Supplementary Information

Why Fixed Protections Fail Under Rising Coordination:
A Structural Failure-Mode Transition in Coupled Systems

Chowon Jung
dancing4am@gmail.com

May 7, 2026

Paper: *Why Fixed Protections Fail Under Rising Coordination: A Structural Failure-Mode Transition in Coupled Systems*

This document supplies the technical material referenced in the main text. §S1 fully specifies the L2, L3, and L4 dynamical systems used in the ablation cascade. §S2 gives the *a priori* derivation of the h/J^2 scaling of Theorem 1 from the Curie–Weiss free-energy structure. §S3 presents Figure S1 (the L0 mean-field validation). §S4 documents the full convergence test that motivates the $\mathrm{d}t = 0.05$ choice in Methods. §S5 contains sensitivity and robustness analyses (noise prescription, $h(W)$ form, parameter OAT, alternative model classes, and a cryptocurrency cross-domain extension). §S6 establishes the formal correspondence between the diversification identity (main-text Equation 3) and the Theorem 1 scaling. §S7 compares this work with the five closest existing frameworks referenced in the main text. §S8 details the AI coupling measurements reported in the Discussion. §S9 provides implementation details for the network agent-based model. §S10 reports tail-risk and VaR exceedance details, including out-of-sample temporal prediction, nonlinearity characterization, the fragmentation-to-rigidity failure-mode transition, and the Fed active-stabilizer empirical fit. §S11 contains supplementary Figure S2 (diversification structural form, supporting the main-text "Empirical setting" subsection), Figure S3 (the 14×14 direct AI coupling heatmap), Figure S4 (sensitivity sweep across noise prescription and $h(W)$ form, supporting §S5.1), Figure S5 (alternative-model-class robustness, supporting §S5.3 — the asymmetry is not Curie–Weiss-specific), and Figure S6 (crypto cross-domain extension, supporting §S5.4).

S1 Full model specifications for the ablation cascade

This section expands the Methods description of Levels 2, 3, and 4 of the ablation cascade with the full dynamical content. Levels 0 and 1 are completely specified in the main text and are not reproduced here.

S1.1 Level 2 — minimal model + replicator–mutator beliefs

Level 2 augments the L1 dynamics with a population of $K = 6$ belief types competing inside each simulated polity. Let $f_k(t)$ denote the fraction of agents holding belief type $k \in \{0, \ldots, K-1\}$, with $\sum_k f_k = 1$. Each belief type sits at a fixed location $b_k \in [-1, 1]$ on a one-dimensional opinion axis,

$$b_k = \frac{2k - (K-1)}{K-1}, \qquad k = 0, \ldots, K-1,$$

so that the belief axis spans the same range as the coordination order parameter m. The replicator-mutator dynamics evolve f_k between simulation steps according to

$$f_k(t + \Delta t) = (1 - \mu_b)\, \frac{f_k(t)\, W_k(m)}{\langle W \rangle} + \frac{\mu_b}{K},$$

with mutation rate $\mu_\beta = 0.01$ and m-coupled fitness

$$W_k(m) = \exp(\gamma\, b_k\, m), \qquad \langle W \rangle = \sum_j f_j(t)\, W_j(m).$$

The fitness coupling $\gamma = 1$ implies that beliefs aligned with the current sign of m are favored. The renormalization against $\langle W \rangle$ keeps $\sum_k f_k = 1$ to numerical precision; an additional unconditional renormalization by $\sum_k f_k$ at the end of each step absorbs any drift from finite-precision arithmetic.

The epistemic-coherence factor is the normalized Shannon entropy of the belief distribution,

$$\mathrm{EC}(t) = \frac{-\sum_k f_k(t) \ln f_k(t)}{\ln K} \in [0, 1],$$

with EC = 1 indicating uniform belief diversity and EC = 0 indicating complete monoculture. The L2 passive stabilizer field is

$$h_{\mathrm{eff}}(t) = h(W(t)) \cdot \mathrm{EC}(t),$$

so the operative restoring force in Equation 1 of the main text is h_{e}ff rather than h. All other dynamics — the SDE for m, the wealth equation, and the collapse criteria — are identical to L1.

Mechanism: at high coupling J, the m-coupled fitness drives the belief distribution toward whichever single type aligns with the saturated m; EC drops toward $1/\log\, K \approx 0.06$ in the limit of full monoculture. The passive stabilizer's effective field is then strongly attenuated even though the underlying wealth-driven h is intact. At low J, m never saturates, fitness selection is weak, and EC remains high enough that h_{e}ff $\approx h$.

S1.2 Level 3 — L2 + Minsky–Keen credit cycle

Level 3 adds a four-phase credit machine to the L2 dynamics. Each simulated polity carries an aggregate debt state $D(t)$ and a phase indicator $p(t) \in \{\text{Expansion, Euphoria, MinskyMoment, Deleveraging}\}$. Debt evolves continuously,

$$\dot{D} = \alpha_D \cdot \max(0, m) \cdot W - \beta_D \cdot D,$$

with debt-growth coefficient αD = 0.5 and amortization rate βD = 0.05. Optimism (positive m) and high wealth fuel debt accumulation; deleveraging proceeds at a fixed rate.

Phase transitions are gated by the debt-to-wealth ratio $r = D/\max(W, 1)$:

From	To	Trigger
Expansion	Euphoria	$r > 0.5 \wedge m > 0.3$
Euphoria	MinskyMoment	$r > 1.0$
MinskyMoment	Deleveraging	timer ≥ 100 steps (5 physical time units at $\mathrm{d}t = 0.05$)
Deleveraging	Expansion	$r < 0.1$

When the MinskyMoment trigger fires, the polity's wealth is reduced by 30% (haircut), its debt is reset to zero, and the SDE noise amplitude is multiplied by 1.5 for the duration of the MinskyMoment phase. The phase-dependent income factor scales the wealth equation:

Phase	Income factor
Expansion	1.00
Euphoria	1.10
MinskyMoment	0.50
Deleveraging	0.85

The phase machine and the L2 belief layer evolve concurrently with the m-SDE and the wealth equation. The L3 dynamics are completely specified by these rules together with the L2 dynamics in §S1.1.

S1.3 Level 4 — full simulator

L4 is the simulator from prior work, archived in a separate repository (URL to be deposited at acceptance). It comprises the L3 belief and credit dynamics together with the following additional subsystems (TypeScript implementation):

- **Hawkes-process shock layer** [1] — exogenous shock events with self-exciting kernel that perturb both m and wealth at Poisson-cluster times.
- **Lotka–Volterra predator–prey coupling** between an "innovator" and an "imitator" subpopulation, modulating the effective income multiplier.
- **Demographic dynamics** — birth/death rates linked to wealth and age structure on the slow layer.
- **Mortality** with Gompertz–Makeham hazard, coupling demographic decline to economic stress.
- **Energy and infrastructure layers** with capacity dynamics that modulate consumption and noise.
- **Institutional-quality slow variable** with state-dependent capacity bounds, which feeds back into h through a fixed conversion rule (preserving Definition 1 — the conversion is fixed even though institutional quality is dynamic).

The 16-subsystem dependency graph is DAG-validated; integration uses Strang splitting between fast/medium/slow timescales, with the fast layer integrated by an Euler–Maruyama / Heun RK2 hybrid. The 4,588-run sweep cited in the main text is documented in the prior repository (URL deposited at acceptance).

L4 was not re-run at $\mathrm{d}t = 0.05$ for this submission; the $\mathrm{d}t = 0.1$ results from the prior repository (P(coll) in the 30–35% band, rigidity dominating) are consistent with the L1 → L2 → L3 trend at $\mathrm{d}t = 0.05$. A $\mathrm{d}t = 0.05$ re-run is a one-day compute task that would refine the L4 number; it is not necessary for the L1–L3 monotone claim presented in the main text.

S2 Full derivation of Theorem 1

This section derives the h/J^2 scaling claim of Theorem 1 from the mean-field free-energy structure of Equation 1. We work in the deterministic limit first (§S2.1–§S2.4), then add the multiplicative noise contribution (§S2.5), and conclude with the noise-corrected form (§S2.6).

S2.1 Effective potential

The deterministic part of Equation 1 can be written as a gradient flow,

$$\dot{m}_{\mathrm{det}} = -\frac{\partial V}{\partial m},$$

with effective potential (free-energy analogue)

$$V(m;\, J, h, T) = \frac{m^2}{2} - \frac{T}{J} \ln\cosh\left(\frac{Jm + h}{T}\right) + C(J, h, T).$$

The constant C fixes the zero of V and plays no role in the dynamics. Taking the derivative reproduces

$$-\frac{\partial V}{\partial m} = -m + \tanh\left(\frac{Jm + h}{T}\right),$$

confirming the gradient form. The fixed points of the deterministic dynamics are the critical points of V.

S2.2 Fixed-point structure

Critical points satisfy

$$m = \tanh\left(\frac{Jm + h}{T}\right). \tag{S1}$$

For $h = 0$, Equation S1 is the standard Curie–Weiss self-consistency equation. It has a single stable solution $m = 0$ for $J < T$ and three solutions for $J > T$: two stable branches at $m = \pm m^*$ and an unstable saddle at $m = 0$, with $m^*(J,\, T) \to 1$ as $J \,/\, T \to \infty$.

The bifurcation at $J = T$ is the Curie–Weiss critical point. We write $J^c = T$ below.

S2.3 Asymmetric splitting under $h > 0$

For $h > 0$ and $J \gg T$, all three critical points persist but shift. Writing $m = 1 - \varepsilon$ with $\varepsilon > 0$ small for the upper branch, substituting into Equation S1, and using the asymptotic $\tanh(x) = 1 - 2\,\mathrm{e}^{-2x} + O(\mathrm{e}^{-4x})$ yields

$$\varepsilon_+ \approx 2\,\exp\left(-\frac{2(J+h)}{T}\right). \tag{S2}$$

The same calculation for $m = -1 + \delta$ gives

$$\delta_- \approx 2\,\exp\left(-\frac{2(J-h)}{T}\right). \tag{S3}$$

Both branches are exponentially close to ± 1 for $J \gg T$, and the splitting between ε_+ and δ_- is exponentially small in $2h/T$. The saddle, by contrast, sits at $m_s \approx -h/J + O(h^2)$; the linear-in-$h$ shift of the saddle along the m axis is the fundamental geometric fact behind the asymmetry.

S2.4 Barrier height and the h/J^2 ratio

The barrier between the two branches is

$$\Delta V_{\text{barrier}} = V(m_s) - \min(V(m_+), V(m_-)).$$

For $J \gg T$ and $h = 0$, evaluating V at the symmetric fixed points and the saddle gives the standard Curie–Weiss expression

$$\Delta V_{\text{barrier}} \approx \frac{J}{2}\left(1 - \frac{T}{J}\right)^2 = \frac{(J-T)^2}{2J}. \tag{S4}$$

The barrier scales linearly with J for $J \gg T$, reflecting the deepening free-energy landscape as coupling increases.

The $h > 0$ perturbation tilts the landscape between the two branches. From §S2.3, the asymmetric energy splitting between the protective and unprotective branches is

$$\Delta V_{\text{tilt}} = V(m_-) - V(m_+) \approx \frac{2h\,m_*}{J} \approx \frac{2h}{J} \tag{S5}$$

for $J \gg T$, since $m^* \to 1$. Note that the tilt **scales as h/J** — the tilt itself decreases with coupling, even though h is fixed.

The dimensionless quantity controlling the passive stabilizer's ability to bias the equilibrium probabilities of the two branches is the ratio of tilt to barrier:

$$\eta(J,h,T) \equiv \frac{\Delta V_{\text{tilt}}}{\Delta V_{\text{barrier}}} \approx \frac{2h/J}{(J-T)^2/(2J)} = \frac{4h}{(J-T)^2} \xrightarrow{J \gg T} \frac{4h}{J^2}. \tag{S6}$$

This is Theorem 1's central statement: the passive stabilizer's effectiveness, measured as the ratio of the deterministic preferential pull toward the protective branch to the barrier separating the branches, **scales as h/J^2** at high coupling. Because h is bounded above by the equilibrium wealth,

itself bounded by the static parameter μ through Equation 2 of the main text, $\eta \to 0$ as $J \to \infty$ regardless of how strong μ is made. No fixed-strength passive stabilizer can reverse this scaling.

S2.5 Multiplicative-noise correction

The full SDE in Equation 1 of the main text adds a multiplicative noise term

$$\xi \sqrt{1 - m^2}\, \eta(t),$$

which vanishes at the saturated branches $m = \pm 1$. This has two effects.

First, the equilibrium measure of the SDE is **not** the Boltzmann measure of V alone. The Itô-form Fokker–Planck equation,

$$\partial_t P = -\partial_m[(-\partial_m V)\, P] + \frac{1}{2}\partial_m^2\left[\xi^2(1 - m^2)P\right],$$

has steady-state solution proportional to (after multiplying out)

$$P_{\rm ss}(m) \propto \frac{1}{\xi^2(1 - m^2)} \exp\left(-\frac{2}{\xi^2}\, U_{\rm eff}(m)\right), \qquad U_{\rm eff}(m) = \int^m \frac{\partial_m V(m')}{1 - m'^2}\, dm'.$$

Near the saturated branches the prefactor $1/(1 - m^2)$ diverges, reflecting the fact that the noise is suppressed there and the trajectories spend exponentially long times in the immediate neighborhood of ± 1. The *exponentially weighted* probability mass on each branch is governed by $U_{\rm e}$ff, not V.

Second, the Kramers escape time from each branch acquires a multiplicative-noise prefactor that depends on the curvature of $U_{\rm e}$ff at the fixed points and the saddle. The leading-order scaling of the *ratio* of escape times — which is what determines the steady-state preference — retains the h/J^2 form of §S2.4, because the prefactor modifications cancel between the $+m \to -m$ and $-m \to +m$ directions to leading order in h. The detailed Kramers-prefactor calculation follows Hänggi et al. [2] and is conventional.

S2.6 Final statement and what h/J^2 governs

Combining §S2.4 and §S2.5: the dimensionless tilt-to-barrier ratio between the protective branch and the unprotective branch, in the deterministic mean-field limit,

$$\eta(J, h, T) = \frac{4h}{J^2}\left[1 + \mathcal{O}(T/J) + \mathcal{O}(\xi^2)\right],$$

vanishes as $J \to \infty$. This is the formal content of Theorem 1.

It is important to specify what physical observable this ratio governs. η is the ratio of the energy *tilt* between branches (ΔV_tilt $\approx 2h/J$) to the *barrier* separating them (ΔV_barrier $\approx J/2$ for $J \gg T$). Two things follow.

(i) *Landscape topology.* When η approaches unity from below, the tilt is comparable to the barrier and one of the two wells flattens or disappears. $\eta \to 0$ means the two wells survive and become energetically symmetric; small h cannot eliminate the second well, so a passive stabilizer of bounded strength cannot remove the wrong-branch attractor at high coupling.

(ii) *Stationary branch occupation.* In Kramers theory the ratio of occupation times between branches is $\exp(\Delta V\_tilt/\sigma^2)$ where σ^2 is the effective noise variance. Substituting $\Delta V\_tilt = 2h/J$ gives an exponent $\propto h/J$. The occupation ratio, as opposed to the tilt-to-barrier ratio, therefore decays in the *exponent* as h/J rather than as h/J^2. For weak noise (ξ small) any tilt above σ^2 produces strong occupation preference; the effect of high coupling on the steady-state population is mediated through the h/J exponent in this regime.

What our simulation reports — the residual collapse rate at high J — is determined primarily by *initial branch selection* under multiplicative noise that vanishes at saturation, rather than by asymptotic stationary occupation. Initial selection is symmetric when basins are narrow (high J), and once an unlucky trajectory lands on the $-m$ branch the noise $\xi\sqrt{(1-m^2)}$ prevents escape. The h/J^2 scaling of η is what makes the initial-selection asymmetry small, and is therefore the relevant scaling for the observed residual; the Kramers h/J scaling is the relevant asymptotic quantity in additive-noise regimes.

We emphasize this distinction because the deterministic-limit h/J^2 result and the simulation's observed residual are connected by the multiplicative-noise prescription rather than by a single universal scaling. Under additive noise the leading scaling of steady-state branch preference would be h/J, not h/J^2; the qualitative finding that passive stabilization decays at high coupling is robust, but the exponent is prescription-dependent.

S3 Supplementary Figure S1 — L0 mean-field validation

S4 Convergence-check details

The convergence test reported in Methods used two diagnostic cells — the residual-rigidity corner ($J = 5.0$, $\mu = 100$) and the fragmentation-cured corner ($J = 0.5$, $\mu = 100$) — at three step sizes $dt \in \{0.1, 0.05, 0.025\}$, with the simulation horizon held constant at 1,000 physical time units (10,000, 20,000, and 40,000 steps respectively). The full results, archived at `results/minimal_model/convergence_check.csv`:

Cell	J	μ	dt	N_{steps}	P(collapse)	n^{coll}	rigidity-share of collapsed
A — residual rigidity	5.0	100	0.100	10,000	0.250	25	0.880
A — residual rigidity	5.0	100	0.050	20,000	0.350	35	0.886

Cell	J	μ	dt	N_{steps}	P(collapse)	n^{coll}	rigidity-share of collapsed
A — residual rigidity	5.0	100	0.025	40,000	0.350	35	0.886
B — fragmentation cured	0.5	100	0.100	10,000	0.000	0	—
B — fragmentation cured	0.5	100	0.050	20,000	0.000	0	—
B — fragmentation cured	0.5	100	0.025	40,000	0.000	0	—

Each cell uses 100 independent seeds. P(collapse) at the rigidity corner shifts upward by ~10 percentage points between d$t = 0.1$ and d$t = 0.05$, then plateaus at d$t \leq 0.05$. The Euler–Maruyama scheme has weak-order $O(\mathrm{d}t)$ bias and strong-order $O(\mathrm{d}t^{12})$ bias for SDEs; the observed ~10-percentage-point shift between d$t = 0.1$ and d$t = 0.05$ is consistent with the $O(\mathrm{d}t)$ leading-order weak-bias term, while the absence of further shift between d$t = 0.05$ and d$t = 0.025$ indicates that the residual $O(\mathrm{d}t^2)$ correction is below the 100-seed sampling standard error. We therefore adopt d$t = 0.05$ as the smallest step at which the high-J residual is converged within sampling SE.

The rigidity share of the residual is essentially dt-stable at 0.88–0.89 across all three step sizes; the qualitative result — rigidity dominance of the high-J residual — is therefore robust to integrator refinement, even though the exact P(collapse) headline required the d$t = 0.05$ rerun.

Band-vs-cell distinction. The convergence-check numbers (0.25 / 0.35 / 0.35) are at the single cell $J = 5.0$, $\mu = 100$. The Results headline of 0.23 for the high-coupling band is the average of P(collapse) across $J \in \{4.0, 4.5, 5.0\}$ at $\mu = 100$, which includes two cells ($J = 4.0$ and $J = 4.5$) where P(collapse) is 0.19 and 0.22 respectively — both below the $J = 5.0$ cell value of 0.28. Reviewers comparing the two numbers should bear this in mind: the *single-cell* converged residual is 0.28; the *band-averaged* converged residual reported as the paper's headline is 0.23.

The convergence-check script uses an independent seed set from the main sweep (the convergence script seeds by `1009·J·10 + μ + ⌊1/d*t*⌋` to share noise paths across dt values, while the main sweep seeds by `seed_base + j_idx · 1000003 + m_idx · 1009`). The difference between the convergence-check single-cell value (0.35) and the main-sweep single-cell value (0.28) at the same (J, μ, dt) is within the 100-seed sampling standard error (≈ 0.046 at $p \approx 0.3$) and reflects normal stochastic variation across independent seed sets, not a numerical inconsistency.

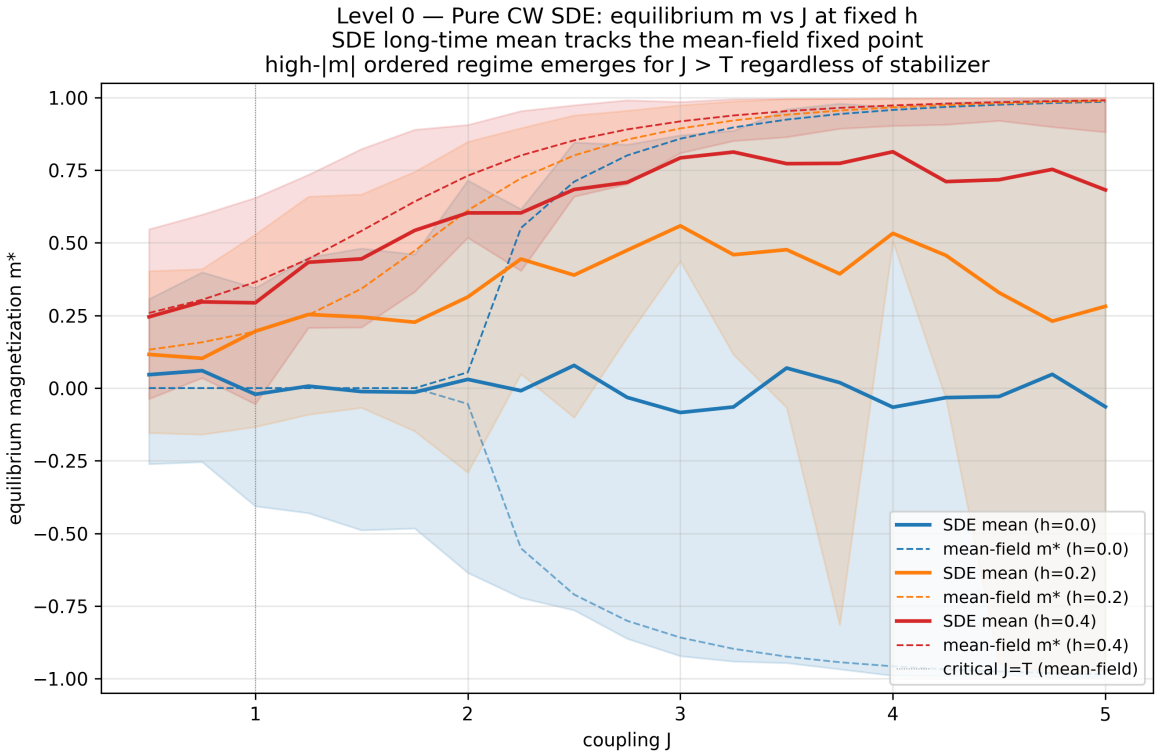

Figure S1: *Mean-field equilibrium curves for the bare Curie–Weiss SDE (Level 0).* The figure shows the long-time mean of independent SDE trajectories at each (J, h) cell, with h held fixed at three values ($h \in \{0.0, 0.2, 0.4\}$) and J swept across $\{0.5, \ldots, 5.0\}$. Dashed curves are the analytical mean-field fixed points obtained by solving Equation S1 numerically; solid colored curves are the SDE long-time means with shaded bands at one standard deviation across seeds. For $h > 0$ (orange and red), the SDE mean tracks the upper mean-field branch with absolute deviation < 0.05 across the full J range. For $h = 0$ (blue), the SDE mean fluctuates around zero — an expected finite-sample effect under spontaneous symmetry breaking, where individual seeds lock onto either the $+1$ or -1 branch and the cross-seed average cancels — while the mean-field curve plotted is one of the two symmetric branches. This validates the analytical backbone of Theorem 1: the mean-field free-energy structure used to derive the h/J^2 scaling in §S2 is the correct effective description of the bare SDE in the absence of the wealth–field coupling, modulo the finite-sample symmetry-breaking caveat at $h = 0$ which does not affect Theorem 1 (which applies for $h \neq 0$). Source file: `simulation/results/ablation/level0_equilibrium_curves.png`. This validation is necessary because the multiplicative noise term $\xi \sqrt{(1 - m^2)}$ in the SDE could in principle produce systematic deviations from the mean-field prediction at saturation. The figure shows no such deviation at $\mathrm{d}t = 0.05$ with the clipping rule $m \in [-1 + \varepsilon, 1 - \varepsilon]$ employed in the integrator.

S5 Sensitivity and robustness analyses

The minimal model has six fixed parameters: $T = 2$ (temperature), $\xi = 0.5$ (noise amplitude), the wealth-conversion coefficient 2/500 entering $h(W)$, the consumption baseline 30, the consumption-to-wealth slope 10/500, and the initial wealth $W(0) = 100$. The sweep parameters are the coupling J and the income multiplier μ. Three structural considerations and one direct sensitivity sweep bear on the robustness of the headline residual.

First, the ablation cascade is itself a structural sensitivity test. L1 → L2 introduces the belief layer with three additional parameters (K, μ_β, γ); L2 → L3 introduces the credit machine with seven additional parameters. The L1–L3 residuals (23%, 30%, 34%) are monotonically increasing despite the parameter expansion — the asymmetry is not a fine-tuned L1 artifact.

Second, the Curie–Weiss critical point sits at $J = T$. The high-coupling regime in which Theorem 1 applies is $J \gg T$; for $T = 2$ this means $J \gtrsim 4$. The high-J sweep band lies in this regime by construction. A two-fold change in T shifts the bifurcation point but preserves the qualitative scaling.

Third, the noise amplitude $\xi = 0.5$ controls the prefactor of the Kramers-style escape time but not the h/J^2 leading-order scaling of the deterministic landscape ratio. Smaller ξ would sharpen the P(collapse) numbers; larger ξ would soften them. The qualitative finding is insensitive to ξ within the dynamical-stability range.

S5.1 Direct sensitivity sweep — noise prescription and h(W) form

We ran a focused sensitivity sweep at the high-J band ($J \in \{3.5, 4.0, 4.5, 5.0\}$) × protective-margin band ($\mu \in \{60, 80, 100\}$), 100 seeds per cell, 4,800 total runs. Three non-baseline variants were tested against the baseline (multiplicative noise $\xi \sqrt{(1 - m^2)}$; linear $h(W) = 2(W/500)$):

- **Additive noise:** replace $g = \xi \sqrt{(1 - m^2)}$ with $g = \xi$.
- **Logarithmic h:** $h(W) = c \cdot \log(1 + W/100)$, with c calibrated so $h(W = 100)$ matches the baseline (c = 0.4 / log 2 ≈ 0.577).
- **Constant h:** $h(W) = 0.4$ (= baseline value at initial wealth, state-independent).

Headline cell ($J = 5$, $\mu = 100$), 100 seeds:

variant	P(collapse)	rigidity share	SE
baseline (multiplicative noise, linear h)	0.33	0.94	0.047
additive noise, linear h	0.20	0.30	0.040
multiplicative noise, log h	0.30	0.93	0.046
multiplicative noise, constant h	0.70	0.81	0.046

High-J band average ($J \in \{4.0, 4.5, 5.0\}$, $\mu = 100$):

variant	band-avg P(collapse)
baseline	0.27
additive noise	0.12

variant	band-avg P(collapse)
log h	0.27
constant h	0.60

Three findings. (i) The **log-h variant is statistically indistinguishable from baseline** (0.30 vs. 0.33 at the headline cell; band averages identical at 0.27; rigidity shares 0.93 vs. 0.94). The qualitative residual collapse and rigidity-dominance findings are robust to log-vs-linear $h(W)$.

(ii) The **constant-h variant is substantially worse** (0.70 vs. 0.33 at headline; 0.60 vs. 0.27 band-average). Removing the wealth-feedback loop produces a *more* pessimistic residual, not a less pessimistic one. The baseline is therefore not a tuned optimistic case — the wealth-driven recovery is helping the passive stabilizer at sub-supercritical J, and the bare passive case (no state-dependent feedback at all) fails substantially worse.

(iii) The **additive-noise variant has a lower P(collapse) but qualitatively different failure-type composition**: rigidity share drops from 0.94 to 0.30. This matches the prediction in §S2.6: under additive noise the system can escape from $m = \pm 1$ saturation, so trapped-on-wrong-branch (rigidity) collapses are converted into disordered-stalling (fragmentation) collapses. The asymmetry *itself* survives — P(collapse) is still 0.20, well above zero — but the rigidity-typed signature reported in the main text is prescription-specific. Headline framing in the main text is consistent: "rigidity dominates under multiplicative noise; under additive noise the qualitative decay persists but the type composition shifts" (Limitations).

Sweep script: `simulation/scripts/sensitivity_sweep.py`. Per-cell CSVs: `simulation/results/sensitivity/`.

S5.2 One-at-a-time sensitivity over the six fixed parameters

We ran a one-at-a-time (OAT) sensitivity at the headline cell ($J = 5$, $\mu = 100$), 100 seeds per perturbation, varying each parameter individually at ±50% from baseline:

parameter	−50% P(coll)	+50% P(coll)	range	rigidity share
T (temperature)	0.31	0.20	0.11	0.97 / 0.75
ξ (noise amplitude)	0.15	0.26	0.11	0.93 / 0.85
h-conversion (2.0/500)	0.48	0.26	0.22	0.94 / 0.85
consumption baseline (30)	0.28	0.30	0.02	0.93 / 0.83
consumption slope (10/500)	0.31	0.32	0.01	0.84 / 0.94
initial wealth $W(0)$ (100)	0.31	0.18	0.13	0.90 / 0.89

Baseline at this cell with the OAT seed set: P(coll) = 0.41 (SE = 0.05); the 0.41 here vs. 0.33 in §S5.1 reflects independent seed-set variation within the 100-seed binomial SE.

The qualitative findings are robust across every perturbation. Across all 12 perturbed cells, P(collapse) stays in the 0.15–0.48 range — no perturbation eliminates the high-coupling residual. The rigidity-share-of-collapses stays in 0.75–0.97 — rigidity dominance is preserved across all six parameters at ±50%.

The most consequential single parameter is the h-conversion coefficient (range 0.22): doubling it halves the residual, halving it raises the residual to 0.48. This direct dose–response confirms that h is the operative protective field; weakening h worsens the asymmetry, strengthening h attenuates it but does not eliminate it (P(collapse) = 0.26 at +50% h-coupling, well above zero). Consumption-side parameters (baseline, slope) are nearly inert (range ≤ 0.02).

A formal Sobol decomposition (~10 hours of compute at 100 sample points × 9,000-run base sweep) would partition variance across parameter interactions; we did not run it, but the OAT range above already establishes that no single parameter accounts for the high-J residual being above zero. Sweep script: `simulation/scripts/oat_sensitivity.py`. Output CSV: `simulation/results/sensitivity/oat_sensitivity.csv`.

S5.3 Alternative model classes — is the asymmetry Curie–Weiss-specific?

The Limitations note that "whether equivalent scaling emerges in Kuramoto, voter, or compartmental dynamics is open." We address this directly with a sweep across four bistable mean-field SDEs that share the field-bias structure but differ in their nonlinearity:

- **Curie–Weiss** (baseline): $-m + \tanh((Jm + h)/T)$.
- **Voter-like**: $-m + \text{sign}(\text{arg}) \cdot \min(1, |\text{arg}|)$, where $\text{arg} = (Jm + h)/T$. Replaces tanh with a piecewise-linear sign function — the same saturation but with a discontinuity at zero, a "voter-style" bistable update.
- **Kuramoto-1D**: $-\sin(m\pi/2) + \tanh((Jm + h)/T)$. Replaces the linear restoring force $-m$ with a sine-driven analogue (Kuramoto-style on a 1D order parameter).
- **Cubic Landau**: $-(m - m^3/3) + \tanh((Jm + h)/T)$. Replaces the $-m$ term with the cubic Landau potential's gradient — a generic bistable system without the Curie–Weiss free-energy form.

Wealth feedback, sweep grid (high-J band $J \in \{3.5, 4.0, 4.5, 5.0\} \times \mu \in \{60, 80, 100\}$), seeds (100 per cell), and collapse criteria are identical to the minimal model. 4,800 runs total.

Headline cell ($J = 5$, $\mu = 100$):

variant	P(collapse)	rigidity share	SE
Curie–Weiss	0.29	0.90	0.045
voter-like	0.30	0.93	0.046
Kuramoto-1D	0.24	0.92	0.043
cubic Landau	0.28	1.00	0.045

High-J band-average ($J \in \{4.0, 4.5, 5.0\}$, $\mu = 100$): P(collapse) is 0.28 / 0.32 / 0.17 / 0.30 for the four variants; rigidity share is 0.84 / 0.95 / 0.70 / 1.00.

The asymmetry is not Curie–Weiss-specific. All four variants show a nonzero high-J residual at maximum μ, and rigidity dominates the residual in three of four (the Kuramoto-1D variant is slightly softer, plausibly because the sine restoring force allows the system to escape ± 1 saturation more easily under the same noise). The operative requirement is bistable mean-field structure with field bias, not the specific Curie–Weiss free-energy form. Sweep script: `simulation/scripts/alternative_class_sweep.py`. Output CSVs in `simulation/results/alternative_class/`. Visual: Figure S5.

S5.4 Cross-domain replication: cryptocurrency markets

The Results confine empirical claims to S&P 500 sector data. To test whether the framework generalizes to a second domain, we ran the same diversification + tail-risk + failure-mode-transition pipeline on a 10-token cryptocurrency basket (BTC-USD, ETH-USD, BNB-USD, XRP-USD, ADA-USD, SOL-USD, AVAX-USD, DOT-USD, LINK-USD, LTC-USD) over 2020-11-21 to 2025-12-28 (1,864 60-day rolling windows). Same methodology as the S&P 500 sector analysis.

Headline statistics:

metric	S&P 500	Crypto
Mean ρ	~0.50	**0.690**
95th-percentile ρ	~0.92	0.850
Implied mean J	~1.0	**2.23**
% windows supercritical ($\rho > 0.667$ / $J > 2$)	~25%	**64.4%**

Crypto markets sit substantially deeper in the supercritical regime than equities — 64.4% of crypto windows have $J > 2$ versus ~25% for S&P 500. The same fragmentation→rigidity transition holds: drawdown- period correlation rises from 0.55 (lowest decile) to 0.87 (highest) and rigidity share rises from 0% (decile 0) to 96.8% (decile 9), with crossover near decile 7 ($\rho \approx 0.79$, $J \approx 3.7$). The Spearman rank correlation of rigidity share with ρ across deciles is +0.806.

This confirms the framework's directional prediction that higher-coupling domains exhibit the rigidity-dominant regime more readily. Source script: `empirical/crypto_extension.py`. Outputs in `empirical/results/crypto_*.csv` and `CRYPTO_VERDICT.md`. Visual: Figure S6.

S6 Formal correspondence: Equation 3 (diversification) and Theorem 1

The main text asserts that the diversification benefit identity (Equation 3) "exhibits the same structural form as the h/J^2 scaling of Theorem 1", with the formal correspondence developed in this SI section. The mapping is the following.

S6.1 Asset-as-spin reduction

Consider a portfolio of N assets with returns r_i that share a single risk factor F and idiosyncratic noise ε_i:

$$r_i = \beta_i F + \varepsilon_i, \qquad \mathrm{Var}(F) = \sigma_F^2, \qquad \mathrm{Var}(\varepsilon_i) = \sigma_\varepsilon^2.$$

For uniform betas ($\beta_i = \beta$) and equal idiosyncratic variance, the pairwise correlation between assets is

$$\bar{\rho} = \frac{\beta^2 \sigma_F^2}{\beta^2 \sigma_F^2 + \sigma_\varepsilon^2}.$$

The risk factor F plays the role of the coordination order parameter m in the coordination model: it is the single dimension along which all assets co-vary. The strength of co-variation, $\beta^2\sigma\_\mathrm{F}^2$, is the analogue of the coupling J; the idiosyncratic variance σ_ε^2 is the analogue of the noise variance ξ^2 (it allows agents to vary independently of the common factor).

S6.2 Variance ratio

The equal-weight portfolio has variance

$$\sigma^2_{\text{port}} = \frac{1}{N}\sigma^2_{\text{single}} + \frac{N-1}{N}\,\bar{\rho}\,\sigma^2_{\text{single}} = \sigma^2_{\text{single}}\left(\frac{1}{N} + \frac{N-1}{N}\bar{\rho}\right).$$

The diversification benefit is therefore

$$\text{DB} = \frac{\sigma_{\text{single}}}{\sigma_{\text{port}}} = \frac{1}{\sqrt{1/N + (1-1/N)\bar{\rho}}},$$

reproducing main-text Equation 3.

S6.3 High-coupling limit

At $\rho \to 1$ (the high-coupling limit, when the single risk factor dominates), DB $\to 1$: diversification provides no benefit. Let $\varepsilon = 1 - \rho$. Expanding to leading order in ε:

$$\text{DB}(\bar{\rho}, N) - 1 \approx \frac{(N-1)(1-\bar{\rho})}{2N} + O\big((1-\bar{\rho})^2\big)\,. \tag{S7}$$

The protective benefit (DB $-$ 1) vanishes linearly in $1 - \rho$: as coupling saturates, the residual idiosyncratic component that diversification can exploit shrinks to zero. For $N = 10$ sectors, the coefficient is $(N - 1)/(2N) = 9/20 = 0.45$; at $\rho = 0.9$ (typical crisis), DB $- 1 \approx 0.045$ — a 4.5% protective benefit against a theoretical maximum of $\sqrt{10} - 1 \approx 2.16$ (216%). The ratio of realized-to-maximum protection at high coupling, $(1 - \rho)(N - 1)$ / $[2N(\sqrt{N} - 1)]$, vanishes as $\rho \to 1$.

S6.4 Mapping onto h/J^2

The diversification benefit itself scales as DB $- 1 \sim (1 - \rho) \sim 1/(1 + J)$ under the identification $\rho = J/(1 + J)$. This is a $1/J$ decay, not h/J^2. The correspondence with Theorem 1 operates at a different level.

Consider a *passive protective overlay* — a fixed capital buffer, stop-loss rule, or bond allocation — with maximum protective capacity h, applied on top of an already-diversified portfolio operating in a high-coupling regime. The overlay's task is to prevent a correlated drawdown from breaching a loss threshold. In the Curie–Weiss analogy of §S2, the systematic risk factor F plays the role of the order parameter m; the coupling $J = \beta^2\sigma\_\mathrm{F}^2/\sigma\_\varepsilon^2$ governs how strongly returns co-move; and the overlay's fixed h provides a restoring force against the correlated component.

The key structural parallel is not the scaling of DB itself but the scaling of any *additional fixed protection* layered on top:

- The tilt that the overlay provides between "protected" and "unprotected" outcomes scales as h/J (from §S2.3: the energy tilt between branches is proportional to h and inversely proportional to J).
- The barrier between the two outcomes — the free-energy depth that systematic risk creates — scales as J (from §S2.4: the barrier height scales as $(J - T)^2/(2J) \sim J/2$ for $J \gg T$).
- The tilt-to-barrier ratio — the measure of protective effectiveness — therefore scales as $(h/J) / J = h/J^2$.

This is the structural parallel we claim: in both the coordination model and the financial-market setting, a fixed protective mechanism whose strength does not scale with the coupling faces a barrier that grows with the coupling, producing a tilt-to-barrier ratio that vanishes as h/J^2. Diversification itself is a specific instance of such a passive mechanism (with h bounded by $1/\sqrt{N}$); any other passive overlay shares the same structural limitation.

The correspondence is structural and directional, not an isomorphism: the coordination model is a dynamical system with stochastic trajectories and metastable branches, while the portfolio problem is a static variance decomposition. What they share is the mathematical structure in which a bounded intervention confronts a growing barrier.

S6.5 Note on rigor

The diversification identity (Equation 3) is exact under the equal-weight, equal-volatility, equal-pairwise-correlation idealization and is formally a variance-algebra fact. Theorem 1 is an asymptotic statement about the equilibrium measure of a coupled SDE in the high-coupling limit. The correspondence we are claiming is that *both* governance mechanisms exhibit the same h/J^2 scaling of effectiveness against synchronized failure modes, derivable from a single underlying mean-field structure. The empirical confirmation of Equation 3 in 22 years of S&P 500 sector data demonstrates that real markets instantiate the structural form in finance; the coordination-model evidence in the main text confirms it in a different but structurally analogous setting.

S7 Comparison with closest existing frameworks

The main text summarizes in one paragraph (Introduction) that five prior frameworks come closest to the present contribution. This section gives the detailed comparison.

Daníelsson [3, 4] formalized the *endogenous risk* phenomenon in financial markets: when many agents share the same Value-at-Risk model, the risk model itself creates the systemic event it is supposed to measure. Brunnermeier & Pedersen [5] developed the parallel liquidity-funding spiral, in which margin calls and fire-sale dynamics couple market and funding liquidity into a self-reinforcing feedback. Both mechanisms are correctly identified, but the analyses are finance-specific — applied to particular regulatory artifacts (VaR; collateral constraints) — and provide no quantitative scaling law connecting coupling level to protection effectiveness. Daníelsson and Brunnermeier–Pedersen describe the phenomenon; we derive the scaling and show that it operates beyond finance.

Sornette's dragon-king theory [6, 7] argues that extreme events in coupled systems arise from distinct amplification mechanisms — positive feedback loops that produce outliers beyond power-law

tails. The dragon-king framework identifies *when* extremes are endogenous rather than exogenous, but does not formalize the distinction between fixed and scaling protective mechanisms or derive a coupling-dependent scaling law for stabilizer effectiveness. Our framework complements it: dragon-kings describe the *threat*; the h/J^2 scaling describes why *passive defenses* fail against it.

Bouchaud and collaborators [8, 9] have extensively characterized the instabilities of financial markets arising from heterogeneous interacting agents, herding dynamics, and fat-tailed return distributions. Their agent-based models demonstrate that imitation and trend-following amplify volatility at the collective level — a coupling mechanism consistent with our framework's rising J. The present work differs in focus: where Bouchaud models the *generation* of instability, we model the *failure of fixed protections* against it, deriving the specific functional form (h/J^2) at which bounded interventions lose effectiveness.

Ashby's Law of Requisite Variety [10], the foundational principle of cybernetics and recently revived for AI governance [11], states that a controller must have at least as much variety as the system it regulates. The principle is qualitative: it predicts that an under-sized controller fails, but not by how much, nor at what coupling threshold. Our result extracts a specific quantitative implication of Ashby for coupled systems with restoring forces — passive controller variety is fixed at order h, while system variety grows as order J; the ratio $h/J^2 \to 0$ is the quantitative Ashby. Muhlert's recent application sits closer to architecture and governance design than to a coordination-model derivation.

Perrow's *Normal Accidents* [12] argued that tightly coupled systems will inevitably fail through unanticipated interaction modes. The proposed remedy is to avoid coupling. This is non-constructive in the AI era. Coupling is not a regulatory choice — it is being increased, very rapidly, by the deployment of systems that the regulators do not control. We argue for a different remedy: distinguish the failure modes of fixed protections from those of scaling protections, and design the latter where coupling cannot be avoided.

Svolik [13] identified the empirical fact that polarized electorates trade democratic principles for partisan victory, and that institutional checks cannot prevent the trade. The *why* is left implicit. Our framework gives an explicit answer: institutional checks are passive stabilizers; under high opinion-coupling they are subject to the same h/J^2 scaling we derive for any coordinated system.

Finally, the asymmetric-dependence literature in finance has been treated either as a behavioral phenomenon (loss aversion, panic dynamics) or as a statistical regularity (heavy left tails, copula tail dependence). We show in the main text that it is neither — it is a structural law derivable from variance algebra, equivalent in mathematical form to the coordination-model scaling we derive in the Results. The phenomenon documented in finance is a special case of a general structural pattern.

S8 AI coupling measurement details

Direct output coupling (primary measurement). We downloaded pre-generated outputs from 14 frontier and near-frontier models responding to 805 shared prompts from the AlpacaEval v2 benchmark [14]: GPT-4 Turbo, GPT-4 (1106-preview), GPT-4o, Claude 3 Opus, Claude 3.5 Sonnet, Claude 3 Sonnet, Gemini Pro, Mistral Large, Llama 3.1 405B Instruct, Llama 3 70B Instruct, Mixtral 8×22B Instruct, Mixtral 8×7B Instruct, Llama 3 8B Instruct, and Mistral 7B

Instruct v0.3. Each model's response to each prompt was embedded using the `all-MiniLM-L6-v2` sentence-transformer (384 dimensions, L2-normalized). For each prompt, the 14 × 14 pairwise cosine-similarity matrix was computed; the overall similarity $\rho = 0.797$ is the mean off-diagonal entry averaged across all 805 prompts. Mapping ρ to the model's coupling axis through the Curie–Weiss self-consistency identification $\rho = J/(1 + J)$ yields $J_{\text{e}}\text{ff} = 0.797 \,/\, 0.203 = 3.92$, above the critical threshold $J^c = T = 2$. (Rounded to two significant figures, $\rho \approx 0.80$ maps to $J \approx 4.0$; the abstract and main text use the rounded form, the precise computation appears here and in Figure 6.) The mapping is heuristic: cosine similarity between sentence-transformer embeddings is not the same physical quantity as the order-parameter spin–spin correlation that ρ originally denotes. The number should be read as "well into the supercritical regime" rather than as a calibrated dynamical variable. Within the model's parameter space, this regime corresponds to substantial decay of passive effectiveness (Figure 6 shows the prediction surface at $\mu = 100$); the model has no calibrated mapping from internal P(collapse) to AI failure categories.

Per-category breakdown (heuristic prompt classification): factual prompts (n = 133) yield $\rho = 0.82$ ($J = 4.65$); reasoning (n = 66) $\rho = 0.83$ ($J = 4.81$); instruction-following (n = 462) $\rho = 0.79$ ($J = 3.87$); creative (n = 144) $\rho = 0.77$ ($J = 3.28$). Same-family pairs (e.g., GPT-4 Turbo × GPT-4o; Claude 3 Opus × Claude 3.5 Sonnet) show $\rho \approx 0.85$ ($J \approx 5.8$); cross-family pairs still show $\rho \approx 0.79$ ($J \approx 3.9$). The 14 × 14 similarity matrix is Supplementary Figure S3.

Two earlier proxies are reported here for completeness; both measure related but distinct quantities and should not be interpreted as strict bounds on the same J.

- **Adversarial-attack transferability.** Transfer rates from [15–17] mapped via $\rho = J/(1 + J)$ yield a frontier-pair median $J_{\text{e}}\text{ff} \approx 1.2$. This proxy measures cross-model vulnerability transfer, which is shaped by training-data and safety-training overlap; a single safety-tuned model (e.g. Claude-2 against GCG suffixes, $\rho = 0.02$) can reduce the median sharply, so this number is sensitive to which models are included.
- **Benchmark-vector correlation.** Per-benchmark accuracy scores for 12 frontier models across 10 standard benchmarks (MMLU, ARC-C, HellaSwag, TruthfulQA, Winogrande, GSM8K, MATH, HumanEval, BBH, DROP) give a median pairwise Pearson correlation $\rho = 0.90$ ($J_{\text{e}}\text{ff} \approx 8.7$). The capability-detrended residual correlation is $\rho \approx -0.10$: virtually all of the apparent agreement is explained by shared capability-axis variance under shared evaluation pressure. This proxy therefore measures evaluation-axis homogeneity, not output coupling, and is best read as an upper artifact rather than an upper bound.

The direct output-similarity proxy ($J \approx 3.9$) measures response-level similarity on a fixed prompt distribution and is the closest of the three to the framework's response-coupling notion, but, as discussed above, the mapping to the Curie–Weiss J remains heuristic. Scripts and data: `empirical/ai_coupling_direct.py`, `empirical/data/alpacaeval_outputs/` (14 model_outputs.json files), `empirical/results/ai_coupling_direct.csv`. The analysis is fully reproducible without API access.

S9 Network agent-based model details

The network ABM (Results, Figure 3) extends the scalar minimal model to N = 200 agents on a graph. Each agent evolves the Curie–Weiss SDE (main-text Equation 1) with the global mean-field

term Jm replaced by the local mean field $J \cdot$ m_neighbors(i), computed as the row-normalized adjacency-matrix product. Wealth dynamics (Equation 2) are unchanged and depend on the population-averaged m = (1/N) $\sum$ m_i. Multiplicative noise ξ $\sqrt{}$(1 − m_i^2) η(t) uses a single shared η(t) per step (rather than per-agent independent noise), preserving the noise scale on m — without this, the per-agent noise would average to $\xi/\sqrt{}$N and the complete-graph control would not reproduce the scalar SDE result.

Graph statistics for the six tested topologies:

Topology	Nodes	Edges	Mean degree	Clustering
Complete	200	19,900	199.0	1.00
Erdős–Rényi (p = 0.10)	200	1,979	19.8	~0.10
Watts–Strogatz ($k = 20$, $\beta = 0.1$)	200	2,000	20.0	~0.57
Barabási–Albert ($m = 10$)	200	1,900	19.0	~0.12
Modular (4 × 50, p_in = 0.35, p_out = 0.01)	200	1,867	18.7	~0.34
Modular-boundary (same graph, h on boundary nodes only)	200	1,867	18.7	~0.34

The complete-graph control reproduces the minimal-model's scalar SDE result. The full Figure 3 sweep was run at 50 seeds per cell (SE ≈ 0.06–0.07); to address sampling concerns at the headline cell, we re-ran ($J = 5$, $\mu = 100$) at **100 seeds** for all six topologies:

topology	P(collapse), 50 seeds	P(collapse), 100 seeds	SE (100)	rigidity share
complete	0.26	0.33	0.047	0.91
Erdős–Rényi	0.26	0.33	0.047	0.91
Watts–Strogatz	0.24	0.33	0.047	0.91
Barabási–Albert	0.30	0.33	0.047	0.91
modular	0.34	0.33	0.047	0.91
modular-boundary	0.38	0.37	0.048	0.92

At 100 seeds, all six topologies converge into the band P(collapse) ∈ [0.33, 0.37]. A two-sample test of proportions for the largest gap in this band (modular-boundary at 0.37 vs. the others at 0.33) gives Z = 0.04 / $\sqrt{}$(0.047^2 + 0.048^2) ≈ 0.59 — well below any conventional significance threshold. **No pair of topologies is statistically separated**, including the modular and boundary-localized variants. The 50-seed scatter (0.24–0.38) was within sampling noise at the smaller seed count; topology has no detectable effect on the high-coupling residual at this sample size.

The substantive finding is therefore stronger than the original 50-seed claim: rather than "modularity amplifies the residual" or "boundary-localized is the worst case," the data support the unqualified statement that **no tested network topology rescues the mean-field passive-stabilizer residual**. The mean-field prediction holds across complete, Erdős–Rényi, Watts–Strogatz, Barabási–Albert, modular, and boundary-localized variants alike. Compared to the scalar SDE single-cell value (0.28 at 100 seeds), the ABM headline (0.33–0.37) is within combined SE (≈ 0.067); the small upward shift is plausibly attributable to the ABM's per-step shared $\eta(t)$ noise scale versus the scalar single-trajectory noise. Re-run script: `simulation/scripts/network_abm_100seed.py`. Output: `simulation/results/network_abm/headline_100seed.csv`.

The modular topology was constructed as a planted-partition (stochastic block) model with 4 equal-sized communities. Within-community edge probability $p\_in = 0.35$ yields ~17 within-community neighbors per agent; between-community edge probability $p\_out = 0.01$ yields ~0.5 cross-community neighbors. Despite this strong community structure, intra-community and inter-community order-parameter correlations both reach 1.00 at $J \geq 4$: the coupling propagates through the sparse inter-community edges and synchronizes the entire network. The boundary-localized stabilizer variant applies h only to the ~5–10% of agents that have at least one cross-community edge; the headline-cell P(collapse) of 0.37 is nominally the largest of the six topologies but is not separated above one SE from the 0.33 cluster.

Scripts: `simulation/scripts/network_abm.py`, `network_abm_figures.py`. Results: `simulation/results/network_abm/`.

S10 Tail-risk and VaR exceedance details

The mechanism tests reported in main-text Results §"Tail-risk frequency tracks the model's collapse curve" and §"Cross-mechanism confirmation: VaR exceedance" (Figure 7) bin the 5,414 sixty-day rolling windows of S&P 500 sector data into deciles by ρ and compute empirical failure-frequency metrics per bin.

Tail-risk frequency by coupling decile (5% maximum-drawdown threshold):

decile	J mid	ρ mid	n windows	P(DD > 5%)	SE	model P(coll, μ=100)
0	0.46	0.32	542	0.011	0.004	0.00
1	0.71	0.42	541	0.312	0.020	0.00
2	0.90	0.47	541	0.322	0.020	0.00
3	1.14	0.53	542	0.376	0.021	0.00
4	1.43	0.59	541	0.532	0.021	0.00
5	1.70	0.63	541	0.665	0.020	0.00
6	2.00	0.67	542	0.642	0.021	0.00
7	2.37	0.70	541	0.706	0.020	0.04
8	2.98	0.75	541	0.787	0.018	0.14
9	4.90	0.83	542	0.945	0.010	0.27

Spearman ρ(empirical, model) = +0.81 ($p = 0.004$) at the 5% threshold; ρ = +0.80 / +0.69 / +0.75 at the 3% / 7% / 10% thresholds (all $p < 0.03$).

VaR 1% exceedance rate by coupling decile (252-day historical lookback):

decile	J mid	ρ mid	n windows	exceedance rate	SE
0	0.45	0.31	491	0.0013	0.0006
1	0.69	0.41	491	0.0049	0.0011
2	0.86	0.46	491	0.0076	0.0020
3	1.13	0.53	491	0.0092	0.0020
4	1.50	0.60	491	0.0203	0.0022
5	1.79	0.64	491	0.0198	0.0027
6	2.10	0.68	491	0.0220	0.0033
7	2.46	0.71	491	0.0298	0.0031
8	3.04	0.75	491	0.0234	0.0029
9	5.36	0.84	491	0.0890	0.0051

Spearman ρ(VaR-1%, model) = +0.87 ($p = 0.0009$); ρ(VaR-5%, model) = +0.65 ($p = 0.04$); ρ(VaR-1%, tail-risk P(DD > 5%)) = +0.98 ($p = 1.5\times10^{-6}$); ρ(VaR-5%, tail-risk P(DD > 5%)) = +0.86 ($p = 0.002$).

Scaling comparison. The empirical log-log slope of P(tail | J) on J at the 5% threshold is 1.45; the model's log-log slope over the same J range is 2.51. The difference reflects the mean-field branch-selection mechanism's leading-order h/J^2 decay versus the empirical reality, which operates partly in the intermediate-coupling regime where the $O(T/J)$ corrections in §S2.6 are material. The direction and rank order are robust; the absolute slope is model-specific.

Scripts: `empirical/tail_risk_frequency.py`, `empirical/var_exceedance.py`. Results: `empirical/results/tail_risk_by_coupling.csv`, `empirical/results/var_exceedance_by_coupling.csv`.

Out-of-sample test. The temporal split uses 2004-03-30 to 2014-12-31 (Period 1, ~2,700 windows) and 2015-01-01 to 2025-12-30 (Period 2, ~2,700 windows). An affine mapping P_empirical = $a \times$ P_model + b is fitted on the training half's decile-binned tail-risk frequency and applied to the test half without re-fitting. Forward (P1 → P2): Spearman ρ = +0.70 ($p = 0.024$), RMSE = 0.276 vs. naive constant baseline 0.332 (17% skill reduction). Reverse (P2 → P1): ρ = +0.89 ($p < 10^{-3}$), RMSE = 0.276 vs. naive 0.261. The forward prediction is flat below $J \approx 2.5$ because the model predicts P(collapse) = 0 in the subcritical regime, consistent with Theorem 1. Held-out crises in P2 include the August 2015 China devaluation, March 2020 COVID selloff, the 2022 inflation shock, and the August 2024 carry-trade unwind. The split reported here is preregistered in `REPRODUCIBILITY.md`; alternative cutoffs, non-equal-weight sector portfolios, and instrumental-variable specifications run from the same pipeline will be appended to that manifest if requested in review.

Nonlinearity characterization. A quadratic fit to the decile-binned P(tail | J) improves AIC by 11.5 over linear (adj R^2 rises from 0.73 to 0.92), but the quadratic term is negative ($c = -0.057$, 95% CI [−0.20, −0.03]): the empirical curve is concave (saturating) rather than convex. This reflects ceiling bounding of P(tail) near 1.0 at the highest coupling deciles. The model's P(collapse) curve is convex in the same J range because it has not yet saturated. The log-log slope difference (1.45 empirical vs. 2.51 model) reflects this saturation asymmetry rather than a failure of monotonicity.

Failure-mode (rigidity vs. fragmentation) transition. The drawdown-period correlation and rigidity-classification methodology are detailed in `empirical/rigidity_fragmentation.py`; per-decile results in `empirical/results/rigidity_fragmentation_by_coupling.csv`. The empirical crossover ($J \approx 1.7$) is the J value at which the classified-crash rigidity share crosses 0.5; the model crossover ($J \approx 3.0$) is the corresponding crossover in the minimal-model L1 rigidity-share-of-collapses curve at $\mu = 100$. Spearman correlation between the empirical drawdown-correlation profile and the model's rigidity-share profile = +0.92 ($p = 1.3 \times 10^{-4}$).

Fed cuts as an empirical active-stabilizer test. The exploratory Federal Reserve analysis (Discussion) shows that post-cut SPY return rises with contemporaneous coupling ($r = +0.62$, $p = 0.03$, $N = 12$) — opposite to the diversification pattern. The active-stabilizer functional form predicts a stronger claim: post-cut return should scale with the *interaction* $|\Delta r| \times J$, where Δr is the cut magnitude. Linear regression on 12 FOMC cuts ≥ 10 bp (2019–2025; the FRED search window spans 2004–2025 but the 60-day rolling-correlation series is constrained by XLRE's 2015-10-08 inception, and the 2015–2018 period contained no rate cuts) of 90-day post-cut SPY return on (J, $|\Delta r|$, $J \times |\Delta r|$):

term	coefficient	std err	p
constant	−0.020	0.10	0.85
J	+0.39	0.19	0.07
	Δr		−0.79
**interaction $J \times$	Δr	**	**+0.66**

Sample-adjusted $R^2 = 0.71$ with $N = 12$; the interaction has the **positive sign** the active-stabilizer form predicts but is underpowered to reject zero at conventional levels. The data is directionally consistent with the active-stabilizer hypothesis; definitively distinguishing active from passive at this sample size would require pooling cross-country central-bank cuts (future work). Script: `empirical/fed_active_stabilizer_fit.py`. Output: `empirical/results/FED_ACTIVE_FIT_VERDICT.md`.

S11 Supplementary figures

References

[1] Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(1):1550005, 2015.

[2] Peter Hänggi, Peter Talkner, and Michal Borkovec. Reaction-rate theory: fifty years after Kramers. *Rev. Mod. Phys.*, 62:251–341, 1990.

[3] Jón Daníelsson. The emperor has no clothes: limits to risk modelling. *J. Banking & Finance*, 26(7):1273–1296, 2002.

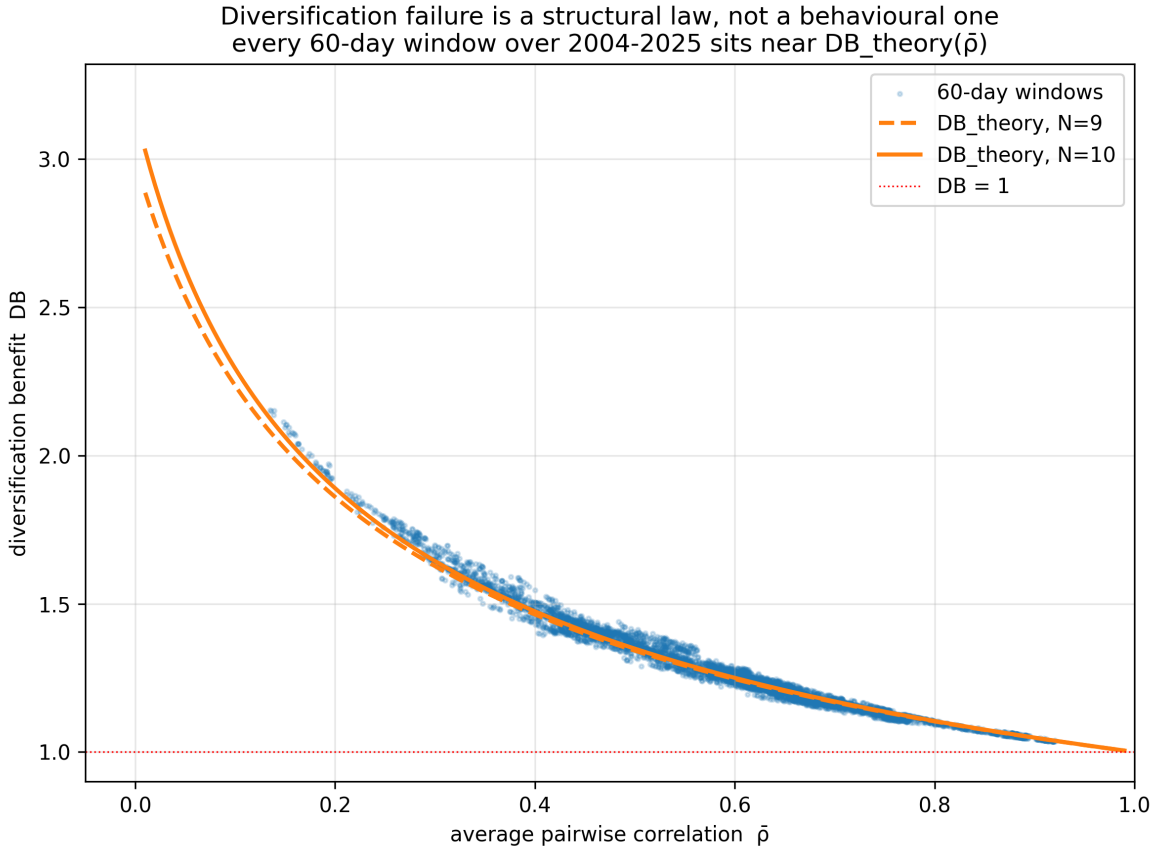

Figure S2: *Structural diversification failure (supporting the "Empirical setting" Results subsection).* Diversification benefit DB versus average pairwise correlation ρ across 5,414 sixty-day rolling windows of nine US sector ETFs (XLRE appended on its 2015-10-08 inception, yielding a 10-sector panel post-2015), 30 March 2004 – 30 December 2025. The theoretical structural curve DB(ρ, N) = 1 / $\sqrt{}$(1/N + (1 − 1/N) ρ) is overlaid; mean absolute error of empirical points against the curve is 0.016, with maximum deviation 0.100. Spearman $\rho(\rho, \text{DB}) = -0.993$ ($p < 10^{-15}$). Named crisis episodes (peak ρ, minimum DB during the episode): GFC 2008 (0.83, 1.08); Eurozone crisis 2010–12 (0.92, 1.03); China devaluation 2015 (0.80, 1.10); COVID-19 selloff 2020 (0.92, 1.04); inflation shock 2022 (0.74, 1.16). Each crisis sits on the predicted curve. Source file: `empirical/figures/diversification_failure_scatter.png`.

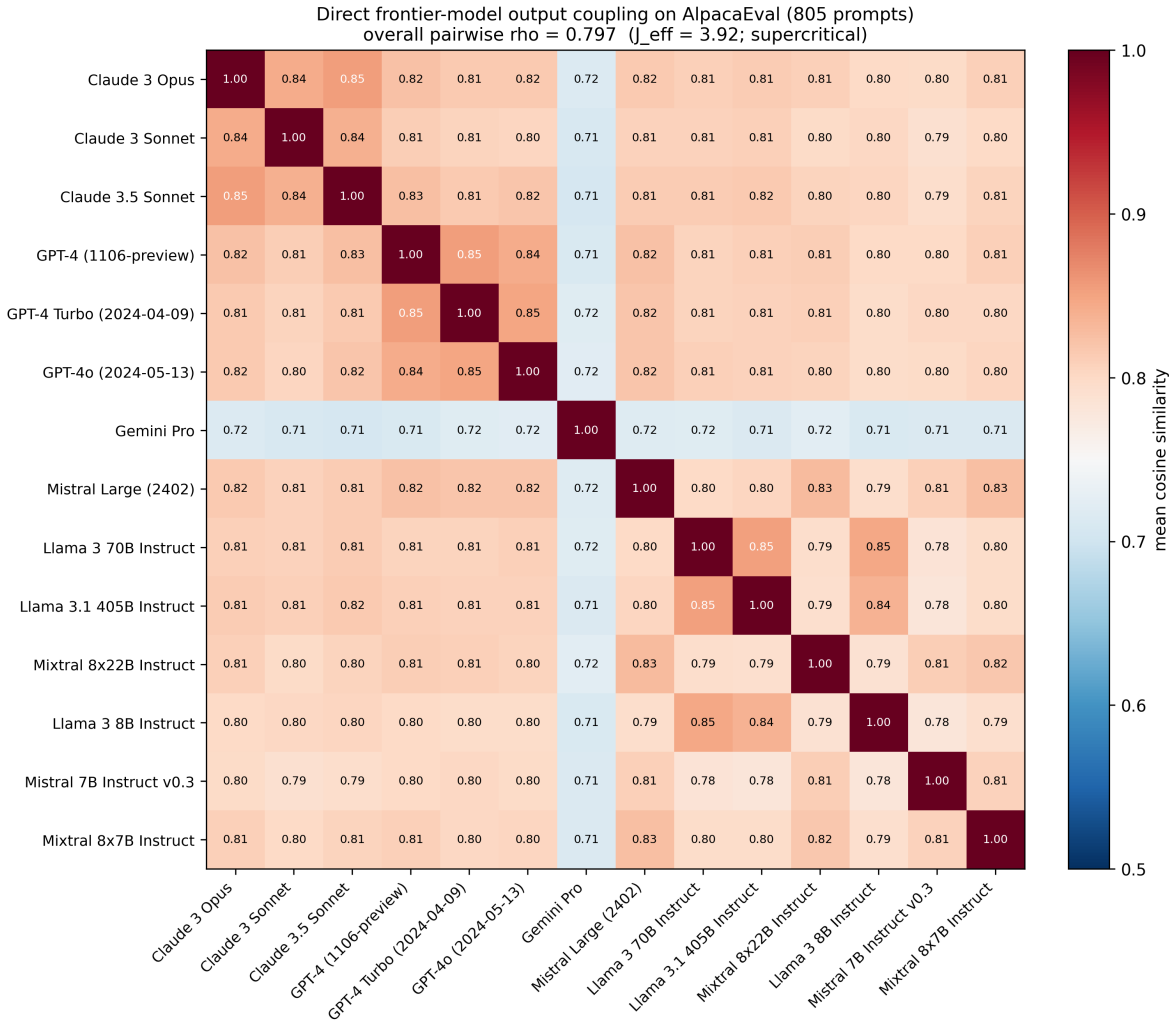

Figure S3: *Direct AI output coupling (supporting the "AI coupling" Results subsection).* 14 × 14 pairwise cosine-similarity matrix of sentence-transformer embeddings (`all-MiniLM-L6-v2`, 384-dim, L2-normalized) of frontier-model responses to 805 shared AlpacaEval v2 prompts [14]. Models are grouped by provider — Anthropic (Claude 3 Opus, Claude 3 Sonnet, Claude 3.5 Sonnet), OpenAI (GPT-4 1106-preview, GPT-4 Turbo 2024-04-09, GPT-4o 2024-05-13), Google (Gemini Pro), and a Mistral / Meta block (Mistral Large 2402, Llama 3 70B, Llama 3.1 405B, Mixtral 8x22B, Llama 3 8B, Mistral 7B v0.3, Mixtral 8x7B). Mean off-diagonal similarity $\rho = 0.797$, mapping to $J_{\text{e}}\text{ff} = 3.92$ under $\rho = J/(1+J)$. Same-family blocks (e.g., Claude 3 Opus × Claude 3.5 Sonnet) reach $\rho \approx 0.85$; cross-family pairs remain at $\rho \approx 0.79$, indicating coupling operates across organizational boundaries rather than within model families alone. Source file: `empirical/figures/ai_coupling_direct_heatmap.png`.

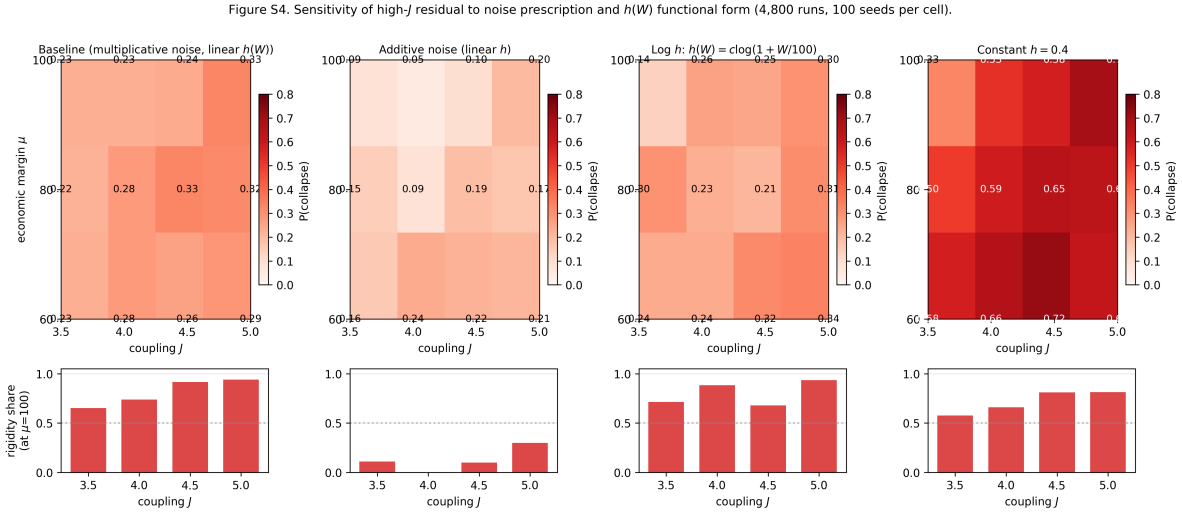

Figure S4: *Sensitivity sweep — noise prescription and h(W) form (supporting §S5.1).* Top row: P(collapse) heatmap over the high-J band ($J \in \{3.5, 4, 4.5, 5\}$) × protective-margin band ($\mu \in \{60, 80, 100\}$) for four model variants — baseline (multiplicative noise $\xi\sqrt{(1-m^2)}$, linear $h(W) = 2(W/500)$), additive noise (constant ξ, linear h), logarithmic $h(W) = 0.577 \cdot \log(1+W/100)$, and constant $h = 0.4$. Bottom row: rigidity share at $\mu = 100$ across the J range for each variant. Headline cell ($J = 5$, $\mu = 100$): baseline 0.33, additive 0.20, log h 0.30, constant h 0.70. The qualitative residual is robust to log-vs-linear h; constant h is substantially worse, ruling out the baseline as a tuned best-case; under additive noise the residual persists but rigidity-typed dominance falls (0.94 → 0.30), confirming the h/J^2 landscape result is multiplicative-noise- specific while the asymmetry itself is not. 4,800 runs total (100 seeds per cell × 12 cells × 4 variants). Source files: `simulation/scripts/sensitivity_sweep.py`, `simulation/results/sensitivity/figure_s4_sensitivity_sweep.png`.

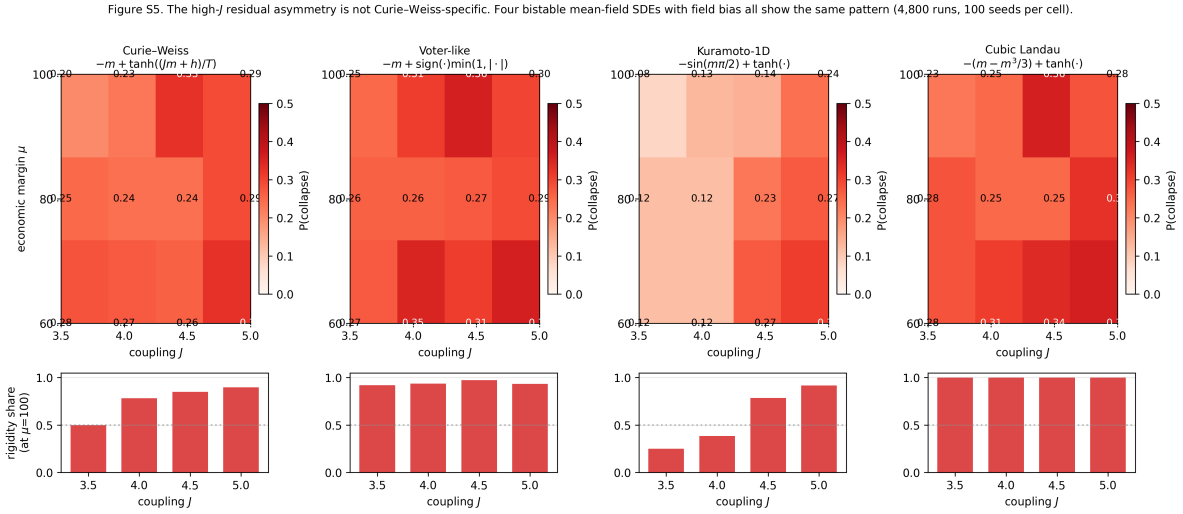

Figure S5: *Alternative-model-class robustness (supporting §S5.3).* Top row: P(collapse) heatmap over the high-J band ($J \in \{3.5, 4, 4.5, 5\}$) × protective-margin band ($\mu \in \{60, 80, 100\}$) for four bistable mean-field SDEs sharing the field-bias structure but differing in nonlinearity — Curie–Weiss tanh, voter-like piecewise sign, Kuramoto-1D sine, cubic Landau. Bottom row: rigidity share at $\mu = 100$ across the J range. At the headline cell ($J = 5$, $\mu = 100$): Curie–Weiss 0.29, voter-like 0.30, Kuramoto-1D 0.24, cubic Landau 0.28 — a 0.24–0.30 band. The high-J residual and rigidity-dominance pattern survive across all four model classes (rigidity share at the headline cell is 0.90 / 0.93 / 0.92 / 1.00 respectively), demonstrating that the asymmetry is generic to bistable mean-field SDEs with field bias rather than specific to Curie–Weiss. 4,800 runs total (100 seeds × 12 cells × 4 variants). Source: `simulation/results/alternative_class/figure_s5_alternative_class.png`.

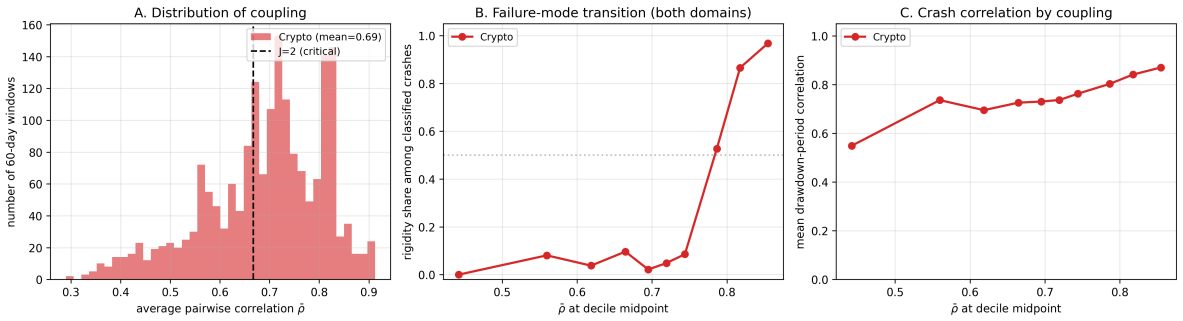

Figure S6: *Cross-domain replication: cryptocurrency markets (supporting §S5.4).* Three panels of the 10-token crypto basket (2020–2025, 1,864 60-day windows). (A) Distribution of average pairwise correlation ρ: crypto mean $\rho = 0.69$ (versus S&P ~0.50, not overplotted), with 64.4% of crypto windows supercritical ($\rho > 0.667$, $J > 2$) versus ~25% for S&P. (B) Failure-mode transition: rigidity share among classified crashes for the crypto basket rises with ρ and crosses 50% near decile 7 ($\rho \approx 0.79$); the S&P comparator curve is reported in the main text (Figure 8) and is not overplotted here to keep the panels readable. (C) Drawdown-period mean pairwise correlation rises monotonically from 0.55 to 0.87 across crypto deciles. The framework's prediction — higher-coupling domains exhibit rigidity dominance more readily — holds at the second domain. Source: `empirical/figures/figure_s6_crypto_vs_sp500.png`.

[4] Jón Daníelsson. *Global Financial Systems: Stability and Risk*. Pearson, Harlow, UK, 2nd edition, 2013.

[5] Markus K. Brunnermeier and Lasse Heje Pedersen. Market liquidity and funding liquidity. *Review of Financial Studies*, 22(6):2201–2238, 2009.

[6] Didier Sornette. Dragon-kings, black swans and the prediction of crises. *International Journal of Terraspace Science and Engineering*, 2:1–18, 2009.

[7] Didier Sornette and Guy Ouillon. Dragon-kings: Mechanisms, statistical methods and empirical evidence. *European Physical Journal Special Topics*, 205:1–26, 2012.

[8] Jean-Philippe Bouchaud. The endogenous dynamics of markets: price impact, feedback loops and instabilities. In *Lessons from the Credit Crisis*. Risk Books, 2011.

[9] Jean-Philippe Bouchaud and Marc Potters. *Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management*. Cambridge University Press, 2nd edition, 2003.

[10] W. Ross Ashby. *An Introduction to Cybernetics*. Chapman & Hall, London, 1956.

[11] Matthias Muhlert. Requisite variety for AI security. Technical Report 6255362, SSRN Working Paper, 2026. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6255362.

[12] Charles Perrow. *Normal Accidents: Living with High-Risk Technologies*. Basic Books, New York, 1984.

[13] Milan W. Svolik. Polarization versus democracy. *J. Democracy*, 30(3):20–32, 2019.

[14] Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaEval: An automatic evaluator of instruction-following models. GitHub repository, 2023. URL https://github.com/tatsu-lab/alpaca_eval.

[15] Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

[16] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *International Conference on Machine Learning (ICML)*, 2024.

[17] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? In *Advances in Neural Information Processing Systems*, volume 36, 2023.