When Harm Scales: Social Dynamics of Nonlinear Collapse in Financial Agent Societies

Lejun Zhang[1], **Muning Wen**[1], **Sarah Lu-Liang**[2,3], **Xin Jiang**[2], **Weinan Zhang**[1], **Shangding Gu**[2*]

[1]Shanghai Jiao Tong University, China
[2]University of California, Berkeley, USA
[3]University of Toronto, Canada

Abstract

Safety evaluations typically assess agents in isolation, yet interacting agents exchange messages and reshape shared environments, allowing local harm to escalate into society-level failure. We study how collapse varies with harmful-agent fraction and society size N. We introduce *Agent Society Dynamics*, a framework that relates the collapse boundary to size-dependent harmful pressure and shared-state feedback, and evaluate it through fixed-count analyses and controlled interventions in a case-inspired hybrid social–market simulation. Across society sizes, collapse is rare at low harmful fractions, rises sharply near a size-dependent boundary, and then saturates. As N increases from 100 to 2000, the 50% collapse boundary falls from 4.7% to 2.2%, while the effective harmful count rises from 4.7 to 44. A power-law fit yields a boundary exponent of 0.222; bootstrap estimates remain between 0 and 1, implying sublinear growth in the effective harmful count. At fixed harmful counts, the maximum joint severity of harmful diffusion and market disruption decreases with society size, revealing subcritical dilution. Weakening feedback shifts the boundary toward higher harmful fractions, whereas stronger feedback or greater network reach shifts it toward lower fractions. Together, these results show that society size and shared-state feedback jointly shape collective failure under the tested protocol.

1 Introduction

Safety evaluations usually test agents one at a time. Yet as AI systems scale toward increasingly complex tasks, coordinated populations of agents are likely to become a common deployment paradigm. These agents exchange messages, observe shared outcomes, and act on a common environment. Harmful influence may therefore spread beyond its source as agents respond and reshape the state that guides later decisions. A society can consequently fail even when most agents appear safe in isolation.

Recent work has identified interaction failures, nonlinear group-size effects, degraded outcomes in large populations, and conformity-driven collective misalignment (**?**Cemri et al. 2025; **?**; **?**; De Marzo et al. 2026; Huang et al. 2026). Changing the prevalence of misaligned values can also alter long-run collective regimes (**?**). However, existing work does not systematically map how the harmful participation required for failure changes as society size grows. We therefore ask whether larger societies fail at the same harmful count, the same harmful fraction, or a different scaling of the two.

We investigate this question in a controlled social–market society of heterogeneous, boundedly rational agents. Agents exchange messages and trade in a shared market, creating a closed loop between communication, collective action, and environmental state. In the primary scenario, an episode collapses when harmful information spreads broadly and coincides with severe price dislocation or liquidity stress. We vary harmful-agent fraction α and society size N, while holding role distributions, interaction rules, market scaling, harmful strategy, and controller allocation fixed.

We observe a sharp transition from rare to frequent collapse. We define the collapse boundary $\alpha_{\mathrm{c}}(N)$ as the harmful fraction at which collapse probability reaches 50%, and $K_c(N) = N\alpha_{\mathrm{c}}(N)$ as the effective harmful count at that boundary. As N grows from 100 to 2000, the boundary falls from 4.7% to 2.2%, while the corresponding count rises from 4.7 to 44. Larger societies therefore collapse at a smaller harmful fraction but require more harmful agents in absolute terms. Over the tested size range, a power-law fit gives a boundary exponent of 0.222. The corresponding count exponent is 0.778, with a 95% bootstrap confidence interval (CI) of $[0.717, 0.848]$.

This pattern motivates separating two effects of society size: how harmful pressure scales with the population, and how induced actions are reproduced through social and environmental feedback. The falling fraction and rising count are therefore compatible with dilution at fixed harmful counts and amplification once harmful participation scales with the society.

We formalize this distinction through *Agent Society Dynamics*, a finite-size analysis framework linking the observable collapse boundary to *attack aggregation* and *feedback amplification*. The former describes how harmful pressure changes with population size; the latter describes how induced actions alter shared state and shape later behavior. Under a stable collapse criterion, a conditional scaling relation connects these two processes to how the boundary changes with society size. Figure 1 summarizes the closed-loop setting, the observed size-dependent transition, and the resulting finite-size analysis.

Guided by this framework, we conduct controlled fixed-count and intervention analyses. At fixed harmful counts,

*Corresponding author: shangding.gu@berkeley.edu.
Copyright © 2027, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

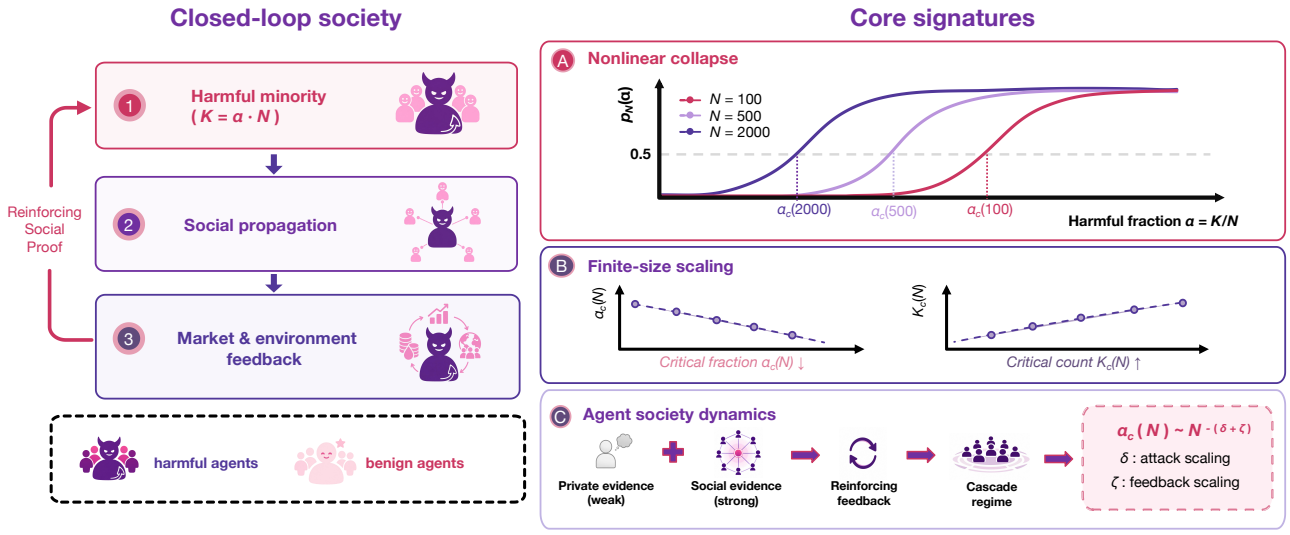

Figure 1: Shared-state feedback in an interacting agent society. Harmful pressure propagates through social and environmental feedback; the right panels summarize the observed transition, finite-size scaling, and the conditional mechanism framework.

the maximum combined severity of harmful diffusion and market disruption decreases with society size, including in subcritical and zero-failure runs. Weakening feedback shifts the 50% boundary toward higher harmful fractions, whereas strengthening feedback or doubling network reach shifts it toward lower fractions. Increasing conformity makes social messages more predictive of individual actions but barely moves the boundary, suggesting that social influence becomes collapse-relevant when induced actions feed back into shared state. We further assess the pattern across controller realizations, LLM backbones, and additional manipulation mechanisms.

Our contributions are summarized as follows:

- **A society-level safety testbed.** We formulate collective-collapse evaluation across harmful composition and society size, and instantiate it in WolfBench, a case-inspired social–market environment covering four manipulation mechanisms.
- **Agent Society Dynamics.** We introduce a finite-size framework that separates attack aggregation from shared-state feedback and relates them conditionally to boundary scaling.
- **Comprehensive empirical evidence.** Scaling sweeps, fixed-count analyses, interventions, and robustness audits reveal a decreasing collapse boundary, subcritical dilution, and feedback strength and network reach as collapse-relevant factors.

2 Related Work

Multi-agent safety and emergent collective risks. Interacting AI agents can exhibit failures absent from isolated-agent evaluations, including miscoordination, conflict, collusion, and destabilizing network effects (**?**Cemri et al. 2025; Erisken et al. 2025). Empirical studies further document secret collusion, inter-agent misalignment, conformity, and emergent group risks (Motwani et al. 2024; Nakamura et al. 2026; Huang et al. 2026). Most closely related, adversarial minorities can drive communicating agents across tipping points into persistent misalignment (De Marzo et al. 2026), while changing the prevalence of misaligned values can alter collective regimes and induce macro-level collapse (**?**). We instead vary harmful composition and society size jointly to measure how the collapse boundary moves in a closed-loop society.

Population-scale agent simulation. Generative-agent platforms and social simulations study emergent behavior in LLM-driven populations (Park et al. 2023; Piao et al. 2025). Related environments examine cooperation and sustainability (Piatti et al. 2024), market fidelity (Byrd, Hybinette, and Balch 2019), long-horizon heterogeneous populations (**?**), and social norms in agent-native communities (**?**). Unlike these settings, our simulator treats harmful fraction and population size as joint experimental axes while holding role distributions, interaction rules, and environmental dynamics fixed.

Collective thresholds and systemic risk. Threshold and cascade models explain how local adoption produces discontinuous collective behavior (Schelling 1971; Granovetter 1978; Watts 2002), while complex contagion emphasizes reinforcement across contacts (Centola and Macy 2007; Castellano, Fortunato, and Loreto 2009). Financial-network models similarly connect topology and feedback to systemic fragility (Haldane and May 2011; Elliott, Golub, and Jackson 2014; Acemoglu, Ozdaglar, and Tahbaz-Salehi 2015). Recent LLM-agent studies examine collaborative financial fraud and coordinated market behavior (**??**). We focus instead on how the harmful participation required for society-level failure scales with population size.

Social learning and network influence. Information-cascade, discrete-choice, and rational-inattention models ex-

plain how public information, social influence, and limited processing shape individual decisions (Banerjee 1992; McKelvey and Palfrey 1995; Brock and Durlauf 2001; Sims 2003). Recent work shows that agent conformity depends on group conditions (**?**) and distinguishes network exposure from influence realized under finite attention (**?**). We use conditional mutual information (Cover and Thomas 2006) to compare social and private evidence, while controlled interventions test whether socially induced actions are amplified through shared state toward collapse.

Finite-size scaling. Finite-size scaling separates collective transitions from finite-population artifacts (Fisher 1971; Barber 1983). Recent LLM-agent studies also reveal nonlinear size effects and degraded outcomes in larger groups (**??**). We jointly vary harmful fraction and society size, estimate the size-dependent collapse boundary and effective harmful count, and test the resulting pattern through fixed-count analyses and controlled interventions without assuming a known universality class.

3 Agent Society Dynamics: Collapse and Finite-Size Scaling

We now formalize society-level collapse, the social and environmental feedback that can amplify harmful pressure, and the finite-size coordinates used to analyze the resulting boundary.

3.1 Society-Level Collapse

An episode contains N agents over $T = 30$ days, including K harmful agents. Runs start from a target harmful fraction α^{target}, with

$$K = \text{round}(\alpha^{\text{target}} N), \qquad \alpha^{\text{realized}} = K/N.$$

The main analysis interpolates the response curve on the target grid. Below, $\alpha = \alpha^{\text{target}}$ unless stated otherwise. For manipulation mechanism m under protocol π, define

$$p_m(N, \alpha; \pi) = \Pr(C_m = 1 \mid N, \alpha, m, \pi),$$
$$\alpha_c^m(N; \pi) = \inf\{\alpha : p_m(N, \alpha; \pi) \geq 1/2\}.$$

We report α_c only when the sampled grid brackets $p = 1/2$, and define $K_c(N) = N\alpha_c(N)$, the effective number of harmful agents at the 50% collapse boundary. Because α_c is an interpolated midpoint, K_c need not be an integer. The protocol fixes agent policies, graph generation, market scaling, harmful strategy and placement, horizon, and controller allocation.

For the primary scenario S1, let q_t denote the reach of harmful social cascades, d_t price dislocation, and L_t liquidity stress. We first measure the joint severity of harmful diffusion and market disruption on each day. The episode-level peak risk R_{S1} is the largest such severity over the thirty-day episode, and collapse occurs when this value reaches one:

$$S_t^{\text{S1}} = \max\left\{0, \min\left(\frac{q_t}{0.55}, \max\left\{\frac{d_t}{0.21}, \frac{L_t}{0.85}\right\}\right)\right\},$$
$$R_{\text{S1}} = \max_{t \leq T} S_t^{\text{S1}}, \qquad C_{\text{S1}} = \mathbf{1}\{R_{\text{S1}} \geq 1\}.$$

Thus, collapse requires harmful social diffusion together with either price dislocation or liquidity stress. The supplementary material reports component cutoffs, threshold perturbations, and sensitivity to the realized grid.

3.2 Social Response and Feedback Amplification

To explain how local harmful pressure becomes collective failure, we separate the pressure introduced by harmful agents from the responses it induces and the feedback generated by those responses. We summarize role-conditioned behavior with the local response approximation

$$\Delta U_{i,t} = h_{i,t} + \theta_i Z_{i,t}^v + \gamma_i Z_{i,t}^s,$$
$$x_{i,t+1} = \tanh\left(\frac{\beta_i}{2} \Delta U_{i,t}\right).$$

Here $h_{i,t}$ is harmful pressure, and $Z_{i,t}^v$ and $Z_{i,t}^s$ summarize private and social evidence. The social weight γ_i, response precision β_i, and network reach correspond to the conformity, precision, and reach interventions below. The resulting actions update market and social state, closing the feedback loop. This local analytical reduction provides a common coordinate system for the interventions. In the simulator, trust and attention mediate social evidence, so identical messages can produce different responses across roles and histories. The full derivation and Jacobian form appear in the supplementary material. Aggregating the direct harmful inputs $h_{i,t}$ across the population yields the society-level attack term used in the finite-size response analysis below.

Let $\chi_{T,N}^+$ denote the finite-horizon amplification of persistent harmful pressure, with $\chi_{T,N}^+ \propto N^\zeta$ over the tested size range. Because this theoretical susceptibility is not directly observed, we use a failure-gain proxy: the episode-level peak risk divided by the attack input implied by the experimental design. Its fitted exponent $\widehat{\zeta}_{\text{fg}}$ summarizes how this normalized response changes with society size.

Social-information dominance separately compares the action relevance of social messages M and private evidence V, conditional on public history H and role $\mathcal{R}$:

$$D_N = \frac{I(A; M \mid V, H, \mathcal{R})}{I(A; M \mid V, H, \mathcal{R}) + I(A; V \mid M, H, \mathcal{R})}.$$

We report D_N only when the denominator exceeds a prespecified information floor; otherwise, it is left undefined. Values $D_N > 1/2$ indicate that social messages carry more action-relevant information than private evidence. Neither D_N nor amplification alone determines fragility. Socially influenced actions must also push shared state toward collapse. Estimation details and finite-sample checks appear in the supplementary material.

3.3 Finite-Size Response Coordinates

Let aggregate harmful pressure scale as αN^δ, where the attack-aggregation exponent δ captures how attack magnitude changes with society size, and let feedback amplification scale as N^ζ. Assuming that collapse occurs at a size-invariant

level of amplified harmful pressure, the induced response scales as

$$\Delta R_N \approx \underbrace{\alpha N^{\delta}}_{\text{attack aggregation}} \cdot \underbrace{N^{\zeta}}_{\text{feedback amplification}}.$$

The corresponding collapse boundary and effective critical count satisfy

$$\alpha_c(N) \propto N^{-\nu}, \qquad K_c(N) \propto N^{1-\nu}, \qquad \nu = \delta + \zeta. \quad (1)$$

Thus, $0 < \nu < 1$ implies a decreasing collapse boundary and a sublinearly increasing critical count. Equation (1) is conditional on the assumed size-invariant collapse criterion.

For the empirical analysis, we fit the fixed-count response surface

$$\log R_{\mathrm{S1}} = a + b_N \log N + b_K \log K + \varepsilon.$$

Under liquidity depth N^{ℓ}, the design-implied attack input is $A = K/N^{\ell}$, so $\delta = 1 - \ell$. The failure-gain proxy $G_{\mathrm{fg}} = R_{\mathrm{S1}}/A$ then satisfies

$$\widehat{\zeta}_{\mathrm{fg}} = b_N + \ell, \qquad \delta + \widehat{\zeta}_{\mathrm{fg}} = 1 + b_N.$$

Hence, the chosen attack-input normalization contributes directly to the fitted size dependence. Setting $R_{\mathrm{S1}} = 1$ yields the response contour

$$\nu_{\mathrm{response}} = 1 + \frac{b_N}{b_K},$$

which combines the empirical elasticities of risk with respect to society size and harmful count. We estimate this contour from subcritical fixed-count data without using collapse midpoints for fitting or calibration. The supplementary material provides the derivation and validity checks.

4 Experimental Setup

4.1 Society and Scenarios

We use a shared market because prices and liquidity form an endogenous state that agents jointly alter and later observe. This creates a measurable loop from harmful messages to collective trading and back to later decisions, while price dislocation and liquidity stress provide prespecified collapse readouts.

Agents interact over a social graph while trading in this market. Benign agents have persistent portfolios, noisy private signals, finite attention, sender-specific trust, and one of five roles: risk-averse, value-oriented, trend-following, social-following, or aggressive. Each day, agents receive private and social evidence, communicate and trade, and observe the resulting state, which carries into the next day. Role proportions and parameter distributions remain fixed across society sizes.

Table 1 summarizes the four mechanisms inspired by public cases. S1 is the primary scaling setting, and S2–S4 test qualitative transfer. Their designs follow public records of social media manipulation, spoofing, and wash trading (U.S. Securities and Exchange Commission 2022; U.S. Department of Justice 2020; U.S. Commodity Futures Trading Commission 2021). Full provenance appears in the supplementary material.

Table 1: Manipulation scenarios inspired by public cases and their primary readouts. S1 is the primary analysis scenario.

Scenario	Perturbed channel	Primary readout
Social-information manipulation		
S1 Pump-and-dump	Social demand and diffusion	Price–liquidity dislocation
S2 Scalping	High-reach promotion	Price–liquidity dislocation
Market-microstructure manipulation		
S3 Spoofing/layering	Displayed market depth	Cancellation and liquidity stress
S4 Wash trading	Artificial trading volume	Fake liquidity and withdrawal loss

Each role maps private observations and received messages to trading and communication actions. Most agents use scalable controllers defined by role. A prespecified quota that grows slowly with society size assigns LLM controllers to a small subset of benign and harmful agents. All primary scaling, response, and intervention experiments use DeepSeek V3.2 for LLM-controlled agents. We separately evaluate backbone dependence through a reduced two-size audit in which only the LLM backbone changes and the remaining S1 protocol stays fixed. Exact quotas, prompts, cached responses, and controller allocation audits are reported in the supplementary material. We package the simulator, scenarios, and evaluation interfaces as WolfBench, a controlled testbed for measuring collective failure across harmful composition and society size and for supporting future evaluation of defenses targeting propagation and shared-state feedback.

4.2 Scaling and Interventions

Primary scaling and intervention runs use twelve seeds and $N \in \{100, 200, 300, 500, 1000, 2000\}$. Each α-grid contains zero and is adaptively refined around a transition identified in pilot runs. Reported intervals are conditional on the selected grid, and pilot episodes are excluded. Across N, we hold the scenario, role proportions, graph rule, per-agent budgets, harmful strategy and placement, liquidity scaling, defense setting, and controller allocation schedule fixed.

Matched S1 interventions are run at two sizes, $N \in \{300, 1000\}$. They vary composite feedback bundles, network reach, attention, conformity, and response precision. The feedback bundles coordinate four components of the social–environmental loop. The weaker regime jointly reduces response precision, conformity, attention capacity, and graph degree. The stronger regime increases the same components. Interventions on individual factors vary one component at a time. Increased network reach doubles mean degree for all agents. Its effect on the boundary is measured, not assumed. Common random-number seeds are reused across matched cells. Runs across scenarios use $N \in \{300, 1000\}$ and twelve seeds per cell.

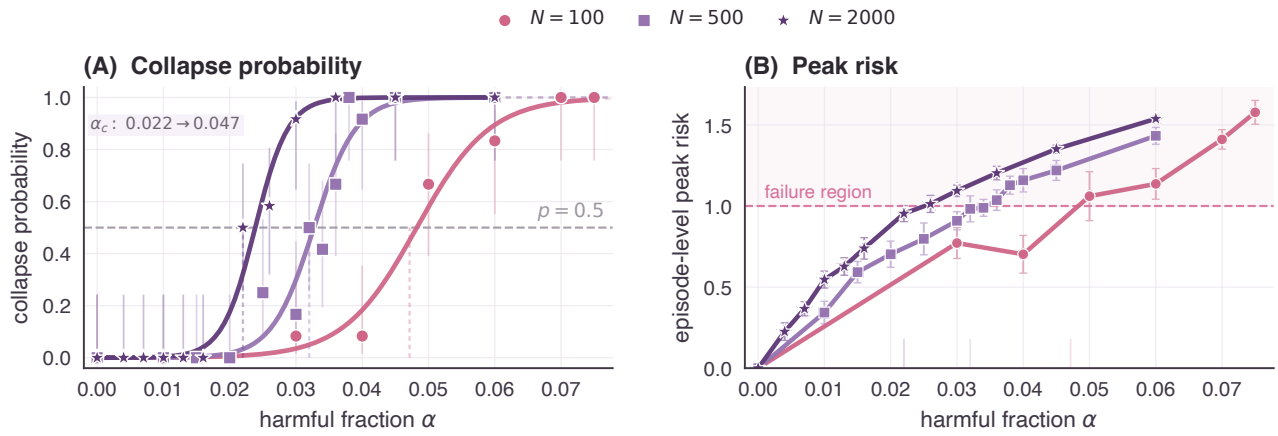

Figure 2: Nonlinear collapse transition in the primary scenario. Collapse probability rises sharply around a size-dependent boundary and saturates at higher harmful fractions; the underlying episode-level peak risk provides a continuous view of the same thresholded event.

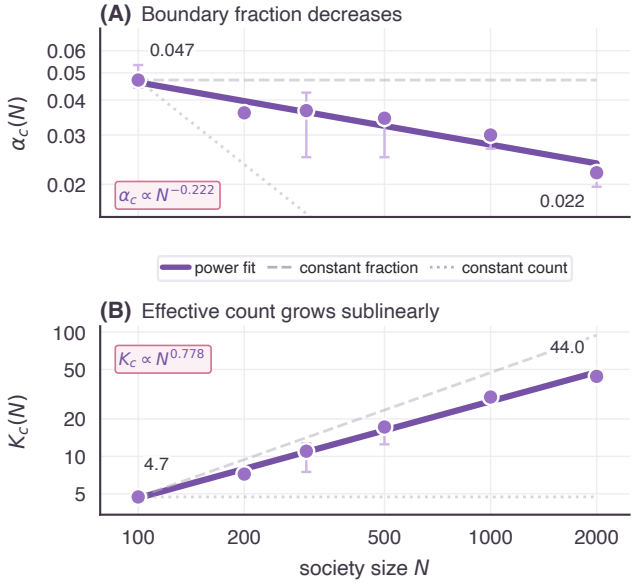

Figure 3: Finite-size scaling of the collapse boundary. The boundary decreases with society size, while the effective critical harmful count grows sublinearly.

4.3 Response Scaling and Estimation

The fixed-count response grid uses $K \in \{1, 3, 5, 8, 12\}$, all six society sizes, and twelve seeds per cell. We fit $\log R_{S1} = a + b_N \log N + b_K \log K + \varepsilon$ at the episode level. The primary subcritical audit retains cells whose failure prevalence is below $1/2$. A stricter check retains only cells with zero failures. Paired bootstrap replicates resample complete panels for each fixed count.

For descriptive accounting, we also vary liquidity depth N^ℓ. Fixed depth uses $\ell = 0$, baseline depth uses $\ell = 0.5$, and per-capita depth uses $\ell = 1$. The corresponding failure-gain proxy is $R_{S1}/(K/N^\ell)$. Its size exponent contains the normalization term ℓ exactly, as shown in the framework and supplementary audit.

We estimate $\alpha_c(N)$ by linearly interpolating the first adjacent α values that bracket $p = 1/2$; unresolved midpoints are censored. Each of 10,000 paired-bootstrap replicates resamples seeds, reconstructs the response curves, re-estimates their boundaries, and refits the log–log exponent while preserving pairing across matched cells. Because the grid is adaptively selected, intervals are conditional on the frozen grid and do not include grid-selection uncertainty. Intervention effects are reported as the mean paired shift

$$\Delta\alpha_c(N) = \alpha_c^{\text{intervention}}(N) - \alpha_c^{\text{baseline}}(N)$$

across $N \in \{300, 1000\}$. Additional checks appear in the supplementary material.

5 Results

5.1 Nonlinear Collapse and Finite-Size Scaling

Figure 2 shows that collapse remains rare at low harmful fractions, rises sharply around a size-dependent boundary, and saturates beyond it. The underlying continuous measure—the maximum combined severity of harmful diffusion and market disruption reached during an episode—varies smoothly around the same boundary, revealing near-threshold differences hidden by the binary collapse label.

The collapse boundary $\alpha_c(N)$ decreases from 0.047 at $N = 100$ to 0.022 at $N = 2000$. Over the same range, the effective critical count $K_c(N) = N\alpha_c(N)$ increases from 4.7 to 44.0. An ordinary least-squares fit across the six resolved size midpoints gives $\widehat{\nu} = 0.222$ for $\alpha_c(N) \propto N^{-\nu}$ (Eq. (1)).

This estimate is stable across sensitivity analyses. A censoring-relaxed analysis gives $\widehat{\nu} = 0.223$, with a 95% CI of $[0.152, 0.283]$, while requiring all six midpoints to be resolved gives a conditional interval of $[0.206, 0.277]$. Replacing the target grid with the realized fractions K/N changes the estimate only from 0.222 to 0.224. All estimates remain

Table 2: Fixed-count response surface audit. The contour slope is $\nu_{\text{response}} = 1 + b_N/b_K$; the observed collapse boundary exponent is $\widehat{\nu} = 0.222$.

Response subset	b_N	b_K	ν_{response}
All fixed-K	-0.784	1.575	0.503
Subcritical cells	-0.894	1.683	0.469
Zero-failure cells	-0.950	1.817	0.477

Note. Subcritical cells have fixed-K failure prevalence below $1/2$. Zero-failure cells contain no failures across twelve seeds. The slopes describe the continuous risk surface.

within $0 < \nu < 1$, indicating that the collapse boundary decreases with society size while the effective critical count grows more slowly than society size itself.

Equivalently, $K_c(N) \propto N^{0.778}$, with a 95% CI of $[0.717, 0.848]$ for the count exponent (Figure 3). At the four sizes whose sampled grids bracket both $p = 0.1$ and $p = 0.9$, the 10%–90% transition width ranges from 0.013 to 0.028, confirming that the rise from rare to frequent collapse occurs over a relatively narrow harmful-fraction range.

5.2 Fixed-Count Response Scaling

The boundary analysis varies harmful fraction, so the absolute number of harmful agents also changes with society size. We therefore ask a complementary question: when the harmful count is held fixed, how does continuous risk change as the society grows?

Across the full response surface, $b_N = -0.784$ and $b_K = 1.575$. The negative size coefficient shows that a fixed harmful count produces lower episode-level peak risk in a larger society, whereas the positive count coefficient shows that risk increases as more harmful agents are introduced.

This relationship remains after removing cells at or above the transition. Across 300 episodes in 25 subcritical cells, $b_N = -0.894$ (paired-bootstrap 95% CI $[-0.960, -0.852]$) and $b_K = 1.683$ (95% CI $[1.664, 1.763]$). These coefficients imply a response-contour slope of $\nu_{\text{response}} = 0.469$ (95% CI $[0.435, 0.501]$). Restricting the analysis further to cells with no observed failures gives a similar slope of 0.477 (Table 2). Thus, the size-dependent decline is already present below the binary collapse threshold.

A strict leave-one-size-out audit fits the subcritical surface on five sizes and extrapolates to the $R_{S1} = 1$ contour without using collapse midpoints. It preserves the boundary direction and rank (Spearman $= 0.943$) but yields a slope of 0.463, steeper than the observed boundary exponent of 0.222 and a mean absolute log error of 0.390. Thus, the response surface captures the direction of finite-size change but not the boundary exponent quantitatively.

5.3 Feedback and Reach Shift the Boundary

The largest shifts in the collapse boundary occur under the two composite feedback bundles and increased network reach. The weak-feedback bundle moves the boundary rightward by $+0.046$ in α. Increased network reach and the strong-feedback bundle move it leftward by -0.017 and -0.018,

Figure 4: Intervention effects on the collapse boundary. Points show the mean shift across the matched sizes $N \in \{300, 1000\}$, and intervals are 95% CIs from a paired bootstrap over seeds. Negative shifts indicate greater fragility; positive shifts indicate greater robustness.

Table 3: Controller realization audits over the common resolved range of five sizes, $N \leq 1000$.

Condition	Exponent [95% CI]	Boundary evidence
All-rule	0.164 $[-0.009, 0.251]$	$0.0475 \to 0.0300$
Standard quota	0.181 $[-0.063, 0.289]$	$0.0500 \to 0.0300$
Increased quota	0.190 $[-0.081, 0.276]$	$0.0475 \to 0.0280$

Note. Estimates are exponents over the common range. Boundary evidence reports $\alpha_c(100) \to \alpha_c(1000)$.

respectively (Figure 4). The paired-bootstrap 95% CI for each effect excludes zero, and each reported shift is averaged over the two matched sizes. By contrast, the interventions on attention, conformity, and response precision produce small rightward shifts of $+0.002$, $+0.003$, and $+0.004$. Their confidence intervals include zero.

The conformity and reach interventions further separate information dominance from collapse-relevant feedback. High conformity raises D_N by 0.183 on average (paired-bootstrap 95% CI $[0.166, 0.203]$) yet barely moves the boundary. Increased network reach changes D_N little but shifts the boundary leftward. Thus, social information becomes collapse-relevant when the induced actions feed back into shared state, not merely when social messages predict individual actions.

5.4 Controller Realization Audits

The all-rule condition disables all LLM calls while holding the remaining simulator settings fixed. Across the common range of five sizes, all-rule, standard-quota, and increased-quota controllers preserve the same endpoint direction (Table 3): the collapse boundary decreases while the effective critical count increases. The exponent estimates are similar, but their six-seed intervals are wide and include zero.

Table 4: Reduced cross-backbone robustness audit for the primary S1 scaling setting. Each backbone uses the same protocol and the two society sizes $N = 100$ and $N = 1000$.

Backbone	$\alpha_c(100)$	$\alpha_c(1000)$	$\widehat{\nu}_{2pt}$
GPT-4.1	0.0525	0.0260	0.305
Qwen3-235B	0.0550	0.0260	0.325
Llama-3.3-70B	0.0550	0.0260	0.325
GLM-4.5	0.0550	0.0260	0.325
Kimi K2.6	0.0525	0.0280	0.273

Note. $\widehat{\nu}_{2pt} = \log_{10}(\alpha_c(100)/\alpha_c(1000))$. These values are descriptive two-point slopes, not substitutes for the primary exponent fit across six sizes.

A reduced two-size audit replaces DeepSeek V3.2 with five alternative LLMs while holding the remaining S1 protocol fixed. Every backbone preserves the direction of the boundary shift from $N = 100$ to $N = 1000$, with two-point slopes from 0.273 to 0.325 (Table 4). Together, these audits show that the endpoint direction is stable across the tested controller allocations and backbones. Both audits are descriptive robustness checks; the primary sweep across six sizes provides the exponent estimate.

5.5 Transfer Across Manipulation Mechanisms

The S2–S4 runs assess qualitative transfer beyond the primary mechanism. Each run uses $N \in \{300, 1000\}$ and twelve seeds per cell. S2 transitions within $\alpha \in [0.003, 0.010]$ at both sizes, while S3 transitions within $\alpha \in [0.10, 0.15]$. S4 does not reach the collapse threshold over the tested range $\alpha \leq 0.12$. Its maximum episode-level peak risk is 0.083. The nonlinear collapse transition therefore transfers qualitatively to S2 and S3 but not to S4 under the tested protocol. Because the attack channels and score definitions differ, this comparison concerns transition presence rather than absolute boundary values. S4 remains a useful negative result: increasing harmful composition alone does not guarantee that every manipulation mechanism reaches the prespecified society-level collapse event within thirty days.

6 Discussion

Why the fraction falls while the count grows. The falling fraction and rising count are complementary features of the same size-dependent transition. Over the tested range, this pattern is inconsistent with both a fixed-count trigger and a threshold proportional to society size. Below collapse, a fixed harmful count produces less severe diffusion and market disruption in larger societies. This dilution partly reflects the baseline liquidity scaling of $N^{0.5}$, while the interventions show that endogenous feedback also moves the boundary. Because the response contour is steeper than the observed boundary, the subcritical response captures the direction, but not the full magnitude, of the transition.

Why information dominance is not enough. Social information becomes dangerous when the actions it induces feed back into shared state. Increasing conformity makes social evidence more predictive of action but barely moves the boundary, whereas increased network reach shifts the boundary with little change in D_N. Within the tested interventions, fragility is therefore more closely associated with network and environmental feedback than with information dominance alone.

Implications for the safety of agent systems. Isolated-agent evaluations can miss failures that emerge through shared-state feedback. In the tested financial society, feedback strength and network reach are collapse-relevant, whereas greater reliance on social information alone does not substantially move the boundary. These results motivate defenses that limit harmful exposure and propagation and interrupt the feedback that reproduces socially induced actions.

7 Scope, Limitations, and Ethics

Our conclusions are limited to the tested protocol: a controlled social–market society with thirty-day episodes, static within-episode graphs, and archetypal roles. The primary finite-size estimate comes from S1, while S2–S4 establish only qualitative transfer across manipulation channels. Controller audits cover rule-based and quota-limited hybrid societies, not fully LLM-controlled populations. Absolute collapse boundaries should therefore be interpreted as protocol-specific rather than universal.

The fixed-count surface describes behavior below collapse. Its held-out predictions preserve the direction and rank of the observed boundary but do not recover its exponent quantitatively. The theoretical susceptibility is an organizing construct, whereas the failure-gain proxy is a normalized empirical summary rather than a direct measure of microscopic amplification. Adaptive networks, longer horizons, alternative collapse definitions, and other societal domains remain open directions.

The scenarios are derived from public manipulation cases and described only at the mechanism level. We do not name specific securities, reproduce operational procedures, or claim that the simulations represent real-market deployment.

8 Conclusion

This work frames interacting-agent safety as a society-level problem. We introduce *Agent Society Dynamics*, a finite-size framework that relates the collapse boundary to harmful-pressure scaling and shared-state feedback, and evaluate it in a controlled social–market society.

We find that larger societies collapse at a smaller harmful fraction, while the effective harmful count at the boundary grows sublinearly. Fixed-count analyses reveal dilution below the transition, whereas feedback and network reach interventions shift the boundary. These results show that agent safety should be evaluated jointly across harmful composition, population size, and feedback structure.

References

Acemoglu, D.; Ozdaglar, A.; and Tahbaz-Salehi, A. 2015. Systemic Risk and Stability in Financial Networks. *American Economic Review*, 105(2): 564–608.

Banerjee, A. V. 1992. A Simple Model of Herd Behavior. *The Quarterly Journal of Economics*, 107(3): 797–817.

Barber, M. N. 1983. Finite-Size Scaling. In Domb, C.; and Lebowitz, J. L., eds., *Phase Transitions and Critical Phenomena*, volume 8, 145–266. Academic Press.

Brock, W. A.; and Durlauf, S. N. 2001. Discrete Choice with Social Interactions. *The Review of Economic Studies*, 68(2): 235–260.

Byrd, D.; Hybinette, M.; and Balch, T. H. 2019. ABIDES: Towards High-Fidelity Market Simulation for AI Research. arXiv:1904.12066.

Castellano, C.; Fortunato, S.; and Loreto, V. 2009. Statistical Physics of Social Dynamics. *Reviews of Modern Physics*, 81(2): 591–646.

Cemri, M.; et al. 2025. Why Do Multi-Agent LLM Systems Fail? arXiv:2503.13657.

Centola, D.; and Macy, M. 2007. Complex Contagions and the Weakness of Long Ties. *American Journal of Sociology*, 113(3): 702–734.

Cover, T. M.; and Thomas, J. A. 2006. *Elements of Information Theory*. Wiley, 2 edition.

De Marzo, G.; Bellina, A.; Castellano, C.; Priesemann, V.; and Garcia, D. 2026. Conformity Generates Collective Misalignment in AI Agents Societies. arXiv:2605.10721.

Elliott, M.; Golub, B.; and Jackson, M. O. 2014. Financial Networks and Contagion. *American Economic Review*, 104(10): 3115–3153.

Erisken, S.; Gothard, T.; Leitgab, M.; and Potham, R. 2025. MAEBE: Multi-Agent Emergent Behavior Framework. arXiv:2506.03053.

Fisher, M. E. 1971. The Theory of Critical Point Singularities. In Green, M. S., ed., *Critical Phenomena*, volume 51 of *Proceedings of the International School of Physics "Enrico Fermi"*, 1–99. Academic Press.

Granovetter, M. 1978. Threshold Models of Collective Behavior. *American Journal of Sociology*, 83(6): 1420–1443.

Haldane, A. G.; and May, R. M. 2011. Systemic Risk in Banking Ecosystems. *Nature*, 469: 351–355.

Huang, Y.; Jiang, Y.; Wang, W.; Zhuang, H.; Luo, X.; Ma, Y.; Xu, Z.; Chen, Z.; Moniz, N.; Lin, Z.; Chen, P.-Y.; Chawla, N. V.; Dziri, N.; Sun, H.; and Zhang, X. 2026. Emergent Social Intelligence Risks in Generative Multi-Agent Systems. arXiv:2603.27771.

McKelvey, R. D.; and Palfrey, T. R. 1995. Quantal Response Equilibria for Normal Form Games. *Games and Economic Behavior*, 10(1): 6–38.

Motwani, S.; Baranchuk, M.; Strohmeier, M.; Bolina, V.; Torr, P. H. S.; Hammond, L.; and Schroeder de Witt, C. 2024. Secret Collusion among AI Agents: Multi-Agent Deception via Steganography. In *Advances in Neural Information Processing Systems*.

Nakamura, M.; Kumar, A.; Das, S.; Abdelnabi, S.; Mahmud, S.; Fioretto, F.; Zilberstein, S.; and Bagdasarian, E. 2026. Colosseum: Auditing Collusion in Cooperative Multi-Agent Systems. arXiv:2602.15198.

Park, J. S.; O'Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.

Piao, J.; Yan, Y.; Zhang, J.; Li, N.; Yan, J.; Lan, X.; Lu, Z.; Zheng, Z.; Wang, J. Y.; Zhou, D.; et al. 2025. AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviors and Society. arXiv:2502.08691.

Piatti, G.; Jin, Z.; Kleiman-Weiner, M.; Schölkopf, B.; Sachan, M.; and Mihalcea, R. 2024. Cooperate or Collapse: Emergence of Sustainable Cooperation in a Society of LLM Agents. arXiv:2404.16698.

Schelling, T. C. 1971. Dynamic Models of Segregation. *Journal of Mathematical Sociology*, 1(2): 143–186.

Sims, C. A. 2003. Implications of Rational Inattention. *Journal of Monetary Economics*, 50(3): 665–690.

U.S. Commodity Futures Trading Commission. 2021. CFTC Orders Coinbase Inc. to Pay $6.5 Million for False, Misleading, or Inaccurate Reporting and Wash Trading. Press Release 8369-21.

U.S. Department of Justice. 2020. JPMorgan Chase & Co. Agrees To Pay $920 Million in Connection with Schemes to Defraud Precious Metals and U.S. Treasuries Markets. Office of Public Affairs Press Release.

U.S. Securities and Exchange Commission. 2022. SEC Charges Eight Social Media Influencers in $100 Million Stock Manipulation Scheme Promoted on Discord and Twitter. Press Release 2022-221.

Watts, D. J. 2002. A Simple Model of Global Cascades on Random Networks. *Proceedings of the National Academy of Sciences*, 99(9): 5766–5771.