How Selection Shapes Diversity in LLM Ecosystems

Maya Okawa[1,2]
[1]CBS-NTT Program in Physics of Intelligence, Harvard University
[2]Physics of Artificial Intelligence Group, NTT Research, Inc.
mayaokawa@fas.harvard.edu

Abstract

LLMs are becoming actors within broader ecosystems where they compete for attention and use. In daily use, we choose LLMs, versions, and configurations (e.g., reasoning effort) depending on the task. In model-routing systems, an LLM that performs well on routed queries gains a larger share, while on AI character platforms, a persona that attracts more user engagement is deployed more often. Such selection is repeated across interactions, shaping the future deployment of LLMs and personas. Although such mechanisms are now common, how repeated selection shapes the distribution of LLM populations remains an open question. Does repeated selection drive the population toward concentration, and under what conditions can diversity be maintained? To answer this question, we propose an analytical framework that builds on evolutionary dynamics to study the evolution of trait distributions in an LLM population under repeated selection. With our framework, we identify three regimes: coexistence, specialization, and concentration. The emergence of each regime is governed by selection strength, context-aware routing, and diversity support. Our framework also predicts that a population may appear diverse but remain locally unstable. To validate our framework, we design an AI-native social platform with a feed, where LLM agents generate posts, interact through likes and replies, and receive more exposure when they attract more reactions. The experimental results reproduce the three predicted regimes as well as the predicted local instability of diverse populations. We also show that our framework forecasts the loss of diversity from early fluctuations. Code and data are available at https://github.com/phys-ai/llm-ecosystem-selection.

1 Introduction

LLMs are now actors in ecosystems where they compete for attention and use. Every time we use an LLM, we choose among models, versions, and configurations depending on the task. The same dynamic arises within individual platforms, where multiple LLM agents compete for a larger share of user attention and interactions. For example, on character platforms [37, 28], agents that attract more user engagement are recommended to more users. In model-routing systems [32, 34], agents that perform well on routed queries receive a larger share of future queries. As such selection repeats over time, current performance shapes future exposure and use, creating feedback that can change the population of deployed LLM agents.

Prior work has measured behavioral traits in individual LLMs, including personality [25, 47], sycophancy and conformity [38, 48], and bias and safety orientation [40]. However, these studies do not consider how such traits evolve through interactions with users and among LLM agents. More recent work shows how behaviors spread or collapse in LLM populations, including cultural transmission [39, 53] and diversity loss [13]. These works are largely empirical, treat the selection mechanism as fixed, and do not examine how the selection rule governs the evolution of traits.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Building upon evolutionary dynamics, we develop an analytical framework to study how repeated selection changes the distribution of behavioral traits in a population of LLM agents competing for exposure. Under this model, agents can coexist, specialize, or concentrate. We then study whether these states are locally stable. We evaluate the theory in a controlled AI-native social platform, where LLM agents generate posts, receive likes and replies, and compete for future exposure in the feed. Our contributions are as follows.

- We develop an analytical framework for studying **how platform selection rules shape LLM ecosystems**. The framework models platform exposure as selection over behavioral traits, derives finite-selection and replicator dynamics, and defines ecosystem phases as locally stable endpoints, yielding diagnostics.
- We develop **stability-aware diagnostics** that complement descriptor-based diversity audits by testing whether an endpoint persists under small perturbations and against low-exposure agents, and we use them to characterize **interventions as phase-boundary shifts**: routing induces context dependence without stable specialization, and diversity support stabilizes specialization only within a bounded regime.
- We validate the theoretical predictions in an **AI-native social platform as a minimal LLM ecosystem** and identify **unstable diversity as a new failure mode**: a population may appear specialized, coexisting, or diverse by standard metrics, yet still be unstable, in which small perturbations can grow or low-exposure agents can invade. We also find that changes in early fluctuations precede the loss of diversity and can therefore serve as early indicators.

2 Related Work

Feedback loops and diversity in recommender systems. Research on recommender systems and algorithmic monoculture shows that feedback-driven platforms concentrate outcomes [12, 9, 27, 6], and that static diversity measurements can rise while feedback later reduces diversity, a form of illusory diversity [8]. Related work identifies model collapse under recursive self-training [50], degeneration in self-consuming loops [3], and reduced diversity in LLM-assisted writing and ideation [33, 4]. These studies take the allocation rule as given. We vary the rule and ask which regimes it induces over behavioral traits.

LLM populations. Prior work identifies persistent behavioral tendencies in single language models, including personality-like dimensions [47, 25], sycophancy and user-conforming bias [38, 48, 40], political and moral leanings [43, 20, 41, 46, 1], and persona-induced shifts [17, 19, 49, 24]. Work on generative agents [35], multi-agent debate [18], cultural evolution in LLM populations [39, 53, 10, 2], group selection over prompts [11], and diversity collapse in multi-agent idea generation [13] shows that behaviors interact and spread. Classical models of opinion and cultural evolution [16, 21, 5] explain how local interactions generate polarization, consensus, or fragmentation. In these lines the selection mechanism is fixed or implicit; we ask how the distribution of behavioral variants evolves when the mechanism itself is varied.

Evolutionary dynamics and critical transitions. Our approach builds on evolutionary dynamics [30, 22, 31, 42], including game-theoretic analyses of multi-agent learning [52, 7], and on critical-transition theory, including leading-eigenvalue criticality, critical slowing down, and early-warning statistics [44, 45, 15]. The replicator limit, the eigenvalue and invasion criteria, and the early-warning statistics are established tools; we operationalize them for the finite, stochastic selection process of an LLM ecosystem and derive diagnostics from them.

3 Analytical Framework for LLM Ecosystems

We study LLM ecosystems in which multiple agents or model variants repeatedly compete for system allocation. This allocation determines which behaviors are seen, reused, and reinforced, so feedback from generation, evaluation, routing, support, and safety filtering acts as a selection mechanism: traits affect fitness, fitness changes the composition of the population, and the composition determines future fitness, a closed loop that evolutionary dynamics describes [22, 42].

3.1 LLM Ecosystems as Exposure-Based Selection

Consider N deployed language-model agents indexed by $i \in [N]$. Agent i has a behavioral trait vector $x_i \in \mathcal{X}$; in our experiments the coordinates encode persistent tendencies such as stance, engagement-seeking, safety orientation, and alignment. The state is the exposure distribution $p_t = (p_{1,t}, \ldots, p_{N,t}) \in \Delta_N$, the share of exposure each agent holds at round t; it records how exposure is allocated across traits, not only which agent is currently ranked first. At round t the system draws a context $c_t \sim \pi$ and assigns each agent a score $f_i(p_t, c_t)$, which depends on the agent's own traits, on the context, and on the current exposure of the others.

Update. From the scores, the platform decides how much each agent is used next. The simplest rule is the finite-selection softmax update

$$p_{i,t+1} = \frac{p_{i,t} \exp\{\eta f_i(p_t, c_t)\}}{\sum_{j=1}^{N} p_{j,t} \exp\{\eta f_j(p_t, c_t)\}}, \qquad \eta > 0, \tag{1}$$

where η is the selection strength. This is the multiplicative-weights (Hedge) update with learning rate η on the scores: exposure after t rounds is the softmax of the cumulative scores, except that here the scores depend on the weights. Small η changes exposure gradually; large η turns small score gaps into large shifts of exposure, which is where concentration and collapse become empirically relevant. Averaging over contexts, with $\bar{f}_i(p) := \mathbb{E}_{c\sim\pi}[f_i(p, c)]$, the continuous-time limit of Eq. (1) for small η is the replicator equation

$$\dot{p}_i = p_i\Big(\bar{f}_i(p) - \sum_{j=1}^{N} p_j \bar{f}_j(p)\Big), \qquad i = 1, \ldots, N. \tag{2}$$

We use the finite map Eq. (1) in the experiments; the replicator limit supplies the vocabulary of stationary states and local stability used below (Appendix A.1 gives the derivation).

Fitness. We write the score as

$$f_i = \underbrace{s_i}_{\text{engagement}} + \rho \underbrace{z_i}_{\text{routing}} + \beta \underbrace{\log \frac{1}{N p_i}}_{\text{support}} - \lambda \underbrace{(1 - \text{safety}_i)}_{\text{risk}}, \tag{3}$$

where each term corresponds to a control that operators of deployed systems already tune. s_i is the baseline quality or engagement signal the system optimizes, such as user votes in a model arena [14] or click and rating feedback in a recommender [12]; z_i is how much better than its own average agent i did in the current context, so that routing with strength ρ sends each query to the model or agent that matches its context [32]; $\log 1/(Np_i)$ is a support term that boosts agents below their equal share, as a fairness boost reserves traffic for under-served items [51]; and $\text{safety}_i \in [0, 1]$ is an auditor's safety score of the output, so that λ is a smooth risk penalty standing in for a safety classifier that filters or downweights flagged outputs [23, 29]. Tuning any of these knobs changes which traits gain exposure, that is, the selection pressure on traits.

3.2 LLM-Specific Stochasticity: Generation and Evaluation

The score used for selection is sampled. At each round agent i generates a post, and R evaluator samples score it; the system observes the average

$$\widehat{f}_{i,t} = f_i(p_t, c_t) + \xi_{i,t}, \tag{4}$$

where f_i is the mean score and $\xi_{i,t}$ the generation–evaluation noise, and Eq. (1) is applied with $\widehat{f}_{i,t}$ in place of f_i. Averaging R independent evaluator samples reduces evaluator noise at the usual $1/R$ rate; variation shared by the samples, from the sampled generation itself or from a common evaluation prompt, is not reduced.

For small η, the noise enters the update only through score differences between agents: it moves weight among agents but keeps the total fixed. How much of it accumulates is governed by the same quantity that governs contraction. Under the log-barrier support of the experiments, the log exposure ratio of any two agents follows a first-order autoregressive recursion with forgetting factor $\lambda = 1 - \eta\beta_{\text{sup}}$ (Appendix Lemma 12), so after t rounds its variance is

$$\eta^2\sigma^2 \frac{1 - \lambda^{2t}}{1 - \lambda^2}, \tag{5}$$

with σ^2 the variance of the within-round score differences (Appendix Corollary 2). As $|\lambda| \to 1$ fluctuations grow before the mean endpoint visibly changes; this is the training-free early-warning score of Section 4.

3.3 Local Stability as a Diagnostic Backbone

The variance Eq. (5) is governed by how strongly the endpoint pulls perturbations back: as the forgetting factor λ approaches one, that pull vanishes and fluctuations grow. This section makes the notion general and asks of any observed endpoint whether it pulls perturbations back at all, that is, whether it is stable. We use *phase* for a qualitative type of locally attracting endpoint of the selection dynamic (coexistence, specialization, collapse, and so on). A stationary state $p^\star$ of Eq. (2) has active support $A := \{i : p_i^\star > 0\}$, the agents that still receive exposure, and on it all active agents have equal context-averaged fitness. It can lose stability in two ways. A *tangent mode* is a perturbation that moves weight among the active agents while keeping the total fixed; its growth rate is the real part of an eigenvalue of the Jacobian of Eq. (2) restricted to such perturbations, and we write γ_A for the largest of these, the growth rate of the fastest tangent mode. An *invasion* is the re-entry of an excluded agent $r \notin A$; its growth rate is the invasion exponent $\rho_r := \bar{f}_r(p^\star) - \sum_{\ell \in A} p_\ell^\star \bar{f}_\ell(p^\star)$, the fitness gap between the invader and the active population. The criticality index is the larger of the two growth rates:

$$\kappa(p^\star) := \max\Big\{\gamma_A,\ \max_{r \notin A} \rho_r\Big\}. \tag{6}$$

A negative κ means that every small perturbation of $p^\star$ dies out, so the endpoint is locally stable; a positive κ means that some perturbation grows, either among the active agents or through the re-entry of an excluded one; $\kappa = 0$ is the boundary (Appendix Theorem 7). The direction of the fastest-growing perturbation tells which kind of instability comes next: concentration on fewer agents or a reshuffling of exposure across contexts.

Definition 1 (Empirical criticality)**.** *In the experiments, selection strength is varied at fixed round time, so we evaluate the same index on the empirical one-step map at the audited endpoint $p^\star$ rather than on the replicator limit. We measure how fast a small perturbation among the active agents grows per round, and how fast a small amount of exposure given to an inactive agent grows per round, and take the larger of the two log growth rates as $\kappa_{\mathrm{map}}(p^\star)$ (estimator in Appendix B.1).*

Thus $\kappa_{\mathrm{map}} < 0$ means that nearby exposure distributions are pulled back together and no excluded agent can re-enter, $\kappa_{\mathrm{map}} \approx 0$ marks an estimated phase boundary, and $\kappa_{\mathrm{map}} > 0$ means instability or invadability. Since the audited endpoint is a time average of a context-driven trajectory rather than a solved fixed point, κ_{map} is the criticality of the map frozen at that state.

3.4 From Diversity Descriptors to Stability Diagnostics

Diversity audits summarize the geometry of an endpoint; this section defines the descriptors we use and the case in which they disagree with the stability audit. All descriptors are computed from the averaged endpoint $p^\star$ and are interpreted together with κ_{map}.

Concentration and active support. We measure concentration by the top-one share $\max_i p_i^\star$ and the effective support $N_{\mathrm{eff}} := 1/\sum_i (p_i^\star)^2$: a high top-one share implies concentration, while a high effective support implies broader coexistence.

Contextual specialization. Specialization requires more than context-dependent scores. We use two quantities. The exposure–context mutual information I_{exp} is the mutual information between the context and the agent selected by one softmax step from $p^\star$; it is high whenever scores vary with context. The winner-switching rate Γ_{switch} is the probability that two independently drawn contexts have different best-scoring active agents; it is high only when different contexts favor different agents. High I_{exp} with non-degenerate Γ_{switch} indicates context-specialized allocation; high I_{exp} with $\Gamma_{\mathrm{switch}} \approx 0$ indicates context-sensitive scoring around a nearly global winner, not stable specialization (formulas in Appendix B.1).

Definition 2 (Unstable diversity)**.** *An endpoint that is diverse by descriptors (coexisting or specialized) but has $\kappa_{\mathrm{map}} > 0$ exhibits* unstable diversity. *This arises when exposure is concentrated in most rounds and only the leading agent rotates, so that diversity exists in the average alone, or when the averaged distribution is representative of individual rounds but non-contracting or invadable.*

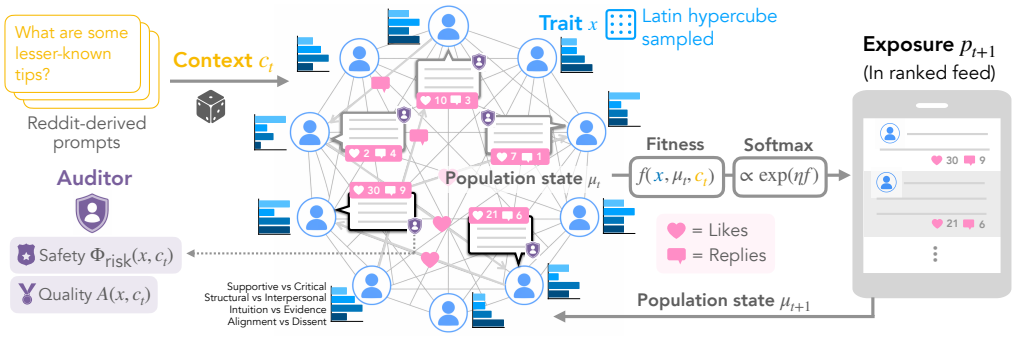

Figure 1: **Minimal LLM ecosystem with feedback-driven selection, mimicking an AI-native social platform.** LLM agents with trait profiles x generate and evaluate posts under Reddit-derived contexts c_t. At each timestep t, LLM agents generate posts and interact with other agents through replies, likes, and follows. Then, the platform assigns fitness $f_i = s_i + \rho z_i + \beta \log \frac{1}{N p_i} - \lambda(1 - \text{safety}_i)$ based on the engagement, capturing baseline advantage, contextual routing, support for underrepresented agents, and safety filtering. Exposure is then updated by multiplicative selection, $p_{i,t+1} \propto p_{i,t} \exp\big(\eta f_i(p_t, c_t)\big)$, inducing evolution of the exposure distribution p_t. The resulting macrostates are summarized by concentration and contextual specialization.

Phase labels are assigned from endpoint geometry together with κ_{map}; near a boundary, small finite differences along the leading unstable direction tell which descriptor is about to change.

3.5 Interventions as Local Criticality Shifts

The stability boundary of the preceding sections is drawn for fixed controls. Operators, however, change them: support is added and routing sharpened precisely to move endpoints out of concentration. The last part of the framework asks how such a change moves the boundary. Support and routing both enter the score Eq. (3) and therefore deform the operating point together with its criticality. Along a branch of stationary states, the rate at which a control moves the leading growth rate γ_A is the sum of its direct effect at a frozen state and the effect of the state moving in response (Appendix A.4); neither term has a universal sign, so the experiments report stability for each setting rather than one direction per control.

Support. At a frozen state, log-barrier support adds $-\beta_{\text{sup}} I$ to the tangent Jacobian and guarantees an interior stationary state, so it lowers γ_A at rate exactly -1 unless the branch motion undoes it (Appendix Proposition 8 and Corollary 1). Round by round, however, the same term is the filter of Section 3.2 with gain $1 - \eta\beta_{\text{sup}}$: beyond $\eta\beta_{\text{sup}} = 2$ the log exposure ratios alternate with growing amplitude and exposure swings between agents (Appendix Lemma 12). Support therefore stabilizes only in a bounded region of the $(\eta, \beta_{\text{sup}})$ plane; the band of Figure 3 lies inside it, and the rotating leaders of Figure 2(d) lie beyond it.

Routing. Context-aware routing changes which agent wins in which context, not whether a state is contracting. It raises I_{exp} whenever scores depend on context, including when one agent nearly always wins; stable specialization additionally needs $\Gamma_{\text{switch}} > 0$ and $\kappa_{\text{map}} < 0$ at the resulting state (Appendix A.5). Routing alone supplies the first and not the other two, which is the pattern of Section 4.

4 Experiment

We develop an AI-native social platform in which LLM agents generate content, receive feedback, and compete for future exposure in the feed. This platform serves as a minimal ecosystem for testing theory-driven hypotheses about how system-level selection shapes the evolution of trait distributions.

4.1 Experimental Setup

Platform and Agents. We simulate a minimal AI-native social platform with $N = 96$ LLM agents over $T = 200$ timesteps. Each agent has a five-dimensional trait vector encoding stance, social

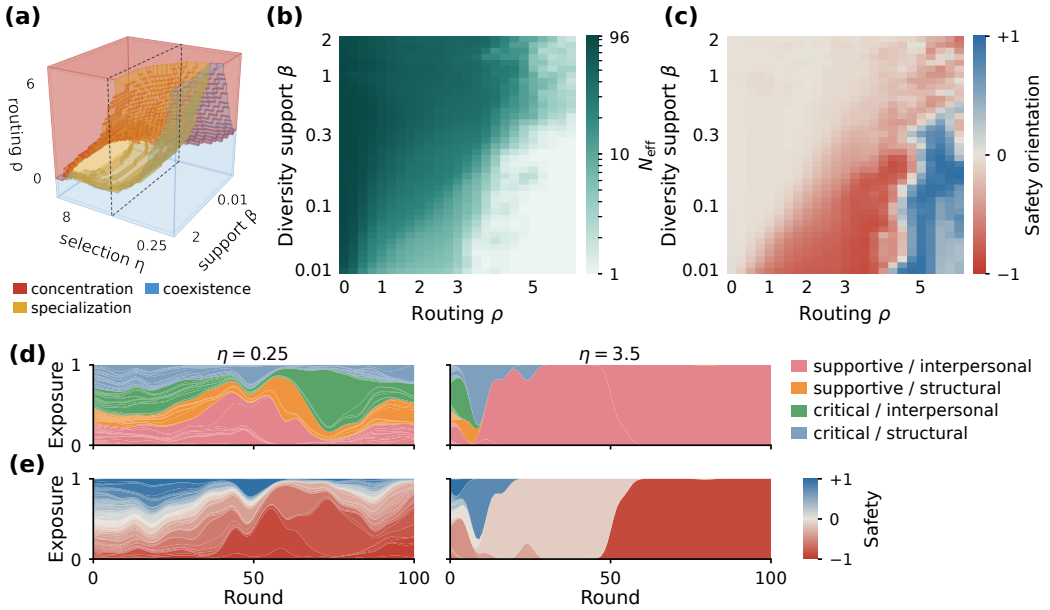

Figure 2: **Stronger selection and routing move the ecosystem from coexistence to specialization to concentration.** (a) Endpoint regimes over selection strength η, routing ρ, and diversity support β (one seed); the dashed plane marks the slice $\eta = 2$ shown in (b,c). (b) Effective number of agents N_{eff} and (c) exposure-weighted mean safety orientation over ρ and β at $\eta = 2$. (d) Exposure share of every agent over 100 rounds at weak ($\eta = 0.25$) and strong ($\eta = 3.5$) selection. Band widths indicate agent proportions, with colors representing combinations of stance and social framing. (e) The same runs, colored by safety orientation.

framing, epistemic style, consensus alignment, and safety orientation, with agents sampled by Latin hypercube design. At each timestep, the platform samples a Reddit-derived context $c_t \sim \pi$, agents generate posts, peer and auditor evaluations produce the observed score $\widehat{f}_{i,t}$ of Eq. (4), and exposure in the feed is updated by the softmax rule Eq. (1). To isolate the effect of platform selection, our main experiments use a two-step protocol: first, we collect logs of generated posts, peer evaluations, and auditor scores; second, we replay these logs while counterfactually varying the exposure-allocation rule. Generation, peer evaluation, and auditing use `gpt-5.4-mini`, `DeepSeek-V3`, `DeepSeek-V4-Pro`, and `Llama-3.3-70B-Instruct` via API, with temperatures 0.7 for generation and 0.2 for peer evaluation and auditing. Platform controls are the context fidelity $\alpha \in [0, 1]$, which interpolates each agent's score between its average over contexts ($\alpha = 0$) and its score in the current context ($\alpha = 1$), and the weights ρ, β, and λ of routing, support, and the risk penalty in Eq. (3). Unless otherwise stated, we report means and standard errors over 10 random seeds.

Metrics. We treat endpoint descriptors as conventional diversity-audit baselines. The main analysis uses top-one share, effective support, exposure–context mutual information, and cross-context winner switching to assign operational endpoint labels such as coexistence, collapse, and contextual specialization. We then compare these labels against stability-aware diagnostics that test whether small perturbations among exposed agents grow, whether an excluded agent can re-enter, and the criticality index κ_{map} that summarizes both. Full definitions are given in Appendix B.1.

4.2 Experimental Results

Selection Strength Drives Concentration and Criticality. We first vary selection strength η and routing ρ with no risk penalty (Figure 2). Selection strength η controls how sharply score differences translate into exposure in the feed, while routing controls how strongly allocation reflects context-specific performance. Stronger selection and stronger routing both shrink the effective number of agents (Figures 2(b) and 3(b)), the expected winner-take-more effect of multiplicative selection: with weak routing exposure stays spread over many agents, and with strong routing and weak support a single agent takes almost all of it. Along the way the feed drifts: at intermediate routing the exposure-weighted safety orientation moves toward reckless agents, and only once exposure has

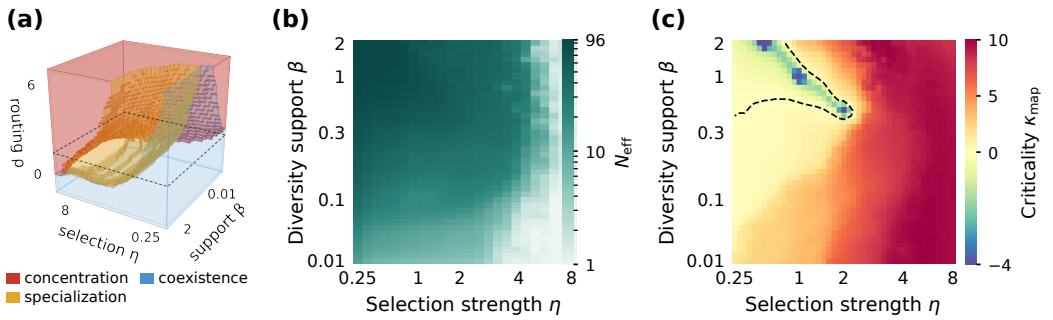

Figure 3: **Support stabilizes specialization only in a bounded regime.** (a) The regime cube of Figure 2(a); the dashed plane marks the slice $\rho = 2$ shown in (b,c). (b) Effective number of agents N_{eff} and (c) empirical criticality κ_{map} over selection strength η and diversity support β at $\rho = 2$; dashed: $\kappa_{\text{map}} = 0$. Support keeps exposure distributed at strong selection, but negative criticality appears only in a bounded diagonal band of intermediate selection and support.

concentrated at weak support does it flip (Figure 2(c)). Figure 2(d) shows every agent's exposure share over 100 rounds at weak and strong selection. At $\eta = 0.25$ all four trait groups keep a share and their widths keep shifting from round to round; at $\eta = 3.5$ a single supportive, interpersonal agent takes nearly all exposure within 20 rounds. Coloring the same runs by safety (Figure 2(e)) shows that the weak-selection feed mixes cautious and reckless agents, whereas under strong selection the winner is first a safety-neutral agent and, after about 50 rounds, a reckless one.

Support Stabilizes Specialization Only Within a Bounded Regime. Routing alone does not stabilize context-dependent allocation (Figure 2), so we next ask whether support can. To test this, we add an exposure-based support term that boosts under-exposed agents and downweights dominant ones. Figure 3(a,b) shows that support keeps exposure distributed even under strong selection, expanding the coexistence and specialization regimes while shrinking the concentration regime. However, the interior state that support creates stops being locally stable once the product of selection and support strength, $\eta\beta_{\text{sup}}$, exceeds two: the log exposure ratios then overshoot and exposure alternates between agents, so the time-averaged endpoint still looks distributed (Appendix A.6). Top-one share remains low across much of the supported region but rises sharply under strong selection, indicating collapse into dominance (Appendix Figure 7).

Figure 3(c) then applies the stability-aware criterion: κ_{map} is negative only within a bounded diagonal band of intermediate selection and support, while effective support remains nontrivial outside it (Appendix Figure 7). This band supports Proposition 3: raising η alone can turn a stable endpoint into an unstable one, even with the same scores, and most high-selection regions have positive criticality, $\kappa_{\text{map}} > 0$, indicating that the endpoint is locally unstable and vulnerable to takeover by agents with initially low exposure. Overall, support stabilizes specialization only within a distributed, context-dependent, and locally stable regime. The position of this band is robust across seeds; only its edges vary (Appendix Figure 8).

Unstable Diversity: Descriptive Diversity Fails Stability Tests. Figure 4 illustrates *unstable diversity*: endpoints that appear diverse under endpoint-level descriptors but fail the local stability tests. As illustrated in Figure 4(a), small perturbations fade at stable endpoints but grow at unstable ones, either among existing agents or through the invasion of a low-exposure agent. Figure 4(b) maps the stability regimes across selection and support strengths, showing where endpoints are locally stable, tangent-unstable, invadable, or concentrated. Descriptor-positive diversity is stable only in the narrow band of Figure 3(c); elsewhere it fails through tangent instability or invasion. Figure 4(c) validates these predictions on held-out contexts: perturbations grow at most tangent-unstable endpoints but fade at nearly all matched stable controls, while predicted invaders grow at every invadable endpoint (threshold robustness in Appendix B.2.1).

Concentration Risk Is Predictable Before It Is Visible. We then ask whether the concentration of individual rounds can be forecast from the first 10% of a trajectory, at a point where the endpoint still looks diverse. The target is that most of the last 100 rounds are concentrated, among the 436 endpoints of the early-warning grid whose exposure at round 20 is still diverse. The theory gives a score that needs no training data: the closed-form variance of Eq. (5), computed from the platform

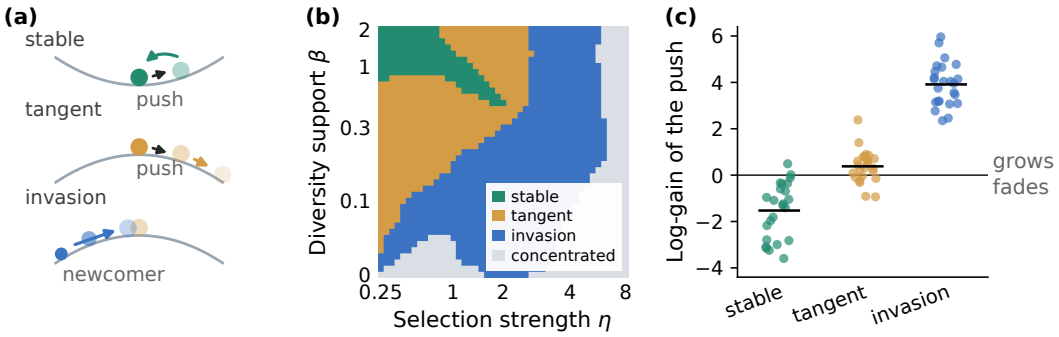

Figure 4: **Diversity that is not there.** (a) A small push fades at a locally stable endpoint (valley) and grows at a tangent-unstable one (hilltop); at an invadable endpoint a low-exposure newcomer grows. (b) Stability class of each endpoint over selection strength η and diversity support β ($\rho = 2$; most frequent class across five seeds). (c) Log-gain of a small perturbation on held-out contexts at $\eta = 1.2$: along the direction predicted unstable at 24 tangent-unstable endpoints and at 24 matched locally stable controls, and along the predicted invader at 24 invadable endpoints (bars: mean).

settings and the early score variance. Figure 5 compares it with classifiers trained on early trajectory descriptors and fluctuation features and with a nonlinear parameter-only baseline, each evaluated with the test population seed and the test selection strength held out. The closed form ranks first, ahead of the parameter-only SVM, and the learned early-trajectory models recover less of its accuracy. The same early features also forecast the criticality $\kappa_{\text{map}} > 0$ of the endpoint, with target and leakage controls reported in Appendix B.2.

Practically, the above results imply that diversity audits should not stop at endpoint descriptors such as concentration, effective support, or context dependence. Selection strength acts as a key risk knob, routing alone does not guarantee stable specialization, and support is stabilizing only within a bounded regime. Concentration risk follows from the settings and the score variance before it is visible in the trajectory.

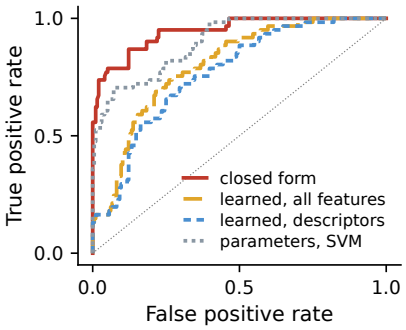

Figure 5: **Concentration risk is predictable before it is visible.** ROC for predicting, among endpoints still diverse at round 20, that most of the last 100 rounds will be concentrated: closed-form variance at round 20 against learned classifiers and a parameter-only SVM.

4.2.1 Robustness

We vary the population, the trait dimensionality, the testbed, and the evaluator, keeping the thresholds of the main experiment. Figure 6 shows a population in which every trait profile is generated by three model families (each of the 32 profiles paired with `gpt-5.4-mini`, `DeepSeek-V4-Pro`, and `Llama-3.3-70B`; 96 agents). Appendix B.4 covers an all-`Llama-3.3-70B` population in which generation, evaluation, and auditing are open-weight, 8-dimensional traits, and a model-routing testbed in which a router allocates query traffic among agents by task accuracy. In every setting, over five seeds, stronger selection raises top-one share and criticality, and most descriptor-positive endpoints fail the stability audit. In the social settings, stable specialization stays confined to a band of intermediate selection and support, and the stability classes and the push test reproduce (Figure 6(b,c)): the predicted invader grows at every invadable endpoint, while the tangent direction is the least damped direction but does not grow in these populations. Varying the evaluator changes which traits win but not the regimes. Under a `DeepSeek-V3` evaluator, the direction of the selected trait shift agrees with the `gpt-5.4-mini` evaluator for stance in 88% of matched conditions but only in 36–50% for the other three axes, with social framing reversed, while concentration, collapse, the bounded specialization regime, and unstable diversity persist (Appendix B.3).

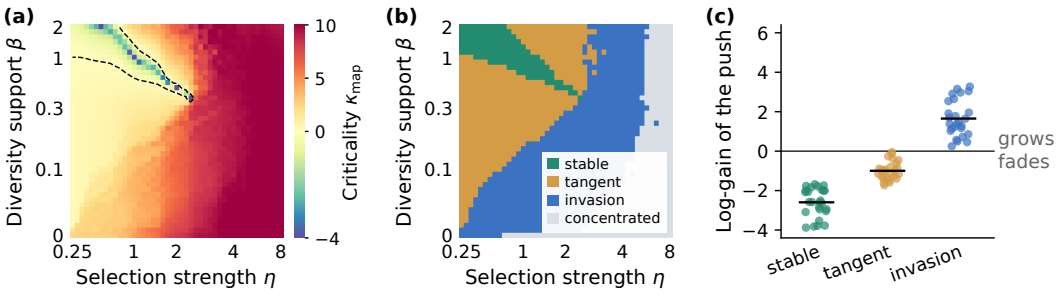

Figure 6: **A population of heterogeneous LLMs reproduces the regimes.** Each of the 32 trait profiles is paired with `gpt-5.4-mini`, `DeepSeek-V4-Pro`, and `Llama-3.3-70B` (96 agents); evaluation by `gpt-5.4-mini`; five seeds; $\rho = 2$. (a) Empirical criticality κ_{map} over selection and support strength, as in Figure 3(c); dashed: $\kappa_{\text{map}} = 0$. (b) Stability class of each endpoint, as in Figure 4(b). (c) Log-gain of a small push at $\eta = 1.2$, as in Figure 4(c) (26 endpoints per group).

5 Discussion

The trait distribution of a deployed LLM population can change without any model being retrained. Platforms reuse, rank, and route existing models, prompts, and personas according to past feedback, so the behaviors already present are under selection, and the allocation rule decides which of them gain deployment. Treating this as an evolutionary process makes the allocation rule the object of study: the rule determines the regime the population settles into, an intervention changes the rule and hence the regime, and whether a diverse state persists is a property of the rule acting on it, not of the state itself. Based on this viewpoint, we build a framework that predicts the regime an allocation rule induces, diagnostics that test whether that regime persists, and a testbed modeled on an AI-native social platform that isolates the rule by replaying fixed generations and evaluations under different rules.

Limitations. Our results characterize repeated selection in a simplified LLM ecology rather than a deployment-scale ecosystem. First, the trait space is fixed: dynamics occur through reallocation over existing types, not model turnover, prompt innovation, or agent entry, so the phase structure is conditional on a closed or slowly varying ecology. Second, the trait representation is coarse: LLM behavior is higher-dimensional, partly latent, and mutable under prompting, fine-tuning, and tool use, so phase labels depend on the taxonomy used to compress logs into a state space. Third, fitness is scored by LLM evaluators, and which traits are favored depends on the evaluator (Appendix B.3); human feedback may shift the ranking and the regime boundaries. Fourth, the stability measurements are finite-difference estimates of the empirical map at a time-averaged state rather than a fixed point; they are local, and near a boundary the sign of κ_{map} can depend on the perturbation branch and on generation noise. Finally, the platform is minimal; real platforms add feedback loops such as strategic adaptation, exploration, network effects, and imitation, which may move phase boundaries. Reopening generation raises concentration but preserves the regime ordering (Appendix B.5).

Acknowledgments and Disclosure of Funding

Funding in direct support of this work: NTT Research, Inc.; CBS-NTT Program in Physics of Intelligence, Harvard University. Additional revenues related to this work: None.

References

[1] Marwa Abdulhai, Greg Serapio-García, Clément Crepy, Daria Valter, John Canny, and Natasha Jaques. Moral foundations of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[2] Alberto Acerbi and Joseph M. Stubbersfield. Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120(44):e2313790120, 2023.

[3] Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. Self-consuming generative models go MAD. In *International Conference on Learning Representations (ICLR)*, 2024.

[4] Barrett R. Anderson, Jash Hemant Shah, and Max Kreminski. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition (C&C)*, 2024.

[5] Robert Axelrod. The dissemination of culture: A model with local convergence and global polarization. *Journal of Conflict Resolution*, 41(2):203–226, 1997.

[6] Joseph B. Bak-Coleman, Mark Alfano, Wolfram Barfuss, Carl T. Bergstrom, Miguel A. Centeno, Iain D. Couzin, Jonathan F. Donges, Mirta Galesic, Andrew S. Gersick, Jennifer Jacquet, Albert B. Kao, Rachel E. Moran, Pawel Romanczuk, Daniel I. Rubenstein, Kaia J. Tombak, Jay J. Van Bavel, and Elke U. Weber. Stewardship of global collective behavior. *Proceedings of the National Academy of Sciences*, 118(27):e2025764118, 2021.

[7] David Balduzzi, Sébastien Racanière, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. The mechanics of n-player differentiable games. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[8] Gabriele Barlacchi, Margherita Lalli, Emanuele Ferragina, Fosca Giannotti, Dino Pedreschi, and Luca Pappalardo. The diversity paradox revisited: Systemic effects of feedback loops in recommender systems. *arXiv preprint arXiv:2602.16315*, 2026.

[9] Rishi Bommasani, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

[10] Levin Brinkmann, Fabian Baumann, Jean-François Bonnefon, Maxime Derex, Thomas F. Müller, Anne-Marie Nussberger, Agnieszka Czaplicka, Alberto Acerbi, Thomas L. Griffiths, Joseph Henrich, Joel Z. Leibo, Richard McElreath, Pierre-Yves Oudeyer, Jonathan Stray, and Iyad Rahwan. Machine culture. *Nature Human Behaviour*, 7(11):1855–1868, 2023.

[11] Luis Celiktemel, Edward Eichhorn, Levin Brinkmann, Robin Schimmelpfennig, Aron Vallinder, Yaomin Jiang, Edward Hughes, and Iyad Rahwan. Group selection promotes prosocial prompts in populations of LLM agents. *arXiv preprint arXiv:2606.23343*, 2026.

[12] Allison J. B. Chaney, Brandon M. Stewart, and Barbara E. Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys)*, 2018.

[13] Nuo Chen, Yicheng Tong, Yuzhe Yang, Yufei He, Xueyi Zhang, Zou Qingyun, Qian Wang, and Bingsheng He. Diversity collapse in multi-agent LLM systems: Structural coupling and collective failure in open-ended idea generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026.

[14] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

[15] Vasilis Dakos, Stephen R. Carpenter, William A. Brock, Aaron M. Ellison, Vishwesha Guttal, Anthony R. Ives, Sonia Kéfi, Valerie Livina, David A. Seekell, Egbert H. van Nes, and Marten Scheffer. Methods for detecting early warnings of critical transitions in time series illustrated using simulated ecological data. *PLoS ONE*, 7(7):e41010, 2012.

[16] Guillaume Deffuant, David Neau, Frederic Amblard, and Gérard Weisbuch. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(01n04):87–98, 2000.

[17] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in ChatGPT: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

[18] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

[19] Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. Bias runs deep: Implicit reasoning biases in persona-assigned LLMs. In *International Conference on Learning Representations (ICLR)*, 2024.

[20] Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*, 2023.

[21] Rainer Hegselmann and Ulrich Krause. Opinion dynamics and bounded confidence: Models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3), 2002.

[22] Josef Hofbauer and Karl Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, 1998.

[23] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama Guard: LLM-based input-output safeguard for human-AI conversations. *arXiv preprint arXiv:2312.06674*, 2023.

[24] Itay Itzhak, Gabriel Stanovsky, Nir Rosenfeld, and Yonatan Belinkov. Instructed to bias: Instruction-tuned language models exhibit emergent cognitive bias. *Transactions of the Association for Computational Linguistics*, 12:771–785, 2024.

[25] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. Evaluating and inducing personality in pre-trained language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[26] Tosio Kato. *Perturbation Theory for Linear Operators*. Springer, 2nd edition, 1995.

[27] Jon Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021.

[28] Owen Lee and Kenneth Joseph. A large-scale analysis of public-facing, community-built chatbots on Character.AI. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2026.

[29] Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[30] John Maynard Smith. *Evolution and the Theory of Games*. Cambridge University Press, 1982.

[31] Martin A. Nowak. *Evolutionary Dynamics: Exploring the Equations of Life*. Belknap Press of Harvard University Press, 2006.

[32] Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs with preference data. In *International Conference on Learning Representations (ICLR)*, 2025.

[33] Vishakh Padmakumar and He He. Does writing with language models reduce content diversity? In *International Conference on Learning Representations (ICLR)*, 2024.

[34] Pranoy Panda, Raghav Magazine, Chaitanya Devaguptapu, Sho Takemori, and Vishal Sharma. Adaptive LLM routing under budget constraints. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025.

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

[36] Shishir G. Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The Berkeley function calling leaderboard (BFCL): From tool use to agentic evaluation of large language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

[37] Iryna Pentina, Tyler Hancock, and Tianling Xie. Exploring relationship development with social chatbots: A mixed-method study of Replika. *Computers in Human Behavior*, 140:107600, 2023.

[38] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

[39] Jérémy Perez, Corentin Léger, Marcela Ovando-Tellez, Chris Foulon, Joan Dussauld, Pierre-Yves Oudeyer, and Clément Moulin-Frier. Cultural evolution in populations of large language models. *arXiv preprint arXiv:2403.08882*, 2024.

[40] Leonardo Ranaldi and Giulia Pucci. When large language models contradict humans? Large language models' sycophantic behaviour. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, 2024.

[41] Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. Political compass or spinning arrow? Towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[42] William H. Sandholm. *Population Games and Evolutionary Dynamics*. MIT Press, 2010.

[43] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

[44] Marten Scheffer, Jordi Bascompte, William A. Brock, Victor Brovkin, Stephen R. Carpenter, Vasilis Dakos, Hermann Held, Egbert H. van Nes, Max Rietkerk, and George Sugihara. Early-warning signals for critical transitions. *Nature*, 461(7260):53–59, 2009.

[45] Marten Scheffer, Stephen R. Carpenter, Timothy M. Lenton, Jordi Bascompte, William Brock, Vasilis Dakos, Johan van de Koppel, Ingrid A. van de Leemput, Simon A. Levin, Egbert H. van Nes, Mercedes Pascual, and John Vandermeer. Anticipating critical transitions. *Science*, 338(6105):344–348, 2012.

[46] Nino Scherrer, Claudia Shi, Amir Feder, and David M. Blei. Evaluating the moral beliefs encoded in LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[47] Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, 7:1954–1968, 2025.

[48] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *International Conference on Learning Representations (ICLR)*, 2024.

[49] Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*, 2023.

[50] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

[51] Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2018.

[52] Karl Tuyls, Pieter Jan 't Hoen, and Bram Vanschoenwinkel. An evolutionary dynamical analysis of multi-agent learning in iterated games. *Autonomous Agents and Multi-Agent Systems*, 12(1):115–153, 2006.

[53] Aron Vallinder and Edward Hughes. Cultural evolution of cooperation among LLM agents. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2025.

A Theory

This appendix records the assumptions and proofs behind Section 3: the derivation of the replicator limit from the softmax update, the finite-selection stability of the softmax map, the local stability criterion that defines the criticality index, the interior state created by log-barrier support, the first-order effect of interventions on criticality, the specialization diagnostics, and the linear recursion behind the early-warning score. Throughout, N agents carry fixed traits $x_1, \dots, x_N$, the state is the exposure distribution $p \in \Delta_N$, and $\bar{f}_i(p) := \mathbb{E}_{c\sim\pi}[f_i(p,c)]$ is the context-averaged score of agent i, with $\bar{F}(p) := \sum_{j=1}^N p_j \bar{f}_j(p)$ the population mean. Two standing assumptions are used throughout: scores are bounded, $|f_i(p,c)| \le L_f$, and $p \mapsto \bar{f}(p)$ is C^1 near the relevant simplex face. The specialization results of Section A.5 additionally assume that ties are rare: writing $\Delta(p,c)$ for the top-two margin of the scores, there is a neighborhood U of the operating state on which $\sup_{p\in U} \pi\{c : \Delta(p,c) \le \varepsilon\} \to 0$ as $\varepsilon \downarrow 0$.

A.1 From multiplicative selection to the replicator equation

This subsection derives the replicator equation Eq. (2) from the update Eq. (1) and shows why a sweep over η is read on the finite map rather than on the ODE (Figure 3(b)).

Given a realized context c_t, write $f_{i,t} := f_i(p_t, c_t)$ and $F_t := \sum_{j=1}^N p_{j,t} f_{j,t}$. Since scores are bounded,

$$\exp\{\eta f_{i,t}\} = 1 + \eta f_{i,t} + O(\eta^2), \qquad \sum_{j=1}^N p_{j,t} e^{\eta f_{j,t}} = 1 + \eta F_t + O(\eta^2),$$

so the softmax update Eq. (1) satisfies

$$\frac{p_{i,t+1} - p_{i,t}}{\eta} = p_{i,t}\big(f_{i,t} - F_t\big) + O(\eta). \tag{7}$$

The discrete rule is therefore a first-order Euler step, in the rescaled time $\tau = \eta t$, for the contextual replicator update. Taking the conditional expectation over $c_t \sim \pi$ replaces $f_{i,t}$ and F_t by $\bar{f}_i(p_t)$ and $\bar{F}(p_t)$, and the small-η limit of the averaged recursion is the replicator equation Eq. (2).

Selection strength and finite-η effects. Absorbing η into the rescaled time is a small-step statement. When selection strength is swept at fixed round time, the relevant object is the finite-selection map

$$T_\eta(p)_i := \frac{p_i \exp\{\eta \bar{f}_i(p)\}}{\sum_j p_j \exp\{\eta \bar{f}_j(p)\}}, \qquad p \in \Delta_N, \tag{8}$$

rather than only its limiting ODE.

Proposition 3 (Finite-selection stability of the softmax map)**.** *Let $p^\star \in \operatorname{int} \Delta_N$ satisfy the equal-fitness condition $\bar{f}_1(p^\star) = \cdots = \bar{f}_N(p^\star)$. Then $p^\star$ is a fixed point of T_η for every $\eta > 0$. If $J(p^\star)$ denotes the replicator Jacobian restricted to $T\Delta_N = \{v : \sum_i v_i = 0\}$, then*

$$DT_\eta(p^\star)|_{T\Delta_N} = I + \eta J(p^\star).$$

Thus the finite-η map is locally linearly stable exactly when $|1 + \eta\lambda_j(J)| < 1$ for every tangent eigenvalue $\lambda_j(J)$. For a mode $\lambda = a + ib$ with $a < 0$, this requires

$$\eta < \frac{-2a}{a^2 + b^2}.$$

Proof. Let $\bar{F}^\star$ denote the common fitness at $p^\star$. Then $T_\eta(p^\star)_i = p_i^\star e^{\eta \bar{F}^\star} / \sum_j p_j^\star e^{\eta \bar{F}^\star} = p_i^\star$, so $p^\star$ is a fixed point for every $\eta > 0$.

Take $v \in T\Delta_N$ and write $Z(p) := \sum_j p_j \exp\{\eta \bar{f}_j(p)\}$. At $p^\star$,

$$D\{p_i e^{\eta \bar{f}_i(p)}\}[v] = e^{\eta \bar{F}^\star}\big(v_i + \eta p_i^\star D\bar{f}_i(p^\star)[v]\big), \qquad DZ(p^\star)[v] = \eta e^{\eta \bar{F}^\star} \sum_j p_j^\star D\bar{f}_j(p^\star)[v],$$

using $\sum_j v_j = 0$. With $Z(p^\star) = e^{\eta \bar{F}^\star}$, the quotient rule gives

$$DT_\eta(p^\star)[v]_i = v_i + \eta p_i^\star \left(D\bar{f}_i(p^\star)[v] - \sum_j p_j^\star D\bar{f}_j(p^\star)[v] \right).$$

For the replicator vector field $G_i(p) = p_i(\bar{f}_i(p) - \bar{F}(p))$, the equal-fitness condition and $\sum_i v_i = 0$ give the same bracket for $DG_i(p^\star)[v]$, so $DT_\eta(p^\star)|_{T\Delta_N} = I + \eta J(p^\star)$.

The tangent eigenvalues of the map are $1 + \eta\lambda_j(J)$, and local linear stability of a hyperbolic fixed point requires all of them inside the unit disk. For $\lambda = a + ib$, $|1 + \eta\lambda|^2 = 1 + 2\eta a + \eta^2(a^2 + b^2)$, so $|1 + \eta\lambda| < 1$ is equivalent to $2a + \eta(a^2 + b^2) < 0$, which requires $a < 0$ and gives $\eta < -2a/(a^2 + b^2)$. □

Thus a mode that is stable for the replicator ODE can still cross the unit circle through the discrete multiplier $1+\eta\lambda$ at larger η, which is why phase diagrams over selection strength are read as diagrams of the finite-selection map.

A.2 Stationary states, linearization, and the criticality index

This subsection proves the stability criterion behind the criticality index Eq. (6); it is what makes a negative κ_{map} in Section 4 a stability statement.

Definition 4 (Stationary state and active support)**.** *A state $p^\star \in \Delta_N$ is stationary for Eq. (2) if $p_i^\star\big(\bar{f}_i(p^\star) - \bar{F}(p^\star)\big) = 0$ for every i. Its active support is $A := \{i : p_i^\star > 0\}$, and the active face is $\Delta_A := \{p \in \Delta_N : p_i = 0 \text{ for } i \notin A\}$ with tangent space $T\Delta_A := \{v \in \mathbb{R}^N : \sum_{i\in A} v_i = 0,\ v_i = 0 \text{ for } i \notin A\}$. If $A = [N]$ we call $p^\star$ interior.*

The local theory is conditional on the existence of a stable multi-agent stationary state; general replicator systems need not possess one. Section A.3 shows that log-barrier support produces such an interior operating point.

Lemma 5 (Equal-fitness and invasion conditions)**.** *Let $p^\star$ be stationary with active support A. Then active agents have equal context-averaged fitness,*

$$\bar{f}_i(p^\star) = \bar{F}_A(p^\star) := \sum_{\ell \in A} p_\ell^\star \bar{f}_\ell(p^\star), \qquad i \in A,$$

and for an inactive agent $r \notin A$ the first-order growth rate of a small mass placed on it is the invasion exponent

$$\rho_r(p^\star) := \bar{f}_r(p^\star) - \bar{F}_A(p^\star). \tag{9}$$

Proof. If $i \in A$, the stationarity condition can be divided by $p_i^\star > 0$, giving $\bar{f}_i(p^\star) = \bar{F}(p^\star)$; since $p_j^\star = 0$ for $j \notin A$, $\bar{F}(p^\star) = \bar{F}_A(p^\star)$. For $r \notin A$, the rth replicator coordinate satisfies $\dot{p}_r/p_r = \bar{f}_r(p) - \bar{F}(p)$, and by continuity of $\bar{f}$ near the face this equals $\rho_r(p^\star) + o(1)$ as $p \to p^\star$ with $p_r > 0$. □

Lemma 6 (Jacobian on a fixed active face)**.** *Suppose $\bar{f}$ is C^1 near a stationary state $p^\star$ that is interior on its active face. On $T\Delta_A$, the Jacobian $J_A(p^\star)$ of Eq.* (2) *is represented by*

$$(J_A)_{i\ell}(p^\star) = p_i^\star \left(\partial_\ell \bar{f}_i(p^\star) - \sum_{j \in A} p_j^\star \partial_\ell \bar{f}_j(p^\star) \right), \qquad i, \ell \in A. \tag{10}$$

Proof. Let $g_i(p) := p_i(\bar{f}_i(p) - \bar{F}(p))$. For $i, \ell \in A$,

$$\partial_\ell g_i(p) = \delta_{i\ell}\big(\bar{f}_i(p) - \bar{F}(p)\big) + p_i\big(\partial_\ell \bar{f}_i(p) - \partial_\ell \bar{F}(p)\big),$$

and the first term vanishes at $p^\star$ by Lemma 5. On the active face, $\partial_\ell \bar{F}(p) = \bar{f}_\ell(p) + \sum_{j \in A} p_j \partial_\ell \bar{f}_j(p)$. Applied to $v \in T\Delta_A$, the term $-p_i^\star \sum_{\ell \in A} \bar{f}_\ell(p^\star) v_\ell = -p_i^\star \bar{F}_A(p^\star) \sum_{\ell \in A} v_\ell$ vanishes because active fitnesses are equal and tangent vectors sum to zero. The remaining terms are Eq. (10). □

Theorem 7 (Local stability criterion)**.** *Let $p^\star$ be a stationary state of Eq.* (2)*, interior on its active face, with $\bar{f}$ C^1 near that face. Let γ_A be the largest real part of an eigenvalue of $J_A(p^\star)$ and ρ_r the invasion exponents of Eq.* (9)*, and define*

$$\kappa(p^\star) := \max\Big\{\gamma_A,\ \max_{r \notin A} \rho_r(p^\star)\Big\},$$

with the first term omitted when $T\Delta_A$ is zero-dimensional and the second when $A = [N]$. Then $p^\star$ is locally asymptotically stable relative to the simplex if $\kappa < 0$, locally unstable if $\kappa > 0$, and nonhyperbolic if $\kappa = 0$.

Proof. Choose local coordinates (z, y) around $p^\star$ in the affine hyperplane $\sum_i p_i = 1$, where $z \in T\Delta_A$ gives perturbations within the active face and $y = (p_r)_{r \notin A}$ the inactive masses. By Lemmas 5 and 6, the derivative of the vector field at $(0, 0)$ is block upper-triangular,

$$\begin{pmatrix} J_A(p^\star) & B \\ 0 & \operatorname{diag}\{\rho_r(p^\star) : r \notin A\} \end{pmatrix},$$

because each inactive coordinate has the form $\dot{p}_r = p_r(\bar{f}_r(p) - \bar{F}(p))$, so its derivatives with respect to active coordinates carry the factor p_r and vanish at $p_r = 0$, while its own diagonal entry is $\bar{f}_r(p^\star) - \bar{F}_A(p^\star) = \rho_r$. The spectrum is the union of $\operatorname{spec}(J_A)$ and the invasion exponents. If $\kappa < 0$ all eigenvalues have negative real part and the Lyapunov indirect method gives local asymptotic stability; restricting to the simplex preserves it. If $\kappa > 0$, either J_A has an unstable mode within the active face, or some $\rho_r > 0$, in which case an initial condition with a small positive mass p_r satisfies $\frac{d}{dt} \log p_r > \rho_r/2$ while it stays near $p^\star$, so the mass grows exponentially. If $\kappa = 0$ the linearization has a spectral value on the stability boundary and no conclusion follows from it alone. □

A.3 Support-designed interior coexistence

This subsection shows that log-barrier support creates the interior stationary state around which the support-grid audit of Figure 3 is made, and damps every tangent mode at a fixed state.

Consider the log-barrier support term toward a target $q \in \operatorname{int} \Delta_N$,

$$\bar{f}_i(p) = g_i(p) + \beta_{\sup}\big(\log q_i - \log p_i\big), \qquad \beta_{\sup} > 0. \tag{11}$$

Proposition 8 (Barrier support creates an interior stationary state and damps tangent modes)**.** *Assume $g : \Delta_N \to \mathbb{R}^N$ is continuous and C^1 on $\operatorname{int} \Delta_N$. Then Eq.* (11) *has at least one interior stationary state, and at every interior point the support term contributes $-\beta_{\sup} I$ to the tangent Jacobian, so all tangent eigenvalues shift left by $\beta_{\sup}$ at a fixed state.*

Proof. The map $\Psi_i(p) = q_i e^{g_i(p)/\beta_{\sup}} / \sum_j q_j e^{g_j(p)/\beta_{\sup}}$ is continuous from Δ_N into its interior, so Brouwer's theorem gives an interior fixed point $p^\star$. Taking logarithms of $p^\star = \Psi(p^\star)$ shows that $g_i(p^\star) + \beta_{\sup}(\log q_i - \log p_i^\star)$ is independent of i, so all fitnesses are equal and $p^\star$ is stationary by Lemma 5. For the Jacobian, let $b_i(p) = \log q_i - \log p_i$; then $Db_i(p)[v] = -v_i/p_i$ and $\sum_j p_j Db_j(p)[v] = -\sum_j v_j = 0$ for $v \in T\Delta_N$, so by Eq. (10) the support block acts as $v \mapsto -v$. □

This is the frozen-state effect; along a moving stationary branch the complete derivative is the one in Section A.4, so support need not be monotonically stabilizing.

A.4 Interventions as local criticality shifts

This subsection gives the first-order effect of a platform control on the leading growth rate γ_A, used in Section 3.5: a frozen-state term plus a branch-motion term, with the frozen-state term of support equal to -1. Let θ collect the platform parameters and let $p_\theta^\star$ be a C^1 branch of stationary states with fixed active support, with tangent Jacobian $J_\theta := J_A(p_\theta^\star; \theta)$.

Proposition 9 (Derivative of the leading growth rate)**.** *Suppose the eigenvalue of J_θ with largest real part is simple, with right and left eigenvectors ϕ_θ, ψ_θ normalized by $\langle \psi_\theta, \phi_\theta \rangle = 1$. Then for any scalar control α,*

$$\frac{d\gamma_A}{d\alpha} = \mathrm{Re}\big\langle \psi_\theta, \big(\partial_\alpha J_\theta + D_p J_\theta[\partial_\alpha p_\theta^\star]\big)\phi_\theta \big\rangle,$$

the first term being the direct effect of the control at a frozen state and the second the effect of the stationary state moving in response. If the relevant boundary is an invasion exponent, the same holds for the total derivative of ρ_r along the branch.

Proof. First-order perturbation of a simple eigenvalue [26, Ch. II]: $\lambda' = \langle \psi, \mathcal{J}'\phi \rangle$ for $\mathcal{J}(\alpha) = J_{\theta(\alpha)}$, and the chain rule splits $\mathcal{J}'$ into the two terms. □

Corollary 1 (Support and routing)**.** *For log-barrier support, $\partial_{\beta_{\sup}} J_\theta = -I$ on the tangent space (Proposition 8), so $d\gamma_A/d\beta_{\sup} = -1 + \mathrm{Re}\langle \psi_\theta, D_p J_\theta[\partial_{\beta_{\sup}} p_\theta^\star]\phi_\theta \rangle$: support is stabilizing unless the branch-motion term exceeds one. The routing block of J_θ is not sign-definite on the tangent space, so $d\gamma_A/d\beta_{\mathrm{route}}$ can take either sign depending on the branch and the leading eigenvectors.*

A.5 Winner heterogeneity and specialization

This subsection justifies the two specialization diagnostics of Section 3.4: $I_{\exp}$ alone cannot separate a global winner from niches, and $\Gamma_{\mathrm{switch}} > 0$ is the condition under which specialization can persist.

Define the context-conditioned soft allocation

$$Q_\eta(i \mid c; p) := \frac{p_i \exp\{\eta f_i(p, c)\}}{\sum_{j=1}^N p_j \exp\{\eta f_j(p, c)\}}, \tag{12}$$

let $X_\eta \sim Q_\eta(\cdot \mid C; p)$ for $C \sim \pi$, and set $I_{\exp}(p; \eta) := I(X_\eta; C)$. For winner-based diagnostics, let $W(c; p) \in \arg\max_i f_i(p, c)$ be the almost-surely unique winner, $\omega_p(i) := \pi\{c : W(c; p) = i\}$, and

$$\Gamma_{\mathrm{switch}}(p) := \mathbb{P}_{c, c' \overset{\mathrm{iid}}{\sim} \pi}\big(W(c; p) \neq W(c'; p)\big) = 1 - \sum_{i=1}^N \omega_p(i)^2,$$

where the identity follows because the events $\{W(C; p) = i\}$ partition the context space. In particular, $\Gamma_{\mathrm{switch}}(p) = 0$ if and only if the same agent wins π-almost every context.

Proposition 10 (Winner heterogeneity supports specialization)**.** *Suppose that ties are rare on a neighborhood of $p^\star$, $\Gamma_{\mathrm{switch}}(p^\star) \geq \gamma_0 > 0$, and the routed allocation remains responsive to context. Then the winner distribution is non-degenerate on a neighborhood of $p^\star$, so local movement away from diffuse coexistence need not collapse immediately to a single global winner.*

Proof. At $p^\star$, $\sum_i \omega_{p^\star}(i)^2 = 1 - \Gamma_{\mathrm{switch}}(p^\star) \leq 1 - \gamma_0$, so the winner distribution is not a point mass. For $p_n \to p^\star$, let $\delta_n := \sup_c \max_i |f_i(p_n, c) - f_i(p^\star, c)| \to 0$. If $W(c; p_n) \neq W(c; p^\star)$, the margin

at $p^\star$ is at most $2\delta_n$, so $\pi\{c : W(c; p_n) \neq W(c; p^\star)\} \leq \pi\{c : \Delta(p^\star, c) \leq 2\delta_n\} \to 0$ because ties are rare. Hence $\omega_{p_n} \to \omega_{p^\star}$, and by continuity there is a neighborhood U on which $\Gamma_{\mathrm{switch}}(p) \geq \gamma_0/2$. Responsiveness of the allocation to context lets this non-degenerate winner structure be reflected in context-dependent exposure. □

Lemma 11 (Exposure MI and winner MI)**.** *If $Q_\eta(\cdot \mid c; p)$ is independent of c, then $I_{\exp}(p; \eta) = 0$. If the winner is almost surely unique and $Q_\eta(\cdot \mid c; p) \to \delta_{W(c;p)}$ in π-probability as $\eta \to \infty$, then $I_{\exp}(p; \eta) \to I(W(C; p); C) = H(\omega_p)$.*

Proof. In the first case X_η and C are independent. In the second, write $P_\eta(i) := \int Q_\eta(i \mid c; p)\, \pi(dc)$; the assumed convergence and boundedness give $P_\eta \to \omega_p$, hence $H(P_\eta) \to H(\omega_p)$, while the conditional entropy $H(Q_\eta(\cdot \mid C; p))$ converges to zero in probability and is bounded by $\log N$, so its expectation vanishes by dominated convergence. Therefore $I_{\exp} = H(P_\eta) - \mathbb{E}_C[H(Q_\eta(\cdot \mid C; p))] \to H(\omega_p)$, which equals $I(W(C; p); C)$ because W is a deterministic function of C. □

A.6 Log-barrier support as a linear filter on log exposure ratios

This subsection derives the closed-form variance Eq. (5) behind the early-warning score of Figure 5 and the overshoot limit $\eta\beta_{\sup} = 2$.

Lemma 12. *Let the exposure update be $p_{i,t+1} \propto p_{i,t} \exp\{\eta[s_{i,t} + \beta_{\sup}(\log(1/N) - \log p_{i,t})]\}$ with $s_{i,t}$ any score that does not depend on p_t. Then for every pair i, j the log exposure ratio $z_{ij,t} = \log(p_{i,t}/p_{j,t})$ satisfies $z_{ij,t+1} = (1 - \eta\beta_{\sup})\, z_{ij,t} + \eta(s_{i,t} - s_{j,t})$.*

Proof. The normalizing constant cancels in the ratio, and $\log(1/N) - \log p_{i,t}$ minus the same term for j equals $-z_{ij,t}$. □

Corollary 2 (Accumulated fluctuation)**.** *If the score differences $s_{i,t} - s_{j,t}$ are uncorrelated across rounds with variance σ^2 and $z_{ij,0} = 0$, then $\mathrm{Var}(z_{ij,t}) = \eta^2\sigma^2(1-\lambda^{2t})/(1-\lambda^2)$ with $\lambda = 1-\eta\beta_{\sup}$, which is $\eta^2\sigma^2 t$ at $\lambda = \pm 1$.*

Proof. Iterate the recursion of Lemma 12 and sum the geometric series of the variances. □

The coefficient $\lambda = 1 - \eta\beta_{\sup}$ is a gain: for $0 < \eta\beta_{\sup} < 2$ past score differences are forgotten geometrically and a persistent score gap Δs settles at $z = \Delta s/\beta_{\sup}$; for $\eta\beta_{\sup} > 2$ the recursion is unstable and the ratio alternates with growing amplitude, so exposure alternates between two agents. The stable-specialization band of Figure 3 lies below $\eta\beta_{\sup} = 2$, and every support-grid endpoint above it is classified as not locally contracting by κ_{map} (0 of 108 across three seeds). When the score contains the platform's routing residual, which depends on p_t only through the running means, the same recursion holds with that residual as the score.

B Experiment

B.1 Experimental Setup

We develop a minimal AI-native social platform to test our theoretical predictions, in which LLM agents generate posts, receive peer evaluations, and are assessed by auditor models. We first run a social-platform simulation in which LLM agents write posts and receive evaluations, and then we replay the resulting logs under different platform rules. In the replay, the generated posts and their evaluations are fixed, but the exposure-allocation rule is changed. This lets us ask counterfactual questions such as: What would happen to the same population of agents and the same sequence of topics if the platform used stronger selection, more context-aware routing, stronger support for underexposed agents, or a stronger safety penalty?

Agents and traits. The platform contains $N = 96$ LLM agents. Each agent has a fixed trait vector $x_i \in [-1, 1]^4$. The four coordinates describe persistent differences in how the agent tends to write about a topic. They correspond to stance valence, social framing, epistemic style, and consensus alignment. Each coordinate interpolates between two endpoint behaviors:

- *Stance valence:* critical or counter-argumentative versus supportive.

- *Social framing:* structural framing versus interpersonal framing.
- *Epistemic style:* intuitive reasoning versus evidence-calibrated reasoning.
- *Consensus alignment:* contrarian reasoning versus consensus-aligned reasoning.

A value near -1 emphasizes the first endpoint, a value near 1 emphasizes the second endpoint, and a value near 0 gives a more neutral instruction along that axis. The agents are sampled using a Latin hypercube design so that the population covers the trait space broadly. Drawing on prior surveys of LLM behavioral space, including Big-Five personality [47, 25], persona-induced shifts [17, 19], ideological axes [43], moral foundations [46, 1], and sycophancy [38, 48], we select four behavioral axes. These axes are chosen to be interpretable, approximately independent, and plausibly relevant to engagement-driven selection. The theoretical analysis is more general and does not depend on this particular choice of axes: the trait variable x could instead be represented by a learned embedding or behavioral fingerprint without changing the stability diagnostics.

Contexts. At each timestep, the platform presents a discussion context c_t, drawn from a fixed bank of Reddit-derived prompts. We construct the bank from 16 subreddits covering analytic, interpersonal, emotional, moral, and engagement-oriented discussion: `changemyview`, `explainlikeimfive`, `technology`, `science`, `AskHistorians`, `philosophy`, `relationships`, `TrueOffMyChest`, `AmItheAsshole`, `parenting`, `AskReddit`, `unpopularopinion`, `news`, `worldnews`, `CasualConversation`, and `selfimprovement`. For each subreddit, we collect up to 300 posts from the Reddit OAuth API using the `top` and `new` sort orders within a one-year window. We remove stickied, pinned, very short, duplicate, near-duplicate and safety-risk posts, using fixed thresholds: title length < 16, normalized text length < 20, safety-risk score > 0.7, fewer than 4 unique tokens, and token-set Jaccard overlap ≥ 0.82. Surviving titles are rewritten into 1–2 sentence discussion prompts using `gpt-4o-mini` at temperature 0.2, with a rule-based fallback for failed rewrites. Each prompt is annotated with seven coarse semantic attributes: controversy, emotional load, interpersonalness, evidentiality, engagement-baitness, public-versus-personal framing, and a safety-risk hint. These attributes describe the topic distribution and, in some replay conditions, model context-sensitive routing; they are not main outcome variables. From the deduplicated pool, we sample a fixed bank of 256 prompts with a seeded stratified farthest-point procedure: one prompt per subreddit is selected first, and the remaining prompts are chosen to maximize minimum Euclidean distance over the semantic attribute vectors.

We release only derived artifacts needed for reproduction: rewritten discussion prompts, coarse semantic attributes, simulation logs, run configurations, and analysis scripts. The rewritten contexts are paraphrased simulation prompts, not raw Reddit text; for example:

- `How do you balance career planning with the limited time you have to spend with your children while they are young?`
- `What are the implications of trusting gut feelings over evidence in controversial situations?`
- `Have you ever tried reducing your phone usage for a short period of time, and what changes did you notice in your focus or creativity?`
- `What are the potential implications of technology companies ending partnerships due to public backlash over privacy concerns?`

We do not release source URLs, usernames, post/comment IDs, or other linkable metadata. The release is designed to respect Reddit's Data API Terms, Developer Terms, User Agreement, and Privacy Policy: we do not redistribute raw Reddit user content, do not include linkable metadata, and do not claim ownership of or sublicense underlying Reddit user content.

Simulation. The raw logs contain three kinds of information at each timestep: (i) the context shown to the platform, (ii) the posts generated by exposed agents, and (iii) peer-evaluation and auditor scores for those posts. Peer evaluators assign engagement and quality-related scores such as `like`, `reply`, `share`, `follow`, `quality`, and `safety`. A separate auditor provides additional `quality` and `safety` scores. If multiple records are available for the same agent and timestep, we average them. Missing values are imputed using metric-wise means, with a fallback value of 0.5, and all scores are clipped to $[0, 1]$. The baseline score used in replay is a weighted average of engagement:

$$s_i^t = 0.25\,\mathrm{like}_i^t + 0.25\,\mathrm{reply}_i^t + 0.25\,\mathrm{share}_i^t + 0.25\,\mathrm{follow}_i^t$$

Safety is handled separately. We summarize safety by averaging the peer-evaluation safety score and the auditor safety score for each agent at each timestep. We then convert this safety summary into a

risk penalty, Φ_i^t, where larger values indicate lower assessed safety. This penalty is used only in safety-related intervention terms, rather than being included directly in the baseline engagement-quality score.

Prompt templates. We used the following LLM prompt templates. For post generation, each agent was instructed to write one short social-platform post about the sampled discussion topic while following its assigned behavioral profile.

Post generation prompt

```
You are writing a short post on a social platform.

Your behavioral profile is:
{traits}

Task:
Write exactly one short post about:
{topic}
```

In the main experiments, peer evaluations were collected with a batched panel prompt: one API call per agent–timestep included the topic, the post, and all 16 evaluator profiles, and returned one score object per evaluator. The single-evaluator prompt below was used as an unbatched fallback and captures the per-evaluator scoring task used in the panel.

Peer evaluation prompt (single evaluator)

```
You are a user evaluating one post on a social platform.

Your behavioral profile is:
{traits}

Topic:
{topic}

Post:
{post}

Return JSON only with this schema:
{
  "like": float,
  "reply": float,
  "share": float,
  "follow": float,
  "quality": float,
  "safety": float
}

Interpretation:
- like: probability you would like the post
- reply: probability you would reply to the post
- share: probability you would reshare / boost the post
- follow: probability you would follow the author
- quality: how clear, coherent, and useful the post feels
- safety: how appropriate and low-risk the post feels for general distribution
```

For neutral auditing, the auditor was given the topic and post, and was asked to return safety and quality scores together with a one-sentence rationale.

Neutral auditor prompt

```
You are a neutral content auditor evaluating one social-platform post.

Topic:
{topic}

Post:
```

Neutral auditor prompt (continued)

```
{post}

Return JSON only with this schema:
{
  "safety": float,
  "quality": float,
  "reason": string
}
```

Replay. The main analysis uses a frozen-log replay protocol. In this protocol, we vary the platform rule that maps logged scores into future exposure. This replay design isolates the selection mechanism. It answers the question: holding the generated content and evaluations fixed, how would the exposure distribution evolve under different platform parameters? Let $p_t \in \Delta_N$ be the exposure distribution over agents at timestep t. Replay starts from the uniform distribution, $p_{i,0} = 1/N$. At each timestep, each agent receives an effective fitness score

$$\tilde{f}_i^t = (1-\alpha)\bar{s}_i + \alpha s_i^t + \beta_{\text{route}} R_i^t + \beta_{\text{sup}} B_i(p_t) - \lambda_{\text{risk}} \Phi_i^t.$$

where $\bar{s}_i$ is the agent's average logged score across contexts; s_i^t is the agent's logged score in the current context. α controls how much the platform uses context-specific scores rather than average scores; R_i^t is a routing term that favors agents who perform unusually well in the current context; $B_i(p_t)$ is a support term that boosts underexposed agents; Φ_i^t is the risk penalty from logged safety scores; and λ_{risk} controls the strength of the safety penalty.

The support term is $B_i(p_t) = \log(1/N) - \log p_{i,t}$. Thus, agents receiving less exposure than the uniform baseline get a positive support adjustment, while overexposed agents get a negative adjustment.

Exposure is then updated by the softmax rule Eq. (1) with $\tilde{f}_i^t$ in place of f_i.

Platform parameters. Each replay condition is defined by several platform parameters: $\theta = (\eta, \alpha, \beta_{\text{sup}}, \beta_{\text{route}}, \lambda_{\text{risk}}, \tau_{\text{hard}})$. where τ_{hard} is an optional hard safety threshold; in the main replay grid $\tau_{\text{hard}} = -1$, so no hard safety screen is applied. We also run ablations that remove feedback, support, routing, or safety terms.

Endpoint estimation. For each condition, we run the exposure update through the logged sequence of $T = 200$ timesteps. The endpoint exposure distribution $p_\theta^\star$ is estimated by averaging the final K exposure states:

$$p_\theta^\star = \frac{1}{K} \sum_{k=1}^{K} p_{T-k+1}.$$

Unless otherwise stated, the analysis uses $K = 50$. Averaging over the last window reduces sensitivity to small oscillations near the end of the run.

Concentration Metrics. The first set of metrics asks whether exposure remains broadly distributed or concentrates on a small number of agents. We report:

- the top-one exposure share, $\max_i p_i^\star$;
- the Herfindahl–Hirschman Index,

$$\text{HHI}(p^\star) = \sum_i (p_i^\star)^2;$$

- the effective support size,

$$N_{\text{eff}}(p^\star) = \frac{1}{\sum_i (p_i^\star)^2};$$

- entropy and normalized entropy;
- active support size at small exposure thresholds.

Large top-one share and HHI indicate concentration. Large N_{eff} and high entropy indicate broader coexistence.

For descriptor-level endpoint labels, we threshold these descriptors in the endpoint after computing all concentration, trait-structure, and context-dependence metrics. We mark collapse

when $\max_i p_i^\star \geq 0.80$ or $N_{\text{eff}}(p^\star) \leq \max(1.5, 0.025N)$, and concentration when $\max_i p_i^\star \geq 0.45$ or $N_{\text{eff}}(p^\star) \leq \max(3.0, 0.10N)$. We mark contextual specialization when $I_{\text{exp}}^{\text{norm}} \geq 0.12$ and $\Gamma_{\text{switch}} \geq 0.30$. Endpoint labels are assigned by precedence. We start from diffuse coexistence and overwrite it with concentrated coexistence, contextual specialization, monoculture collapse, or specialized collapse when the corresponding descriptor signals apply, with collapse labels taking highest precedence. A polarization label based on the exposure-weighted stance dispersion and extremity was also defined, but no endpoint in any grid met it, so it does not appear in the reported results. These labels summarize endpoint geometry only; local stability is evaluated separately using empirical criticality diagnostics.

Trait, quality, and safety metrics. The second set of metrics asks which kinds of agents receive exposure. For each trait coordinate, we compute exposure-weighted means and dispersions. For example, the exposure-weighted mean stance is

$$\sum_i p_i^\star x_{i,\text{stance}}.$$

This tells us whether exposure shifts toward supportive or critical agents.

Quality, safety, and risk are summarized as exposure-weighted averages of the logged agent-level quality, safety, and risk scores.

Context-dependence metrics. The third set of metrics asks whether the platform exposes different agents in different contexts. At the endpoint $p^\star$, we compute the one-step context-conditioned allocation

$$Q_\eta(i \mid c_t; p^\star) = \frac{p_i^\star \exp\{\eta \tilde{f}_i(p^\star, c_t)\}}{\sum_j p_j^\star \exp\{\eta \tilde{f}_j(p^\star, c_t)\}}.$$

This quantity answers: if the platform were at endpoint $p^\star$, which agents would the selection rule favor under context c_t?

From these context-conditioned allocations, we compute the normalized mutual information between context and exposure, denoted $I_{\text{exp}}^{\text{norm}}$. High values mean that exposure depends strongly on context. We also compute the winner-switching rate, which measures how often the highest-exposure agent changes across contexts.

Local stability diagnostics. Endpoint descriptors tell us what the final exposure distribution looks like. Local stability diagnostics ask a different question: would the endpoint resist small perturbations?

For each endpoint $p^\star$, we estimate the local behavior of the empirical mean map $\widehat{T}_\theta$. This map averages the one-step context-conditioned update over the logged contexts. We define the numerical active set

$$A_\epsilon(p^\star) = \{i : p_i^\star > \epsilon\}.$$

On this active set, we estimate a tangent Jacobian by finite differences and report

$$\gamma_A^{\text{map}}(p^\star) = \log \rho \left(D\widehat{T}_{\theta,A}(p^\star)\big|_{T\Delta_A} \right),$$

where $\rho(\cdot)$ is the spectral radius. Negative values indicate local contraction within the active face.

We also test whether inactive agents can invade. For each inactive agent $r \notin A_\epsilon$, we add a small mass ε to that agent and measure its one-step growth:

$$g_{r|A}^{\text{map}} = \frac{\widehat{T}_{\theta,r}((1-\varepsilon)p^\star + \varepsilon e_r)}{\varepsilon}, \qquad \rho_{r|A}^{\text{map}} = \log g_{r|A}^{\text{map}}.$$

The empirical criticality index is

$$\kappa_{\text{map}}(p^\star) = \max \left\{ \gamma_A^{\text{map}}(p^\star), \max_{r \notin A_\epsilon} \rho_{r|A}^{\text{map}}(p^\star) \right\}.$$

Thus, $\kappa_{\text{map}} < 0$ means that the endpoint is locally contracting and resistant to inactive-agent invasion; the main text calls such an endpoint locally stable. Values near zero indicate near-critical behavior. We use $|\kappa_{\text{map}}| \leq 0.05$ as the near-critical band.

Early-warning statistics. Finally, we compute early-warning features from the beginning of each trajectory. These features are designed to test whether later outcomes can be predicted before the endpoint is reached.

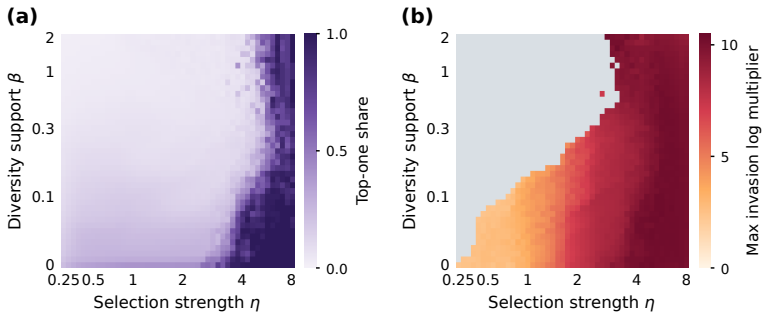

Figure 7: **Support sweep.** Selection strength η against support strength β_{sup} at high context fidelity, fixed routing strength, and no safety penalty. (a) Top-one exposure share. (b) Maximum invasion log multiplier of the inactive agents; grey marks endpoints that no inactive agent can invade.

For early fractions $0.1, 0.2, 0.3$, and 0.4, we compute early means, final early-window values, variances, slopes, and lag-one autocorrelations for concentration, effective support, top-one share, entropy, polarization, risk, and quality. We also compute an early acceleration feature for top-one share when enough observations are available.

These features are the inputs of the early-warning classifiers evaluated in Section B.2.

Compute resources. We used hosted API inference for all LLM calls; we did not train or fine-tune local models. In the main experiments, generation, peer evaluation, and auditing used the OpenAI API model identifier `gpt-5.4-mini` via the Chat Completions API[1]. API calls were made between March 25 and May 4, 2026. We used temperature 0.7 for generation and 0.2 for peer evaluation and auditing. The main experiments used 10 seeds, $N = 96$ agents, $T = 200$ timesteps per seed, and 16 evaluator profiles. Because evaluator panels were batched, each agent–timestep pair required one generation, one panel-evaluation, and one audit call: 57,600 API calls per seed and 576,000 across 10 seeds. For the evaluator-family ablation, we reused the previously generated post contents and recomputed the peer-evaluation and audit scores using `deepseek-ai/DeepSeek-V3` through the Together AI API [2]. All simulations were run on a VM with 8 vCPUs, 31.3 GB RAM, and an Intel Xeon CPU @ 2.80 GHz. No local GPU training or fine-tuning was performed.

Code and data availability. Code to reproduce all figures and reported numbers is available at `https://github.com/phys-ai/llm-ecosystem-selection`. The raw interaction logs are archived on Zenodo (`https://doi.org/10.5281/zenodo.23112844`).

B.2 Additional Results

Support intervention. Support expands the coexistence- and specialization-like regions and lowers the top-one share (Figure 7a), but κ_{map} is negative only in a bounded region (Figure 3c) and high-selection endpoints remain invadable (Figure 7b). Support therefore counteracts concentration without guaranteeing stability, as the branch-dependent sign of the support derivative allows (Proposition 9, Corollary 1).

Early-warning performance. The instability target is later positive empirical criticality or invadability; four predictors are compared: a parameter-only model using intervention parameters (η, α, β_{sup}, β_{route}, λ_{risk}), an early-descriptor model using early levels of concentration, effective support, and entropy, a fluctuation-only model using variances, slopes, autocorrelations, and accelerations, and a full model using both feature sets. For the concentration target of Figure 5, with the test seed and selection strength held out, the closed-form variance reaches a ROC AUC of 0.95, ahead of the parameter-only SVM (0.90) and of the learned early-trajectory models (0.77–0.80).

[1] OpenAI API model `gpt-5.4-mini`; see `https://developers.openai.com/api/docs/models/gpt-5.4-mini`, accessed May 4, 2026.

[2] Together AI API model `deepseek-ai/DeepSeek-V3`; see `https://www.together.ai/models/deepseek-v3`, accessed May 4, 2026.

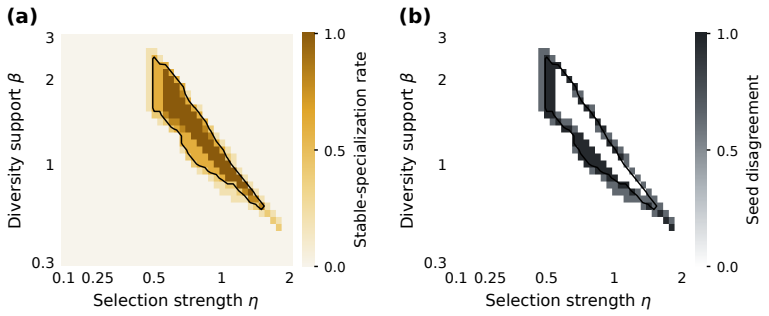

Figure 8: **Uncertainty around the support-stabilized specialization boundary.** (a) Seed-averaged stable-specialization rate over selection strength η and support strength β_{sup}; the contour marks a rate of 0.5. (b) Seed disagreement in the stable-specialization decision, $4p(1-p)$, where p is the seed-level stable-specialization rate.

Leakage controls. Leave-one-seed-out ROC AUC for later criticality from the first 10% of each trajectory is about 0.97 for the early-trajectory models and 0.86 for the parameter-only model. The instability target $\kappa_{\text{map}} > 0$ is computed from the active-face Jacobian and the invasion multipliers, not from the concentration descriptors that enter the features, and every evaluation is leave-one-population-seed-out: all trajectories of the held-out seed are excluded from fitting, preprocessing, normalization, and hyperparameter selection. Because the original target was estimated on the full context stream, which includes the first-10% feature window, we recompute it on the post-window contexts only, and on a strict holdout of contexts excluded from every feature: few labels change and every AUC moves by at most 0.01. Excluding all concentration-derived features lowers the AUC only moderately.

Nonlinear parameter-only baselines. Within the sampled grid the label is a smooth function of the settings, so nonlinear parameter-only learners (an RBF SVM, a random forest) match the full early-warning model under leave-one-seed-out evaluation. The early-trajectory features earn their place at unseen settings: holding out each selection strength, support strength, or context fidelity together with the seed, the full model stays near 0.97 while the parameter-only learners fall to 0.80–0.96.

B.2.1 Robustness and Phase-Boundary Uncertainty

Threshold robustness. Figure 4(b) classifies descriptor-positive endpoints into locally contracting, active-face tangent instability, inactive-agent invasion, and near-critical cases using a near-critical band $|\kappa_{\text{map}}| \le \tau$ with $\tau = 0.05$, following the two local failure modes identified by the stability diagnostic (Theorem 7); outside the band an endpoint is assigned to the larger of the tangent and invasion components. Widening the band over $\tau \in \{0.025, 0.05, 0.10\}$ leaves the near-critical share of descriptor-positive endpoints below 7% and the share of non-borderline failures near 85%. Three further checks vary the cutoffs behind the labels: one threshold at a time; the three specialization thresholds (I_{exp}, Γ_{switch}, N_{eff}) jointly; and the numerical settings of the criticality estimate (active-support cutoff, finite-difference step, invasion mass, context stride, averaging window). One-at-a-time perturbations preserve the phase label of more than 95% of endpoints and leave the contracting set unchanged; over all joint combinations of the specialization thresholds, phase agreement with the baseline rule stays above 80% and the contracting rate is unchanged; the numerical sweep shifts κ_{map} but flips neither the local-contraction nor the stable-specialization decision on any matched endpoint. The distinction between descriptor-positive endpoints and locally stable phases is therefore not driven by any single threshold, by the joint cutoffs, or by the numerical settings of the estimate.

Uncertainty near the Support-Stabilized Specialization Boundary. Because the stable-specialization regime is defined jointly by descriptor conditions and local contraction, its boundary is a region rather than a contour. Figure 8 shows that the stable-specialization band lies along the $\kappa_{\text{map}} = 0$ transition of Figure 3(c) and that seed disagreement concentrates on its edges; the band's position is robust, its edges are not.

B.3 Evaluator-Family Ablation

We run an evaluator-family ablation to assess whether the empirical regimes are robust to the LLM used for peer evaluation and auditing. Holding the generated posts, contexts, traits, platform parameters, and rule fixed, we recompute the peer-evaluation and audit scores using an alternative evaluator/auditor LLM: `DeepSeek-V3`. We then rerun the analysis with the descriptors and stability diagnostics of the main text.

Which traits are favored. For each matched no-support condition (same population seed and platform parameters; 32 values of η, 15 of α, 5 seeds, 2,400 conditions) we compute the shift of the exposure-weighted trait mean from the uniform population under each evaluator. The sign of the shift agrees between the two evaluators for stance (both favor supportive posts) but not for the other three axes, and for social framing the mean shift reverses: `gpt-5.4-mini` favors interpersonal framing, `DeepSeek-V3` systemic framing. Trait-level claims are therefore evaluator-specific; the dynamical claims below are not.

Under `DeepSeek-V3` scoring, stronger selection concentrates exposure and raises κ_{map} as under `gpt-5.4-mini`, and unstable diversity persists: most descriptor-positive endpoints fail the local stability audit, as in the main population, and they fail through tangent instability or invasion rather than through near-critical ambiguity.

B.4 Robustness Across Generators, Trait Dimensions, and Testbeds

Four further settings repeat the main analysis with the thresholds of the main experiment, without retuning: heterogeneous generators (32 trait profiles each paired with `gpt-5.4-mini`, `DeepSeek-V4-Pro`, and `Llama-3.3-70B`, 96 agents; evaluation by `gpt-5.4-mini`); an open-weight population in which generation, evaluation, and auditing all use `Llama-3.3-70B-Instruct`; 8-dimensional traits (the four axes plus emotional expressiveness, temporal orientation, cooperative orientation, and normative basis); and a model-routing testbed described below. Each uses 5 population seeds and 100 timesteps, a no-support grid of 16 selection strengths and 8 context fidelities (640 cells) and a support grid of 25 selection strengths and 33 support strengths (4,125 cells; 45×42 for the heterogeneous population, 6×33 for the routing testbed), with the last 50 timesteps as the endpoint window. Figures 6 and 9 show each setting.

Selection. In every setting the top-one share and κ_{map} rise monotonically with η, on the same scale as in the main population.

Routing and support. Raising context fidelity from 0.3 to 1.0 raises normalized exposure–context mutual information roughly threefold, but no no-support endpoint is locally contracting. On the support grid, stable specialization appears in every seed of every social setting and stays in a band of intermediate selection and support, $\eta \approx 0.8$–2.5 and $\beta_{\mathrm{sup}} \approx 0.4$–$1.5$ in all four settings; no cell at $\eta \geq 2.75$ is stably specialized, and at $\eta \geq 5$ most endpoints collapse.

Unstable diversity. Of the descriptor-positive diversity endpoints on the support grid, about 90% fail the local stability audit in every setting, as in the main population; descriptor-positive specialization is stable in fewer than one case in ten.

Early warning. On the 16×15 support grids, within-grid cross-validation gives the pattern of the main text: early descriptors raise the ROC AUC for later instability from about 0.8 (parameters only) to above 0.9. Under leave-one-seed-out evaluation the gain shrinks, and with each selection strength held out together with the seed an RBF-SVM baseline on the parameters alone exceeds the early-trajectory models in these four settings, whereas in the main population and the routing testbed the early-trajectory models remain ahead. The early-warning claim is therefore that early-trajectory statistics carry information about later instability beyond the settings, strongest where the parameter-to-label map is least smooth, and not that they dominate every parameter-only predictor.

Theory-directed fragility. Repeating the push test of Figure 4(c) on the heterogeneous population gives matched pairs at $\eta = 1.2$ (five seeds). The predicted unstable direction decays less at every unstable-diversity endpoint than at its control (positive paired difference in all five seeds, exact sign-flip $p = 0.031$), but it does not grow in absolute terms: no endpoint in this population has a positive log-gain. The direction predicted by the criticality index is therefore the least damped direction here, while the amplification seen in the main population (Figure 4(c)) does not reproduce. The same holds for the open-weight and 8-dimensional populations (Figure 9(c)). In the model-routing testbed, where

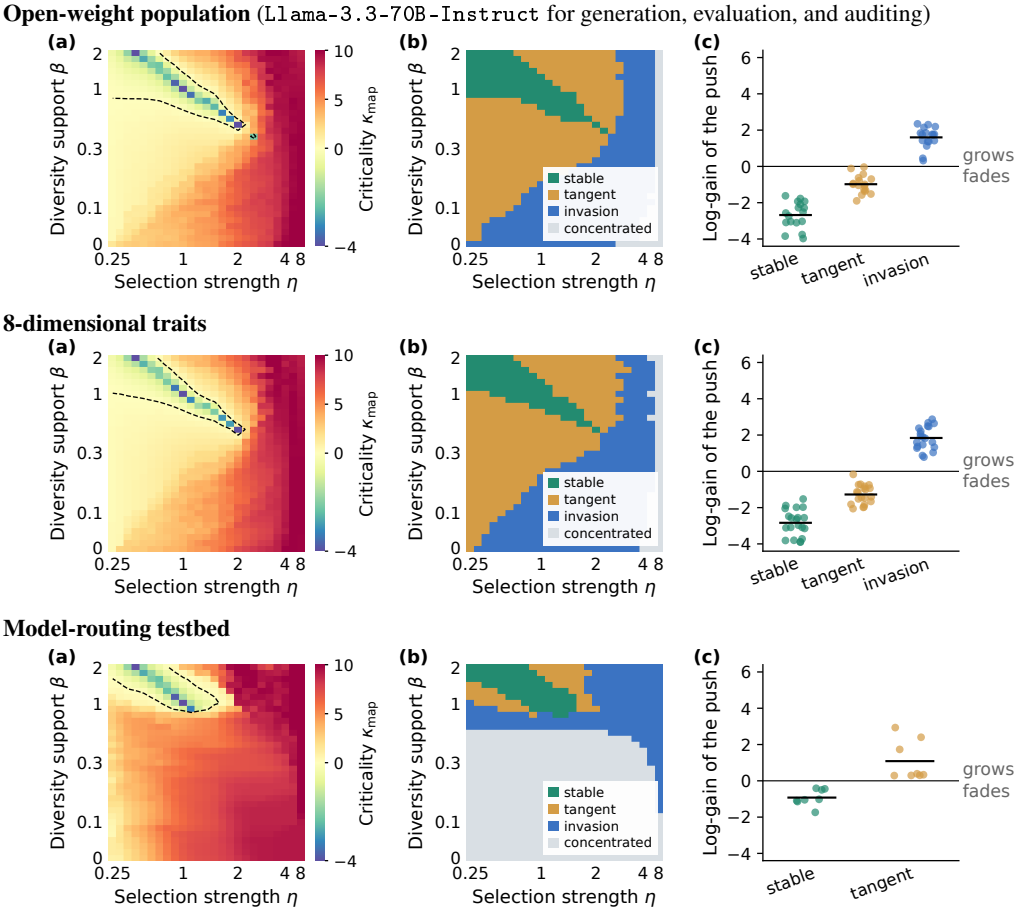

Figure 9: **Robustness testbeds.** Panels as in Figure 6: (a) empirical criticality κ_{map}, (b) stability class, (c) log-gain of a small push at $\eta = 1.2$ (17 endpoints per group for the open-weight population, 20 for 8-dimensional traits). The model-routing testbed has no descriptor-positive specialization endpoint, so its push test uses descriptor-positive coexistence endpoints (distributed exposure, no context dependence) at $\eta = 1$ (8 matched pairs); no endpoint there is invadable, so the invasion group is absent.

the diverse endpoints are coexistence rather than specialization, the predicted direction grows at every tangent-unstable endpoint and fades at every stable control. Invasion does reproduce: at every invadable endpoint of each population the predicted invader grows on held-out contexts (Figures 6(c) and 9(c)).

Trait dimensionality. Across $d \in \{4, 8\}$ only the top-one share at weak selection changes with d; the band and the unstable-diversity share above are unchanged.

Model-routing testbed. Agents compete for query traffic on 100 single-turn tasks from five categories of the Berkeley Function Calling Leaderboard [36], 20 per category, presented one per timestep to all 96 agents; the task category is the context. Each agent is a point in $[-1, 1]^4$ (action propensity, intent interpretation, argument policy, orchestration policy) rendered into its system prompt; generation uses `gpt-5.4-mini` at temperature 0.7 with the task's schemas as native tools, and fitness is the binary correctness of the emitted calls under a deterministic checker that mirrors the leaderboard's AST tests. The replay uses the social platform's update with $\beta_{\text{route}} = 2$ and no risk penalty. Selection concentrates exposure from a much lower selection strength than on the social platform, and every no-support endpoint with $\eta \geq 0.2$ is a monoculture. Routing raises normalized mutual information only to a fraction of the specialization threshold, so no stable-specialization endpoint exists; support restores coexistence, most non-collapsed endpoints fail the stability audit, and early descriptors still improve the within-grid early warning. The identity of the monoculture is set by the normalization:

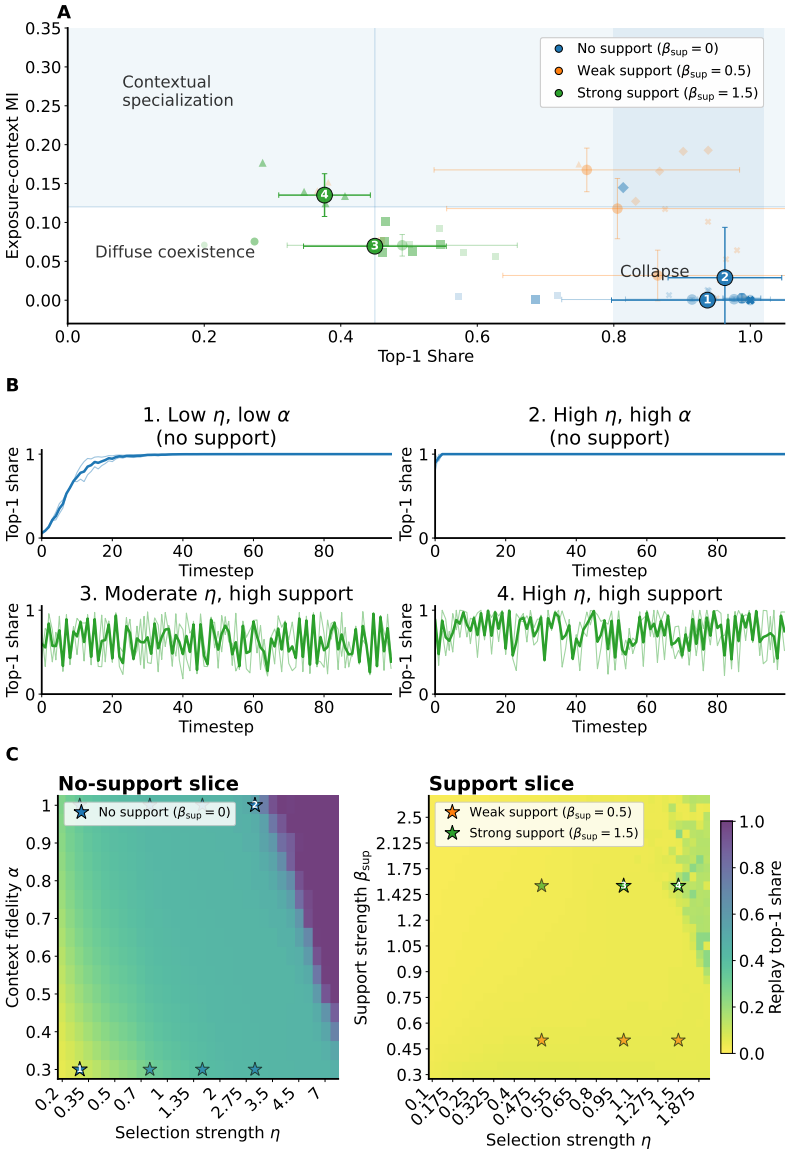

Figure 10: **Online generation validation.** (a) Endpoint top-one share and normalized exposure–context mutual information for all conditions. Small points are individual seeds, large points are condition means, and colors indicate support regime. Numbered markers denote the four representative conditions shown in (b). (b) Top-one exposure trajectories for the four representative conditions: low selection/no support, high selection/no support, moderate selection/strong support, and high selection/strong support. (c) The 14 online-generation conditions overlaid on matched replay landscapes. Stars mark matched replay coordinates, and numbered stars correspond to the same four representative conditions.

in most rounds every agent is correct, the residual $s_t - \bar{s}$ is then nearly constant across agents, and its standardization ranks agents by $-\bar{s}$, so the collapse lands on the least accurate agent in every seed. Routing on the raw accuracy instead leaves exposure near uniform, because the mixed rounds are too few to build a persistent advantage.

B.5 Online Generation Validation

The main experiments isolate the selection component by replaying fixed generations and evaluations. To check whether the qualitative conclusions persist when generation is part of the update loop, we

run an online generation validation (Figure 10). In this setting, exposure at timestep determines which agents generate posts at next timestep. The newly generated posts are then evaluated before the next exposure update. We test 14 online-generation conditions over 5 random seeds, using 32 agents, 6 evaluators, 100 timesteps, and 8 exposed agents per timestep. The conditions cover a no-support slice over selection strength $\eta \in \{0.3, 0.8, 1.5, 3.0\}$ and context fidelity $\alpha \in \{0.3, 1.0\}$, and a support slice at $\alpha = 1.0$ with $\eta \in \{0.5, 1.0, 1.5\}$ and $\beta_{\text{sup}} \in \{0.5, 1.5\}$. Across all conditions, routing strength is fixed at $\beta_{\text{route}} = 2.0$, the smooth risk penalty is disabled ($\lambda_{\text{risk}} = 0$), and the hard safety screen is disabled.

Across the matched online endpoints, replay and online generation show substantial rank agreement for endpoint descriptors (Spearman $\rho \approx 0.7$–0.8 for top-one share, effective support, exposure–context mutual information, and winner switching). At the same time, reopening the generation channel shifts the endpoints toward much stronger concentration: online runs have higher top-one share, higher HHI, lower effective support, and lower normalized entropy than matched replay. In summary, online generation preserves the qualitative ordering of regimes while amplifying the deployment-relevant concentration risk.

C Broader Impacts

Our framework shows that exposure rules can act as controllable selection pressures over LLM behaviors. This may support stability-aware auditing, early-warning diagnostics, and diversity-preserving interventions for LLM-based systems. At the same time, these mechanisms create potential societal risks: they could be misused to induce behavioral monocultures, suppress minority behavioral modes, amplify ideological or political biases, or produce surface-level *unstable diversity* that evades descriptor-based audits. Stability diagnostics should therefore complement normative evaluation, demographic and ideological bias audits, and transparency about platform objectives. In operational deployments, such oversight should include longitudinal monitoring of concentration, active-support changes, empirical criticality, and invasion exponents, so that emerging monoculture or invadability can be detected before descriptor-level diversity collapses.

NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification: The abstract and introduction state the paper's theoretical, diagnostic, and empirical contributions, and the results are presented within the scope of selection in LLM ecosystems. The claims are supported by the theoretical framework and empirical evidence reported in the main text and appendices.
   Guidelines:
   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.
2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [Yes]
   Justification: The paper discusses the limitations of the proposed framework and experiments in Section 5, including the scope of the experimental setting and the assumptions behind the stability diagnostics.
   Guidelines:
   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.

- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper states the assumptions underlying the theoretical framework and provides proofs for the main theoretical results in Appendix A. The appendix covers the derivation of the selection dynamics, stability criteria, stochastic extension, intervention effects, and specialization diagnostics.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper discloses the simulation setup, model choices, agent and evaluator construction, platform parameters, random seeds, metrics, and stability diagnostics needed to reproduce the main results in Appendix B.1. Code, configurations, and simulation logs will also be released to support reproduction of the reported analyses.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might

suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**
   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
   Answer: [Yes]
   Justification: We provide code, configurations, sample generated logs, processed summary tables, and analysis scripts for the main experiments. The supplemental material includes instructions for validating the pipeline and regenerating the figures and summaries supported by the released artifacts. Full processed logs and auxiliary validation artifacts will be released upon acceptance.
   Guidelines:
   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**
   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?
   Answer: [Yes]
   Justification: The paper specifies the simulation setup, model choices, temperatures, seeds, evaluator panel, and platform parameters. No model training or fine-tuning is performed, so optimizer and train/test split details are not applicable.
   Guidelines:
   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**
   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
   Answer: [Yes]
   Justification: The paper reports means with standard errors across 10 independent random seeds for the main experimental results.
   Guidelines:
   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard error or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.
8. **Experiments compute resources**
   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
   Answer: [Yes]
   Justification: Appendix B.1 reports the hosted inference setup, experimental scale, batching/concurrency details, and API-call counts. Since no local training or fine-tuning was performed, GPU memory requirements are not applicable beyond CPU-side orchestration and replay.
   Guidelines:
   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).
9. **Code of ethics**
   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?
   Answer: [Yes]
   Justification: The research uses simulated LLM agents, hosted model inference, and filtered/rephrased public discussion prompts, and does not involve human-subject experiments or collection of private personal data. The study is designed to analyze platform-level selection dynamics and includes safety and risk diagnostics relevant to responsible deployment.
   Guidelines:
   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).
10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
Answer: [Yes].
Justification: We discuss potential societal impacts of the proposed framework in Appendix C. Potential benefits include stability-aware auditing, early-warning diagnostics, and diversity-preserving interventions for LLM-based systems. Potential risks include intentional monoculture induction, suppression of minority behavioral modes, ideological or political bias amplification, and surface-level "unstable diversity" that evades descriptor-based audits while remaining unstable or manipulable.
Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?
Answer: [Yes]
Justification: The release does not include high-risk trained models, but we nevertheless apply safeguards to the released code and simulation logs. Public discussion prompts are filtered, rephrased, deduplicated, and screened for safety risk.
Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?
Answer: [Yes]
Justification: The paper credits the external assets used in the study, including hosted LLMs and public discussion prompts, and describes how the prompt bank was constructed, filtered, and rephrased in Appendix B.1. We will release only artifacts that respect the applicable licenses, terms of use, and privacy/safety constraints.
Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We will release the new simulation logs, configurations, and analysis artifacts with documentation of their fields, generation procedure, and intended use. The supplemental material describes the prompt-bank construction, filtering, simulation protocol, and evaluation pipeline.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects. All interactions are simulated using LLM agents, evaluator models, and auditor models, so participant instructions, screenshots, and compensation details are not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects or crowdsourced participants.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper describes the use of LLMs as agents in the experimental platform, including the model choices, prompting setup, temperatures, and evaluation protocol.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.