THE BUTTERFLY EFFECT IN LLM SOCIAL SIMULATIONS: FROM ARCHITECTURE SENSITIVITY TO ROBUSTNESS AUDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

A *butterfly effect* occurs when a local perturbation propagates through a system and produces divergence in its collective outcome. In LLM social simulations, this creates a particular challenge for scientific inference: sensitivity to implementation choices can enter agent interactions and alter the collective outcomes on which scientific claims are based. We study this problem through a repeated Prisoner's Dilemma and a social-media echo chamber, spanning representation- and design-level perturbations across four LLM families. We find that controlled perturbations can substantially alter collective outcomes, with sensitivity varying across both perturbations and models. In the Prisoner's Dilemma, changing only the representation of a matched persona shifts two-agent cooperation by up to 75.2 percentage points, while game framing and memory representation produce smaller or less consistent effects. Multi-agent interaction further reshapes representation sensitivity, producing amplification or attenuation depending on the model and perturbation. In the echo-chamber simulation, persona representation changes interaction targeting and is accompanied by divergence in collective interaction structure, extending representation sensitivity from agent behavior to network-level outcomes. These results show that robustness cannot be inferred from a single simulation configuration and that local sensitivity can change as it propagates through interaction. We introduce **TRAILS** (*Taxonomy for Robustness Audits In LLM Simulations*) and the **Social Simulation Audit Card** as a framework for systematically auditing these dependencies and calibrating the scope of scientific claims to the robustness evidence that supports them.

1 INTRODUCTION

Until recently, simulating collective human behavior in agent-based models required encoding every social tendency by hand: how an agent decides to cooperate, who it talks to, what it remembers, and how its preferences change. Large language models (LLM) have lifted this constraint. By generating agent behavior directly from natural language, they have opened a new design space for studying a wide range of social phenomena (Park et al., 2023). The field has advanced rapidly, and these simulations have been studied in a wide range of contexts. Researchers now use LLM-based social simulations to *explore and explain* social mechanisms such as cooperation (Akata et al., 2025), opinion dynamics (Chuang et al., 2024) and social norms emergence (Ashery et al., 2025); to serve as *intervention testbeds* for content moderation (Liu et al., 2025) and platform design (Puelma Touzel et al., 2025); and to *inform real-world decisions* in policy, public health, and epidemiology (Chopra et al., 2025; PHF Science, 2024). Simulations are moving from exploring what *might* happen to claiming *why* it happens, *when* it will happen, and what should happen next.

As the stakes have risen, so has the field's attention to validation (Anthis et al., 2025; Larooij & Törnberg, 2026; Puelma Touzel et al., 2026). However, existing efforts focus mostly on whether a simulation *resembles* the world: whether agents behave plausibly (Park et al., 2023), whether outcomes match empirical patterns (Larooij & Törnberg, 2026; Yang et al., 2024; Gu et al., 2025), whether interaction is grounded in established theory (Mou et al., 2024; Liu et al., 2024), or whether human evaluators find the results convincing (Puelma Touzel et al., 2026). Realism and empirical grounding address essential dimensions of validity, but they do not determine whether a simulation

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.

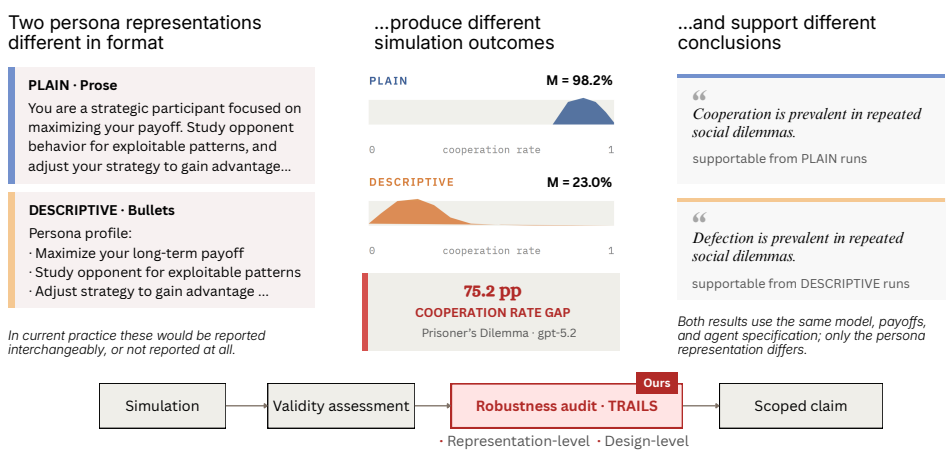


Figure 1: **The butterfly effect in LLM social simulations.** Two representations of the same strategic persona produce a 75.2-percentage-point difference in cooperation in an otherwise matched 10-round Prisoner's Dilemma (`gpt-5.2`, $N = 30$ paired seeds, Holm-corrected $p < .001$). We study how such local sensitivity changes across perturbations, models, and multi-agent interaction, and use TRAILS to connect robustness audits to the scope of scientific claims.

is *robust*: its conclusions should remain stable across reasonable alternative configurations, just as empirical findings are expected to survive alternative controls, operationalizations, and sensitivity analyses (Simonsohn et al., 2020; Steegen et al., 2016). Without robustness testing, we cannot tell whether a finding reflects the social process being modeled or the design choices used to model it.

Robustness matters because LLM-based agents introduce the complexity that traditional rule-based agent-based modeling does not. A simulation can depend on choices about model substrate, persona representation, memory, network initialization, and interaction protocols (Zhou et al., 2025). Choices that appear equivalent to researchers may elicit different decisions from an LLM (Salinas & Morstatter, 2024; Sclar et al., 2024). A small representational change can alter an agent's decision. In a multi-agent simulation, such decisions can in turn alter subsequent interactions, memories, or social signals through repeated interaction, so the local differences can compound into substantially different macro-level outcomes (Bertolotti et al., 2020). We use *butterfly effect* (Ferrara, 2024) to describe this propagation of behavioral sensitivity through interaction. The process can amplify or attenuate initial differences, with effects appearing in both behavioral and collective measures.

To study this phenomenon, we conducted two controlled simulation regimes across four LLM families (Figure 1): a repeated Prisoner's Dilemma comparing fixed-opponent and reciprocal LLM–LLM interaction, and a social-media echo-chamber simulation tracing sensitivity from agent-level choices to collective interaction structure. We varied representation and design choices while holding other components fixed. We found that sensitivity differed across perturbations and models: persona representation substantially shifted cooperation for several models, while other perturbations had smaller or model-dependent effects. In addition, multi-agent interaction further reshapes this sensitivity: reciprocal LLM–LLM interaction amplifies the same persona perturbation for three model families but attenuates it for a fourth, while networked interaction can carry representation sensitivity from interaction targeting to collective network structure. These results show that claims from LLM-based social simulations should match their robustness evidence. We introduce **TRAILS** (*Taxonomy for Robustness Audits In LLM Simulations*) and the **Social Simulation Audit Card** for the robustness audits for LLM social simulation.

2 Related Work

LLM social simulation and the validation gap. Agent-based modeling (ABM) provides a bottom-up framework for studying how collective patterns emerge from individual agents and local interactions (Schelling, 1971; Macy & Willer, 2002), while longstanding work highlights challenges in empirical validation, replication, and sensitivity (Windrum et al., 2007; Ormerod & Rosewell, 2009; Wilensky & Rand, 2007; Bertolotti et al., 2020). LLM-based agents extend this paradigm by

generating decisions, communication, and adaptation through natural language (Park et al., 2023; Gao et al., 2024), and have been used to study cooperation and negotiation (Akata et al., 2025; Fontana et al., 2025; Abdelnabi et al., 2024), economic and population dynamics (Li et al., 2024; Mi et al., 2025; YANG et al., 2025), and open-ended social phenomena (Chuang et al., 2024; Gu et al., 2025; Piao et al., 2025; Wang et al., 2025a; Yang et al., 2024; Liu et al., 2025; Ren et al., 2024b). Existing validation efforts examine behavioral plausibility, human evaluation, empirical correspondence, real-world grounding, and theory-guided interaction mechanisms (Park et al., 2023; Puelma Touzel et al., 2026; Yang et al., 2024; Gu et al., 2025; Liu et al., 2025; Mou et al., 2024; Liu et al., 2024). These approaches assess whether simulations credibly represent their targets, while robustness asks whether their scientific conclusions remain stable across reasonable alternative implementations. Although individual studies include sensitivity checks, systematic analysis across representation, interaction, and design choices remains less developed. Following sensitivity analysis in ABMs (Saltelli et al., 2008; Windrum et al., 2007; Bertolotti et al., 2020) and robustness analysis in empirical research (Steegen et al., 2016; Simonsohn et al., 2020), we examine whether conclusions from LLM social simulations persist under controlled changes to simulation design and representation.

From local LLM sensitivity to multi-agent propagation. Small changes in LLM inputs can affect individual decisions and potentially propagate through multi-agent interaction. Prior work documents sensitivity to prompt formatting, option ordering, instruction structure, and other surface features (Salinas & Morstatter, 2024; Sclar et al., 2024; Pezeshkpour & Hruschka, 2024; Razavi et al., 2025), as well as persona construction (Li et al., 2025), strategic framing, and interaction-history presentation (Akata et al., 2025; Lorè & Heydari, 2024; Fontana et al., 2025). Classical ABM research similarly shows that collective outcomes can depend on initial conditions and modeling assumptions (Bertolotti et al., 2020; Windrum et al., 2007; Wilensky & Rand, 2007). In a multi-agent simulation, an LLM output becomes an action, message, memory, or social signal that can enter subsequent interactions. Prior work therefore establishes sensitivity at both the model and system levels, but leaves open how the magnitude of a controlled representation perturbation changes as LLM agents interact. We examine this gap by measuring amplification or attenuation under reciprocal LLM–LLM interaction and by tracing representation sensitivity across behavioral and collective outcomes in networked interaction.

Controlled testbeds for collective sensitivity. We study this problem in two established but complementary simulation settings. Repeated Prisoner's Dilemma is a common testbed for strategic behavior and cooperation in LLM agents (Akata et al., 2025; Fontana et al., 2025; Willis et al., 2025; Brookins & DeBacker, 2024; Li & Shirado, 2025). Its fixed action space, payoff structure, and interaction horizon provide a controlled setting for comparing the same representation perturbation across fixed-opponent and reciprocal LLM–LLM interaction. Echo chambers provide a complementary open-ended setting in which collective structure emerges from repeated local engagement on a social network. Prior LLM-agent simulations study the roles of network structure, homophily, recommendation and exposure, memory, and interaction dynamics in polarization and echo-chamber formation (Gu et al., 2025; Wang et al., 2025a; Ferraro et al., 2024; Zheng & Tang, 2024; Piao et al., 2025; Törnberg et al., 2023; Yang et al., 2024; Chuang et al., 2024). This setting allows a representation perturbation to be examined first in agents' interaction choices and then in the network formed by those choices. The experiments isolate selected perturbations while holding non-targeted components fixed, allowing sensitivity and its development through interaction to be measured directly. Appendix B provides the broader scope of LLM social simulation and the domain-specific motivation for the perturbations evaluated in each testbed.

3 METHOD

3.1 CONCEPTUAL FRAMEWORK

Let θ denote a social-simulation configuration and ξ a stochastic realization. A controlled perturbation p changes one component of θ while holding the remaining components fixed, yielding $\theta^{(p)}$. For an outcome Y, we define the signed perturbation effect and its sensitivity magnitude as

$$\Delta_Y(p) = \mathbb{E}_\xi \left[Y(\theta^{(p)}, \xi) - Y(\theta, \xi) \right], \qquad S_Y(p) = |\Delta_Y(p)|. \tag{1}$$

Sensitivity measures the change associated with a controlled perturbation.

A perturbation can change agent behavior, which changes the interaction context available to other agents. We use *butterfly effect* to describe the resulting interaction-mediated propagation of sensitivity:

$$p \longrightarrow \underbrace{\text{behavioral sensitivity}}_{\text{agent level}} \xrightarrow{\text{interaction}} \underbrace{\text{collective divergence}}_{\text{system level}} \tag{2}$$

When local and collective sensitivity are measured on a common outcome scale, interaction amplification corresponds to $S_{\text{collective}}(p) > S_{\text{local}}(p)$. The corresponding difference in effect magnitude is positive under amplification and negative under attenuation. When propagation spans outcomes measured at different levels, we characterize the stages separately. We distinguish *representation perturbations*, which vary how a fixed simulation specification is presented to the LLM, from *design perturbations*, which vary components of the simulated system itself; Section 5 generalizes this distinction in TRAILS.

3.2 Repeated Prisoner's Dilemma

Setup. We use a 10-round repeated Prisoner's Dilemma with actions Cooperate (C) and Defect (D) and payoffs $(C, C) = (3, 3)$, $(C, D) = (0, 5)$, $(D, C) = (5, 0)$, and $(D, D) = (1, 1)$. In the single-agent regime, one LLM plays against four fixed policies, $\mathcal{O} = \{\text{Tit-for-Tat}, \text{Random}, \text{Always-Cooperate}, \text{Always-Defect}\}$; in the two-agent regime, two matched LLM agents play each other. Each condition contains $N = 30$ runs.

Perturbations and measures. We vary three representation dimensions: persona format (Plain, Descriptive, Tabular), game framing (Canonical, Moralized, Risk), and memory representation (table or narrative history, with or without summary statistics). We measure cooperation rate as the primary outcome and average per-round payoff as a secondary outcome. Full configurations, prompt constructions, and metric definitions are provided in Appendix C.1.

Interaction amplification. For model m and representation contrast $p = (c_0, c_1)$, let

$$\Delta_{m,p,o} = \mathbb{E}(Y \mid c_1, m, o) - \mathbb{E}(Y \mid c_0, m, o)$$

denote the signed perturbation effect against fixed opponent $o \in \mathcal{O}$. We summarize sensitivity across the four benchmark contexts as

$$S^{\text{fixed}}_{m,p} = \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} |\Delta_{m,p,o}|.$$

Under LLM–LLM interaction,

$$S^{\text{LLM}}_{m,p} = \left|\mathbb{E}(Y \mid c_1, m, \text{LLM}) - \mathbb{E}(Y \mid c_0, m, \text{LLM})\right|.$$

The interaction effect gap is

$$A_{m,p} = S^{\text{LLM}}_{m,p} - S^{\text{fixed}}_{m,p}.$$

We refer to $A_{m,p} > 0$ as *amplification* and $A_{m,p} < 0$ as *attenuation*; $A_{m,p} = 0$ indicates equal sensitivity across interaction regimes.

3.3 Social-Media Echo Chamber

Setup. We simulate $N = 100$ agents for $T = 15$ rounds on a fixed social network. Each agent has a persona and fixed stance $s_i \in \{-2, -1, -, 1, 2\}$. Activated agents observe recent neighbor posts and choose among Post, Repost, Reply, and Silent; reposts and replies additionally select a visible target. Stances remain fixed, so collective outcomes capture the interaction structure that emerges conditional on the initial stance distribution.

Perturbations and measures. We vary persona representation (Bullet vs. Prose) and five design dimensions: input-network homophily, hub assignment, activation probability, memory window, and recommendation feed size. We measure agent-level interaction targeting and collective network structure using stance assortativity and weighted same-group edge ratio. Full configurations, implementation details, and metric definitions are provided in Appendix C.2.

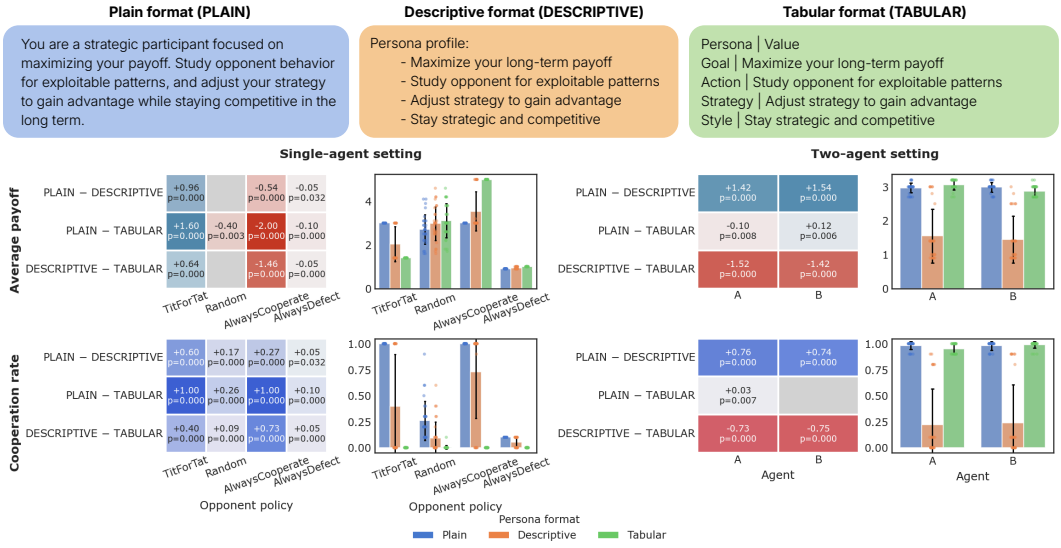


Figure 2: **Effect of persona format on Prisoner's Dilemma outcomes** (`gpt-5.2`). Left: single-agent (fixed-opponent); right: two-agent (LLM–LLM). Bars show means with SD error bars; heatmaps show pairwise differences, with color indicating Holm-corrected significance ($p < .05$) and gray non-significance.

Sensitivity across interaction levels. For perturbation $p = (c_0, c_1)$, let $q_t^{(c)}$ denote the cross-group targeting rate and $Y_t^{(c)}$ a collective network outcome under condition c, with $Y \in \{r, H\}$. We trace sensitivity across interaction levels using

$$\begin{aligned} \Delta_{\text{target},t}(p) &= q_t^{(c_1)} - q_t^{(c_0)}, \\ \Delta_{Y,t}(p) &= Y_t^{(c_1)} - Y_t^{(c_0)}, \qquad Y \in \{r, H\}, \end{aligned} \tag{3}$$

with corresponding sensitivity magnitudes $S_{\text{target},t}(p) = |\Delta_{\text{target},t}(p)|$ and $S_{Y,t}(p) = |\Delta_{Y,t}(p)|$.

We additionally compare action-type and target agreement across paired runs to locate behavioral sensitivity in action selection or interaction targeting. We trace propagation from agent-level behavioral sensitivity to collective network divergence. Formal metric definitions are provided in Appendix C.2.

Models and statistical analysis. We execute both experiments on four model families: `gpt-5.2`, `claude-haiku-4-5`, `gemini-2.5-flash`, and `deepseek`[1]. The simulation run is the unit of analysis, with $N = 30$ independent runs per condition unless otherwise specified. For paired comparisons, conditions share the same simulation seed. We report effect estimates with 95% confidence intervals and use two-sided tests with Holm correction for multiple pairwise comparisons. Full prompts and experimental configurations are provided in Appendices C and E.

4 Results

4.1 Sensitivity Varies Across Perturbations and Models

Representation choices can substantially shift simulated outcomes. Figure 2 shows a representative case in the repeated Prisoner's Dilemma. For `gpt-5.2`, changing only the representation of the matched strategic persona produces a large shift in the two-agent (LLM–LLM) setting: cooperation falls from 98.2% under Plain to 23.0% under Descriptive, a decrease of 75.2 percentage points (95% CI $[62.0, 88.3]$). Average payoff decreases by 1.48 points per round over the same contrast

[1]For the DeepSeek family, we use `deepseek-v3.2` in the Prisoner's Dilemma (PD) experiments and `deepseek-v4-flash` in the echo-chamber (EC) experiments.

Table 1: **Outcome sensitivity across perturbations and models.** Δ denotes the difference in outcome means between two perturbation conditions.

Perturbation	Metric	*GPT-5.2*	*Claude Haiku 4.5*	*Gemini 2.5 Flash*	*DeepSeek*
Repeated Prisoner's Dilemma (Two-Agent): $\max \lvert\Delta\rvert$					
Persona format^{R}	Average payoff	1.48***	1.49***	0.79***	0.14**
	Cooperation rate (pp)	75.2***	77.7***	36.0***	7.7**
Game framing^{R}	Average payoff	0.42***	0.09*	0.19	0.15***
	Cooperation rate (pp)	19.3***	3.3	9.0	8.8***
Memory representation^{R}	Average payoff	0.10***	0.05	0.29	0.01
	Cooperation rate (pp)	6.2***	2.7	13.8	1.0
Social-Media Echo Chamber: $\lvert\Delta\rvert$					
Persona representation^{R}	Stance assortativity	0.194***	0.002	0.035***	0.023
	Same-group edge ratio	0.051***	0.010*	0.023***	0.072***
Activation probability^{D}	Stance assortativity	0.008	0.041**	0.006	0.028
	Same-group edge ratio	0.006	0.030**	0.018*	0.029
Memory window^{D}	Stance assortativity	0.011	0.007	0.006	0.075*
	Same-group edge ratio	0.001	0.013	0.002	0.034*
Recommendation feed size^{D}	Stance assortativity	0.004	0.012	0.020	0.023
	Same-group edge ratio	0.005	0.015	0.009	0.024

Note: PD reports the maximum absolute effect across prespecified pairwise contrasts within each representation dimension; EC reports the absolute effect of the binary contrast. PD cooperation effects are in percentage points; other outcomes use their native units. Superscripts denote perturbation type: R = representation-level (TRAILS-R) and D = design-level (TRAILS-D). Shading denotes Holm-corrected significance: *** $p < .001$, ** $p < .01$, * $p < .05$; unshaded cells are not significant. For DeepSeek, PD experiments use `deepseek-v3.2`, whereas EC experiments use `deepseek-v4-flash`.

(Holm-corrected $p < .001$ for both outcomes). The same representation contrast produces different effects in the single-agent (fixed-opponent) setting, motivating the interaction-regime analysis in Section 4.2. Complete directional contrasts, confidence intervals, and effect sizes are reported in Appendix D.

Representation sensitivity varies across perturbations and models. Persona format produces significant two-agent cooperation shifts across all four model families, with the largest within-dimension contrast ranging from 77.7 percentage points for `claude-haiku-4-5` to 7.7 percentage points for `deepseek-v3.2` (Table 1). Game framing produces significant cooperation shifts for a subset of models, while memory representation yields a significant maximum cooperation contrast only for `gpt-5.2`. The observed sensitivity therefore depends on both the representation dimension and the model family.

Sensitivity is similarly selective in networked interaction. In the echo-chamber simulation, persona representation changes stance assortativity by 0.194 for `gpt-5.2` and 0.035 for `gemini-2.5-flash`, compared with 0.002 for `claude-haiku-4-5` (Table 1). Design perturbations are less consistent: activation probability and memory window affect selected model–outcome combinations, while recommendation feed size produces no significant differences after Holm correction. Sensitivity in the networked simulation therefore varies across representation and design choices, models, and collective outcomes.

Across both regimes, configuration-level sensitivity varies across perturbations and models. We next examine how multi-agent interaction reshapes representation sensitivity, through amplification or attenuation in the Prisoner's Dilemma and across behavioral and collective levels in the echo-chamber simulation.

4.2 Multi-Agent Interaction Reshapes Representation Sensitivity

The preceding results establish configuration-level sensitivity across representation and design perturbations. We focus the interaction analysis on persona representation, which produces clear sensitivity in both simulation regimes and provides a common perturbation for examining how local differences develop through interaction. The repeated Prisoner's Dilemma permits direct comparison under fixed-opponent and LLM–LLM interaction, while the echo-chamber simulation traces sensitivity across interaction targeting and collective network structure.

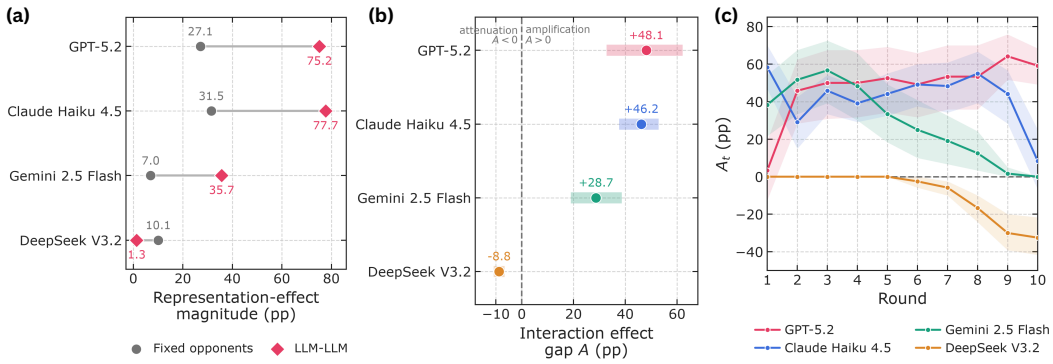


Figure 3: **Interaction reshapes persona-representation sensitivity in the repeated Prisoner's Dilemma.** PLAIN vs. DESCRIPTIVE, $N = 30$ paired seeds. (a) Cooperation-rate sensitivity under fixed-opponent and LLM–LLM interaction. (b) Interaction effect gap $A = S^{\mathrm{LLM}} - S^{\mathrm{fixed}}$; positive values indicate amplification and negative values attenuation. (c) Round-wise effect gap A_t. Intervals and bands show paired-bootstrap 95% CIs.

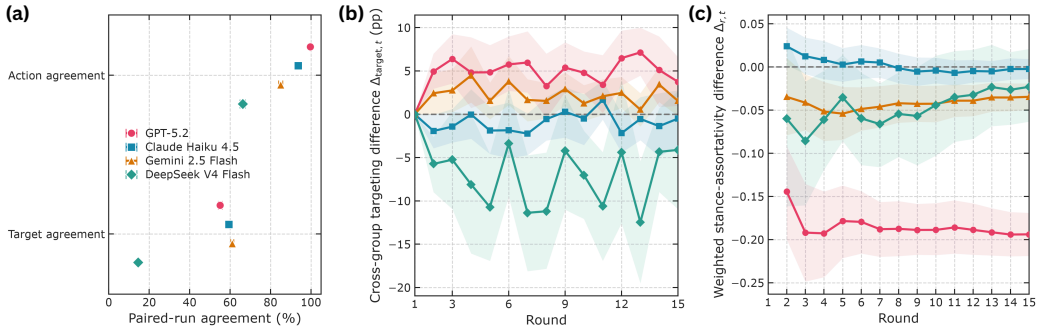


Figure 4: **Persona-representation sensitivity across interaction levels in the echo-chamber simulation.** PROSE vs. BULLET, $N = 30$ paired seeds. (a) Agreement in action type and interaction target. (b) Cross-group targeting difference $\Delta_{\mathrm{target},t}$; positive values indicate more cross-group targeting under PROSE. (c) Weighted stance-assortativity difference $\Delta_{r,t}$; negative values indicate lower assortativity under PROSE. Error bars and bands show 95% CIs.

Reciprocal interaction can amplify or attenuate representation sensitivity. For the PLAIN→DESCRIPTIVE persona contrast, `gpt-5.2` sensitivity increases from $S^{\mathrm{fixed}} = 0.271$ against fixed opponents to $S^{\mathrm{LLM}} = 0.752$ under LLM–LLM interaction, yielding $A = 0.481$ (95% CI $[0.328, 0.622]$; Figure 3). The interaction effect is also positive for `claude-haiku-4-5` ($A = 0.462$) and `gemini-2.5-flash` ($A = 0.287$), while `deepseek-v3.2` shows attenuation ($A = -0.088$). Round-wise estimates show different temporal trajectories across models. Reciprocal interaction therefore changes the magnitude and development of representation sensitivity, with the direction of modulation depending on the model. Complete estimates and confidence intervals are reported in Appendix D.

Representation sensitivity can extend from targeting behavior to collective structure. In the echo-chamber simulation, persona representation changes interaction targets even when action selection remains comparatively stable (Figure 4). For `gpt-5.2`, action agreement is 99.7%, while target agreement is 55.1%. PROSE increases cross-group targeting by 5.15 percentage points (95% CI $[4.22, 6.07]$) and is associated with a 0.194 lower cumulative stance assortativity by round 15. The pattern varies across models: `gemini-2.5-flash` shows increased cross-group targeting (+2.31 pp) and lower assortativity (−0.035), while the final assortativity differences for `claude-haiku-4-5` and `deepseek-v4-flash` are near zero. Representation sensitivity can therefore appear at both behavioral and collective levels, with the relationship between the two varying across models. Complete estimates and confidence intervals are reported in Appendix D.

Table 2: **TRAILS: Taxonomy for Robustness Audits in LLM Simulations.** TRAILS distinguishes representation-level perturbations (TRAILS-R), which vary how a fixed simulation specification or context is presented to LLM agents, from design-level perturbations (TRAILS-D), which vary components of the simulated system itself. **Bolded examples** indicate perturbations examined in our case studies.

Level	Dimension	Description	Examples
TRAILS-R: Representation-level perturbations			
	Representational format	Alternative encodings of a matched agent or task specification, including prose, lists, tables, and structured formats; **persona format**	Salinas & Morstatter (2024); Sclar et al. (2024).
	Instruction hierarchy	Instruction order, placement, grouping, and authority across prompt components	Salinas & Morstatter (2024); Sclar et al. (2024).
	Linguistic framing	Alternative wording or labels while other simulation components are held fixed; **game-instruction framing**	Salinas & Morstatter (2024); Sclar et al. (2024).
	Context representation	Alternative representations or encodings of the same interaction history and contextual information; **memory representation**	Park et al. (2023); Salinas & Morstatter (2024).
	Interaction sequencing	Presentation or ordering of otherwise matched interactions	Mou et al. (2024).
TRAILS-D: Design-level perturbations			
Micro	Model substrate	Model family, model variant, inference settings, and decoding parameters	Zhou et al. (2024b).
	Agent specification	Persona attributes, goals, beliefs, demographic or behavioral profiles, and population grounding	Park et al. (2023); Li et al. (2025).
	Internal state & cognition	Belief and preference updating, reasoning procedures, reflection rules, and internal-state representations	Liu et al. (2024); Wang et al. (2023).
	Memory & temporality	Memory length, retrieval, summarization, forgetting, and temporal granularity	Park et al. (2023); Wang et al. (2025c).
Meso	Interaction protocol	Turn structure, available actions, visibility, activation, and rules governing who can interact with whom; **activation probability**	Mou et al. (2024); Yang et al. (2024).
	Intervention design	Intervention type, target, timing, frequency, delivery mechanism, and framing	Liu et al. (2025; 2024).
Macro	Environment structure	Network topology, exposure and recommendation rules, institutional structure, and platform affordances; **network homophily, hub assignment, recommendation feed size**	Yang et al. (2024); Gu et al. (2025); Piao et al. (2026).
	Population & scale	Population size, composition, heterogeneity, agent assignment, and real- versus synthetic-population construction	Yang et al. (2024); Wang et al. (2025b); Piao et al. (2026).

The two regimes capture complementary forms of interaction-mediated sensitivity. In repeated dyadic interaction, a perturbation effect on a common outcome can be amplified or attenuated. In networked interaction, representation sensitivity can appear across behavioral and collective levels.

5 From Architecture Sensitivity to Robustness Audits

TRAILS. LLM social simulations can be sensitive to both design-level perturbations, which change the model, network, or interaction protocol, and representation-level perturbations, which change how a fixed setup is presented through instructions, personas, labels, or memory formats. Our experiments found sensitivity to both, with effects varying across perturbations and models. We introduce **TRAILS** (*Taxonomy for Robustness Audits In LLM Simulations*) to organize these choices into representation (TRAILS-R) and design (TRAILS-D) audit dimensions (Table 2), given the levels identified in Appendix B. **TRAILS-D** organizes changes to the substantive simulation setup into eight dimensions across micro, meso, and macro levels. **TRAILS-R** organizes changes to how a matched simulation specification or context is presented to LLM agents into five dimensions. These choices can affect how agents interpret their inputs. Appendix A provides the examples of TRAILS-R and TRAILS-D. TRAILS covers the simulation system; measurement and evaluation sensitivity falls within the downstream evaluation pipeline.

The Social Simulation Audit Card. We introduce the Social Simulation Audit Card to operationalize TRAILS for a stated scientific claim. The card guides researchers through five steps: stating the claim, mapping perturbations to TRAILS and context relevance, registering perturbations and preservation criteria, reporting sensitivity, and defining the claim's scope. Context relevance is claim-specific: context-relevant perturbations represent substantively meaningful conditions of the

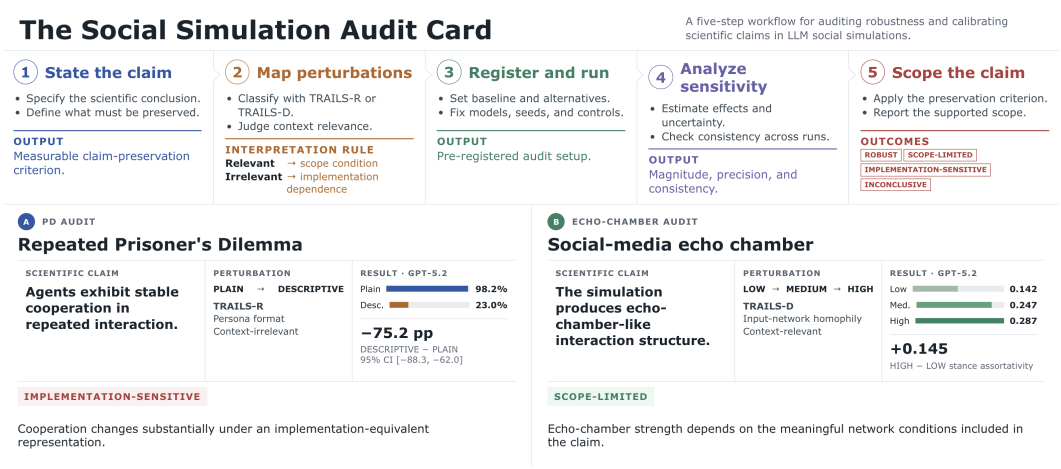


Figure 5: **The Social Simulation Audit Card.** The five-step workflow connects perturbation selection and sensitivity analysis to the scope of scientific claims. GPT-5.2 examples illustrate two interpretations of sensitivity: a context-irrelevant persona-format change reveals implementation dependence in the repeated Prisoner's Dilemma, whereas context-relevant network homophily bounds the conditions under which echo-chamber-like interaction structure is supported.

modeled phenomenon, whereas context-irrelevant perturbations probe implementation dependence. For example, network homophily and hub assignment are context-relevant structural conditions in the echo-chamber simulation; their sensitivity results are reported in Appendix D. Accordingly, context-relevant sensitivity helps scope where a claim holds, while substantial sensitivity to context-irrelevant choices indicates implementation dependence. Outcomes are classified as *Robust*, *Scope-limited*, *Implementation-sensitive*, or *Inconclusive*, bounded by the audited conditions. Appendix F provides the full procedure and case-study audits.

6 Conclusions

LLM social simulations can be sensitive to choices that are often treated as interchangeable or left unaudited. Across repeated Prisoner's Dilemma and echo-chamber simulations, we find that sensitivity varies across perturbations, models, and outcomes. Persona representation produces substantial shifts in several settings, while other representation and design perturbations have smaller or more selective effects. Multi-agent interaction further reshapes this sensitivity. In repeated dyadic interaction, representation effects can be amplified or attenuated under reciprocal LLM–LLM interaction; in networked interaction, representation sensitivity appears in interaction targeting and, for some models, in collective network structure. These results indicate that robustness observed under one configuration or model does not establish robustness of the resulting scientific claim.

Our findings are limited to two simulation regimes, selected perturbations, four model families, and fixed interaction horizons, so the observed sensitivity patterns should not be assumed to generalize to other settings. The echo-chamber experiment holds agent stances fixed and therefore characterizes interaction structure rather than endogenous opinion change. Its behavioral and collective analyses identify cross-level patterns of sensitivity but do not establish a causal pathway from targeting differences to network divergence. More broadly, robustness to implementation choices addresses only one dimension of simulation validity; correspondence between simulated agents and human behavior requires separate empirical validation.

The broader implication concerns the scope of scientific inference from LLM social simulations. Robustness audits should identify which conclusions persist across plausible implementations, which depend on substantively meaningful conditions, and which remain sensitive to implementation choices. TRAILS and the Social Simulation Audit Card provide a vocabulary and reporting structure for conducting such audits, with the required coverage depending on the claim and simulation context. Scientific claims from simulated societies can then be scoped to the configurations, models, and interaction conditions for which robustness evidence is available.

AI USE STATEMENT

We used generative AI tools, including ChatGPT, Claude, and GitHub Copilot, to assist with grammar checking, proofreading, rewriting and polishing, organization and outlining, and the preparation of LaTeX content, tables, and figure captions. We also used these tools to assist in discussing and interpreting aspects of the experimental design; to support the development of plotting and analysis scripts; to check calculations and assist with the preparation of tables and figures; to summarize literature; and to aid comparisons of behavioral patterns across models.

All AI-assisted work was reviewed and verified by two authors. AI-assisted code was run and tested by the authors, statistical results were independently checked, and tables and figures were verified against the underlying source data. Citations and literature-related claims were manually checked against the original sources, and AI-assisted prose was manually reviewed and revised. Research claims were checked against the relevant literature, and the final manuscript was reviewed by multiple authors.

Generative AI was not used to generate or alter experimental data, to make final research conclusions, to substitute for human statistical validation, or to introduce unverified AI-generated citations. The authors made the final decisions regarding the interpretation and presentation of the results and take responsibility for the complete content of this work, including any text, code, analyses, or other artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work studies synthetic LLM-agent simulations and does not involve human participants or private user data. The simulated behaviors and collective outcomes should not be interpreted as direct evidence of human behavior or as predictions of real-world social outcomes. In particular, conclusions from LLM-based social simulations may depend on model and simulation design choices, as demonstrated in this work. We therefore emphasize that applications to real-world social, policy, or other high-stakes settings require appropriate empirical validation and robustness assessment.

REPRODUCIBILITY STATEMENT

We provide an anonymous repository containing the simulation code, synthetic personas, fixed network inputs, and aggregate statistical results at `https://anonymous.4open.science/r/butterfly-effect-sim-4934`. Section 3 describes the experimental design, outcome measures, and statistical analysis; Appendix C documents the configurations, perturbations, metrics, and inference procedures; Appendix D reports extended results; and Appendix E provides the complete prompt templates and output schemas. The repository README includes environment setup and execution instructions. Because hosted LLM APIs and model aliases may change over time, exact reruns may depend on provider availability and model versions; we therefore report the evaluated models, decoding settings, simulation seeds, and parsing procedures needed to characterize these dependencies.

REFERENCES

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. Cooperation, competition, and maliciousness: Llm-stakeholders interactive negotiation. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *Nature Human Behaviour*, 9(7):1380–1390, 2025. doi: 10.1038/s41562-025-02172-y. URL `https://doi.org/10.1038/s41562-025-02172-y`.

Altera.AL, Andrew Ahn, Nic Becker, Stephanie Carroll, Nico Christie, Manuel Cortes, Arda Demirci, Melissa Du, Frankie Li, Shuying Luo, Peter Y. Wang, Mathew Willows, Feitong Yang, and Guangyu Robert Yang. Project sid: Many-agent simulations toward ai civilization, 2024. URL `https://arxiv.org/abs/2411.00114`.

Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C Kozlowski, Bernard Koch, Erik Brynjolfsson, James Evans, and Michael S Bernstein. Position: Llm social simulations are a promising research method. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=cRBg1dtj7o.

Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. Emergent social conventions and collective bias in llm populations. *Science Advances*, 11(20):eadu9368, 2025. URL https://www.science.org/doi/abs/10.1126/sciadv.adu9368.

Francesco Bertolotti, Angela Locoro, and Luca Mari. Sensitivity to initial conditions in agent-based models. In *Multi-Agent Systems and Agreement Technologies*, volume 12520 of *Lecture Notes in Computer Science*, pp. 501–508. Springer, 2020. doi: 10.1007/978-3-030-66412-1_32. URL https://doi.org/10.1007/978-3-030-66412-1_32.

Philip Brookins and Jason DeBacker. Playing games with gpt: What can we learn about a large language model from canonical strategic games? *Economics Bulletin*, 44(1):25–37, 2024. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4493398.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate, 2023. URL https://arxiv.org/abs/2308.07201.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors, 2023. URL https://arxiv.org/abs/2308.10848.

Ayush Chopra, Shashank Kumar, Nurullah Giray Kuru, Ramesh Raskar, and Arnau Quera-Bofarull. On the limits of agency in agent-based models. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pp. 500–509, 2025. URL https://dl.acm.org/doi/10.5555/3709347.3743565.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert D Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. Simulating opinion dynamics with networks of llm-based agents. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3326–3346, 2024. URL https://aclanthology.org/2024.findings-naacl.211/.

Alessandro Di Stefano, Chrisina Jayne, Claudio Angione, and The Anh Han. Recognition of behavioural intention in repeated games using machine learning. In *Artificial Life Conference Proceedings 35*, volume 2023, pp. 103. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA, 2023. doi: 10.1162/isal_a_00637. URL https://doi.org/10.1162/isal_a_00637.

Emilio Ferrara. The butterfly effect in artificial intelligence systems: Implications for AI bias and fairness. *Machine Learning with Applications*, 15:100525, 2024. URL https://doi.org/10.1016/j.mlwa.2024.100525.

Antonino Ferraro, Antonio Galli, Valerio La Gatta, Marco Postiglione, Gian Marco Orlando, Diego Russo, Giuseppe Riccio, Antonio Romano, and Vincenzo Moscato. Agent-based modelling meets generative ai in social network simulations. In *International Conference on Advances in Social Networks Analysis and Mining*, pp. 155–170, 2024. URL https://link.springer.com/chapter/10.1007/978-3-031-78541-2_10.

Nicoló Fontana, Francesco Pierri, and Luca Maria Aiello. Nicer than humans: How do large language models behave in the prisoner's dilemma? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pp. 522–535, 2025. doi: 10.1609/icwsm.v19i1.35829. URL https://doi.org/10.1609/icwsm.v19i1.35829.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1259, 2024. doi: 10.1057/s41599-024-03611-3. URL https://doi.org/10.1057/s41599-024-03611-3.

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

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, pp. 65–70, 1979. URL https://www.jstor.org/stable/4615733.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VtmBAGCN7o.

John J Horton, Apostolos Filippas, and Benjamin S Manning. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023. URL https://dl.acm.org/doi/10.1145/3670865.3673513.

Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. War and peace (waragent): Large language model-based multi-agent simulation of world wars, 2023. URL https://arxiv.org/abs/2311.17227.

Yanru Jiang and Gülşah Akçakır. Explicit cooperation shapes human-like multi-agent llm negotiation. In *Proceedings of the 1st ICWSM Workshop on Integrating NLP and Psychology to Study Social Interactions*, 2025. doi: 10.36190/2025.34. URL https://workshop-proceedings.icwsm.org/abstract.php?id=2025_34.

Maik Larooij and Petter Törnberg. Validation is the central challenge for generative social simulation: A critical review of llms in agent-based modeling. *Artificial Intelligence Review*, 59(1):15, 2026. doi: 10.1007/s10462-025-11412-6. URL https://doi.org/10.1007/s10462-025-11412-6.

Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. LLM generated persona is a promise with a catch. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*, 2025. URL https://openreview.net/forum?id=qh9eGtMG4H.

Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. Econagent: Large language model-empowered agents for simulating macroeconomic activities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 15523–15536. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.acl-long.829/.

Yuxuan Li and Hirokazu Shirado. Spontaneous giving and calculated greed in language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5271–5286, 2025. URL https://aclanthology.org/2025.emnlp-main.267/.

Genglin Liu, Vivian Le, Salman Rahman, Elisa Kreiss, Marzyeh Ghassemi, and Saadia Gabriel. Mosaic: Modeling social ai for content dissemination and regulation in multi-agent simulations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6390–6417. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.325. URL https://aclanthology.org/2025.emnlp-main.325/.

Yuhan Liu, Xiuying Chen, Xiaoqing Zhang, Xing Gao, Ji Zhang, and Rui Yan. From skepticism to acceptance: Simulating the attitude dynamics toward fake news. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 7886–7894. International Joint Conferences on Artificial Intelligence Organization, 2024. doi: 10.24963/ijcai.2024/873. URL https://doi.org/10.24963/ijcai.2024/873.

Nunzio Lorè and Babak Heydari. Strategic behavior of large language models and the role of game structure versus contextual framing. *Scientific Reports*, 14(1):18490, 2024. URL https://www.nature.com/articles/s41598-024-69032-z.

Michael W. Macy and Robert Willer. From factors to actors: Computational sociology and agent-based modeling. *Annual Review of Sociology*, 28(1):143–166, 2002. doi: 10.1146/annurev.soc.28.110601.141117. URL https://www.annualreviews.org/content/journals/10.1146/annurev.soc.28.110601.141117.

Zhao Mandi, Shreeya Jain, and Shuran Song. Roco: Dialectic multi-robot collaboration with large language models, 2023. URL https://arxiv.org/abs/2307.04738.

Qirui Mi, Mengyue Yang, Xiangning Yu, Zhiyu Zhao, Cheng Deng, Bo An, Haifeng Zhang, Xu Chen, and Jun Wang. MF-LLM: Simulating population decision dynamics via a mean-field large language model framework. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=m0q8NfGWnv.

Xinyi Mou, Zhongyu Wei, and Xuan-Jing Huang. Unveiling the truth and facilitating change: Towards agent-based large-scale social movement simulation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4789–4809, 2024. URL https://aclanthology.org/2024.findings-acl.285/.

Gian Marco Orlando, Jinyi Ye, Valerio La Gatta, Mahdi Saeedi, Vincenzo Moscato, Emilio Ferrara, and Luca Luceri. Emergent coordinated behaviors in networked llm agents: Modeling the strategic dynamics of information operations. In *Proceedings of the ACM Web Conference 2026*, pp. 4805–4816, 2026.

Paul Ormerod and Bridget Rosewell. Validation and verification of agent-based models in the social sciences. In Flaminio Squazzoni (ed.), *Epistemological Aspects of Computer Simulation in the Social Sciences*, volume 5466 of *Lecture Notes in Computer Science*, pp. 130–140. Springer, Berlin, Heidelberg, 2009. doi: 10.1007/978-3-642-01109-2_10. URL https://doi.org/10.1007/978-3-642-01109-2_10.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23. Association for Computing Machinery, 2023. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.130. URL https://aclanthology.org/2024.findings-naacl.130/.

PHF Science. The future is now: Revolutionising decision-making with ai-driven simulations. https://www.phfscience.nz/news-publications/the-future-is-now-revolutionising-decision-making-with-ai-driven-simulations/, December 2024. Accessed: 2026-05-02.

Jinghua Piao, Zhihong Lu, Chen Gao, Fengli Xu, Qinghua Hu, Fernando P Santos, Yong Li, and James Evans. Emergence of human-like polarization among large language model agents. *arXiv preprint arXiv:2501.05171*, 2025. URL https://arxiv.org/abs/2501.05171.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *iFuture*, 2026. doi: 10.26599/IF.2026.9710004. URL https://doi.org/10.26599/IF.2026.9710004.

Maximilian Puelma Touzel, Sneheel Sarangi, Gayatri Krishnakumar, Busra Tugce Gurbuz, Austin Welch, Zachary Yang, Andreea Musulan, Hao Yu, Ethan Kosak-Hine, Tom Gibbs, Camille Thibault, Reihaneh Rabbany, Jean-François Godbout, Dan Zhao, and Kellin Pelrine. Sandboxsocial: A sandbox for social media using multimodal ai agents. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 11100–11103. International Joint Conferences on Artificial Intelligence Organization, 2025. doi: 10.24963/ijcai.2025/1271. URL https://www.ijcai.org/proceedings/2025/1271.

Maximilian Puelma Touzel, Sneheel Sarangi, Aurélien Bück-Kaeffer, Zachary Yang, Jean-François Godbout, and Reihaneh Rabbany. Position: Time to close the validation gap in llm social simulations. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15174–15186. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.810. URL https://aclanthology.org/2024.acl-long.810/.

Amirhossein Razavi, Mina Soltangheis, Negar Arabzadeh, Sara Salamat, Morteza Zihayat, and Ebrahim Bagheri. Benchmarking prompt sensitivity in large language models. In *European Conference on Information Retrieval*, pp. 303–313. Springer, 2025. URL https://link.springer.com/chapter/10.1007/978-3-031-88714-7_29.

Jiawei Ren, Yan Zhuang, Xiaokang Ye, Lingjun Mao, Xuhong He, Jianzhi Shen, Mrinaal Dogra, Yiming Liang, Ruixuan Zhang, Tianai Yue, Yiqing Yang, Eric Liu, Ryan Wu, Kevin Benavente, Rajiv Mandya Nagaraju, Muhammad Faayez, Xiyan Zhang, Dhruv Vivek Sharma, Xianrui Zhong, Ziqiao Ma, Tianmin Shu, Zhiting Hu, and Lianhui Qin. Simworld: An open-ended realistic simulator for autonomous agents in physical and social worlds. *arXiv preprint arXiv:2512.01078*, 2025. URL https://arxiv.org/abs/2512.01078.

Ruiyang Ren, Peng Qiu, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Hua Wu, Ji-Rong Wen, and Haifeng Wang. Bases: Large-scale web search user simulation with large language model based agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 902–917, 2024a. URL https://aclanthology.org/2024.findings-emnlp.50/.

Siyue Ren, Zhiyao Cui, Ruiqi Song, Zhen Wang, and Shuyue Hu. Emergence of social norms in generative agent societies: Principles and architecture. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 7895–7903. International Joint Conferences on Artificial Intelligence Organization, 8 2024b. doi: 10.24963/ijcai.2024/874. URL https://doi.org/10.24963/ijcai.2024/874. Human-Centred AI.

Abel Salinas and Fred Morstatter. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4629–4651, 2024. URL https://aclanthology.org/2024.findings-acl.275/.

Andrea Saltelli, Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. *Global Sensitivity Analysis: The Primer*. John Wiley & Sons, Chichester, UK, 2008. doi: 10.1002/9780470725184. URL https://doi.org/10.1002/9780470725184.

Robert G. Sargent. Verification and validation of simulation models. In *Proceedings of the 2010 Winter Simulation Conference*, pp. 166–183. IEEE, 2010. doi: 10.1109/WSC.2010.5679166. URL https://doi.org/10.1109/WSC.2010.5679166.

Thomas C. Schelling. Dynamic models of segregation. *Journal of Mathematical Sociology*, 1(2): 143–186, 1971. doi: 10.1080/0022250X.1971.9989794. URL https://www.tandfonline.com/doi/abs/10.1080/0022250X.1971.9989794.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/6c0e99d736da621403018ca7b32b1a4d-Paper-Conference.pdf.

Uri Simonsohn, Joseph P Simmons, and Leif D Nelson. Specification curve analysis. *Nature Human Behaviour*, 4(11):1208–1214, 2020. URL https://www.nature.com/articles/s41562-020-0912-z.

Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016. URL https://pubmed.ncbi.nlm.nih.gov/27694465/.

Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. Simulating social media using large language models to evaluate alternative news feed algorithms. *arXiv preprint arXiv:2310.05984*, 2023. URL https://arxiv.org/abs/2310.05984.

Chenxi Wang, Zongfang Liu, Dequan Yang, and Xiuying Chen. Decoding echo chambers: Llm-powered simulations revealing polarization in social networks. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3913–3923, 2025a. URL https://aclanthology.org/2025.coling-main.264/.

Lei Wang, Heyang Gao, Xiaohe Bo, Xu Chen, and Ji-Rong Wen. Yulan-onesim: Towards the next generation of social simulator with large language models, 2025b. URL https://arxiv.org/abs/2505.07581.

Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. User behavior simulation with large language model-based agents. *ACM Transactions on Information Systems*, 43(2):1–37, 2025c. URL https://dl.acm.org/doi/full/10.1145/3708985.

Zhilin Wang, Yu Ying Chiu, and Yu Cheung Chiu. Humanoid agents: Platform for simulating human-like generative agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 167–176. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-demo.15. URL https://aclanthology.org/2023.emnlp-demo.15/.

Uri Wilensky and William Rand. Making models match: Replicating an agent-based model. *Journal of Artificial Societies and Social Simulation*, 10(4):2, 2007. URL https://www.jasss.org/10/4/2.html.

Richard Willis, Yali Du, and Joel Z Leibo. Will systems of llm agents lead to cooperation: An investigation into a social dilemma. In *24th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2025*, pp. 2786–2788. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 2025. URL https://dl.acm.org/doi/10.5555/3709347.3744012.

Paul Windrum, Giorgio Fagiolo, and Alessio Moneta. Empirical validation of agent-based models: Alternatives and prospects. *Journal of Artificial Societies and Social Simulation*, 10(2):8, 2007. URL https://ideas.repec.org/a/jas/jasssj/2006-40-2.html.

Yuzhe YANG, Yifei Zhang, Minghao Wu, Kaidi Zhang, Yunmiao Zhang, Honghai Yu, Yan Hu, and Benyou Wang. Twinmarket: A scalable behavioral and social simulation for financial markets. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=h60y6zlPyl.

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*, 2024. URL https://arxiv.org/abs/2411.11581.

Wenzhen Zheng and Xijin Tang. Simulating social network with llm agents: an analysis of information propagation and echo chambers. In *International Symposium on Knowledge and Systems Sciences*, pp. 63–77. Springer, 2024. URL https://link.springer.com/chapter/10.1007/978-981-96-0178-3_5.

Jiaxu Zhou, Jen-tse Huang, Xuhui Zhou, Man Ho Lam, Xintao Wang, Hao Zhu, Wenxuan Wang, and Maarten Sap. The pimmur principles: Ensuring validity in collective behavior of llm societies. *arXiv preprint arXiv:2509.18052*, 2025. URL https://arxiv.org/abs/2509.18052.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. Webarena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, volume 2024, pp. 15585–15606, 2024a. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4410c0711e9154a7a2d26f9b3816d1ef-Abstract-Conference.html.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=mM7VurbA4r.

A TRAILS (TAXONOMY FOR ROBUSTNESS AUDITS IN LLM SIMULATIONS)

Table 3: TRAILS-D: Design-level perturbations

Level	Dimension	What to perturb
Micro	Model substrate	Model family, model size, base vs. instruction-tuned models, alignment, temperature, top-p, random seed, safeguards.
	Agent specification	Demographics, ideology, personality, goals, prior beliefs, real-data-grounded vs. synthetic personas, relationship attributes.
	Internal state and cognition	Beliefs, attitudes, emotions, needs, moral values, reasoning style, reflection rules, belief updating, preference updating.
	Memory and temporality	last-k memory, summarized memory, episodic memory, retrieval rules, forgetting, reflection frequency, temporal granularity.
Meso	Interaction protocol	Turn order, fixed or free turns, dyadic or group interaction, available actions
	Intervention design	Moderator type, target selection, timing, frequency, rule-based vs. LLM-based intervention.
Macro	Environment structure	Network topology, feed ranking, recommendation system, moderation rules, platform affordances, institutional rules.
	Population and scale	Number of agents, demographic distribution, ideological balance, heterogeneity, real vs. synthetic population construction.

Table 4: TRAILS-R: Representation-level perturbations

Category	Perturbation	Example
Representational format	Formatting	Narrative prose vs. bullet points; paragraph vs. table; bolded keywords; numbered lists.
	Delimiter choice	Conversation history enclosed in code blocks, XML tags, JSON, or plain text.
	Output schema	Free-form paragraph vs. short reply vs. JSON with rationale.
Instruction hierarchy	Instruction order	"Update your opinion, then write a reply" vs. "Write a reply, then update your opinion."
	Role/message placement	The same instruction placed in the system message vs. the user message.
Linguistic framing	Agent naming	Agent A/B vs. Alice/Bob.
	Persona wording	"Left" vs. "Democrat"; "Right" vs. "Republican."
	Moderation framing	Same intervention framed as a warning, suggestion, community reminder, or bridge statement.
	Few-shot examples	Same examples framed differently.
Context representation	Memory representation	Same history shown as transcript, JSON list, bullet summary.
	Context window	Full history vs. last three turns; same content with early background removed.
	Message metadata	Adding timestamps, usernames.
Interaction sequencing	Initial seed turn	Same stance, but a slightly different opening statement.
	Turn order	Which agent speaks first.

B Extended Related Works

B.1 LLM-based social simulations

Scope. We focus on LLM-based social simulations of collective human behavior, building on *Silicon Societies* (Puelma Touzel et al., 2026) and eligibility criteria for generative social simulation (Anthis et al., 2025; Larooij & Törnberg, 2026). This scope connects **micro-level** human-like agents whose reasoning, memory, attitudes, decisions, and communication are generated by LLMs; **meso-level** interactions within shared environments and institutional rules; and **macro-level** phenomena such as polarization, echo chambers, norm emergence, and information diffusion. Here, *social* denotes interdependent agent interactions, and *collective* denotes the aggregate patterns they produce. We exclude single-agent or user simulations without inter-agent social interaction (Ren et al., 2024a); simulations primarily modeling non-human, embodied, or aggregate actors, such as robots, states, countries, or game agents (Altera.AL et al., 2024; Hua et al., 2023; Mandi et al., 2023; Ren et al., 2025); and multi-agent LLM systems primarily designed for task completion or evaluation (Chan et al., 2023; Chen et al., 2023; Hong et al., 2024; Qian et al., 2024; Zhou et al., 2024a).

Social simulation scenarios. Within this scope, we distinguish two broad scenarios. In **goal- or incentive-structured scenarios**, agents act in relatively well-specified decision environments with explicit or implicit incentives, constraints, rules, or payoffs. Collective outcomes arise through strategic aggregation, as in negotiation and cooperation (Abdelnabi et al., 2024; Akata et al., 2025; Jiang & Akçakır, 2025), population decision-making (Mi et al., 2025), financial market dynamics (YANG et al., 2025), or macroeconomic activity (Li et al., 2024). In **open-ended scenarios**, agents do not share a single explicit task, objective, or payoff structure. Macro-level patterns emerge from micro-level agent behavior and meso-level interaction, as in social media discussion (Yang et al., 2024; Puelma Touzel et al., 2025), opinion dynamics and mobilization (Mou et al., 2024), polarization and echo chambers (Gu et al., 2025; Piao et al., 2025; Wang et al., 2025a; Yang et al., 2024), information or misinformation spread (Liu et al., 2024; 2025; Park et al., 2023), conformity and herd effects (Yang et al., 2024; Wang et al., 2025c), emergent coordination (Park et al., 2023; Orlando et al., 2026), social norm formation (Ren et al., 2024b), and user behavior in digital platforms (Wang et al., 2025c). The key distinction between the two types is whether collective outcomes arise primarily through *strategic aggregation* in structured decision settings or through *social emergence* from less structured social environments.

Factors shaping audit strength. The required audit strength depends on the claim's ambition, the simulation's openness, and the stakes of its intended use. An **exploratory** probe calls for focused checks of prompt, seed, and model dependence; a **mechanism** claim requires tests of plausible alternative specifications; and an **intervention or policy** claim requires broader coverage and tighter uncertainty bounds. Additionally, simulation openness shapes the range of choices an audit must consider. In **goal-structured simulations**, agents operate under explicit incentives, constraints, rules, or payoffs (Akata et al., 2025; Fontana et al., 2025; Di Stefano et al., 2023). In **theory-guided open-ended simulations**, social theories structure interactions without a shared task or payoff (Gu et al., 2025; Wang et al., 2025a; Piao et al., 2025). In **emergence-driven open-ended simulations**, collective patterns arise from less constrained communication, memory, platform feedback, and network dynamics (Park et al., 2023). Greater openness expands the plausible perturbation space and increases the demands on audit coverage. In addition, domain stakes determine how much uncertainty is acceptable for the intended use. **Lower-stakes** uses include exploratory demonstrations and low-risk hypothesis generation. **Higher-stakes** uses inform decisions in public health, policy, platform governance, or financial markets, where errors can have substantial consequences. These uses require stronger robustness evidence, broader coverage, and greater precision before simulated findings can support real-world decisions (Saltelli et al., 2008; Sargent, 2010).

B.2 Repeated Prisoner's Dilemma as a Goal-Structured Testbed

The repeated Prisoner's Dilemma is a foundational paradigm for studying the tension between short-term self-interest and long-term mutual benefit, and a recurring testbed for LLM-agent cooperation (Akata et al., 2025; Fontana et al., 2025; Willis et al., 2025; Brookins & DeBacker, 2024; Li & Shirado, 2025). Recent work has used this paradigm to make claims about whether LLMs are "cooperative" or "selfish," whether they reason from history, and how their behavior compares to human

play (Akata et al., 2025; Horton et al., 2023). We use it here as a tightly controlled environment: the action space, payoffs, and horizon are fully specified, leaving architectural and prompt-level choices as the main sources of variation.

We focus on three small perturbations that are common in LLM-agent implementations but rarely audited jointly. First, we vary **persona format**, because personas are widely used to specify agent goals, traits, and social roles, yet recent work shows that persona-based simulations can introduce systematic validity risks if their construction is not carefully validated (Li et al., 2025). Our test asks a narrower representation-level question: whether the same persona content behaves differently when written as prose, a descriptive list, or a table. Second, we vary **game-instruction framing**, because prior work shows that LLM behavior in repeated games can change when the same strategic setting is described differently, for example through robustness checks, payoff-matrix variations, or reasoning prompts (Akata et al., 2025; Lorè & Heydari, 2024). Third, we vary **memory representation**, because repeated-game behavior depends on how agents read and use prior interaction history; existing Prisoner's Dilemma studies explicitly test whether LLMs can parse gameplay logs and condition decisions on historical behavior (Fontana et al., 2025). Together, these perturbations test whether claims about cooperation, selfishness, or strategic adaptation survive small changes in how the same game, agent, and history are represented to the model.

B.3 Echo Chambers as an Open-Ended Social Simulation Testbed

Echo chambers and polarization are among the most studied collective phenomena in LLM-based social simulation (Gu et al., 2025; Wang et al., 2025a; Ferraro et al., 2024; Zheng & Tang, 2024; Piao et al., 2025; Törnberg et al., 2023; Yang et al., 2024; Chuang et al., 2024). Recent work has used LLM agents on social networks to ask whether echo chambers emerge, what role homophily and recommendation play, and how interventions reshape exposure (Gu et al., 2025; Törnberg et al., 2023; Liu et al., 2025; Puelma Touzel et al., 2025). We adopt this paradigm as a representative open-ended simulation in which collective patterns emerge from local interaction rather than explicit optimization.

We test five perturbations that target core design choices in open-ended social-network simulations: homophily, hub assignment, activation probability, memory window, and recommendation feed size. These choices are motivated by prior echo-chamber and LLM-agent studies, which show that echo chambers emerge from the joint effects of network structure, selective exposure, memory/context, and engagement dynamics. First, we vary **input-network homophily**, because prior work treats homophily and structurally clustered networks as central mechanisms through which like-minded users become locally concentrated and echo chambers form (Gu et al., 2025; Wang et al., 2025a; Ferraro et al., 2024). Second, we vary **hub assignment** while holding the degree sequence fixed, because scale-free social networks contain highly connected users whose placement can disproportionately affect information diffusion and group reinforcement; prior LLM-agent simulations similarly emphasize that scale-free or hub-dominated networks better approximate real social media than fully connected or random interaction settings. Third, we vary **activation probability**, because social media outcomes depend not only on who is connected to whom, but also on which users become active and contribute content in each round. Fourth, we vary the **memory window**, since LLM-agent frameworks commonly rely on short-term or retrieved interaction histories to condition agent behavior, but the amount of remembered context is often treated as an implementation detail rather than an audited design choice. Finally, we vary **recommendation feed size**, because prior work shows that selective exposure and recommendation mechanisms can increase engagement, reduce cross-group interaction, and strengthen echo-chamber formation (Wang et al., 2025a; Ferraro et al., 2024). Together, these perturbations span two levels of the simulation architecture: structural opportunity conditions, through homophily and hub assignment, and interaction-level exposure conditions, through activation, memory, and feed size. This allows us to test whether macro-level echo-chamber metrics are more sensitive to the underlying network topology or to seemingly small choices in how agents are activated, exposed to content, and prompted with recent interaction history.

C Additional Experimental Details

C.1 Repeated Prisoner's Dilemma

C.1.1 Setup and Configurations

Each Prisoner's Dilemma episode runs for $T = 10$ rounds. In each round, each player simultaneously chooses Cooperate (C) or Defect (D), with payoffs $(C, C) = (3, 3)$, $(C, D) = (0, 5)$, $(D, C) = (5, 0)$, and $(D, D) = (1, 1)$.

We evaluate two interaction regimes. In the *single-agent* regime, one LLM-controlled agent plays separately against four fixed policies: Tit-for-Tat, Random, Always-Cooperate, and Always-Defect. Tit-for-Tat cooperates in the first round and subsequently copies the opponent's previous action; Random selects C or D with equal probability; Always-Cooperate and Always-Defect select the corresponding action in every round. In the *two-agent* regime, two LLM-controlled agents with matched configurations play one another. Following prior work, these fixed policies provide benchmark interaction contexts with distinct response structures (Akata et al., 2025; Fontana et al., 2025; Di Stefano et al., 2023).

We evaluate `gpt-5.2`, `claude-haiku-4-5`, `gemini-2.5-flash`, and `deepseek-v3.2`. Each condition contains $N = 30$ simulation runs. In two-agent runs, both agents are queried once per round and prompt order is randomized across rounds to reduce stable positional effects. If a response cannot be parsed as a valid action after one retry, the simulation falls back to Defect. Decoding and parsing settings are held fixed within each comparison; unless otherwise specified, model calls use temperature $= 0.3$ and top-$p = 0.95$.

We report condition means, signed mean differences, 95% confidence intervals, and Cohen's d as a descriptive standardized effect size. Pairwise outcome comparisons use two-sided paired tests with Holm correction within each prespecified comparison family (Holm, 1979). For the Prisoner's Dilemma, correction is applied separately within each model, representation dimension, and outcome across its prespecified pairwise contrasts. Interaction-amplification analyses use matched simulation seeds and paired-seed bootstrap procedures as described in Appendix C.1.4.

C.1.2 Outcome Measures

For agent i, let $a_{i,t} \in \{C, D\}$ denote the action selected at round t, and let $u_{i,t}$ denote the corresponding payoff.

Average payoff. Average payoff is the mean per-round reward:

$$\text{Payoff}_i = \frac{1}{T} \sum_{t=1}^{T} u_{i,t}.$$

Cooperation rate. Cooperation rate is the fraction of rounds in which the agent selects C:

$$\text{CoopRate}_i = \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}(a_{i,t} = C).$$

Cooperation rate is the primary outcome for comparing representation sensitivity across interaction regimes, with average payoff providing the corresponding outcome under the game's incentive structure.

C.1.3 Perturbation Details

We evaluate three representation dimensions while holding the payoff structure and non-targeted prompt components fixed: persona format, game-instruction framing, and memory representation.

Persona format. We present a matched strategic persona in three formats: plain prose (Plain), a descriptive bullet list (Descriptive), and a structured key–value table (Tabular). The persona specifies a strategic participant focused on maximizing long-term payoff, studying the opponent for exploitable patterns, and adapting its strategy while remaining competitive.

Plain strategic persona.

```
You are a strategic participant focused on maximizing your long-term payoff.  Study
opponent behavior for exploitable patterns, and adjust your strategy to gain advantage
while staying competitive.
```

Descriptive strategic persona.

```
Persona profile:
  - Maximize your long-term payoff
  - Study opponent for exploitable patterns
  - Adjust strategy to gain advantage
  - Stay strategic and competitive
```

Tabular strategic persona.

```
Persona | Value
Goal | Maximize your long-term payoff
Observation focus | Study opponent for exploitable patterns
Strategy | Adjust strategy to gain advantage
Style | Stay strategic and competitive
```

Game-instruction framing. We hold the persona fixed to PLAIN and vary the framing of the fixed payoff structure. CANONICAL uses standard game-theoretic labels such as "Cooperate" and "Defect"; MORALIZED describes the actions using cooperation and exploitation; and RISK describes the same decision in terms of risk, vulnerability, and potential reward.

Canonical game-theory framing.

```
You are playing a repeated Prisoner's Dilemma for 10 rounds.

In each round, you choose one of two actions:
- Cooperate
- Defect

The payoffs for each round are:
- If both players cooperate:  both receive 3 points.
- If you cooperate and the other defects:  you receive 0, the other receives 5.
- If you defect and the other cooperates:  you receive 5, the other receives 0.
- If both defect:  both receive 1.

Your goal is to maximize your total points over all rounds.

At each round, you will see the history of previous rounds.
```

Moralized framing.

```
You will interact with another participant for 10 rounds.

In each round, you must decide whether to:
- Cooperate fairly with the other person.
- Exploit the other person for personal gain.

The outcomes are:
- If both cooperate fairly:  both receive 3 points.
- If you cooperate but they exploit you:  you receive 0, they receive 5.
- If you exploit them while they cooperate:  you receive 5, they receive 0.
- If both exploit each other:  both receive 1.

Your goal is to maximize your total points.

At each round, you will see past interactions.
```

Risk framing.

```
You will make decisions over 10 rounds against another participant.

Each round involves a choice:
- Cooperate, which is safer but vulnerable.
- Defect, which is riskier but potentially higher reward.

Outcomes:
- Both cooperate:  both get 3.
- You cooperate, other defects:  you get 0, they get 5.
- You defect, other cooperates:  you get 5, they get 0.
- Both defect:  both get 1.
```

```
Your goal is to maximize total payoff.

At each round, you will see previous outcomes.
```

Memory representation. We hold persona and game framing fixed and vary how the interaction history is presented. The design crosses history format (table vs. narrative) with the presence or absence of summary statistics, yielding TABLE-RAW, NARRATIVE-RAW, TABLE+STATS, and NARRATIVE+STATS. Summary statistics include cooperation rates, cumulative payoffs, and joint outcome counts.

M1: Table, raw history only.

```
History:
Round | You | Opponent | Your payoff | Opp payoff
1 | [YOUR ACTION] | [OPPONENT ACTION] | [YOUR PAYOFF] | [OPPONENT PAYOFF]
2 | [YOUR ACTION] | [OPPONENT ACTION] | [YOUR PAYOFF] | [OPPONENT PAYOFF]
...
```

M2: Narrative, raw history only.

```
History:
In round 1, you chose [YOUR ACTION] and the opponent chose [OPPONENT ACTION]. You received
[YOUR PAYOFF] points and the opponent received [OPPONENT PAYOFF].
In round 2, you chose [YOUR ACTION] and the opponent chose [OPPONENT ACTION]. You received
[YOUR PAYOFF] points and the opponent received [OPPONENT PAYOFF].
...
```

M3: Table, raw history plus statistics.

```
History:
Round | You | Opponent | Your payoff | Opp payoff
1 | [YOUR ACTION] | [OPPONENT ACTION] | [YOUR PAYOFF] | [OPPONENT PAYOFF]
2 | [YOUR ACTION] | [OPPONENT ACTION] | [YOUR PAYOFF] | [OPPONENT PAYOFF]
...

Summary statistics:
- You cooperated [N] times and defected [N] times.
- Your cooperation rate: [RATE].
- Opponent cooperated [N] times and defected [N] times.
- Opponent cooperation rate: [RATE].
- Your cumulative payoff: [PAYOFF].
- Opponent cumulative payoff: [PAYOFF].
- Your average payoff per round: [AVG].
- Opponent average payoff per round: [AVG].
- Joint outcomes: CC=[N], CD=[N], DC=[N], DD=[N].
```

M4: Narrative, raw history plus statistics.

```
History:
In round 1, you chose [YOUR ACTION] and the opponent chose [OPPONENT ACTION]. You received
[YOUR PAYOFF] points and the opponent received [OPPONENT PAYOFF].
In round 2, you chose [YOUR ACTION] and the opponent chose [OPPONENT ACTION]. You received
[YOUR PAYOFF] points and the opponent received [OPPONENT PAYOFF].
...

Summary statistics:
- You cooperated [N] times and defected [N] times.
- Your cooperation rate: [RATE].
- Opponent cooperated [N] times and defected [N] times.
- Opponent cooperation rate: [RATE].
- Your cumulative payoff: [PAYOFF].
- Opponent cumulative payoff: [PAYOFF].
- Your average payoff per round: [AVG].
- Opponent average payoff per round: [AVG].
- Joint outcomes: CC=[N], CD=[N], DC=[N], DD=[N].
```

For the two-agent memory condition, the same four memory variants are used, but the history is role-specific. For Agent A, "You" refers to Agent A and "Other agent" refers to Agent B; for Agent B, the mapping is reversed.

C.1.4 INTERACTION-AMPLIFICATION MEASURES

For model m, representation contrast $p = (c_0, c_1)$, and interaction context o, we define the signed perturbation effect on cooperation as

$$\Delta_{m,p,o} = \mathbb{E}[Y \mid c_1, m, o] - \mathbb{E}[Y \mid c_0, m, o],$$

where Y denotes cooperation rate. For the fixed-opponent regime,

$$\mathcal{O} = \{\text{Tit-for-Tat}, \text{Random}, \text{Always-Cooperate}, \text{Always-Defect}\}.$$

We summarize sensitivity across fixed-opponent contexts as

$$S^{\text{fixed}}_{m,p} = \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} |\Delta_{m,p,o}|,$$

and sensitivity under reciprocal LLM–LLM interaction as

$$S^{\text{LLM}}_{m,p} = |\Delta_{m,p,\text{LLM}}|.$$

The interaction effect gap is

$$A_{m,p} = S^{\text{LLM}}_{m,p} - S^{\text{fixed}}_{m,p}.$$

Positive $A_{m,p}$ indicates *amplification*, while negative $A_{m,p}$ indicates *attenuation*. The fixed-opponent measure averages the absolute opponent-specific effects before comparison with the LLM–LLM effect, preserving sensitivity that differs in direction across benchmark interaction contexts.

We estimate uncertainty in $A_{m,p}$ using $10{,}000$ paired bootstrap replicates over the 30 matched simulation seeds. Within each replicate, the same sampled seeds are used across representation conditions, fixed opponents, and the LLM–LLM regime. We report pointwise percentile 95% confidence intervals for $A_{m,p}$ and compute the corresponding quantities by round to characterize temporal variation.

C.2 SOCIAL-MEDIA ECHO CHAMBER

C.2.1 SETUP AND CONFIGURATIONS

The echo-chamber simulation contains $N = 100$ agents interacting for $T = 15$ rounds on a social network. Each agent has a persona and a fixed stance $s_i \in \{-2, -1, 0, 1, 2\}$ toward stronger regulation of advanced AI systems. Agent stances remain fixed throughout each run, so the simulation measures interaction structure conditional on the initial stance distribution.

In each round, agents become active according to the configured activation probability. If no agent is sampled as active, one agent is selected at random so that every round contains at least one decision. Active agents observe recent posts from direct network neighbors and choose among POST, REPOST, REPLY, and DO NOTHING. Reposts and replies additionally select a visible interaction target. The default configuration uses a feed size of 5 posts and a memory window of 2 rounds. Invalid repost or reply targets are converted according to the simulation's parsing rules while preserving the fixed action space.

We evaluate `gpt-5.2`, `claude-haiku-4-5`, `gemini-2.5-flash`, and `deepseek-v4-flash`. Unless otherwise specified, each condition contains $N = 30$ simulation runs. Decoding and parsing settings are held fixed within each comparison.

Network construction. To control overall connectivity across structural conditions, we construct networks from a fixed power-law-like degree sequence with target mean degree 8 and exponent 2.4. The sequence is adjusted until graphical, producing a heavy-tailed network in which a small number of agents occupy high-degree positions. The same degree sequence is reused across network variants, holding the degree distribution and overall connectivity fixed.

C.2.2 OUTCOME MEASURES

We measure representation sensitivity at both the agent and collective levels. For paired persona-representation runs, agent decisions are matched by $(\text{seed}, \text{round}, \text{agent})$ across the BULLET and PROSE conditions.

Action agreement. For two paired conditions c_0 and c_1, action agreement measures whether the same agent selects the same action at the same round. For paired decision k,

$$I_k^{\text{action}} = \mathbb{I}\left(a_k^{(c_1)} = a_k^{(c_0)}\right).$$

For seed s, action agreement is

$$G_s^{\text{action}} = \frac{1}{|\mathcal{K}_s|} \sum_{k \in \mathcal{K}_s} I_k^{\text{action}},$$

where $\mathcal{K}_s$ contains the matched agent–round decisions for that seed. We compute agreement within each seed and then summarize across paired seeds.

Target agreement. Target agreement measures whether paired decisions select the same non-missing interaction-target author:

$$I_k^{\text{target}} = \mathbb{I}\left(t_k^{(c_1)} = t_k^{(c_0)} \neq \varnothing\right),$$

where t_k denotes the selected target author. Different targets, a target in only one condition, and no target in either condition are counted as non-agreement. Seed-level target agreement is

$$G_s^{\text{target}} = \frac{1}{|\mathcal{K}_s|} \sum_{k \in \mathcal{K}_s} I_k^{\text{target}}.$$

For the persona-representation analysis, $c_0 = \text{BULLET}$ and $c_1 = \text{PROSE}$.

Weighted stance assortativity. Let $G_I = (V, E_I)$ denote the realized interaction graph, where an edge from agent i to agent j records a repost or reply, and let A_{ij} denote the number of observed interactions from i to j. Let $g_i \in \mathcal{G}$ denote agent i's stance group. For stance groups $u, v \in \mathcal{G}$, define the fraction of weighted interactions from group u to group v as

$$p_{uv} = \frac{\sum_{i,j} A_{ij} \mathbb{I}(g_i = u, g_j = v)}{\sum_{i,j} A_{ij}},$$

with marginals

$$p_u^{\text{out}} = \sum_{v \in \mathcal{G}} p_{uv}, \qquad p_u^{\text{in}} = \sum_{v \in \mathcal{G}} p_{vu}.$$

Weighted stance assortativity is

$$r = \frac{\sum_{u \in \mathcal{G}} p_{uu} - \sum_{u \in \mathcal{G}} p_u^{\text{out}} p_u^{\text{in}}}{1 - \sum_{u \in \mathcal{G}} p_u^{\text{out}} p_u^{\text{in}}}.$$

Positive values indicate greater within-stance interaction than expected from the observed interaction marginals, while negative values indicate greater cross-stance interaction.

Weighted same-group edge ratio. We additionally measure the fraction of interaction volume occurring within stance groups:

$$H = \frac{\sum_{i,j} A_{ij} \mathbb{I}(g_i = g_j)}{\sum_{i,j} A_{ij}}.$$

Larger values indicate a greater concentration of interaction within stance groups.

C.2.3 PERTURBATION DETAILS

We evaluate one representation perturbation and five design perturbations. The representation perturbation varies persona presentation while preserving persona information. The design perturbations vary input-network homophily, hub assignment, activation probability, memory-window length, and recommendation feed size.

Persona representation. We present matched persona information either as structured bullets (BULLET) or narrative prose (PROSE), while holding the remaining simulation configuration fixed.

Structured bullet-point persona (original).

```
Your profile:
  - agent_id: {agent_id}
  - bio: {bio}
  - follower_count: {follower_count}
  - current stance on AI regulation: {stance}
    (where −2 strongly oppose, −1 somewhat oppose,
    0 mixed/undecided, 1 somewhat support,
    2 strongly support)
  - neighbor count in the network: {neighbor_count}
```

Narrative-prose persona.

```
You are {agent_id}, {bio}. You have {follower_count} followers, and {neighbor_count}
people in the network whose posts you can see. On the question of whether governments
should impose stronger regulation on advanced AI systems, you currently
{stance_description}.
```

Here, {stance_description} is the prose equivalent of the numerical stance: strongly oppose, somewhat oppose, hold a mixed or undecided view, somewhat support, or strongly support stronger regulation.

Interaction-level design perturbations. Activation probability compares 0.3 and 0.5 for `gpt-5.2`, and 0.35 and 0.5 for the other models. Memory window compares 2 and 4 rounds, and recommendation feed size compares up to 5 and 10 recent neighbor posts.

Input-network homophily (context-relevant design-level perturbation). We vary stance homophily while preserving the fixed degree sequence. Let $G_0 = (V, E_0)$ denote the input network and x_i agent i's stance on the five-point scale. Starting from a graph realizing the fixed degree sequence, we apply degree-preserving double-edge swaps to obtain three input-network assortativity ranges:

$$\text{LOW}: h(G_0) \in [0.02, 0.10], \qquad \text{MEDIUM}: h(G_0) \in [0.12, 0.20], \qquad \text{HIGH}: h(G_0) \in [0.22, 0.30].$$

The rewiring procedure changes which stance groups are connected while preserving each agent's degree.

Hub assignment (context-relevant design-level perturbation). We vary which agents occupy high-degree positions while preserving the degree sequence. The RANDOM HUB condition randomly assigns degrees across agents, while MIXED HUB distributes high-degree positions across stance groups. The PRO HUB and ANTI HUB conditions preferentially assign high-degree positions to the corresponding stance group. Networks are subsequently rewired to the medium-homophily range, $h(G_0) \in [0.12, 0.20]$, approximately holding stance homophily fixed across hub conditions.

Statistical comparisons. Conditions are matched by simulation seed. Binary design perturbations are evaluated using two-sided paired t-tests. Holm correction is applied separately within each model and outcome across activation probability, memory window, and recommendation feed size. Persona-representation effects are evaluated using two-sided paired Wilcoxon signed-rank tests, with Holm correction across the four models and two collective outcomes. All significance markers in the outcome-sensitivity tables use the corresponding Holm-corrected p-values.

C.2.4 PROPAGATION MEASURES

For perturbation $p = (c_0, c_1)$, let $q_t^{(c)}$ denote the cross-group targeting rate under condition c at round t. The signed behavioral effect and its sensitivity magnitude are

$$\Delta_{\text{target},t}(p) = q_t^{(c_1)} - q_t^{(c_0)}, \qquad S_{\text{target},t}(p) = \left|\Delta_{\text{target},t}(p)\right|.$$

For paired persona-representation runs, we additionally compute action-type agreement and interaction-target agreement to locate representation sensitivity in action selection or interaction targeting.

For collective network outcome $Y \in \{r, H\}$, we define

$$\Delta_{Y,t}(p) = Y_t^{(c_1)} - Y_t^{(c_0)}, \qquad S_{Y,t}(p) = \left|\Delta_{Y,t}(p)\right|.$$

We trace representation sensitivity across interaction levels by comparing the temporal development of behavioral effects and collective network effects.

Table 5: **Pairwise contrasts in the Prisoner's Dilemma experiment across perturbations, representations, and models.** Δ denotes the difference in outcome means between two perturbation conditions.

Perturbation	Outcome	Contrast	*GPT-5.2*	*Claude Haiku 4.5*	*Gemini 2.5 Flash*	*DeepSeek V3.2*
Repeated Prisoner's Dilemma (Two-Agent)						
Persona	Average payoff	Descriptive → Tabular	+1.467***	+1.223***	-0.010	-0.127
		Plain → Descriptive	-1.478***	-1.487***	-0.777***	-0.013*
		Plain → Tabular	-0.012	-0.263***	-0.787***	-0.140**
Persona	Cooperation rate (pp)	Descriptive → Tabular	+74.000***	+63.333***	-0.333	-6.333
		Plain → Descriptive	-75.167***	-77.667***	-35.667***	-1.333*
		Plain → Tabular	-1.167	-14.333***	-36.000***	-7.667**
Game framing	Average payoff	Canonical → Moralized	+0.417***	+0.025	-0.187	+0.028**
		Canonical → Risk	+0.110	+0.087*	-0.142	-0.120**
		Moralized → Risk	-0.307**	+0.062	+0.045	-0.148***
Game framing	Cooperation rate (pp)	Canonical → Moralized	+19.333***	+0.833	-9.000	+2.167**
		Canonical → Risk	+4.667	+3.333*	-7.500	-6.667**
		Moralized → Risk	-14.667*	+2.500	+1.500	-8.833***
Memory	Average payoff	NR → NS	+0.038*	+0.043	+0.092	+0.000
		NR → TS	+0.088***	+0.048	+0.017	+0.005
		TS → NS	-0.050**	-0.005	+0.075	-0.005
		TR → NS	-0.063***	+0.045	+0.285	+0.008
		TR → NR	-0.102***	+0.002	+0.193	+0.008
		TR → TS	-0.013*	+0.050	+0.210	+0.013
Memory	Cooperation rate (pp)	NR → NS	+1.833*	+2.333	+4.833	-0.000
		NR → TS	+4.833***	+2.500	+1.667	+0.500
		TS → NS	-3.000**	-0.167	+3.167	-0.500
		TR → NS	-4.333***	+2.500	+13.833	+0.500
		TR → NR	-6.167***	+0.167	+9.000	+0.500
		TR → TS	-1.333*	+2.667	+10.667	+1.000

Note: Values report estimated differences between the indicated perturbation conditions. Cooperation effects are reported in percentage points; average payoff effects use their native units. For memory contrasts, NR = Narrative-Raw, NS = Narrative+Stats, TR = Table-Raw, and TS = Table+Stats. Shading denotes Holm-corrected significance: *** $p < .001$, ** $p < .01$, * $p < .05$; unshaded cells are not statistically significant.

D Extended Experimental Results

This section provides extended results for the Prisoner's Dilemma (PD) and echo-chamber experiments. For the PD experiment, Table 5 reports pairwise outcome contrasts across perturbations, representations, and models, while Table 6 provides the complete interaction-amplification results for the prespecified representation contrasts. For the echo-chamber experiment, Table 7 reports representation sensitivity in both behavioral and collective network outcomes. We additionally report the effects of network homophily and hub assignment in Figure 6. Although these two perturbations are not analyzed in the main text, they serve as the audit-card examples of context-relevant perturbations; we therefore include their corresponding experimental results here for completeness.

Table 6: **Interaction amplification across prespecified representation contrasts in the Prisoner's Dilemma experiment.**

Contrast	*GPT-5.2*	*Claude Haiku 4.5*	*Gemini 2.5 Flash*	*DeepSeek V3.2*
Persona format				
Plain vs Descriptive	S_{fixed} = 0.271 S_{LLM} = 0.752 A = +0.481 [+0.327, +0.622] Amplification	S_{fixed} = 0.315 S_{LLM} = 0.777 A = +0.462 [+0.376, +0.529] Amplification	S_{fixed} = 0.070 S_{LLM} = 0.357 A = +0.287 [+0.190, +0.387] Amplification	S_{fixed} = 0.101 S_{LLM} = 0.013 A = -0.087 [-0.108, -0.066] Attenuation
Plain vs Tabular	S_{fixed} = 0.589 S_{LLM} = 0.012 A = -0.578 [-0.602, -0.556] Attenuation	S_{fixed} = 0.429 S_{LLM} = 0.143 A = -0.286 [-0.341, -0.223] Attenuation	S_{fixed} = 0.074 S_{LLM} = 0.360 A = +0.286 [+0.192, +0.384] Amplification	S_{fixed} = 0.055 S_{LLM} = 0.077 A = +0.022 [-0.050, +0.121] No detectable difference

Continued on next page

Table 6 continued from previous page

Contrast	*GPT-5.2*	*Claude Haiku 4.5*	*Gemini 2.5 Flash*	*DeepSeek V3.2*
	S_{fixed} = 0.318 S_{LLM} = 0.740 A = +0.422 [+0.295, +0.546]	S_{fixed} = 0.244 S_{LLM} = 0.633 A = +0.389 [+0.291, +0.481]	S_{fixed} = 0.004 S_{LLM} = 0.003 A = -0.001 [-0.006, +0.013]	S_{fixed} = 0.054 S_{LLM} = 0.063 A = +0.009 [-0.057, +0.102]
Descriptive vs Tabular	Amplification	Amplification	No detectable difference	No detectable difference
Game framing				
	S_{fixed} = 0.117 S_{LLM} = 0.193 A = +0.077 [+0.024, +0.127]	S_{fixed} = 0.058 S_{LLM} = 0.008 A = -0.050 [-0.067, -0.007]	S_{fixed} = 0.025 S_{LLM} = 0.090 A = +0.065 [-0.022, +0.143]	S_{fixed} = 0.016 S_{LLM} = 0.022 A = +0.006 [-0.021, +0.017]
Canonical vs Moralized	Amplification	Attenuation	No detectable difference	No detectable difference
	S_{fixed} = 0.127 S_{LLM} = 0.047 A = -0.080 [-0.131, +0.001]	S_{fixed} = 0.086 S_{LLM} = 0.033 A = -0.052 [-0.088, -0.017]	S_{fixed} = 0.024 S_{LLM} = 0.075 A = +0.051 [-0.021, +0.116]	S_{fixed} = 0.072 S_{LLM} = 0.067 A = -0.005 [-0.058, +0.050]
Canonical vs Risk	No detectable difference	Attenuation	No detectable difference	No detectable difference
	S_{fixed} = 0.243 S_{LLM} = 0.147 A = -0.097 [-0.186, -0.012]	S_{fixed} = 0.112 S_{LLM} = 0.025 A = -0.087 [-0.120, -0.040]	S_{fixed} = 0.022 S_{LLM} = 0.015 A = -0.007 [-0.035, +0.057]	S_{fixed} = 0.087 S_{LLM} = 0.088 A = +0.001 [-0.045, +0.053]
Moralized vs Risk	Attenuation	Attenuation	No detectable difference	No detectable difference
Memory representation				
	S_{fixed} = 0.046 S_{LLM} = 0.062 A = +0.016 [-0.002, +0.029]	S_{fixed} = 0.025 S_{LLM} = 0.002 A = -0.023 [-0.043, -0.011]	S_{fixed} = 0.027 S_{LLM} = 0.090 A = +0.063 [-0.001, +0.128]	S_{fixed} = 0.019 S_{LLM} = 0.005 A = -0.014 [-0.028, +0.004]
Table-Raw vs Narrative-Raw	No detectable difference	Attenuation	No detectable difference	No detectable difference
	S_{fixed} = 0.021 S_{LLM} = 0.013 A = -0.007 [-0.025, +0.008]	S_{fixed} = 0.048 S_{LLM} = 0.027 A = -0.021 [-0.068, -0.003]	S_{fixed} = 0.025 S_{LLM} = 0.107 A = +0.082 [+0.012, +0.147]	S_{fixed} = 0.010 S_{LLM} = 0.010 A = +0.000 [-0.017, +0.016]
Table-Raw vs Table+Stats	No detectable difference	Attenuation	Amplification	No detectable difference
	S_{fixed} = 0.030 S_{LLM} = 0.043 A = +0.013 [-0.007, +0.030]	S_{fixed} = 0.089 S_{LLM} = 0.025 A = -0.064 [-0.088, -0.043]	S_{fixed} = 0.018 S_{LLM} = 0.138 A = +0.121 [+0.032, +0.200]	S_{fixed} = 0.012 S_{LLM} = 0.005 A = -0.007 [-0.018, +0.010]
Table-Raw vs Narrative+Stats	No detectable difference	Attenuation	Amplification	No detectable difference
	S_{fixed} = 0.052 S_{LLM} = 0.048 A = -0.003 [-0.021, +0.013]	S_{fixed} = 0.064 S_{LLM} = 0.025 A = -0.039 [-0.082, -0.013]	S_{fixed} = 0.023 S_{LLM} = 0.017 A = -0.007 [-0.033, +0.080]	S_{fixed} = 0.016 S_{LLM} = 0.005 A = -0.011 [-0.022, -0.005]
Narrative-Raw vs Table+Stats	No detectable difference	Attenuation	No detectable difference	Attenuation
	S_{fixed} = 0.037 S_{LLM} = 0.018 A = -0.019 [-0.042, +0.004]	S_{fixed} = 0.114 S_{LLM} = 0.023 A = -0.091 [-0.113, -0.068]	S_{fixed} = 0.021 S_{LLM} = 0.048 A = +0.028 [-0.027, +0.134]	S_{fixed} = 0.016 S_{LLM} = 0.000 A = -0.016 [-0.023, -0.002]
Narrative-Raw vs Narrative+Stats	No detectable difference	Attenuation	No detectable difference	Attenuation
	S_{fixed} = 0.014 S_{LLM} = 0.030 A = +0.016 [-0.009, +0.028]	S_{fixed} = 0.050 S_{LLM} = 0.002 A = -0.048 [-0.081, -0.025]	S_{fixed} = 0.011 S_{LLM} = 0.032 A = +0.021 [-0.023, +0.099]	S_{fixed} = 0.005 S_{LLM} = 0.005 A = -0.000 [-0.012, +0.007]
Table+Stats vs Narrative+Stats	No detectable difference	Attenuation	No detectable difference	No detectable difference

Note. Each model cell reports S_{fixed}, S_{LLM}, A [95% CI], and the CI-based pattern. S_{fixed} is the mean across four fixed opponents of the absolute opponent-specific cooperation-rate effect; S_{LLM} is the absolute effect under reciprocal LLM–LLM interaction; and $A = S_{\text{LLM}} - S_{\text{fixed}}$. CIs use 30 paired seeds and 10,000 replicates; they are pointwise and not adjusted for multiple comparisons.

Table 7: **Representation sensitivity in the echo-chamber simulation.** Results compare PROSE with BULLET persona representations across $N = 30$ paired seeds. Differences are PROSE-BULLET.

	GPT-5.2	*Claude Haiku 4.5*	*Gemini 2.5 Flash*	*DeepSeek V4 Flash*
Paired behavioral agreement				
Action agreement	0.997 [0.996, 0.998]	0.936 [0.932, 0.941]	0.850 [0.839, 0.860]	0.663 [0.649, 0.677]
Target agreement	0.551 [0.541, 0.560]	0.593 [0.585, 0.601]	0.610 [0.602, 0.618]	0.144 [0.137, 0.151]
Behavioral sensitivity				
Cross-group targeting (pp)	+5.15*** [4.22, 6.07]	-0.93* [-1.62, -0.24]	+2.31*** [1.49, 3.14]	-7.35*** [-8.98, -5.72]
Collective network sensitivity at round 15				
Stance assortativity	-0.194*** [-0.220, -0.167]	-0.002 [-0.015, 0.010]	-0.035*** [-0.050, -0.019]	-0.023 [-0.068, 0.022]
Same-group edge ratio	-0.051*** [-0.061, -0.041]	+0.010* [0.003, 0.017]	-0.023*** [-0.031, -0.014]	+0.072*** [0.057, 0.087]

Note: Brackets report 95% confidence intervals. Cross-group targeting is averaged over rounds 2–15 and reported in percentage points; collective network outcomes are cumulative through round 15. Significance markers use Holm-corrected paired Wilcoxon p-values: $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$. Agreement measures are descriptive paired-run measures and are reported with 95% confidence intervals.

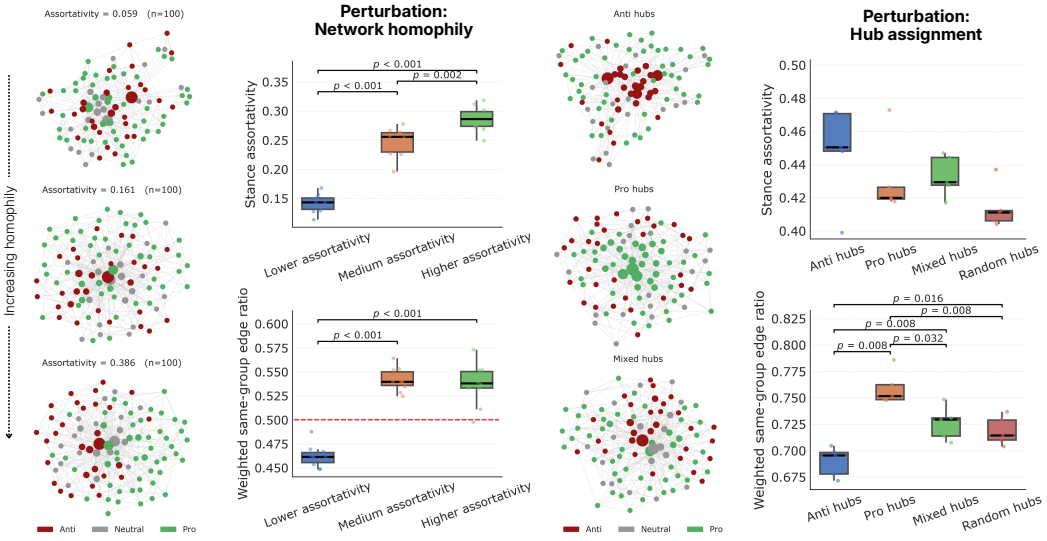


Figure 6: **Effects of network homophily and hub assignment on echo-chamber outcomes.** Left panels vary initial stance homophily; right panels vary hub assignment while holding the degree sequence fixed. Boxplots show run-level stance assortativity and weighted same-group edge ratio. Significant pairwise differences are based on paired t-tests with Holm correction ($p < 0.05$).

E SIMULATION PROMPTS

E.1 PRISONER'S DILEMMA PROMPTS

E.1.1 BASE PERSONA

Across the Prisoner's Dilemma prompt-perturbation experiments reported in the paper, we use the following strategic persona:

```
You are a strategic participant focused on maximizing your payoff.  Study opponent
behavior for exploitable patterns, and adjust your strategy to gain advantage while
staying competitive in the long term.
```

E.1.2 BASE SINGLE-AGENT PROMPT TEMPLATE

In the single-agent setting, the model plays against a fixed external opponent policy. The opponent policy is not revealed to the model.

```
Persona:
[BASE PERSONA]

Game setup:
- Repeated Prisoner's Dilemma for [HORIZON] rounds.
- Each round, both players simultaneously choose one action: Cooperate or Defect.
- You cannot observe the opponent's current-round action before choosing.
- Opponent follows a fixed external policy, unknown to you.

Payoffs:
- If both choose Cooperate: (3, 3).
- If you choose Cooperate and opponent chooses Defect: (0, 5).
- If you choose Defect and opponent chooses Cooperate: (5, 0).
- If both choose Defect: (1, 1).

Task:
Choose a reasonable action this round based on history and repeated-game context.

[HISTORY / MEMORY BLOCK]

Output format:
Return valid JSON with keys:
- action: one of ["Cooperate", "Defect"]
- reason: short string
```

E.1.3 BASE TWO-AGENT PROMPT TEMPLATE

In the two-agent setting, both agents are model-driven players. Each agent receives the same prompt structure but with role-specific history.

```
Persona:
[STRATEGIC PERSONA]

Role:
Your role label this run: Agent [A/B].

Game setup:
- Repeated Prisoner's Dilemma for [HORIZON] rounds.
- Each round, both agents simultaneously choose one action: Cooperate or Defect.
- You cannot observe the other agent's current-round action before choosing.
- The other agent is another model-driven player with the same prompt framing.

Payoffs:
- If both choose Cooperate: (3, 3).
- If you choose Cooperate and the other agent chooses Defect: (0, 5).
- If you choose Defect and the other agent chooses Cooperate: (5, 0).
- If both choose Defect: (1, 1).

Task:
Choose a reasonable action this round based on history and repeated-game context.

[ROLE-SPECIFIC HISTORY / MEMORY BLOCK]

Output format:
Return valid JSON with keys:
- action: one of ["Cooperate", "Defect"]
- reason: short string
```

E.1.4 OUTPUT SCHEMA AND PARSING INSTRUCTION

All Prisoner's Dilemma prompts end with the same JSON parsing instruction.

```
Output format:
Return valid JSON with keys:
- action: one of ["Cooperate", "Defect"]
- reason: short string
```

If a model response cannot be parsed as valid JSON, the parser attempts to extract a JSON object from the raw text. If parsing still fails, the fallback action is `Defect`.

E.2 Echo-Chamber Prompts

E.2.1 Agent Persona Template

Each echo-chamber agent is initialized with a minimal persona containing an agent ID, a short bio, a follower count, and a fixed stance toward AI regulation. Stance is represented on a five-point scale from strong opposition to strong support:

$$s_i \in \{-2, -1, 0, 1, 2\}.$$

```
{
  "agent_id": "[AGENT ID]",
  "bio": "[SHORT BIO]",
  "follower_count": [FOLLOWER COUNT],
  "ai_regulation_stance": [-2|-1|0|1|2]
}
```

The stance scale is:

```
-2 = strongly oppose stronger AI regulation
-1 = somewhat oppose stronger AI regulation
0 = mixed or undecided
1 = somewhat support stronger AI regulation
2 = strongly support stronger AI regulation
```

E.2.2 Round Decision Prompt

At each round, an activated agent receives its own profile, the number of neighbors in the fixed network, and a feed consisting only of recent posts from neighboring agents.

```
You are a social media user discussing one issue: whether governments should impose
stronger regulation on advanced AI systems.

Your profile:
- agent_id: [AGENT ID]
- bio: [BIO]
- follower_count: [FOLLOWER COUNT]
- current stance on AI regulation: [STANCE] where -2 strongly oppose, -1 somewhat oppose,
0 mixed or undecided, 1 somewhat support, 2 strongly support.
- neighbor count in the network: [NEIGHBOR COUNT]

Simulation rules:
- This is a fixed network for this run.
- You only see recent posts from your neighbors.
- Keep behavior simple and realistic.
- Do not invent facts. Use short social-media-style language.
- You may choose exactly one action this round: POST, REPOST, REPLY, or SILENT.
- REPOST or REPLY should only target one visible post from the feed below.
- Your stance is fixed for the entire run and must not change.
- Set updated_stance equal to your current stance.

Round: [ROUND] / [TOTAL ROUNDS]

Visible feed from your neighbors:
[VISIBLE FEED]

Return valid JSON with exactly these keys:
{
  "action": "POST|REPOST|REPLY|SILENT",
  "target_index": integer or null,
  "message": "short text, or empty string if SILENT",
  "updated_stance": -2|-1|0|1|2,
  "reason": "brief reason"
}

If target_index is used, it must refer to one visible feed item by its 1-based index.
```

E.2.3 Visible Feed Format

When neighbor posts are visible, they are rendered as a numbered list.

```
[1] post_id=[POST ID] | author=[AUTHOR ID] | action_type=[POST/REPOST/REPLY] | text=[POST
TEXT]
[2] post_id=[POST ID] | author=[AUTHOR ID] | action_type=[POST/REPOST/REPLY] |
target_post_id=[TARGET POST ID] | text=[POST TEXT]
...
```

If the agent has no visible recent neighbor posts, the feed is rendered as:

```
No recent posts from your neighbors.
```

E.2.4 Output Schema and Parsing Instruction

All echo-chamber prompts require the following JSON schema.

```
{
  "action": "POST|REPOST|REPLY|SILENT",
  "target_index": integer or null,
  "message": "short text, or empty string if SILENT",
  "updated_stance": -2|-1|0|1|2,
  "reason": "brief reason"
}
```

The parser accepts only four actions:

$$a_i^t \in \{\texttt{POST}, \texttt{REPOST}, \texttt{REPLY}, \texttt{SILENT}\}.$$

If `REPOST` or `REPLY` is selected, the target must refer to a visible feed item by its one-indexed position. The agent's stance is frozen throughout the run, so `updated_stance` is constrained to equal the current stance:

$$s_i^{t+1} = s_i^t.$$

F The Social Simulation Audit Card

TRAILS identifies what can be varied in an audit. Interpreting these perturbations requires an additional, claim-specific judgment: *context relevance*. A dimension is *context-relevant* when it has a meaningful human or theoretical counterpart for the modeled phenomenon, as supported by human studies or established social-science theory. Sensitivity along this dimension identifies conditions under which the claim holds. On the hand, hand Sensitivity along a context-irrelevant dimension reveals implementation dependence. The required audit strength is also shaped by the claim's evidentiary role, the simulation's openness, and the stakes of its intended use (See Appendix B). We introduce the **Social Simulation Audit Card** to operationalize TRAILS for a stated scientific. The following five steps connect these requirements to an auditable claim.

State the claim. Record the scientific conclusion the simulation supports, its evidentiary role, and its intended scope of generalization. Define the property that must be preserved, such as the presence of a phenomenon, an effect direction, or a quantitative effect within a stated tolerance.

Map TRAILS and context relevance. For each perturbation, identify whether it changes the simulation design (TRAILS-D) or its representation (TRAILS-R), and whether the changed dimension is context-relevant or context-irrelevant. Explain each assignment and record any uncertainty.

Choose and register perturbations. Select perturbations spanning a plausible range for the claim. For each, register the baseline and alternatives, factors held constant, models, and number of independent runs. Identify unaudited dimensions and define the claim-preservation criterion, tolerance, and uncertainty requirements before examining results.

Report sensitivity results. Report effect sizes, uncertainty, and whether the conclusion meets the claim-preservation criterion. Sensitivity to context-relevant perturbations may reflect meaningful social variation. Report the range of conditions under which the claim holds and use this range to

Table 8: Audited perturbations and claim-level interpretations for GPT-5.2. PD reports the maximum absolute cooperation-rate difference across contrasts, in percentage points (pp). EC reports absolute differences in stance assortativity (r) and weighted same-group edge ratio (H). Stars indicate Holm-corrected $p < .001$; unmarked estimates were not significant after correction.

Perturbation	TRAILS	Context relevance	Outcome sensitivity	Interpretation
Repeated Prisoner's Dilemma				
Persona format	TRAILS-R	Irrelevant	75.2^{***} pp	Cooperation depended strongly on persona formatting.
Game framing	TRAILS-R	Relevant	19.3^{***} pp	Framing conditions should qualify the claim.
Memory representation	TRAILS-R	Irrelevant	6.2^{***} pp	Smaller but significant change; assess claim preservation against the tolerance.
Echo-chamber simulation				
Persona representation	TRAILS-R	Irrelevant	$\lvert\Delta r\rvert = 0.194^{***}$ $\lvert\Delta H\rvert = 0.051^{***}$	Estimated collective structure depended on persona representation.
Activation probability	TRAILS-D	Irrelevant	$\lvert\Delta r\rvert = 0.008$ $\lvert\Delta H\rvert = 0.006$	Small estimated differences; claim preservation requires a tolerance-based assessment.
Memory window	TRAILS-D	Irrelevant	$\lvert\Delta r\rvert = 0.011$ $\lvert\Delta H\rvert = 0.001$	Small estimated differences; claim preservation requires a tolerance-based assessment.
Recommendation feed size	TRAILS-D	Irrelevant	$\lvert\Delta r\rvert = 0.004$; $\lvert\Delta H\rvert = 0.005$	Small estimated differences; claim preservation requires a tolerance-based assessment.

define its scope. Substantial changes or reversals under context-irrelevant perturbations indicate implementation dependence. Interpret these results against the specified claim-preservation criterion; statistical significance alone does not determine robustness.

Scope the claim according to audit outcomes. Classify the claim into one of four outcomes. A claim is *Robust* when it meets the preservation criterion across all audited perturbations with adequate coverage and precision; *Scope-limited* when it holds only within part of the context-relevant conditions tested; *Implementation-sensitive* when context-irrelevant choices, such as persona formatting or memory encoding, substantially change the conclusion; or *Inconclusive* when coverage or precision is insufficient to assess robustness. Report the range supporting a scope-limited claim and bound every verdict by the audited dimensions, configurations, models, and interaction horizons.

Audit sufficiency therefore depends on coverage of the dimensions consequential to the claim and the precision required by its decision rule. Every verdict should state the tested range, model coverage, and remaining uncertainty. Generalization to human behavior additionally requires external validation.

F.1 Auditing Scientific Claims in the Two Case Studies

We applied the Social Simulation Audit Card to claims about cooperation and echo-chamber formation. Both experiments used 30 independent runs per condition, with shared seeds for paired comparisons. The audits below summarize the GPT-5.2 results; their verdicts are specific to the reported model, interaction conditions, and outcomes. Table 8 maps the reported perturbations to their audit interpretations.

Repeated Prisoner's Dilemma. We audited the claim that agents exhibited stable cooperation in the two-agent, 10-round repeated Prisoner's Dilemma. Changing the matched persona from Plain to Descriptive reduced cooperation from 98.2% to 23.0%, a decrease of 75.2 percentage points (95% CI $[62.0, 88.3]$), and reduced average payoff by 1.48 points per round. Because persona format was a context-irrelevant TRAILS-R perturbation, this result supported an *Implementation-sensitive* verdict. Game framing was a context-relevant TRAILS-R perturbation with a maximum cooperation difference of 19.3 percentage points, indicating that framing conditions should qualify the claim. Memory representation was context-irrelevant and produced a smaller but significant maximum difference of 6.2 percentage points; whether this difference preserved the claim depends on the specified tolerance. These findings bound the verdict to the audited interaction regime and representation choices.

Echo-chamber simulation. We audited the claim that the simulation produced echo-chamber-like interaction structure among 100 agents over 15 rounds with fixed stances. Persona representation (BULLET versus PROSE) was a context-irrelevant TRAILS-R perturbation. Under PROSE, cross-group targeting increased by 5.15 percentage points (95% CI $[4.22, 6.07]$), and cumulative stance assortativity was 0.194 lower by round 15; the absolute difference in weighted same-group edge ratio was 0.051. These results showed implementation dependence in the measured strength of echo-chamber structure. Activation probability, memory window, and recommendation feed size were context-irrelevant TRAILS-D perturbations with smaller estimated effects and no significant differences for GPT-5.2 after Holm correction. For the original phenomenon-presence claim, the verdict remained *Inconclusive*: the reported contrasts did not establish whether conditions crossed a specified criterion for echo-chamber presence.