary intervention test. Removing the conflict-response channel reduces the volume interaction from 1.414 to 1.059, whose interval includes one, making conflict-conditioned response the largest tested contributor to volume amplification. In contrast, deleting the explicit depth-preference or complete depth channel does not remove the negative depth interaction, consistent with the exposure and target traces above. Finally, the interaction is not tied to the hand-specified response equation. Training-matched additive and saturating activation forms retain volume interactions of 1.158–1.207, and a separately learned reply-topology kernel reproduces the effect at 1.234 $[1.227, 1.240]$. Full activation-form intervals and the disconnected-ranking null control are reported in Appendix Q, while the learned-kernel analysis is reported in Appendix O.6.

Table 1: Mechanism-deletion ablations under paired frozen-intent replay. Intervals are 95% post-level stratified bootstrap intervals.

Setting	Posts	P effect B_0	P effect B_1	Volume RoR [95% CI]	Depth DiD [95% CI]	Strict support
All channels	25	0.992	1.403	1.414 [1.348, 1.479]	-2.748 [-3.911, -1.633]	Yes
No conflict channel	25	1.005	1.064	1.059 [0.988, 1.119]	-3.338 [-4.551, -2.170]	No
No heat channel	25	1.007	1.118	1.110 [1.085, 1.136]	-2.295 [-3.279, -1.351]	Yes
No depth exponent	25	1.217	1.387	1.140 [1.097, 1.185]	-2.055 [-2.512, -1.581]	Yes
No complete depth channel	25	1.008	1.096	1.087 [1.048, 1.126]	-2.619 [-3.576, -1.680]	Yes

Answer: Shallow Swarm arises from two linked pathways: conflict-responsive behavior increases participation, while ranking-conditioned exposure and target selection redirect that activity toward shallower reply structures.

5.3 RQ3: Does Rehydration Support a Credible Fixed-Semantics Contrast?

Hypothesis: shared randomness alone should not recover the fixed-semantics target once post-treatment semantics adapt to branch state; sealing semantic opportunities before treatment should recover that target, reduce rerun dispersion, and preserve usable semantic opportunities under dynamic routing.

We first evaluate Rehydration in a structural causal model where the fixed-semantics effect is known analytically. As reported in Appendix O.2 (Table 12), at semantic shift 0.25, common-seed adaptive replay differs from the fixed-semantics target by 0.287 with 0.0% coverage, whereas Rehydration has mean error -0.001 and 95.2% coverage. Independent live replay differs from the same target by 0.285 with 38.6% coverage. Thus, shared randomness alone does not recover the fixed-semantics estimand once post-treatment semantics remain policy-dependent; semantic freezing changes the estimand rather than merely reducing replay variance. This known-truth experiment verifies the declared estimand and its implementation; it does not imply that adaptive semantic response is irrelevant. The same distinction matters on held-out Reddit threads. As shown in Figure 1a, across 25 threads and 5 seeds, Rehydration lowers intervention-effect dispersion on 24/25 threads, with a median live-to-frozen variance ratio of 4.89 [3.33, 9.61]. The mean adaptive-minus-fixed log-volume contrast is -0.035 [-0.079, 0.011], with no thread-level mean sign reversal. Thus, adaptive and fixed semantics remain distinct estimands even when they agree in direction on this panel. Panel selection and uncertainty procedures are detailed in Appendix O.2, with dispersion results in Table 13. Structural invariance alone does not guarantee semantic compatibility. Following the sampling and rating protocol in Appendix O.2, we ask 90 independent raters to evaluate 600 blinded dynamic parent–reply pairs. The complete semantic-audit results in Appendix H show that parent-conditioned frame rendering is acceptable in 77.7% [72.1, 83.2] of ratings, compared with 17.7% [14.6, 20.7] for verbatim transfer. This comparison tests whether a frozen semantic opportunity can still be rendered coherently after the policy selects a different parent; the verbatim condition is a literal-transfer baseline rather than a live-generation quality ceiling. Finally, the frozen opportunity pool remains adequate at the capacity used in the controlled experiments. The full $512\times$ audit over all source posts and four factorial arms records zero pool exhaustion and requires no additional language-model calls or tokens (Appendix O.3).

Answer: Rehydration recovers the declared fixed-semantics target, whereas shared randomness alone does not. It also reduces replay dispersion while preserving usable semantic opportunities under dynamic routing.

5.4 RQ4: Where Is the Mechanism Robust, and Where Does It Fail?

Hypothesis: the behavior–ranking interaction should persist across independently frozen semantic catalogs and unseen simulator posts, while explicit boundary tests should reveal regimes in which it weakens or reverses.

We first vary the frozen semantic opportunities. Across independently generated OpenAI, DeepSeek, and Qwen catalogs, the behavior–ranking interaction remains positive in volume and negative in depth: volume RoRs range from 1.188 to 1.237, and every depth interval excludes zero. Thus, the primary interaction is not specific to a single frozen semantic catalog.

The interaction also improves prediction on unseen simulator posts. As shown in Table 2, including the interaction reduces log-volume MAE from 0.204 to 0.052 and mean-depth MAE from 3.994 to 3.811 across 104 held-out posts. The improvement is substantially stronger for volume than for depth.

Table 2: Held-out prediction on 104 posts (312 post–seed blocks). Lower MAE is better; positive gains favor the interaction model.

Outcome	Additive MAE	Interaction MAE	MAE gain [95% CI]	Interaction wins
Volume (log1p)	0.204	0.052	0.152 [0.138, 0.165]	93.6%
Mean leaf depth	3.994	3.811	0.183 [0.079, 0.293]	59.3%

The mechanism is nevertheless not universal across simulator settings. Figure 4 summarizes both its local robustness and its failure boundaries. In the controlled simulator, all 24/24 bounded perturbations around the primary configuration retain the joint higher-volume, lower-depth direction. By contrast, a deliberately broad scenario–community stress grid reveals regions in which the joint pattern disappears or reverses: only 108/615 cells satisfy the Shallow Swarm criterion. This global grid is a boundary map rather than an estimate of prevalence. Consistent with this picture, all seven settings predicted to be inactive fail the two-outcome criterion, whereas 8 of 9 predicted active settings pass, and an additional disjoint stress panel exposes two wide-viewport depth reversals. These failures delimit the Shallow Swarm regime rather than treating it as a simulator-wide law. Finally, held-out Reddit provides partial external correspondence. Early conflict positively predicts subsequent volume after adjustment, with partial slope $10.375\ [3.815, 16.982]$. The corresponding depth intervals include zero, however, so the external data support the activation-to-volume pathway but not the full shallow-structure component.

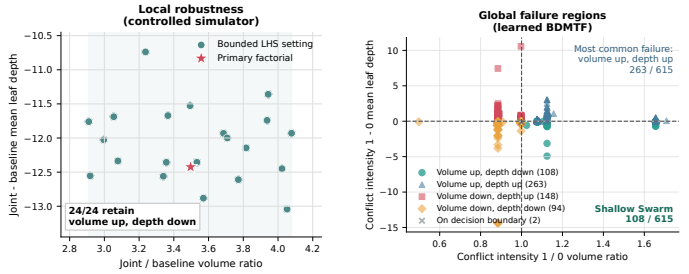


Figure 4: Mechanism robustness and boundary mapping. Left: all 24/24 bounded perturbations retain the joint direction. Right: 108/615 broad-grid cells support Shallow Swarm; the grid maps boundaries, not prevalence.

Answer: the interaction is stable across independently frozen semantic catalogs and improves held-out simulator prediction, but the full Shallow Swarm regime has explicit failure boundaries. Held-out Reddit supports the volume pathway, while the shallower-depth effect remains supported primarily within the simulator.

5.5 Discussion

Rehydration makes the semantic reference distribution an explicit experimental input. Analytic recovery, paired replay, catalog–coordinate decomposition, and human compatibility ratings audit the fixed-semantics contrast; mechanism deletion, pathway traces, boundary mapping, and semantic-pool sensitivity test its implementation. Details appear in Appendices H, O, and R. Within that contrast, conflict-conditioned response supplies volume, while ranking-conditioned exposure and targeting redirect activity toward shallower branches. Held-out Reddit supports the conflict-to-volume pathway, while evidence for the depth component remains inconclusive. Lemmy provides a sparse-outcome diagnostic, while Hacker News tests interface reuse after bounded adaptation. Transport evidence, identification assumptions, and claim boundaries appear in Appendices G and J.

6 Conclusion

Rehydration enables fixed-semantics counterfactual replay in BDMTF by sharing pre-treatment semantic opportunities across policy branches. Mechanism deletion, robustness, and held-out tests reveal *Shallow Swarm*: conflict-responsive behavior increases participation, while ranking-conditioned exposure redirects activity toward shallower replies. The interaction persists across activation forms and semantic catalogs but has explicit failure boundaries. Together, BDMTF and Rehydration separate structural effects from policy-induced semantic adaptation, enabling reproducible mechanistic tests in generative-agent social simulation.

AI USE STATEMENT

In this work, we used generative AI tools for generating and annotating frozen semantic opportunities used as experimental inputs in our simulator. We have not used generative AI tools for theoretical or conceptual development, hypothesis or mathematical-claim formulation, experimental design, method implementation, data analysis, or interpretation of results, and the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for no tasks with recommended disclosure. We have reviewed all AI-assisted work. All uses of generative AI were restricted to generating and annotating the frozen semantic opportunities described in the methodology and experimental sections. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

BDMTF is a research and diagnostic framework, not a prescription for ranking or moderation. Behavioral variables describe discourse-expressed interaction tendencies and must not be interpreted as clinical or personality diagnoses. Public data are analyzed in aggregate, and released intervention panels omit raw identifiers where possible. The semantic audit uses anonymous judgments of pre-existing reply candidates, provides a content warning and a stop option, and retains no direct participant identifiers. It does not randomize platform exposure or estimate a human behavioral treatment effect. Deployment and behavioral experiments require independent ethics review, informed consent where applicable, and safeguards against discriminatory profiling or manipulation.

REPRODUCIBILITY STATEMENT

Algorithmic assumptions and the replay proof appear in Appendix L; parameter provenance, data splits, metrics, uncertainty procedures, and complete auxiliary results appear in Appendices N–O. The supplementary code package includes frozen configurations, result macros, manifests, and offline replay entry points for reproducing every reported table and figure. The anonymous repository is available at `https://anonymous.4open.science/status/Rehydration-3BC0`.

REFERENCES

Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.

Brantly Callaway and Pedro H. C. Sant'Anna. Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230, 2021. doi: 10.1016/j.jeconom.2020.12.001.

Qi Cao, Huawei Shen, Keting Cen, Wentao Ouyang, and Xueqi Cheng. Deephawkes: Bridging the gap between prediction and understanding of information cascades. In *Proceedings of the 2017 ACM Conference on Information and Knowledge Management*, pp. 1149–1158, 2017.

Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. Quarantined! examining the effects of a community-wide moderation intervention on reddit. *ACM Transactions on Computer-Human Interaction*, 29(4), 2022. doi: 10.1145/3490499.

Justin Cheng, Lada Adamic, P. Alex Dow, Jon M. Kleinberg, and Jure Leskovec. Can cascades be predicted? In *Proceedings of the 23rd International Conference on World Wide Web*, pp. 925–936, 2014.

Federico Cinus, Marco Minici, Corrado Monti, and Francesco Bonchi. The effect of people recommenders on echo chambers and polarization. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):90–101, 2022. doi: 10.1609/icwsm.v16i1.19275.

Francesco Fabbri, Maria Luisa Croci, Francesco Bonchi, and Carlos Castillo. Exposure inequality in people recommender systems: The long-term effects. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):194–204, 2022. doi: 10.1609/icwsm.v16i1.19284.

Giacomo Fidone, Lucia Passaro, and Riccardo Guidotti. Evaluating online moderation via LLM-powered counterfactual simulations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 38451–38459, 2026. doi: 10.1609/aaai.v40i45.41186.

Agam Goyal, Yian Wang, Eshwar Chandrasekharan, and Hari Sundaram. From plausible to causal: Counterfactual semantics for policy evaluation in simulated online communities. *arXiv preprint arXiv:2604.03920*, 2026. PoliSim@CHI 2026.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, volume 36, pp. 51991–52008, 2023. doi: 10.52202/075280-2264.

Genglin Liu, Vivian T. 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. 6390–6417, 2025. doi: 10.18653/v1/2025.emnlp-main.325.

Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, pp. 2145–2148, 2020. doi: 10.1145/3340531.3412152.

Lev Muchnik, Sinan Aral, and Sean J. Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651, 2013. doi: 10.1126/science.1240466.

Joon Sung Park, Lindsay Popowski, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–18, 2022.

Joon Sung Park, Joseph 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.

Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420, 2001.

James M. Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992. doi: 10.1097/00001648-199203000-00013.

Matthew J. Salganik, Peter Sheridan Dodds, and Duncan J. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856, 2006. doi: 10.1126/science.1121066.

Cosma Rohilla Shalizi and Andrew C. Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods and Research*, 40(2):211–239, 2011.

Liyang Sun and Sarah Abraham. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199, 2021. doi: 10.1016/j.jeconom.2020.09.006.

Gabor Szabo and Bernardo A. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010.

Stijn Vansteelandt and Rhian M. Daniel. Interventional effects for mediation analysis with multiple mediators. *Epidemiology*, 28(2):258–265, 2017. doi: 10.1097/EDE.0000000000000596.

Yuping Wang, Savvas Zannettou, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, and Gianluca Stringhini. A multi-platform analysis of political news discussion and sharing on web communities. In *arXiv preprint arXiv:2103.03631*, 2021. doi: 10.48550/arXiv.2103.03631.

Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, Prateek Gupta, Shuyue Hu, Zhenfei Yin, Guohao Li, Xu Jia, Lijun Wang, Bernard Ghanem, Huchuan Lu, Chaochao Lu, Wanli Ouyang, Yu Qiao, Philip Torr, and Jing Shao. OASIS: Open agent social interaction simulations with one million agents. *arXiv preprint arXiv:2411.11581*, 2025. doi: 10.48550/arXiv.2411.11581. URL https://arxiv.org/abs/2411.11581v5. Version 5, revised March 23, 2025.

Qingyuan Zhao, Murat A. Erdogdu, Hera Y. He, Anand Rajaraman, and Jure Leskovec. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1513–1522, 2015.

A Implementation Details and Extended Mechanistic Definitions

A.1 Behavioral Propensity Extraction

The profiling pipeline combines lexicon-guided weak priors with ordinal semantic scoring. Representative and high-conflict samples are maintained as separate channels so that high-signal excerpts are not treated as representative of the full interaction history. Prosocial evidence continuously attenuates surface conflict scores under the fixed scoring rule. The resulting variables characterize observable communication tendencies rather than latent psychological traits.

For an ordinal score $\theta \in [0, 5]$, the bounded drive is

$$D(\theta) = \sigma\left(\frac{\theta - \mu}{\mu}\right), \qquad \mu = 2.5, \tag{3}$$

where $\sigma(\cdot)$ denotes the logistic sigmoid. Population distributions are estimated separately by community and frozen before paired simulation.

A.2 Information Barrier

Semantic-opportunity generation is completed before treatment assignment. Generation prompts exclude treatment labels, likes, replies, ranking positions, and other live or post-treatment platform signals. The frozen opportunity store and request store are opened read-only during replay, with no write-back path. Consequently, Phase II cannot trigger new semantic generation or replenish exhausted opportunities.

Each cached semantic opportunity is associated with provenance metadata, including the model identifier, prompt hash, decoding configuration, response hash, and a generation seed when exposed or supported by the underlying API. Cached opportunities therefore permit complete offline replay without additional API calls. This provenance guarantee applies to the frozen semantic pool; the live state-conditioned generation used in the adaptive-replay diagnostic is evaluated separately and may include fresh API sampling without a fixed decoding seed.

A.3 Visibility Operators

Let F_c^+ and F_c^- denote the positive and negative feedback counts associated with comment c, respectively. Three central ranking scores are

$$\mathrm{score}_{\mathrm{best}}(c) = F_c^+ - F_c^-, \tag{4}$$

$$\mathrm{score}_{\mathrm{top}}(c) = F_c^+, \tag{5}$$

$$\mathrm{score}_{\mathrm{cont}}(c) = \min(F_c^+, F_c^-). \tag{6}$$

The simulator additionally implements chronological and hot orderings. Position bias is applied after ranking, and deterministic tie breaking ensures reproducible ordering for identical simulator states.

A.4 Opportunity Uniqueness and Tree Validity

Each indexed semantic opportunity token can be realized at most once within a branch. Multiple indexed tokens may share the same immutable semantic template under the prespecified capacity scheme. Template multiplicity is fixed before treatment assignment, shared across policy branches, and does not condition on treatment labels or post-treatment outcomes.

The capacity multiplier determines the maximum number of branch-local realizations available for each immutable semantic template. Capacity sensitivity is evaluated from $1\times$ to $256\times$, with a separate $512\times$ full-coverage audit. Exhausted requests become logged no-ops and never trigger semantic regeneration or replenishment; detailed capacity and exhaustion results are reported in Appendix O.3.

Every reply must target either the root or an existing earlier event. Cycles, duplicate event keys, and replies after a locking event are rejected. Unit tests cover tree reconstruction, ranking determinism, target visibility, lock enforcement, removal visibility, dropout, pool-consumption behavior, and paired-seed stability.

B Community-Level Calibration Statistics

B.1 Activation Thresholds

Community activation baselines for AskReddit, aww, funny, science, and worldnews are 0.30, 0.73, 0.71, 0.52, and 0.48, respectively. Agent thresholds combine the community baseline with individual propensity. Learned activation uses only chronologically prior event histories and estimates aggregate event risk.

B.2 Early Engagement Normalization

Heavy-tailed engagement is normalized through community training medians. Every test post is transformed with statistics available before its timestamp. This prevents target-period activity from leaking into heat and traffic features.

C Ablation Analysis of Motivational Channels

Channel ablations remove conflict, consensus, heat, external traffic, counterspeech, moderation, or dropout terms while holding semantic intents, paired seeds, and all other policies fixed. Removing the conflict channel primarily reduces volume amplification. Removing heat weakens attention concentration and preserves depth. The combined response is therefore not attributable to a single additive term.

Trait ablations compare coupled, independent, constructive-aggressive, and shuffled assignments. All four preserve the principal higher-volume, lower-depth direction, with volume ratios 3.489–3.522 and leaf-depth changes -12.432 to -11.727.

D Parameter Sensitivity Analysis

The local Latin-hypercube design varies conflict weight, bandwagon weight, activation, target depth, viewport, leader fraction, and traffic near the principal configuration. All 24 settings support the Shallow Swarm direction. Table 3 separates this confirmatory local neighborhood from the broader global boundary search.

The global grid covers leader fractions $0, 0.1, 0.2, 0.4$; viewports $5, 10, 20,$ and unbounded; populations $50, 100, 250, 500$; horizons $72, 144, 288$; five ranking policies; and correction and exit mechanisms. Cells outside the Shallow Swarm quadrant are retained as informative boundary conditions.

D.1 Held-Out Response and Local Robustness

Held-out intervention-response prediction results are reported in Table 2 in the main text. Table 4 consolidates the factorial benchmark, local sensitivity range, and trait-assignment audit.

E Profiling Methodology Details

E.1 Hybrid Extraction Pipeline

Lexicon frequencies are weak surface priors for conflict-amplifying, callous, strategic, and self-focused language. A representative channel estimates baseline style; a high-signal channel preserves rare conflict expressions. Ordinal semantic scoring receives both channels and returns a bounded score, confidence, and brief evidence. No diagnostic labels are requested.

E.2 Three-Family Measurement

The real-Reddit semantic analysis uses one OpenAI-family model, one DeepSeek-family model, and one Qwen-family model. Consensus labels train a local frozen scorer. The held-out posts do not train

Table 3: Two-tier mechanism phase analysis. The confirmatory local tier perturbs the principal operating neighborhood and reports scenario-level ranges. The global tier uses a different learned simulator and intervention contrast to locate phase boundaries; its unweighted cell share is not an estimate of real-world prevalence. The two tiers answer local robustness and global boundary questions, respectively.

Evidence region	Unit	N	Statistic	Shallow Swarm	V_1/V_0	ΔD
Confirmatory local	Paired scenario	24	Range	24 (100.0%)	[2.910, 4.078]	[-13.154, -10.738]
Global reference	Scenario–community	5	Median	1 (20.0%)	1.076	+0.026
Bounded one-factor	Scenario–community	165	Median	35 (21.2%)	1.076	+0.026
Boundary one-factor	Scenario–community	45	Median	9 (20.0%)	1.076	+0.028
Multifactor stress	Scenario–community	300	Median	48 (16.0%)	1.076	+0.136
Multifactor boundary	Scenario–community	100	Median	15 (15.0%)	1.076	+0.066
Global stress total	Scenario–community	615	Median	108 (17.6%)	1.076	+0.086

Table 4: Confirmatory robustness of the Shallow Swarm mechanism. Sensitivity ranges are extrema across 24 Latin-hypercube settings; trait ranges are extrema across four assignment schemes.

Analysis	Design	Volume ratio	Δ mean leaf depth	Directional support
Factorial benchmark	1,500 paired blocks	3.497	-12.424	1,496/1,500 blocks
Local sensitivity	24 LHS settings	2.910–4.078	-13.154 to -10.738	24/24 settings
Trait decoupling	4 trait assignments	3.489–3.522	-12.432 to -11.727	99.67–100.00%

that scorer. This experiment evaluates measurement consistency for observed comments; it does not claim that every open-vocabulary label is invariant across model families. The separate frozen-intent audit covers 500 persona-conditioned tasks for each family. Exact agreement is reported alongside prevalence-sensitive kappa statistics, and strict replay invariance is separated from a deliberately re-coupled negative control. Table 5 reports these agreement measures and distinguishes the strict information-barrier check from the coupled negative control.

F Formalization of Macroscopic Viral Flow

F.1 Base Potential

The external arrival channel is

$$\Delta N_t = \eta_{\text{eff}} \, \text{LogPotency}(S_0) M(H_t) C(H_t) S(H_t) \delta^t \epsilon, \tag{7}$$

with $\eta_{\text{eff}} = \kappa\eta$, nominal $\eta = 200$, and $\kappa = 0.4$. Early potency uses logarithmic normalization to represent heavy-tailed attention and perceptual saturation.

F.2 Dynamic Feedback

Momentum and conflict multipliers are

$$M(H_t) = 1 + \gamma_m (N_{\text{likes},t} + 3N_{\text{replies},t}), \tag{8}$$

$$C(H_t) = 1 + \gamma_c \frac{N_{\text{replies},t}}{N_{\text{likes},t} + N_{\text{replies},t} + \varepsilon}. \tag{9}$$

The survival gate takes values in $\{0.05, 0.2, 1.0, 1.3\}$ according to early momentum and cumulative activity. Each component is exposed to ablation and sensitivity analysis.

F.3 Decay and Numerical Stability

The controlled configuration uses $\delta = 0.998$ and $\epsilon \sim U(0.8, 1.2)$. A hard per-step cap prevents numerical divergence. The cap is recorded in the configuration and is held constant across paired conditions.

Table 5: Three-family semantic-intent agreement and Rehydration audit. Strict replay is an implementation-level information-barrier check; the coupled result is a negative-control diagnostic.

Check	Estimate	Scope
Reply-decision unanimity	90.4%	500 tasks
Polarity unanimity	76.0%	500 tasks
Toxicity interval α	0.794	500 comments
Counterspeech unanimity	80.0%	500 comments
Strict replay max difference	0.000	50 paired posts
Coupled median size range	5.4%	diagnostic

G Cross-Platform Pair Construction and Adapter Diagnostics

Hacker News adapter protocol. Hacker News contributes 5,000 complete cascades in a chronological 3,000/1,000/1,000 train/validation/test split. The zero-shot condition applies the Reddit parameterization without target outcomes. The bounded adapter retains the BDMTF state–action contract and searches only ranking, viewport, depth-preference, and arrival parameters on the Hacker News training and validation partitions. The 1,000-cascade test partition is opened after selection. Direct transfer has normalized Wasserstein distance 0.957. A fair target default uses the target profile, the base Structure configuration, and the same keyed seed schedule as every adapter; its distance is 0.424. Selection on 50 validation cascades reaches 0.422. Thus target-profile estimation provides 55.7% relative improvement over direct transfer, while Structure selection adds 0.4% over the common-seed default. Table 6 places these adapter results beside the equal-budget generative baselines under the same target-data constraint.

Empirical resampling is strongest for unconditional distribution matching, Hawkes is next, and BDMTF outperforms the branching process. The adapter experiment therefore evaluates whether intervention coordinates remain executable after target estimation, rather than claiming a universal generative-fidelity advantage.

Exact-URL HN–Lemmy matching. URL canonicalization removes tracking parameters, normalizes host and scheme, and resolves redirects where available. Pairs are one-to-one and are selected without cascade outcomes; structural evaluation additionally requires a complete reply tree on both platforms. The protocol-frozen pilot contains 44 complete-tree pairs. The confirmation search starts from 20,011 Hacker News root candidates and yields 758 exact-URL pairs, of which 724 have complete trees on both platforms. On the confirmation set, bounded target-platform adaptation improves normalized MAE over direct Hacker News transfer by 0.066 [0.052, 0.080]. Exact content matching reduces observed story-identity differences but does not identify a platform causal effect.

Same-author Reddit–Voat continuity. The Reddit–Voat analysis links 560 same-pseudonymous authors across platforms and compares pre- and post-migration semantic similarity. The approximately 0.049 gain is positive in both migration samples. This is an intent-continuity diagnostic rather than an exact-story comparison: a shared pseudonym does not establish identity, and the design does not estimate the causal effect of moving to Voat.

Transport boundary. Together, the three designs support reuse of the intervention coordinates and state–action contract after limited target-platform estimation. They reject the stronger claim that Reddit parameters transfer unchanged. The evidence is restricted to threaded discussion platforms and does not establish zero-shot universality for feed-first or short-video systems. Table 7 summarizes which construction supports each of these progressively weaker transport claims.

H Model-Family Stability and Semantic Audit

What the agreement statistics measure. The three model families receive the same persona, post context, output schema, and annotation rubric. The audit comprises 1500 intent generations and 75 batched annotations, yielding 1575 valid calls. Reply decisions have 90.4% three-way exact agreement, broad polarity has 76.0% agreement, toxicity reaches interval Krippendorff $\alpha = 0.794$, and counterspeech has 80.0% agreement. These statistics measure consistency among model outputs. Agreement can arise from a shared construct, a shared prompt schema, or correlated model behavior.

Table 6: Equal-target-budget HN fidelity. Every row uses the same 3,000 target training cascades and no validation outcomes; lower NWD is better.

Model	HN train	Median NWD ↓
Empirical bootstrap	3,000	0.387
Hawkes	3,000	0.413
BDMTF target default	3,000	0.424
Branching process	3,000	0.565

Table 7: Cross-platform designs answer different portability questions. Pair construction and target adaptation use no target-test outcomes.

Design	Confirmation sample and supported claim
Reddit → HN	1,000 temporal test cascades; bounded parameter adaptation
HN–Lemmy	724 exact-URL complete-tree pairs; content-controlled interface reuse
Reddit–Voat	560 same-pseudonym authors; semantic continuity diagnostic

We therefore use these fields as operational semantic coordinates and reserve open-vocabulary emotion and topic fields for descriptive analysis.

What strict replay measures. Strict replay excludes model-specific reply, polarity, toxicity, and counterspeech summaries from all structural transition kernels. Given shared posts, seeds, and platform dynamics, its zero maximum structural difference is expected by construction. It is reported as a deterministic reproducibility check that the information barrier is implemented correctly, not as a model- robustness result. The barrier-off/barrier-on variance experiment in Section 5.3 is the scientific test of intervention-estimate stability.

Q_{ref} **effect sensitivity.** Strict equality does not test whether a structural effect changes with the reference proposal distribution. We therefore use the cached 1,500 responses to construct three independently frozen catalogs over the same 50 posts. Each family contrast holds its catalog fixed and applies a prespecified map from that catalog's reply and polarity prevalence to activation and conflict coordinates. This diagnostic jointly changes semantic content and the semantic-to-policy map. It defines $\tau^{\mathrm{FS}}_{\mathrm{link}}(Q_{\mathrm{ref}})$ and does not isolate a content-only Q_{ref} effect.

As shown in Table 8, all three family point estimates increase volume and toxicity, but the Qwen volume interval and every leaf-depth interval include zero. The result supports stability of the toxicity direction and bounded variation in volume under the declared joint mapping; it does not establish a content-only depth result. The strict zero-difference row remains only an information-barrier check.

Dynamic-parent audit. Frozen replies are stored as abstract communicative frames and rendered only after the Target policy selects a parent. Structural invariance under surface rerendering is enforced by test. Semantic compatibility is evaluated by 90 independent human raters on 600 blinded parent–reply pairs sampled from 51,653 eligible replays and balanced over community, factorial cell, and root/shallow/deep target strata. Each item receives three ratings under randomized candidate order. Frame rendering versus verbatim transplantation yields acceptable-response rates 77.7% versus 17.7%, relevance 2.95 versus 1.63, coherence 3.23 versus 3.07, mismatch 0.1% versus 60.3%, and strict-valid rate 8.1% versus 4.5%. Preference counts are 1,212 frame, 126 verbatim, and 462 ties, giving a non-tie frame win rate of 90.6% [88.9, 92.0]. Nominal Krippendorff agreement for preference is $\alpha = 0.557$; frame relevance and coherence agreement are $\alpha = 0.272$ and $\alpha = 0.487$. Frame mismatch is too rare for a stable chance-corrected coefficient (one positive rating), with 99.8% exact three-rater agreement. This human semantic audit is distinct from the deterministic renderer

Table 8: Reference-distribution sensitivity of the semantic-linked policy contrast on matched posts. Values are family-specific paired point estimates; post-cluster intervals are provided in the supplement artifacts.

Q_{ref} family	Model	Volume ratio	Δ leaf depth	Δ toxicity
Deepseek	deepseek-v4-flash	1.163×	+0.130	+0.157
Openai	gpt-4o-mini	1.129×	+0.028	+0.156
Qwen	qwen3.7-plus	1.063×	+0.090	+0.164

non-feedback check. Two-way cluster-robust intervals account for repeated judgments within both items and anonymous rating assignments. The 3/5 acceptability rule is a minimum execution criterion; the 4/5 strict rule remains a deliberately demanding secondary stress test. Frame acceptability is balanced across the declared treatment coordinates: conflict minus baseline is +1.6 pp [-1.9, +5.2], controversy minus best is +0.8 pp [-2.7, +4.3], deep minus shallow replies is +1.8 pp [-3.1, +6.8], and root minus shallow replies is +2.9 pp [-2.0, +7.7]. The semantic variables used by the simulator remain operational policy inputs; the paper does not interpret them as latent psychological traits.

I Small-Sample TBBT Moderation Supplement

Eight quarantine or month-long post-removal interventions have observable within-community activity; five pass the source-overlap, pre-period-fit, trend-equivalence, predictive-skill, and donor-weight gates. Using only pre- intervention records and previously fixed parameters, the fixed mechanism map obtains 100.0% direction accuracy and 3.15 percentage-point MAE, compared with 16.60 for zero effect. Its paired MAE improvement is 13.45 [6.38, 18.02] points, with 80.0% point-estimate coverage. This five-event panel is reported as a secondary cross-community stress test: event eligibility and donor gates are auditable, but the evidence is retrospective. The primary chronological external benchmark is the larger 103-event Lemmy aggregate split, whose test outcomes are excluded from fitting and model selection; its event-path component is reported separately as a sparse-window diagnostic.

J Evidence Scope and Reproducibility

Observational identification. The Lemmy risk-set contrasts rely on correctly ordered treatment times, overlap, stable unit treatment, adequate pretrend balance, and no unobserved time-varying confounding after calibration. The shared control pool is formed from 21 pre-intervention features and audited for balance, overlap, and weight concentration. These diagnostics make the assumptions inspectable but do not turn the design into a randomized experiment. The chronological replay test uses no test-period outcomes for fitting, yet its direction accuracy remains a predictive intervention-fidelity measure rather than an individual causal effect.

Estimand scope. The replay estimand concerns executable simulator policies under semantic potential drawn from the prespecified Q_{ref} distribution and held fixed across arms. It is complementary to the adaptive-semantic simulator effect rather than a claim that post-treatment semantic response is bias. The observational intervention panels test aggregate platform responses, while model-family agreement tests operational semantic stability. Accordingly, the reported claims do not depend on personality diagnosis or on individual-level human treatment-response assumptions.

Reproducibility. Runs are keyed by data and split hashes, configuration, code version, post, condition, model identifier, and seed. Event paths are stored separately from summary statistics, and reported tables and figures are generated from the result artifacts. API credentials are read only from environment variables; cached semantic intents permit offline replay of simulation, evaluation, and report generation.

K Extended Comparison with Agent Simulators

Prediction is not intervention. Popularity prediction and diffusion models estimate whether, when, and how far content will propagate from temporal or structural observations Szabo & Huberman (2010); Cheng et al. (2014); Zhao et al. (2015); Cao et al. (2017). Their compact stochastic representations also provide strong generative baselines. However, matching a cascade distribution does not by itself define what would change under a ranking, behavioral, or moderation intervention. BDMTF therefore separates distributional fidelity from intervention-response fidelity: the former asks whether simulated cascades resemble observed cascades, while the latter asks whether a declared policy contrast produces the correct change in outcomes.

Generative social simulation and the closest policy simulators. Generative Agents and CAMEL established language-model agents with memory, planning, personas, and multi-agent communication

Park et al. (2023); Li et al. (2023). Social Simulacra uses generated communities to explore social-computing design changes Park et al. (2022); OASIS exposes large-scale platform, recommendation, and network mechanisms Yang et al. (2025); and MOSAIC evaluates content-distribution and moderation strategies while comparing agent and human engagement choices Liu et al. (2025). These systems make rich intervention experiments possible, but their online text generation normally remains part of the transition process. Once an intervention changes an agent's visible state, it can also change the prompt and generated text, so structural exposure and semantic adaptation vary together.

COSMOS is the closest counterfactual policy simulator: it executes factual and counterfactual moderation branches in parallel and controls decoding with a common seed Fidone et al. (2026). Its branches nevertheless condition on branch-specific sensory state and moderation memory. It therefore estimates a total simulated response that includes adaptive semantic change. BDMTF targets a different, narrower estimand: before treatment assignment, it freezes the candidate semantic intents available to each agent; treated and control runs then replay the same semantic opportunities through different behavioral and platform mechanisms. The two designs are complementary. COSMOS studies policy response with language adaptation, whereas BDMTF isolates structural response under fixed semantic potential. Table 9 makes this estimand distinction explicit and aligns it with the semantic controls and validation targets reported by each simulator.

Counterfactual semantics, validation, and transport. Recent work formalizes the assumptions needed to interpret policy effects in simulated online communities, distinguishing simulator-conditional causal claims from claims about human platforms Goyal et al. (2026). BDMTF operationalizes a restricted instance of this program: the frozen-intent information barrier defines a reference-distribution fixed-semantics target, paired replay estimates it, and repeated live-versus-frozen trials measure the contribution of branch-specific post-treatment semantic adaptation relative to that target. Adaptive-semantic replay remains a valid simulator estimand; Rehydration answers the complementary structural-policy question.

Observed social-platform effects additionally face exposure, homophily, selection, timing, and overlap assumptions Aral et al. (2009); Shalizi & Thomas (2011). We use moderation natural experiments and modern difference-in-differences diagnostics as external intervention tests rather than treating simulator contrasts as real-world causal estimates Chandrasekharan et al. (2022); Callaway & Sant'Anna (2021); Sun & Abraham (2021). Cross-platform tests pose a related portability problem because content and platform architecture can change simultaneously Wang et al. (2021); we therefore distinguish zero-target transfer, limited target adaptation, and matched-content evaluation. The novelty is not the use of agents or simulated interventions alone, but an executable fixed-semantics replay protocol that shares semantic potential across branches, exposes estimable structural policy interfaces, and tests their predictions against held-out observed interventions.

L Extended Rehydration Algorithm and Proof

Rehydration separates semantic proposal generation from counterfactual state transitions. A generator first returns immutable semantic templates $\bar{z} = (\text{act}, \text{stance}, \text{basis}, \text{tone}, \text{slots})$, which are abstract frames rather than replies tied to future parents. A prespecified capacity map expands each template into indexed opportunity tokens $z = (\bar{z}, k)$; the index changes availability, not content. Let $\chi_i(z, u, a, v) \in \{0, 1\}$ be a compatibility predicate fixed before treatment assignment. It checks whether intent z belongs to agent u, supports action type a, and satisfies the declared static schema for target v. The frozen object contains a dialogue act, stance, argument basis, tone, and typed parent-reference slots. The predicate cannot inspect the treatment label, live engagement counts, or ranking position. Dynamic target availability is determined by the platform-policy channel; after target selection, a deterministic renderer $R(z, v)$ fills the slots from the selected parent. The rendered string is excluded from subsequent transition features. Writing $z = (q, s)$ distinguishes the operational coordinate q, which the transition kernel reads, from the rendered surface s, which cannot feed back into activation, targeting, or timing.

This separation gives two distinct robustness questions. Swapping only the surface distribution tests renderer invariance, whereas changing the frozen action and polarity distribution changes Q^q_{ref} and

Table 9: Counterfactual design and validation targets in closely related social-agent simulators. The entries characterize the reported target and evidence; they do not rank general simulator quality or treat empirical validation as a binary property.

Method	Policy contrast	Cross-arm semantic control	Target quantity	Reported external evidence
OASIS (2025)	Configured platform, recommendation, and scale scenarios	No paired proposal coupling specified	Emergent system response under configured environments	Reproduction of established social phenomena
MOSAIC (2025)	Three content-moderation strategies	No paired realized-proposal coupling reported	Simulated dissemination and engagement response	Engagement alignment with 204 human participants
COSMOS (2026)	Parallel factual and counterfactual moderation branches	Common decoding seed; prompts vary with branch memory and sensory state	Adaptive simulated moderation response, including semantic change	Profile and toxicity realism calibration
BDMTF (ours)	Paired executable policies over the same semantic opportunities	Identical candidate proposals frozen before treatment	Fixed-semantics structural effect	Chronologically held-out real-platform intervention responses

Cross-arm semantic control concerns realized proposal objects, not merely a shared topic, initial state, or random seed. Once treatment changes a prompt, memory, or visible state, common-seed decoding may generate different semantic content. "Held-out intervention response" denotes an outcome recorded after a real platform intervention and excluded from model fitting. It is intentionally distinguished from human action alignment, observational calibration, and the reproduction of known phenomena.

can change the fixed-semantics estimand. We report the latter as reference-coordinate sensitivity and evaluate surface compatibility with blinded judgments.

Algorithm 1 specifies both phases. Phase I generates and seals the pre-treatment semantic pool Z_i. Phase II initializes two policy branches from the same context and replays them with the same immutable pool and a keyed random-number schedule. Each branch has its own used-token mask, so a policy can change whether and when an opportunity is realized without altering the opportunities that were initially available.

The keyed stream makes a random variable addressable by $(i, r, t, u, \mathrm{operation})$. Consequently, an unused draw in one branch does not shift every later draw relative to the other branch. The semantic frame is immutable, the parent-conditioned surface is deterministic, and the transition kernel reads the frame rather than the surface. An opportunity token can be realized at most once within each branch; several indexed tokens may share one immutable template. Thus a capacity multiplier adds declared opportunities without adding generated texts, tokens, or model calls. If no unused compatible intent remains, the attempted action becomes a no-op and an exhaustion event is recorded; the simulator never requests a replacement from the language model. Exhaustion rates and larger-pool sensitivity distinguish a substantive finite-opportunity effect from an undersized semantic cache.

Proof of Proposition 1. The information-barrier condition in Section 3.1 makes the conditional distribution of Z invariant to π, and the coupled construction reuses the same realized pool, so $Z(\pi_1) = Z(\pi_0) = Z$. Because Phase II makes no calls to Q_{ref} and the renderer is excluded from the transition kernel, no post-treatment proposal enters either trajectory. Taking the expectation of the paired difference therefore gives the fixed-semantics effect in Equation 1.

Let $Y_b = M(\pi_b, Z, \epsilon_b)$, $\sigma_b^2 = \mathrm{Var}(Y_b)$, and $C = \mathrm{Cov}(Y_1, Y_0)$. For one paired replay block,

$$\mathrm{Var}(Y_1 - Y_0)_{\mathrm{paired}} = \sigma_1^2 + \sigma_0^2 - 2C, \tag{10}$$

which follows from $\mathrm{Var}(Y_1 - Y_0) = \mathrm{Var}(Y_1) + \mathrm{Var}(Y_0) - 2\,\mathrm{Cov}(Y_1, Y_0)$. Independent replay sets the cross-arm covariance to zero. Under the additional assumption that paired and independent replay have the same marginal outcome distribution under each policy, its variance is $\sigma_1^2 + \sigma_0^2$. Hence pairing has no greater variance when $C \geq 0$ and strictly lower variance when $C > 0$. The construction

Algorithm 1 Rehydration with frozen intents and paired structural replay

Require: $X_{1:n}$, Q_{ref}, π_0, π_1, seeds $\omega_{1:R}$, horizon T
Ensure: Paired effects $\widehat{\Delta}_{ir}$ and exhaustion logs

Phase I: condition-blind proposal generation

for $i = 1, \dots, n$ **do**
  $\bar{Z}_i \leftarrow Q_{\mathrm{ref}}(\cdot \mid X_i, \xi_i)$, excluding treatment and live-state fields
  $Z_i \leftarrow \{(\bar{z}, k) : \bar{z} \in \bar{Z}_i, k = 1, \dots, K_{\bar{z}}\}$ using prespecified capacities
  attach fixed compatibility schema χ_i; seal and hash Z_i

Phase II: paired structural replay

for $i = 1, \dots, n$ and $r = 1, \dots, R$ **do**
  $\epsilon_{ir} \leftarrow \textsc{KeyedStream}(\omega_r, i)$ ▷ addressed by time, agent, and operation
  for $b \in \{0, 1\}$ **do**
    $S_0 \leftarrow \textsc{Initialize}(X_i)$; $A_{ir}^{(b)} \leftarrow \varnothing$ ▷ branch-specific used set
    for $t = 0, \dots, T-1$ **do**
      $S_{t+1} \leftarrow S_t$ ▷ carry state through no-op steps
      $O \leftarrow \textsc{Viewport}(S_t, \pi_b, \epsilon_{ir}[t, \mathrm{rank}])$
      $(u, a, v) \leftarrow \textsc{PolicyStep}(S_t, O, \pi_b, \epsilon_{ir}[t, \mathrm{action}])$
      if $a =$ no-op **then**
        continue
      $C \leftarrow \{z \in Z_i \setminus A_{ir}^{(b)} : \chi_i(z, u, a, v) = 1\}$
      if $C = \varnothing$ **then**
        record PoolExhausted; **continue** ▷ no refill or regeneration
      $z \leftarrow \textsc{Select}(C, \epsilon_{ir}[t, u, \mathrm{intent}])$
      $x \leftarrow \textsc{Render}(z, v)$ ▷ deterministic surface emission; no model call
      $\textsc{Emit}(x)$
      $S_{t+1} \leftarrow \textsc{Apply}(S_t, (u, a, v, z), \pi_b, \epsilon_{ir}[t, u, \mathrm{transition}])$
      $A_{ir}^{(b)} \leftarrow A_{ir}^{(b)} \cup \{z\}$
    $M_{ir}^{(b)} \leftarrow m(S_{1:T})$
  $\widehat{\Delta}_{ir} \leftarrow M_{ir}^{(1)} - M_{ir}^{(0)}$
return $\{\widehat{\Delta}_{ir}\}$, pool-exhaustion diagnostics

targets the prespecified reference distribution and does not reclassify the adaptive effect in Section 3.1 as bias. □

The variance result is conditional on the sign of C; common random numbers are not guaranteed to help when the induced covariance is negative. When multiple replay seeds reuse the same Z_i, uncertainty estimation must retain discussion-level grouping rather than treating all (i, r) pairs as independent observations.

M Extended Policy-Interface Conditions

The four policy interfaces define coordinates of the simulator transition kernel, not merely software abstractions. Given fixed initial context X and semantic pool Z, one simulation step factorizes as

$$\begin{aligned} &K_\pi(O_t, e_t, u_t, a_t, v_t, c_t \mid S_t, X, Z) \\ &\quad = K_R(O_t \mid S_t; \theta_R) K_A(e_t, u_t, a_t \mid S_t, O_t; \theta_A) \\ &\qquad \times K_T(v_t \mid S_t, O_t, u_t, a_t; \theta_T) K_C(c_t \mid S_t, u_t, a_t, v_t; \theta_C). \end{aligned} \tag{11}$$

Here O_t is the exposed viewport; e_t indicates whether an event occurs; u_t, a_t, and v_t are the actor, action, and target; and c_t is a corrective or moderation response. The selected frozen intent and these outputs determine S_{t+1} through the update function in Section 3.1.

Interface coordinates and intervention channels are orthogonal. Parameters inside an interface are tagged by ownership, $\theta_j = (\theta_j^B, \theta_j^P)$: for example, Activation can combine an agent-response coefficient θ_A^B with platform-supplied exposure or external-arrival coefficients θ_A^P. A platform intervention replaces only the declared P components, even when those components enter more than one transition kernel. Table 10 links each interface to its kernel output, intervention lever, and observable calibration target.

Table 10: Policy interfaces as intervention coordinates. Each coordinate has a declared transition quantity, manipulable parameters, and observable calibration targets.

Interface	Kernel output	Example intervention	Calibration target
Activation	event/actor/action	arrival, repeat, exit	timing, survival
Target	reply parent	root share, depth bias	parent/depth shares
Ranking	viewport order	new, hot, best, width	rank/exposure logs
Correction	corrective state	counterspeech, remove, lock	moderation events

An intervention on coordinate $j \in \{A, T, R, C\}$ replaces θ_j^0 with θ_j^1 while holding θ_{-j}, X, Z, and the paired random schedule fixed. Its controlled coordinate effect is

$$\tau_j = \mathbb{E}\big[M(\theta_j^1, \theta_{-j}; X, Z, \epsilon) - M(\theta_j^0, \theta_{-j}; X, Z, \epsilon)\big]. \tag{12}$$

Joint replacements define interaction effects such as Equation 2. Every reported experiment therefore specifies the changed coordinate, the fixed coordinates, and the support over which the contrast is evaluated. This is what makes the interfaces estimable intervention coordinates rather than interchangeable application-programming classes.

State–action contract. For platform p, an adapter g_p maps native event logs into a common contract

$$\mathcal{K} = (\mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{E}), \tag{13}$$

consisting of platform states, exposed viewports, available actions, and timestamped state-transition events. The contract distinguishes replies, votes, abstention, counterspeech, removal, and locking, and preserves reply parentage and event order. An action unavailable on a platform is represented as unsupported; it is not silently imputed.

Platform-parameter-only transport from source platform s to target platform t keeps the contract and behavioral channel fixed and replaces only declared platform parameters:

$$M_t = M(B_s, P_t; X_t, Z_t, \epsilon). \tag{14}$$

This operation is justified only when all of the following conditions hold:

1. **Contract equivalence:** g_s and g_t preserve the meaning of the state variables, reply edges, timestamps, and actions needed by the four kernels.
2. **Conditional behavioral invariance:** over common support, the retained behavioral components of the activation and target kernels satisfy $K^B_{A,t} = K^B_{A,s}$ and $K^B_{T,t} = K^B_{T,s}$ up to prespecified target-validation tolerances.
3. **Support overlap:** target states, viewport positions, depths, and actions lie inside the support on which retained kernels were estimated.
4. **Semantic compatibility:** the frozen intents and compatibility predicate use action and target types that exist under the target contract.
5. **Measurement alignment:** exposure, moderation, and event timing have comparable definitions and are available before target-test outcomes are opened.

These conditions are testable restrictions, not consequences of using a shared API. Contract coverage and support are audited directly; retained kernels are checked on target training and validation partitions; platform parameters are then fitted without target-test outcomes. If an action primitive has no semantics-preserving mapping, if behavioral invariance fails, or if support is inadequate, the state–action contract must be expanded or the behavioral kernel re-estimated. Such a result is reported as contract adaptation rather than platform-parameter-only transport. Accordingly, the portability claim is limited to threaded discussion platforms satisfying these conditions, not arbitrary feed-based social networks.

Connectivity. A connectivity graph is an optional input to the transition kernels, not a substitute for the state–action contract. Reply trees, co-participation networks, and follower relations remain distinct objects. When follower data are unobserved, connectivity is explicitly null; a co-participation graph is not relabeled as a follower graph to manufacture a transport feature.

N Extended Agent Policies and Mechanism Deletions

The agent model specifies how observable thread signals are mapped to simulated actions. Its variables are operational response coefficients rather than latent psychological attributes. We distinguish two parameter regimes. The *prespecified controlled regime* uses theory-guided coefficients for factorial intervention and mechanism-deletion experiments. These coefficients are fixed against predeclared calibration targets before the confirmatory analyses and are intended for controlled mechanism intervention rather than independent behavioral estimation. The *data-estimated regime* fits arrival, reply-targeting, depth-decay, author-reuse, and correction parameters using chronological training cascades only and freezes them before held-out evaluation.

N.1 Parameter Provenance and Freeze Boundary

Table 11 records the provenance and outcome-use boundary of each parameter family. The complete controlled configuration is sealed by a source hash before sensitivity, held-out, cross-platform, and real-intervention analyses. Data-estimated models are frozen after training and before held-out evaluation partitions are opened. No parameter is updated after observing a mechanism-deletion or held-out intervention outcome.

Accordingly, agreement with prespecified reference targets is interpreted as a within-simulator mechanism check, whereas held-out fidelity and intervention-response prediction provide separate empirical evaluations.

N.2 Activation and Action Policies

Let $D_{\mathrm{conf}}(u)$, $D_{\mathrm{pros}}(u)$, and $D_{\mathrm{att}}(u)$ denote bounded communication-history summaries for conflict responsiveness, prosocial response, and attention responsiveness, respectively. The controlled activation rule is

$$I_{u,t} = I_{\mathrm{base}} + \alpha D_{\mathrm{conf}}(u)\psi_{\mathrm{con}}(t) + \gamma D_{\mathrm{pros}}(u)\psi_{\mathrm{cons}}(t) + \beta D_{\mathrm{att}}(u)\psi_{\mathrm{heat}}(t) + \epsilon_{u,t}, \tag{15}$$

with $I_{\mathrm{base}} = 0.20$, $\alpha = 1.8$, $\gamma = 0.6$, $\beta = 1.0$, and $\epsilon_{u,t} \sim U(-0.1, 0.1)$. These coefficients are prespecified mechanism parameters rather than coefficients estimated from held-out cascades.

The three signals are computed from visible approval, disapproval, replies, and the declared community-scale normalizer $m_{\mathrm{early},c}$. Activation requires both $I_{u,t} > \tau_u$ and a separate event-hazard draw; crossing the threshold alone does not realize an action.

For an eligible action, the controlled probability of negative polarity is

$$P(a_{u,t} = \mathrm{negative}) = \sigma(\kappa_{\mathrm{pol}} D_{\mathrm{conf}}(u)\psi_{\mathrm{con}}(t)), \qquad \kappa_{\mathrm{pol}} = 5. \tag{16}$$

The conflict-responsive intervention therefore changes declared response-policy coefficients and activation behavior; it does not assign latent psychological types to authors.

N.3 Targeting and External Traffic

For a visible candidate at depth d, the explicit depth-preference component assigns weight

$$w_{u,t}(d) \propto d^{\lambda_u}, \qquad \lambda_u \in [0, 3.5], \tag{17}$$

before the separately specified depth-fatigue, collapse, root-bias, and viewport constraints are applied. The data-estimated alternative replaces these prespecified quantities with estimates obtained from chronological training cascades.

The external-arrival process is

$$\Delta V_{\mathrm{ext}}(t) = \eta_{\mathrm{eff}}\, \mathrm{LogPotency}(S_0) M(t) C(t) S(t) \delta^t \xi_t, \tag{18}$$

where $\eta_{\mathrm{eff}} = 80$ and $\delta = 0.998$ in the controlled regime. External reactions may alter visibility state but cannot create, modify, or replenish semantic opportunities.

Table 11: Parameter provenance and freeze boundary. Prespecified values define the controlled mechanism; training-estimated values support independent fidelity analyses.

Parameter family	Representative values	Origin	Outcome-use boundary
Impulse and polarity	$I_{\text{base}} = 0.20$, $\alpha = 1.8$, $\beta = 1.0$, $\gamma = 0.6$, $\kappa_{\text{pol}} = 5$	Theory-specified controlled constants	Fixed for all reported contrasts; not estimated from test outcomes
Community activation scale	$\tau_{\text{base},c}$ and $m_{\text{early},c}$	Community-level summary calibration	Frozen before held-out evaluation; not re-estimated on test posts
Action, depth, and traffic	Treatment multipliers, depth routing, $\eta_{\text{eff}} = 80$, $\delta = 0.998$	Controlled reference-target calibration	Reference summaries informed these values; used for mechanism analysis, not independent validation
Data-estimated transition policies	Arrival decay, root share, depth decay, author reuse, correction rates	Chronological training split (285 cascades)	No reference targets and no validation/test outcomes used for fitting
Platform adapters	Ranking, viewport, exposure, and target-platform transition parameters	Target training and validation partitions	Target test partition opened only after adapter selection

N.4 Mechanism Deletion and Local Robustness

Local robustness is distinguished from pathway deletion. A 24-setting nine-dimensional Latin-hypercube design varies conflict, heat, consensus, activation, action, depth, viewport, leader share, and traffic around the frozen controlled configuration. All 24 settings retain the higher-volume, lower-depth direction, with volume ratios from 2.910 to 4.078 and leaf-depth changes from -13.154 to -10.738. This establishes local robustness around the declared configuration rather than simulator-wide universality.

A stronger mechanistic test removes complete dynamic pathways without refitting. Table 1 evaluates 25 fixed posts (five per community) under paired frozen-opportunity replay with two seeds and post-level stratified bootstrap intervals.

Removing conflict-conditioned activation, polarity, reply, targeting, and external-feedback pathways reduces the volume ratio of ratios from 1.414 to 1.059. The paired comparison corresponds to a 1.335-fold attenuation [1.230, 1.458]. Removing heat and the explicit depth mechanisms also attenuates volume. These estimates are four-cell behavior–ranking interactions and should not be conflated with the older joint conflict-plus-ranking versus baseline contrast, which targets a different estimand.

The negative depth interaction remains after setting the displayed depth exponent to zero and after removing the complete explicit depth channel. Instrumented replay shows that ranking changes the exposed candidate set differently under conflict-responsive behavior, after which generic parent selection propagates this difference into selected-parent depth. The residual depth effect therefore does not require a single hand-coded depth exponent. A separately learned reply-topology model provides an additional test of whether the interaction survives replacement of the hand-specified routing structure (Appendix O.6).

O Extended Experimental Protocols

O.1 Data, Splits, and Metrics

The controlled simulator uses 500 posts from r/AskReddit, r/aww, r/funny, r/science, and r/worldnews; 482 pass parent, cycle, and depth checks. The broader Reddit corpus contains 24 communities, 12085 complete cascades, and 5579919 comments. Hacker News contains 5000 complete cascades with a 3,000/1,000/1,000 chronological split. No post occurs in more than one split, and author history is truncated at the target-post timestamp.

Baselines are empirical cascade resampling, a negative-binomial branching process, a Hawkes process, theory-specified BDMTF, and data-estimated BDMTF. Structural outcomes include size, maximum depth, mean leaf depth, depth and width distributions, branch factor, root-reply share, concentration, arrival time, duration, and author reuse. Distributional fidelity uses normalized Wasserstein distance, with KS and energy distance as secondary checks. Intervention analyses report paired effects, cluster-bootstrap intervals, direction accuracy, and interval coverage. For reply tree G,

$$D_{\text{leaf}}(G) = \frac{1}{|\mathcal{L}(G)|} \sum_{v \in \mathcal{L}(G)} d(v). \tag{19}$$

The observed Reddit mean is 2.22; controlled simulator depths are reported separately.

O.2 Identification and Model-Family Checks

The known-ground-truth SCM uses the outcome equation

$$Y(p, z, u) = \tau p + \theta z + \eta p z + (\sigma + \rho p) u, \tag{20}$$

with $E[z] = E[u] = 0$, so the population fixed-semantics effect is exactly τ. Independent live replay redraws semantic and structural innovations; common-seed adaptive replay shares them but adds the same policy-dependent semantic shift; Rehydration uses the pre-treatment z in both branches. Two outcomes, three shift values, 2,000 replications, and 250 units per replication are evaluated with bias, MSE, and 95% interval coverage. The zero-shift cell is a negative control; all methods are approximately unbiased there. Table 12 summarizes the resulting error, MSE, and coverage: Rehydration recovers the fixed-semantics target, whereas both live replay designs retain error when semantics adapt after treatment.

The motivating RQ1 thread contains 53 comments and uses a 48.5-comment median external-arrival target. Online and frozen modes share the post, population, intervention, simulator seeds, prompt family, model, sampling settings, and token budget. Its randomization test permutes within 20 paired seeds.

The multi-thread confirmation uses five held-out posts from each of the five Reddit communities. Before API generation, unique posts are selected nearest the 0.10, 0.30, 0.50, 0.70, and 0.90 cascade-size quantiles in each test split; the resulting sample and SHA256 digest are sealed before any intervention effect is computed. Each thread uses five paired seeds. The prespecified aggregate statistics are the sign count of lower frozen-replay variance and the median thread-level live-to-frozen variance ratio. Intervals resample threads, and the sign test excludes ties. Individual ratios are audit details because five seeds provide limited thread-level precision. The only reversal occurs for r/funny post `1nowo6f`, with a variance ratio of 0.30; it remains in every aggregate result. The contrast compares baseline behavior with best ranking against conflict-responsive behavior with controversy ranking. State-conditioned generation requests one intent batch from each branch's 12-step pilot snapshot for each repeat, then replays using the resulting intent pool and reply/antagonism signals. The frozen condition reuses one condition-blind batch per thread. Pairing applies to simulator seeds; API requests do not specify a decoding seed.

For the same sealed panel, we also report $\widehat{\tau}_{\text{adapt}} - \widehat{\tau}_{\text{FS}}$ rather than treating dispersion as a substitute for estimand comparison. Thread means are computed over the five paired seeds and uncertainty resamples threads. The mean gap is -0.035 [-0.079, 0.011] log-volume units; the median absolute gap is 0.072 [0.061, 0.113], with 0 direction reversals. This diagnostic quantifies the semantic-adaptation component for the tested simulator task without identifying a human mediation effect. Table 13 reports the corresponding thread-level dispersion summary for the sealed panel.

Dynamic-parent compatibility uses frozen abstract frames and deterministic parent-slot rendering. A structural-invariance test runs verbatim and rendered surfaces with identical keyed streams and

Table 12: Estimand and implementation verification in the analytic SCM. Errors are measured relative to the known fixed-semantics target; lower MSE and nominal 95% coverage are preferred.

Method	Error to τ^{FS} ↓	MSE to τ^{FS} ↓	Coverage of τ^{FS} ↑
Independent live	0.285	0.0975	38.6%
Common-seed adaptive replay	0.287	0.0836	0.0%
Rehydration	-0.001	0.0008	95.2%

Table 13: Held-out threads with lower frozen-replay volume-effect dispersion, and median live-to-frozen variance ratios.

Community	Improved	Threads	Median ratio
AskReddit	5	5	3.32
aww	5	5	5.43
funny	4	5	4.67
science	5	5	3.62
worldnews	5	5	22.33
All	24	25	4.89

verifies equality of every cascade metric. The human semantic audit samples 600 pairs from 51,653 eligible replays, stratified by community, factorial condition, and root/shallow/deep target. 90 anonymous raters each score 20 pairs, giving 1,800 ratings and exactly three independent ratings per item. Candidate order is randomized. The rubric records 1–5 relevance and coherence, binary obvious mismatch, and frame/verbatim/tie preference. Minimum acceptability requires relevance ≥ 3, coherence ≥ 3, and no mismatch; strict validity raises both score thresholds to 4. Confidence intervals use an intercept-only paired model with two-way cluster-robust covariance for item and anonymous assignment. Agreement statistics are reported without post-outcome rater exclusions. The interface presents a content warning and permits stopping at any point; assignment labels contain no direct participant identifier.

O.3 Frozen-Pool Capacity

Each counterfactual branch owns an independent no-replacement consumption session. A capacity-convergence panel uses 25 posts, three seeds, four factorial arms, and multipliers 1, 4, 16, 64, and 256 (1,500 runs). Exhausted requests become logged no-ops. At 1×, supportive requests do not exhaust while antagonistic requests exhaust in 92.3–97.8% of attempts; the 256× panel has zero exhaustion. Because that panel alone cannot establish coverage, we freeze 512× and separately replay all 500 source posts over four arms (2,000 runs). This full-coverage audit also has 0.00% exhaustion and yields 3.700× joint volume with leaf-depth change -12.432.

The multiplier is a maximum number of branch-local realizations per immutable abstract intent, not a request for newly generated text. The controlled catalog contains 22,974 empirical-comment frames (4,486–4,693 per community; 6.6 MB). Thus 512× exposes 11,762,688 finite replay opportunities but incurs zero additional LLM calls and zero additional tokens. Five community workers complete the full audit in approximately 18.6 wall-clock minutes on the reported machine. The selected capacity is a coverage setting, not a fitted outcome parameter. Table 14 shows how exhaustion disappears as the finite replay capacity increases.

Across OpenAI-, DeepSeek-, and Qwen-family outputs, reply decisions have 90.4% three-way agreement and broad polarity has 76.0% agreement. Strict replay has maximum structural difference 0.000 because model-specific signals are excluded by construction; this is a reproducibility check, not evidence of semantic validity. Complete audit statistics are in Appendix H.

Table 14: Frozen-pool capacity sensitivity on the convergence panel. Maximum exhaustion is over the four factorial arms.

Capacity	Max exhausted	Joint volume	Δ leaf depth
1×	83.4%	2.258×	-8.787
4×	77.1%	1.788×	-9.200
16×	54.0%	2.111×	-12.516
64×	19.0%	3.204×	-12.988
256×	0.0%	3.694×	-13.095

O.4 Empirical Depth Scale

Raw simulator depth is not interpreted directly in Reddit units. For each community c, the chronological development split defines

$$s_c = \frac{\overline{D}^{\mathrm{real}}_{c,\mathrm{dev}}}{\overline{D}^{\mathrm{sim,base}}_{c,\mathrm{dev}}}, \qquad \widetilde{D}_i = s_c D^{\mathrm{sim}}_i. \tag{21}$$

Neither test outcomes nor intervention outcomes enter s_c. The calibrated baseline differs from the observed test mean by 15.8%. Positive rescaling audits units while preserving the sign of every depth contrast.

O.5 Lemmy Intervention Replay

Risk-set matching yields 503 non-reused intervention–control pairs. Shared synthetic-control weights use 21 pre-intervention trajectory features and no treatment outcomes. Locks and removals are followed by changes of -1.681 [-2.866, -0.582] replies, -0.990 [-1.595, -0.400] active authors, and -0.611 [-0.793, -0.387] maximum depth. Event replay fits on 301 earlier controls, selects parameters and conformal radii on 99 interventions, and evaluates the 103 later interventions reported in the main text. Majority sign, intervention-type mean, additive Ridge, and same-type historical-neighbor baselines use only the 301 training interventions; Ridge regularization and all conformal radii are selected on the 99 validation interventions. No baseline reads a test-period outcome before final scoring.

For an eight-day horizon H, let $C(d)$ be cumulative replies through day d, T arrival minutes, D reply depths, r root-reply share, and h the Herfindahl concentration of reply parents. Observed and predicted paths are scored by

$$e_{\mathrm{count}} = \frac{H^{-1}\sum_{d=1}^{H} |\log(1+C_o(d)) - \log(1+C_p(d))|}{\log(1+s_C)}, \tag{22}$$

$$e_{\mathrm{time}} = \frac{W_1(T_o,T_p)}{1440H}, \qquad e_{\mathrm{depth}} = \frac{W_1(D_o,D_p)}{s_D}, \tag{23}$$

$$e_{\mathrm{root}} = |r_o - r_p|, \qquad e_{\mathrm{HHI}} = |h_o - h_p|, \tag{24}$$

$$E_{\mathrm{path}} = \frac{1}{5}\sum_j \mathrm{clip}_{[0,1]}(e_j). \tag{25}$$

Here s_C and s_D are 95th-percentile count and depth scales estimated from training paths only. If exactly one path is empty, all distributional components equal one; if both are empty, they equal zero. Equal component weights are fixed before test outcomes are scored.

The saved test windows contain only 4 observed replies across 3/103 nonempty interventions. This sparsity makes empty-path handling dominate the composite: an always-empty predictor obtains 0.029, compared with 0.044 for the path-fitted BDMTF and 0.070 for empirical-nearest replay. The comparison is therefore retained as a coverage and failure-regime audit, not as a positive event-path ranking result. Whether an empty window represents genuine silence or incomplete event collection must be resolved before an active-trajectory fidelity claim is made. Table 15 consolidates this comparison and makes the always-empty failure baseline explicit.

O.6 Held-Out Topology Fidelity and Learned Interaction

The topology audit uses 100 chronological test posts, 20 from each empirical size quintile and 16–22 per community. A context-size model and local reply kernel are fitted only on 285 training cascades.

Table 15: Sparsity audit for the held-out Lemmy event-path benchmark. Errors are the equal-weight E_{path} above; lower is better. "Empty" always predicts no post-intervention replies.

Subset	N	Nonempty	Events	BDMTF	Empirical	Empty
All	103	3	4	0.044	0.070	0.029
Locks	39	1	1	0.026	0.087	0.026
Removals	64	2	3	0.055	0.060	0.031

The kernel combines the 15 nearest training cascades by generated log size and produces a valid reply tree from learned depth, branching, and root-reply distributions. Ten seeds per post yield 1,000 baseline trees. No test outcome enters model fitting, operator selection, or stopping, and no output is rescaled after simulation. Table 16 compares the resulting held-out topology metrics with their prespecified fidelity thresholds.

Figure 5 shows that the node-depth CDF is close in its central mass, but the learned model assigns small probability to an overlong tail. This is most visible in large cascades and r/AskReddit and explains the failed maximum-depth criterion.

For the learned interaction test, the frozen behavioral operator changes conflict-responsive activation and depth fatigue, while the platform operator changes only best to controversy ranking. The complete 2×2 design uses the same 100 posts, ten paired seeds, and 4,000 post–seed–cell records. Intervals use a 5,000-draw post-cluster bootstrap. Table 17 reports the resulting behavioral-by-platform contrasts on volume and raw depth.

The multiplicative volume interaction is 1.234 [1.227, 1.240], and the raw depth interaction is -0.313 [-0.360, -0.264]. These estimates replicate the controlled direction under a separately learned topology; they do not establish a total human causal effect.

P Boundary Maps and Aggregate Intervention Baselines

P.1 Local Robustness and Global Failure Regions

The local panel perturbs the declared controlled mechanism, whereas the global panel searches a separately learned simulator. They answer different questions and are not pooled into a single robustness statistic. Figure 4 contrasts these two scopes: the left panel tests stability near the declared operating point, whereas the right panel locates failure and reversal regions across the broader learned-model grid.

Among 507 non-supporting cells, the most common response is higher volume with greater depth (263 cells). The map is a boundary analysis rather than an alternative estimate of the local robustness rate.

Figure 6 complements the broad stress map with a targeted two-coordinate test varying conflict-response strength and ranking viewport. A starred cell requires the volume RoR interval to lie above one and the depth DiD interval to lie below zero. The pattern shows that neither a responsive population under a narrow exposure set nor a wide viewport without conflict response is sufficient. Thus, this panel locates the activation boundary, whereas Figure 4 characterizes robustness and failure modes over broader perturbations.

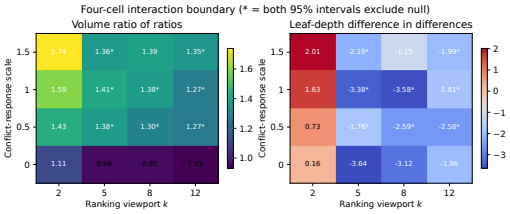


Figure 6: Four-cell boundary; stars mark cells with volume RoR CI > 1 and leaf-depth DiD CI < 0.

Table 16: Held-out topology fidelity without post-hoc scaling. NWD is normalized Wasserstein distance; values at most one meet the prespecified scalar threshold.

Metric	Reddit	Learned	NWD
Mean leaf depth	1.989	2.530	0.633
Maximum depth	6.640	14.888	2.318
Root-reply share	0.571	0.471	0.437
Branching factor	1.541	1.614	0.289
Node-depth distribution	–	–	0.238

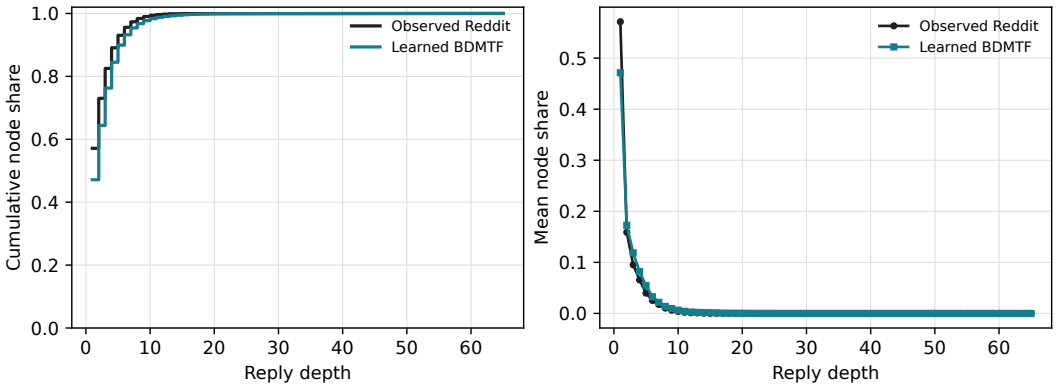


Figure 5: Observed and learned reply-depth profiles on held-out Reddit posts. The left panel shows cumulative node mass; the right panel exposes the small but overlong simulated deep tail.

Table 17: Behavioral-by-platform interaction in the learned topology. Depth is reported in raw simulated units without calibration or scaling.

Outcome	B_0P_0	B_1P_0	B_0P_1	B_1P_1	Interaction
Log volume	5.187	5.330	5.187	5.542	0.210
Mean leaf depth	2.530	2.196	2.530	1.882	-0.313

P.2 Held-Out Lemmy Aggregate Prediction

Table 18 separates aggregate prediction from the event-path objective in the main text. Direction accuracy is secondary because the majority-sign and BDMTF rows coincide.

Q Reference Distributions and Activation Function Forms

Shared comparison design. The comparison uses all 50 posts with complete, prompt-matched caches from OpenAI, DeepSeek, and Qwen. Each post has ten cached participant decisions per family. The same baseline/conflict-responsive behavior and best/controversy ranking define four policy cells. Three paired seeds produce 1,800 reference comparisons and 1,800 activation-form comparisons; 200 rank-disconnected runs check the null contract. Population size, horizon, and viewport remain 50, 72 steps, and five positions. The new panels use random streams keyed by post, seed, step, actor, and operation, preventing cross-operation draw shifts. Only recorded early-engagement fields initialize the root; missing fields contribute zero. These are controlled-policy comparisons, not a forecasting evaluation of the observed cascades.

Frozen reference mapping. Let r be the fraction of cached decisions that propose a reply and a the antagonistic fraction among those replies. The condition-blind mapping is

$$f_r = \mathrm{clip}_{[0.5,1.5]} \frac{r + 0.05}{1.0 + 0.05}, \qquad f_a = \mathrm{clip}_{[0.5,1.5]} \frac{a + 0.05}{0.1 + 0.05}. \tag{26}$$

For every factorial cell, f_r multiplies its action multiplier and f_a multiplies its conflict coefficient and polarity slope. The mapping and reference anchors are fixed independently of the comparison outcomes. Cached reply intents are assigned to their ten participants; unmatched actor/polarity

Table 18: Held-out Lemmy intervention prediction with training-only baselines. MAE is primary. BDMTF additionally emits event paths, whose sparse-window audit appears in Table 15.

Model	MAE ↓	Direction ↑
Zero effect	1.923	8.4%
Majority sign	1.839	73.1%
Type mean	2.181	73.1%
Additive Ridge	1.208	69.6%
Historical nearest	1.612	72.2%
BDMTF event replay	1.634	73.1%

requests use the common empirical background. Model-cache intents account for 26.2–27.5% of realized replies. The maximum grouped exhaustion rate is 0.000% at reuse capacity 512. Reuse creates no new model-generated text. Each model/prompt has one recorded generation, so the uncertainty intervals quantify post and replay variation conditional on these cached catalogs. Because the family-specific pool and (f_r, f_a) vary together, this panel is a semantic-linked implementation sensitivity rather than a content-only Q_{ref} effect. Table 19 reports the corresponding factorial contrasts for each cached reference distribution.

Activation alternatives. Write the original non-noise signal as $s = \sum_j w_j d_j e_j$, where d_j is a behavioral drive and e_j an exposure signal. The alternatives are

$$s_{\mathrm{add}} = \sum_j w_j(\bar{d}_j e_j + \bar{e}_j d_j - \bar{d}_j \bar{e}_j), \tag{27}$$

$$s_{\mathrm{sat}} = c \tanh(s/c). \tag{28}$$

Drive and exposure means and $c = \max(0.1, \mathrm{median}(s))$ use baseline decision states from five chronologically selected training posts per community. A single intercept per community and form matches the original baseline gate-pass rate on those same training states. The 25 calibration posts are disjoint from the 50 comparison posts. The alternatives retain the original noise and threshold gate; targeting, ranking, traffic, and pool rules are unchanged. No comparison volume, depth, or treatment contrast enters the intercept fit. Additivity of the pre-gate signal does not impose additivity of the resulting nonlinear cascade. Table 20 compares these calibrated forms under the shared empirical intent catalog.

All three forms retain positive joint volume and negative joint depth contrasts. The additive and multiplicative depth-interaction intervals span zero, whereas their volume-interaction intervals remain above one. Joint direction and depth non-additivity are therefore evaluated separately. Figure 7 displays the paired interaction estimates and their uncertainty relative to the no-interaction lines.

Uncertainty and records. Intervals use 5,000 community-stratified post bootstrap draws, retaining all seeds and variant pairs. Complete contrasts and nominal intervals accompany the event-level Parquet files and SHA256 manifests. The disconnected Ranking control routes both labels to best throughout the transition; its zero contrast is a software-contract check.

R Independent Semantic Pools and Mechanism Phase Geometry

R.1 Independent Semantic-Catalog Factorials

The reference-distribution experiment uses 50 posts, three independently generated semantic catalogs, eight paired seeds, and all four behavior–platform cells, for 4,800 runs. Posts, policy coordinates, estimators, and the strict two-outcome rule are shared across families. The prespecified semantic-to-policy link maps each catalog's reply and antagonistic shares into action and conflict scales, so the comparison jointly varies catalog content and its declared response coordinates. Every arm completes with zero pool exhaustion and no missing factorial cell.

The equal-family mixture has volume interaction 1.219 [1.209, 1.229] and depth interaction -1.082 [-1.489, -0.679]. All three family-specific intervals satisfy the same positive-volume and negative-depth criterion. Table 21 reports the family-specific factorial estimates underlying this mixture summary.

R.2 Catalog–Coordinate Decomposition

A second frozen experiment crosses pooled and family-specific catalogs with pooled and family-specific operational coordinate maps on the same 50 posts and two paired seeds (4,000 runs). With

Table 19: Same-policy factorial contrasts under three cached leader-reference distributions. Volume entries exponentiate mean paired log contrasts; depth entries are additive contrasts.

Variant	Joint volume	Volume interaction	Δ leaf depth	Depth interaction
OpenAI	3.524	1.203	-12.723	-1.805
DeepSeek	3.518	1.213	-12.732	-1.751
Qwen	3.563	1.206	-12.831	-2.290

Table 20: Activation-form factorial contrasts with a shared empirical intent catalog. Baseline gate rates are matched using training states.

Variant	Joint volume	Volume interaction	Δ leaf depth	Depth interaction
Multiplicative	3.502	1.205	-12.185	-0.767
Additive	3.650	1.158	-12.042	-0.831
Saturating	3.505	1.207	-12.162	-0.874

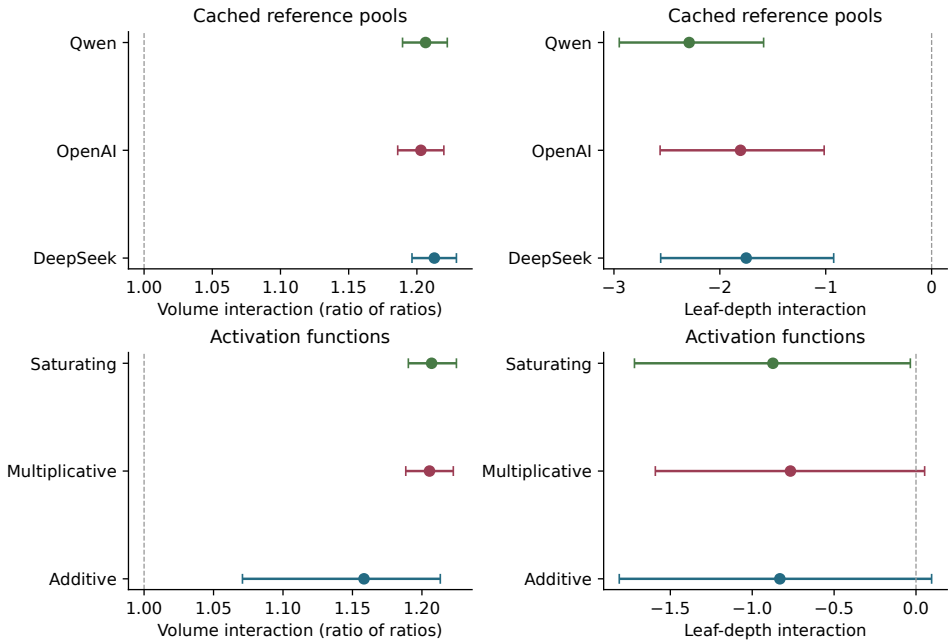


Figure 7: Paired interactions with community-stratified post-cluster 95% intervals. Reference-pool and activation panels vary different inputs. Dashed lines denote no interaction.

Table 21: Behavior–platform interaction under independently frozen semantic reference distributions. Volume uses a ratio of ratios; depth uses a difference in differences.

Family	P effect B_0	P effect B_1	Volume RoR [95% CI]	Depth DiD [95% CI]	Support
DeepSeek	0.987	1.218	1.234 [1.220, 1.247]	-1.103 [-1.584, -0.621]	Yes
OpenAI	0.982	1.215	1.237 [1.224, 1.250]	-1.129 [-1.666, -0.600]	Yes
Qwen	0.992	1.178	1.188 [1.170, 1.205]	-1.014 [-1.500, -0.525]	Yes

Q^q_{ref} fixed, replacing all model-family surface catalogs has maximum absolute changes 0.000 in log volume RoR and 0.000 in depth DiD. These exact zeros verify the declared transition contract: rendered surface text does not feed back into structure. Replacing only Q^q_{ref} produces the nonzero shifts in Table 22; this is the scientific reference-policy sensitivity. All map-only volume-factor intervals include the null value one. For depth, the DeepSeek map produces a positive change from the pooled-map depth DiD, whereas the OpenAI and Qwen change intervals include zero. These columns report changes relative to the pooled-map interaction, not the family-specific interactions after mapping.

R.3 Discovery Boundary and Frozen Interior Phase

The discovery map varies conflict-response strength and ranking viewport over 16 cells. All four zero-response cells and all three additional viewport-2 cells fail the joint rule, whereas 8/9 cells in the declared active region pass. This geometry links the interaction to both a responsive population and a sufficiently expressive exposure set.

Figure 6 shows this targeted boundary; starred cells have volume ratio-of-ratios intervals above one and leaf-depth difference-in-differences intervals below zero.

A disjoint 125-post stress panel then tests a broad active grid. All volume intervals remain above one, while wide-viewport settings expose depth reversals. A separately specified interior grid is evaluated on a third post set. Its initial seeds support 12/16 cells with negative depth point estimates in all cells. A variance audit selects the four borderline coordinates before a single independent-seed precision extension; all four pass using seeds 4–9. The 12 initial cells and four extended cells form a 16/16 composite summary, but only the selected four receive the additional seeds; it is not a uniform 16-cell replication. Figure 8 shows the broad stress grid and the pre-extension interior phase, while Table 23 reports the independent-seed estimates for the four frozen borderline cells.

Table 22: Decomposition of model-family reference sensitivity. Catalog-only columns hold operational coordinates fixed and report changes from the pooled surface catalog. Map-only columns hold the pooled surface catalog fixed and compare each family map f with the pooled coordinate map: volume reports $\mathrm{RoR}_f/\mathrm{RoR}_{\mathrm{pooled}}$ (null 1), and depth reports $\mathrm{DiD}_f - \mathrm{DiD}_{\mathrm{pooled}}$ (null 0). The map-only depth column is a change in the interaction, not the resulting interaction itself.

Family	Catalog-only Δ log RoR	Map-only RoR factor vs. pooled [95% CI]	Catalog-only Δ depth DiD	Map-only Δ depth DiD vs. pooled [95% CI]
DeepSeek	0.000	0.998 [0.969, 1.029]	0.000	1.430 [0.516, 2.351]
OpenAI	0.000	1.017 [0.992, 1.043]	0.000	0.195 [-0.415, 0.802]
Qwen	0.000	1.016 [0.989, 1.043]	0.000	0.302 [-0.670, 1.314]

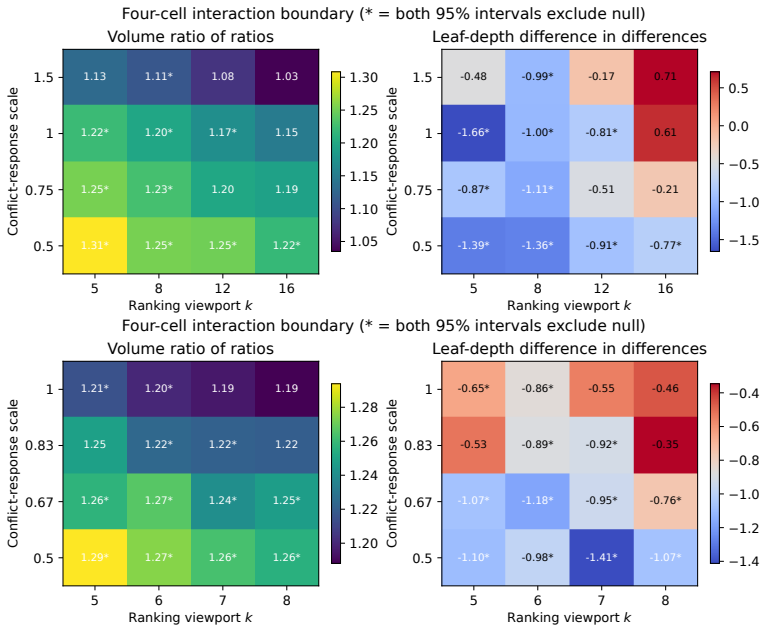


Figure 8: Top: broad disjoint-post stress grid. Bottom: the independently specified interior core phase before the precision extension.

R.4 Pathway and Volume-Matching Diagnostics

The channel-deletion panel removes complete implemented pathways rather than single coefficients. Removing conflict-conditioned response reduces the volume interaction to an interval containing one. Exposure and target traces show a negative behavior-by-ranking interaction in visible-node depth and selected-parent depth. Finally, truncating every factorial cell to the first common number of comments preserves a depth interaction of -1.692 [-2.906, -0.409], showing that the topology shift is not induced by comment count alone.

Table 23: Independent-seed precision estimates for the four frozen core-phase boundary cells. Intervals use only the six new seeds.

Conflict	Viewport	Posts	Volume RoR [95% CI]	Depth DiD [95% CI]	Support
0.83	5	125	1.240 [1.229, 1.251]	-0.762 [-1.162, -0.372]	Yes
0.83	8	125	1.210 [1.199, 1.221]	-0.771 [-1.145, -0.395]	Yes
1	7	125	1.189 [1.177, 1.201]	-0.871 [-1.230, -0.497]	Yes
1	8	125	1.184 [1.173, 1.196]	-0.713 [-1.078, -0.337]	Yes