Memetic Drift in Multi-Agent LLMs: Scaling Laws for Consensus Under Pluralistic Uncertainty

Anonymous Authors[1]

Abstract

Pluralistic alignment often requires AI systems to aggregate or deliberate over multiple defensible perspectives, but consensus alone does not reveal whether an outcome reflects collective reasoning, systematic bias, or chance. We study this ambiguity in near-tie discrete-choice settings, where several options are defensible and no external reward, ground truth, or payoff initially selects among them. We introduce Quantized Simplex Gossip (QSG), a minimal null model in which agents maintain simplex-valued belief states, communicate quantized samples, and adapt locally to one another's outputs; QSG traces consensus formation to mutual in-context learning, a regime we call memetic drift. The model yields early-drift identities and mean-field scaling laws in population size, communication bandwidth, adaptation strength, and internal uncertainty, and predicts a drift–selection crossover under weak asymmetries. We evaluate these predictions in a contextualized municipal budget-plan prompt with neutral plan codes, neutral naming games with GPT-4o and Claude Haiku 4.5, and QSG simulations, providing a diagnostic baseline for distinguishing evidence aggregation from amplified sampling noise.

1. Introduction

Multi-agent LLM systems are increasingly studied as interacting populations for debate, collaboration, voting, and committee-style recommendation (Du et al., 2024; Zhang et al., 2024; Chuang et al., 2024; Brockers et al., 2025; Zhao et al., 2024; Kaesberg et al., 2025; Becker et al., 2025), and are also studied for or proposed in domains such as law, finance, healthcare, policy, and scientific discovery (Watson et al., 2025; Xiao et al., 2024; Sreedhar et al., 2025; Kim et al., 2024; Gottweis et al., 2025). These interaction protocols are especially relevant to pluralistic alignment, where systems must represent, deliberate over, or aggregate multiple defensible human perspectives. In many such systems, agreement is treated as an operational success signal. But consensus is an ambiguous observable: a population may agree because it aggregated evidence, because its members share a systematic bias, or because early stochastic samples were amplified through interaction. The practical question is therefore not only whether a population reaches agreement, but what mechanism produced that agreement. We focus on near-tie discrete-choice settings, where this ambiguity is especially sharp because several options are defensible and no external reward, ground truth, or payoff strongly selects one option at the outset.

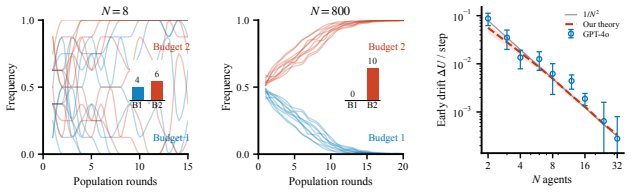
As a concrete example, consider a committee recommendation setting in which a municipal planning committee must choose one of several budget plans, while the staff memo states that the plans are equally strong under the available evidence. In such a setting, a final recommendation can look like collective agreement even if the winning option partly reflects early stochastic sampling. Figure 1 shows this phenomenon in a contextualized budget-plan task: small populations exhibit visible run-to-run variability, while larger populations can amplify even weak blank-memory asymmetries. This task is not intended as a benchmark of municipal planning quality; it is a controlled near-tie setting for diagnosing the mechanism by which agreement forms.

This example highlights the diagnostic problem that motivates the paper. A final vote alone cannot distinguish evidence aggregation from path-dependent amplification, because both can end in the same apparent consensus. What changes across population size is the balance between neutral sampling drift and weak selection from measured blank-memory asymmetries. QSG provides the corresponding null model: before interpreting agreement as collective reasoning, we ask how much agreement should be expected from quantized communication, uncertainty, and local adaptation alone.

The mechanism we isolate is mutual in-context learning. In standard single-agent in-context learning, an agent updates from examples drawn from an external distribution

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.



(a) Drift regime ($N = 8$). *(b)* Selection regime ($N = 800$). *(c)* Theory vs. GPT-4o.

Figure 1. **Overview: drift and weak-selection amplification in committee budget-plan recommendation tasks. Our theory predicts early polarization drift in collective opinion dynamics.** A population of N GPT-4o agents repeatedly recommends one of two near-tie municipal budget plans, shown as Budget 1/2. (a) At $N = 8$, ten runs remain visibly seed-dependent, with final majorities split 4 vs. 6 out of 10. (b) At $N = 800$, the weak asymmetry is amplified more consistently; all ten runs end with the same majority code by round 20. (c) Theory and experiment agree: QSG predictions (Sec. 2) match observed early drift across nine N values, with no parameters fit to the drift curve.

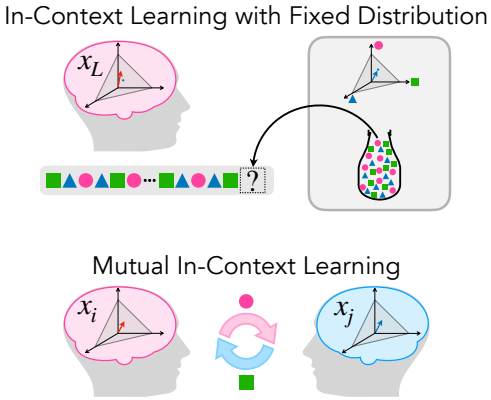


Figure 2. **Individual vs. mutual in-context learning.** Top: standard in-context learning, where a single agent updates from i.i.d. tokens drawn from a stationary external distribution. Bottom: mutual in-context learning, where agents update from one another's sampled outputs, so the population becomes its own evolving data source.

(Akyürek et al., 2023; von Oswald et al., 2023; Dai et al., 2023; Park et al., 2025a;b). In a multi-agent population, by contrast, the learning signal is generated by other agents' sampled outputs (Fig. 2). One agent's arbitrary early sample can therefore become another agent's evidence, and repeated local adaptation can amplify that sample into population-level agreement. Thus, in the multi-agent setting, the population becomes its own evolving data source. This makes final consensus mechanistically underdetermined: the same agreement can arise from genuine evidence aggregation, shared priors, or repeated reuse of sampled outputs. We call the resulting neutral coordination regime *memetic drift*, by analogy with neutral drift in evolutionary dynamics (Kimura, 1968; 1983; Moran, 1958; Clifford & Sudbury, 1973; Holley & Liggett, 1975; Liggett, 1999).

This framing makes the role of the naming game precise. The naming game is not meant to stand in for all collaborative reasoning tasks. Instead, it is a controlled limit in which semantic confounds, objective payoffs, and ground-truth correctness are removed, so that stochastic drift can be isolated (Steels, 1995; Baronchelli et al., 2006; 2008; Ashery et al., 2025; Flint et al., 2025). The naming-game experiments below are therefore not the motivating application, but the clean control for the same early-drift mechanism observed in the committee task.

The committee and naming settings therefore play complementary roles: one anchors the phenomenon in a semantically meaningful near-tie decision, while the other removes semantics, payoffs, and correctness to isolate the same discrete-choice drift mechanism.

To quantify this mechanism, we introduce Quantized Simplex Gossip (QSG), a minimal null model of simplex-valued internal states, quantized communication, and local adaptation. QSG predicts how early drift depends on population size, communication bandwidth, internal uncertainty, and effective adaptation strength. It also predicts a drift–selection crossover: when weak asymmetries are present, larger populations or higher-bandwidth communication can make those asymmetries more decisive, while in smaller or lower-bandwidth populations the outcome remains more seed-dependent.

Specifically, we make the following contributions:

1. **Mutual in-context learning and memetic drift as a near-tie diagnostic.** We identify a sampling-driven mechanism by which consensus can arise in weak-signal multi-agent LLM settings, trace it to mutual in-context learning, and frame the resulting neutral regime as memetic drift: stochastic amplification of early sampled outputs under approximate neutrality.
2. **Quantized Simplex Gossip (QSG).** We introduce QSG, a minimal and analytically tractable null model for simplex-valued beliefs, quantized communication, and local adaptation in discrete-choice coordination.
3. **Scaling laws for early drift and weak selection.** We derive drift identities and mean-field scaling laws showing how early polarization depends on population size, bandwidth, adaptation strength, and uncertainty, and how weak asymmetries enter through a drift–selection crossover.
4. **Empirical and simulation validation.** We test QSG in a contextualized committee budget-plan prompt with neutral plan codes, neutral naming games with GPT-4o and Claude Haiku 4.5, QSG simulations, and Top-m bandwidth interventions. In the committee task, probe-estimated adaptation and uncertainty predict early po-

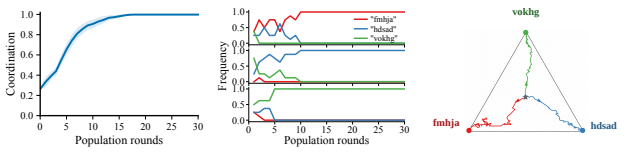
larization drift; these estimates are not fit to the drift curve.

2. Modeling: Quantized Simplex Gossip (QSG)

Both empirical settings can be viewed as repeated discrete-choice coordination protocols: a population of N agent memory states repeatedly selects among K candidate codes for a shared decision or referent. QSG models this shared interaction structure, not the full semantics of either task. Unlike naming-game protocols that condition updates on communicative success (Steels, 1995; Baronchelli et al., 2006; 2008; Ashery et al., 2025), QSG removes this channel: the listener adapts to the speaker's sampled output regardless of agreement, with simplex-valued uncertainty making sampling noise explicit. Each interaction selects an ordered speaker–listener pair; because the listener only records sampled outputs, coordination arises through mutual in-context learning rather than explicit reward. Figure 3 shows symmetry breaking in the clean-control naming-game setting: even with neutral synthetic labels and no correctness signal, independent runs break symmetry toward different eventual labels.

The central modeling challenge is the mismatch between continuous internal beliefs and discrete communication. Even under neutrality, sampling a discrete message from a continuous distribution injects stochasticity that can destabilize symmetric states and drive consensus. We capture this with **Quantized Simplex Gossip (QSG)**. Each agent holds a distribution $x_i \in \Delta^{K-1}$, the speaker sends a quantized message (Hard/Top-m/Soft), and the listener updates toward it with adaptation rate α (Martins, 2008). Figure 4 schematizes the QSG interaction protocol, with random speaker–listener pairing, quantized communication, and listener adaptation.

Agent state and neutrality. Fix $K \geq 2$ candidate codes and a population of $N \geq 2$ agents. Agent $i \in \{1, \ldots, N\}$ maintains an internal probability distribution $x_i \in \Delta^{K-1}$, where $\Delta^{K-1} \triangleq \{x \in \mathbb{R}^K : x_k \geq 0, \sum_{k=1}^K x_k = 1\}$. The population state is $X = (x_1, \ldots, x_N) \in (\Delta^{K-1})^N$. Throughout, token labels are *neutral and exchangeable*: there is no intrinsic fitness difference, external reward signal,



(a) Mean coordination. *(b)* Representative trajectories. *(c)* Winner-conditioned paths.

Figure 3. **Symmetry breaking in an LLM naming game** (GPT-4o, $N = 24$, $K = 3$). (a) Mean coordination across trials. (b) Representative label-frequency trajectories. (c) Winner-conditioned paths.

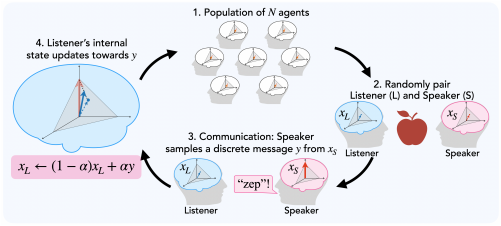


Figure 4. **Quantized Simplex Gossip (QSG): interaction protocol.** 1. A population of N agents, each with an internal state $x_i \in \Delta^{K-1}$. 2. An ordered speaker–listener pair is selected uniformly at random. 3. For a shared discrete choice, the speaker samples and sends a message y from x_S in the quantized regimes (Hard or Top-m); Soft transmits the full distribution. 4. The listener updates toward the received message with adaptation rate α, $x_L \leftarrow (1-\alpha)x_L + \alpha y$.

or prior bias associated with any convention index $k \in \{1, \ldots, K\}$. This is a theoretical neutrality assumption, not a claim that real LLM prompts are exactly neutral. In empirical LLM experiments, exact exchangeability is not guaranteed even for random codes. We therefore treat the experiments as *near-neutral*: we measure blank-memory priors and interpret residual asymmetries as weak selection forces. This distinction matters because large populations can amplify even small measured asymmetries.

We represent each agent by a continuous distribution x_i because LLM token generation induces a distribution over discrete tokens, and uncertainty in that distribution is a central control variable. Communication, however, is tokenized: agents exchange discrete outputs or short summaries, not real-valued internal representations. This separation between continuous internal state and quantized message is the minimal mechanism that produces endogenous stochastic sampling under neutrality.

Quantized communication. At each interaction, we sample an *ordered* pair (S, L) of distinct agents uniformly from all $N(N-1)$ speaker–listener pairs. The speaker S samples a message y from its internal distribution x_S and transmits it to the listener. This message is the only communicated object; it is a discrete sample or short list drawn from x_S. We parameterize the amount of information per interaction by an effective bandwidth $m \in \{1, 2, \ldots\} \cup \{\infty\}$. Let $\{e_k\}_{k=1}^K$ denote the standard basis in $\mathbb{R}^K$.

- *Hard* ($m = 1$): sample $k^\star \sim \mathrm{Cat}(x_S)$, i.e. $\Pr(k^\star = k) = (x_S)_k$, and transmit $y = e_{k^\star}$.
- *Top-m* ($m < \infty$): sample $k_1, \ldots, k_m \overset{\text{iid}}{\sim} \mathrm{Cat}(x_S)$ and transmit the empirical message

$$y^{(m)} = \frac{1}{m} \sum_{j=1}^{m} e_{k_j}. \tag{1}$$

- *Soft* ($m = \infty$): transmit the full distribution $y = x_S$

deterministically.

Under all three regimes, the conditional mean is identical, $\mathbb{E}[y \mid x_S] = x_S$, while the conditional variance of the message decreases as m increases (scaling as $1/m$ in the Top-m family). Soft is randomized gossip averaging (Boyd et al., 2006); Hard resembles a voter/Moran copying process (Moran, 1958; Clifford & Sudbury, 1973; Holley & Liggett, 1975; Liggett, 1999). Top-m is a multi-label / multi-bit variant: messages land on a higher-dimensional face of the simplex rather than a vertex.

The bandwidth parameter m controls how much evidence the listener receives per interaction: a single discrete choice (one token) or a short transcript providing multiple draws. Top-m treats these variants uniformly by controlling message variance while keeping $\mathbb{E}[y \mid x_S]$ fixed. Soft ($m = \infty$) is an analytic baseline that removes quantization noise; it does not assume literal transmission of hidden states, but serves as an infinite-bandwidth reference against which drift can be identified.

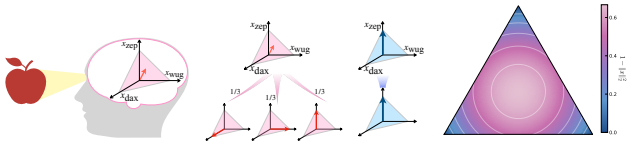
Listener update and in-context adaptation. After receiving a message y, the listener moves a single step toward it. Work on in-context learning interprets the forward pass as an implicit online update (Akyürek et al., 2023; von Oswald et al., 2023; Dai et al., 2023; Park et al., 2025a). Recent work also suggests that in-context behavior can reflect a mixture of distinct strategies rather than a single monolithic algorithm (Wurgaft et al., 2025). Motivated by that view, we model each interaction as a single in-context–style adaptation step with rate $\alpha \in (0, 1]$:

$$x'_L \;=\; (1-\alpha)x_L + \alpha y, \tag{2}$$

while the speaker and all other agents remain unchanged. Equation (2) is a minimal abstraction of this viewpoint: the listener performs one adaptation step toward a target distribution encoded by the received message. The convex-combination form is the simplest contractive update on the simplex that isolates adaptation rate α while remaining analytically tractable. In the Soft case, this update reduces to DeGroot-style opinion averaging (DeGroot, 1974; Friedkin & Johnsen, 1990) with a simplex-valued opinion $x_i \in \Delta^{K-1}$; in the quantized regimes, it is the corresponding update toward a sampled message rather than a full belief state. Small α yields weak per-interaction adaptation (slow accumulation of influence), whereas larger α amplifies the effect of each sampled message on the listener's belief state.

As a null model, QSG rests on four explicit assumptions: simplex-valued agent states, well-mixed speaker–listener pairing, quantized message channels, and first-order listener adaptation (Appendix A.1).

Control parameters. The controls are population size N, adaptation rate α, and communication bandwidth m, with label count K fixed per experiment.



(a) Probabilistic naming. (b) Sampling noise from quantization. (c) Drift strength on the simplex.

Figure 5. **Probabilistic naming and sampling-noise geometry.** (a) A referent is represented internally as a distribution over candidate labels. (b) Agent states lie on the simplex: near-uniform, high-entropy states generate larger sampling noise under quantization, whereas peaked, low-entropy states generate less. (c) Sampling-driven drift strength on the simplex, proportional to $1 - \|x\|_2^2$, is maximal near the center and vanishes at the vertices.

3. Analysis and Scaling Laws

We analyze QSG at the population level to ask when coordination is shaped mainly by selection versus stochastic drift, using macroscopic observables defined directly from the population state. The drift mechanism we isolate is not mutually exclusive with selection, but adds to any systematic biases.

Macroscopic observables. We analyze the QSG dynamics directly on the simplex using the update rules in Sec. 2 (Eqs. (1) and (2)). This lets us track how sampling variance perturbs the mean dynamics near symmetry. The primary order parameter is the population mean distribution, analogous to magnetization in statistical mechanics.

$$\bar{x} \coloneqq \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad U \coloneqq \|\bar{x}\|_2^2 \in \left[\tfrac{1}{K}, 1\right], \qquad V \coloneqq \sum_{i=1}^{N} \|x_i - \bar{x}\|_2^2. \tag{3}$$

Here U is the **polarization**, equal to $1/K$ at perfect symmetry and 1 at consensus, while V is the **disagreement energy**, measuring how dispersed agents remain around the population mean. At the perfectly symmetric initialization, sampling noise is the only driver of motion; more generally, drift and selection can coexist, and these observables isolate the drift component. Figure 5 visualizes the simplex geometry and the sampling-noise–driven drift strength that underlies the analysis.

Soft exchange preserves the mean in expectation and contracts disagreement. In Soft QSG, the speaker transmits $y = x_S$ and only the listener updates, $x'_L = (1-\alpha)x_L + \alpha x_S$. The population mean evolves as $\bar{x}' = \bar{x} + \frac{\alpha}{N}(x_S - x_L)$, so uniform ordered-pair sampling gives $\mathbb{E}[\bar{x}' \mid X] = \bar{x}$. Hence $\bar{x}$ is preserved in expectation (equivalently, each

coordinate ($\bar{x}_k(t)$) is a bounded martingale), but it is *not* generally invariant along individual trajectories. Moreover, the disagreement energy $V := \sum_{i=1}^{N} \|x_i - \bar{x}\|_2^2$ contracts in expectation:

$$\mathbb{E}[\Delta V \mid X] = -\frac{2\alpha}{N-1}\left(1 - \alpha + \frac{\alpha}{N}\right) V \;\le\; 0. \quad (4)$$

Hence, if $x_i(0) = \mathbf{1}/K$ for all i, then $x_i(t) = \mathbf{1}/K$ for all t (no symmetry breaking). This provides the neutral baseline. Full-distribution exchange smooths disagreement but does not create spontaneous convention formation under neutrality. Against this baseline, any symmetry breaking in the quantized regimes must come from the extra sampling variance injected by communication itself.

Hard sampling injects an extra variance term. Hard and Soft share the *same conditional mean* (since $\mathbb{E}[e_{k^\star} \mid x_S] = x_S$), but Hard injects additional sampling variance.

Theorem 1 (Hard sampling increases polarization via sampling variance)**.** *Consider QSG with adaptation rate $\alpha \in (0, 1]$. Let $\bar{x}$ be the population mean and define the polarization potential $U := \|\bar{x}\|_2^2$. Conditioned on the current state X, the expected one-step change in U satisfies*

$$\mathbb{E}[\Delta U \mid X]_{\text{hard}} = \mathbb{E}[\Delta U \mid X]_{\text{soft}} + \frac{\alpha^2}{N^2}\,\mathbb{E}\big[1 - \|x_S\|_2^2 \mid X\big], \quad (5)$$

where the expectation is over the random choice of (S, L) and (for Hard) the sample $k^\star$. The additional term in (5) is the sampling variance; it is nonnegative, and is strictly positive iff $\mathbb{E}_S[1 - \|x_S\|_2^2] > 0$, equivalently if at least one agent has nonzero internal uncertainty. In particular, at the perfectly symmetric initialization $x_i = \mathbf{1}/K$, we have $\mathbb{E}[\Delta U \mid X]_{\text{soft}} = 0$ but $\mathbb{E}[\Delta U \mid X]_{\text{hard}} = \frac{\alpha^2}{N^2}(1 - \frac{1}{K}) > 0$, so symmetry is noise-unstable under Hard sampling.

Proof. See Appendix A.2.

This variance injection drives symmetry breaking in the neutral model under symmetric initialization and adds to any selection effects. To see what Hard adds beyond Soft, note that under Soft exchange the symmetric state $\bar{x} = \mathbf{1}/K$ is a fixed point and drift arises only from heterogeneity ($x_S \neq x_L$), so U grows only insofar as agents disagree. Hard/Top-m adds variance even when $x_S = x_L$, so the symmetric interior state becomes stochastically unstable and trajectories are pushed toward consensus. Equivalently, $\mathbb{E}[\Delta U \mid X]_{\text{soft}} = \frac{2\alpha^2}{N^2(N-1)} V$, while Hard adds the positive term in Eq. (5).

Mean-field approximation and consensus time. Under a mean-field ansatz where agents remain similar ($x_i \approx p$ for all i), we have $\|x_S\|_2^2 \approx U$, yielding $dU/dt = \alpha^2(1 - U)/(mN^2)$ as an approximation to the expected trajectory. Starting from the symmetric state $U(0) = 1/K$, mean-field gives two useful summaries: the polarization trajectory $U(t)$ and the time $t_{\text{cons}}(U_\star)$ to reach a target consensus threshold $U_\star$:

$$U(t) = 1 - \left(1 - \frac{1}{K}\right)\exp\left(-\frac{\alpha^2 t}{mN^2}\right), \quad (6)$$

$$t_{\text{cons}}(U_\star) \approx \frac{mN^2}{\alpha^2}\log\left(\frac{1 - 1/K}{1 - U_\star}\right). \quad (7)$$

The characteristic time is $t_{\text{char}} \sim mN^2/\alpha^2$ interaction steps, or equivalently $\tau_{\text{char}} \sim mN/\alpha^2$ population rounds ($\tau = t/N$), with Hard corresponding to $m = 1$. Mean-field predicts an approximate single-parameter collapse of $U(t)$ across N when time is rescaled by mN^2/α^2. For a fixed consensus threshold $U_\star \in (1/K, 1)$, the logarithmic factor depends only on $(K, U_\star)$, so $t_{\text{cons}} \propto N^2$ in steps (equivalently $\tau_{\text{cons}} = t_{\text{cons}}/N \propto N$ in population rounds). Physically, larger populations and higher-bandwidth messages weaken the stochastic push of any single interaction, whereas stronger in-context adaptation accelerates the approach to consensus. Figure 6a compares simulated Hard trajectories to the mean-field curve; trajectories track the mean-field prediction, consistent with the variance-injection mechanism in Theorem 1.

Top-m reduces the symmetry-breaking drift as $1/m$. To quantify bandwidth, we model the candidate list as m i.i.d. samples from x_S and transmit their empirical distribution. This preserves the mean but reduces sampling variance by $1/m$.

The empirical message in Eq. (1) is unbiased, with $\mathbb{E}[y^{(m)} \mid x_S] = x_S$ and $\mathrm{Cov}(y^{(m)} \mid x_S) = m^{-1}(\mathrm{diag}(x_S) - x_S x_S^\top)$. Thus $\mathbb{E}[\|y^{(m)} - x_S\|_2^2 \mid x_S] = m^{-1}(1 - \|x_S\|_2^2)$, i.e., variance scales as $1/m$.

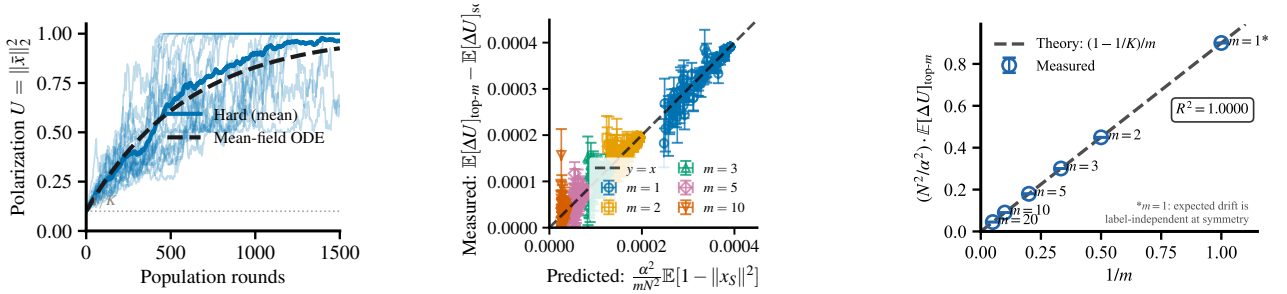
Theorem 2 (Top-m drift term scales as $1/m$)**.** *Consequently, under QSG with Top-m empirical communication $y^{(m)}$, the polarization drift satisfies*

$$\mathbb{E}[\Delta U \mid X]_{\text{top}m} = \mathbb{E}[\Delta U \mid X]_{\text{soft}} + \frac{\alpha^2}{m\,N^2}\,\mathbb{E}\big[1 - \|x_S\|_2^2 \mid X\big]. \quad (8)$$

At the perfectly symmetric initialization $x_i = \mathbf{1}/K$, this yields $\mathbb{E}[\Delta U \mid X]_{\text{top}m} = \frac{\alpha^2}{mN^2}\left(1 - \frac{1}{K}\right)$.

Thus, increasing bandwidth weakens symmetry-breaking drift linearly in $1/m$. *Proof.* See Appendix A.3.

Fig. 9 visualizes how the drift term depends on uncertainty and bandwidth, and Fig. 6b tests Eq. (8) directly by estimating one-step conditional expectations from shared snapshot states and comparing the measured excess drift to the predicted variance-injection term (Appendix B.2.3 summarizes the estimator and parameters). Panel (c) of Fig. 6 shows the corresponding $1/m$ scaling at the symmetric initialization.



(a) Mean-field trajectories. *(b)* Measured vs. predicted drift. *(c)* $1/m$ scaling at symmetry.

Figure 6. **QSG simulations validate the mean-field approximation and drift identities.** (a) Hard-sampling trajectories and their ensemble mean track the mean-field solution in Eq. (6) ($N = 24$, $K = 10$, $\alpha = 0.2$). (b) One-step excess drift from shared snapshot states X, plotted against the predicted variance-injection term; dashed reference line $y = x$ ($N = 24$, $K = 10$, $\alpha = 0.5$). (c) Symmetric $1/m$ corollary. Under $x_i = \mathbf{1}/K$, the normalized excess drift follows the theoretical line $(1 - 1/K)(1/m)$ ($N = 24$, $K = 10$, $\alpha = 0.5$).

The drift identities above are expectation-level results, not a finite-time absorption theorem for all α. Appendix A.4 gives the absorption result in the hard-copying limit $\alpha = 1$ and explains why, for $\alpha < 1$, run-level winner selection should be treated as an empirical and diffusion-approximation phenomenon. That neutral winner symmetry is the reference point for the weak-asymmetry drift–selection crossover in Appendix A.6, controlled by $\Gamma_h = (mN/\alpha)h$.

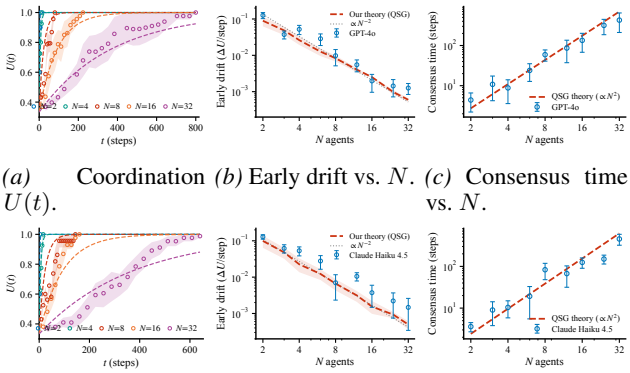
These results point to a single mechanism linking population size, bandwidth, adaptation strength, and internal uncertainty. Quantized communication injects sampling variance, and that variance controls both the speed of neutral consensus formation and the extent to which weak biases are amplified. That mechanism yields testable scalings, including $1/N^2$ and $1/m$ early drift and a mean-field consensus time, which we evaluate next in LLM populations.

4. Experimental Validation

We evaluate QSG in three complementary settings: a contextualized near-tie committee prompt, neutral naming games that remove semantic and payoff confounds, and a Top-m intervention testing whether larger communication bandwidth suppresses early drift.

Contextualized near-tie committee recommendation. The committee budget-plan task in Fig. 1 shares the K-way discrete-choice structure modeled by QSG while embedding the alternatives in a realistic recommendation prompt. Agents are told that the available proposals are equally strong under the current evidence and must recommend one plan code. There is no objective best plan, external payoff, coordination reward, or feedback signal. The goal is not to evaluate planning quality, but to test whether QSG's early-drift mechanism remains predictive in a contextualized near-tie setting.

We operationalize internal uncertainty using probe-estimated label distributions. For an agent distribution x, the sampling-variance term predicted by QSG is proportional to $1 - \|x\|_2^2$. We also measure one-step adaptation rather than treating it as a free population-level fit. Given a listener state x_L, a received message y, and the post-interaction listener state x'_L, all estimated by measurement-only probes, we estimate $\widehat{\alpha}_{\text{eff}} = \langle x'_L - x_L,\ y - x_L \rangle / \|y - x_L\|_2^2$. This estimate is computed from single-interaction pre/post changes rather than fit to the population-level drift curve. In the matched $K = 3$ committee task, the measured value $\alpha_{\text{eff}} = 0.725$, together with probe-estimated uncertainty, predicts early polarization drift across population sizes; predicted and observed drift have correlation $r = 0.985$ across nine N values (Fig. 1(c)). Appendix B.3.1 reports the plug-in estimator



(a) Coordination $U(t)$. *(b)* Early drift vs. N. *(c)* Consensus time vs. N.

(d) Coordination $U(t)$. *(e)* Early drift vs. N. *(f)* Consensus time vs. N.

Figure 7. **LLM scaling tests of QSG predictions across two model families.** Top row: GPT-4o; bottom row: Claude Haiku 4.5. (a,d) Coordination trajectories $U(t)$ at varying N, from one measurement-only probe sample per agent per probe time. (b,e) Measured early drift (32 draws) vs. plug-in QSG predictions using probe-estimated α_{eff} and re-probed uncertainty; gray guides show the pure N^{-2} reference scaling, not the plug-in predictions. (c,f) Conditional consensus times to the $U \geq 0.9$ threshold (same operational probe statistic), with dashed cN^2 QSG references.

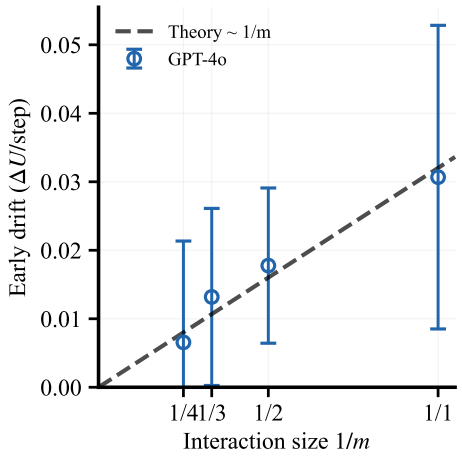


Figure 8. **Top-m scaling in GPT-4o.** Early drift decreases as bandwidth m increases, equivalently increasing with $1/m$.

and adaptation diagnostics.

Neutral naming games. We then test the same scaling predictions in LLM populations using the Neutral Naming Drift (NND) protocol. In each run, N agents repeatedly name a fixed referent r using K neutral synthetic labels, with no external reward or ground truth. Interactions are ordered speaker–listener updates: the speaker emits one label, only the listener updates, and measurement-only probes estimate U, early drift $\Delta U/\text{step}$, and time to consensus. Appendix B.3 gives the probe cadence, estimators, decoding controls, and prompt templates.

Figure 7 is consistent with the QSG scaling predictions in both GPT-4o and Claude Haiku 4.5. The raw $K = 3$ $U(t)$ trajectories are summarized by the mean-field form $U(t) = 1 - (1 - 1/K)\exp(-\alpha^2 t/(mN^2))$ with a single effective α shared across N. The 32-draw early-drift panels use probe-estimated α_{eff} (0.982 for GPT-4o and 0.927 for Claude Haiku 4.5) plus measured speaker uncertainty, and the consensus-time panels grow close to the $t_{\text{cons}} \sim mN^2/\alpha^2$ reference (Eq. (7)).

Bandwidth intervention. For Top-m, the speaker emits exactly m labels, repeats are allowed, and the transmitted list is treated as an unranked empirical multiset, matching Eq. (1). The GPT-4o sweep in Fig. 8 is consistent with monotone suppression of early drift as bandwidth increases. Appendix B.3.2 gives the prompt and measurement details.

5. Discussion, Limitations, and Design Guidance

QSG should be read as a null model for weak-signal, near-tie consensus formation, not as a complete model of multi-agent reasoning. Its main implication is that agreement is an outcome to be explained, not by itself evidence of collective reasoning or information aggregation. In near-tie settings, high uncertainty $1 - \|x\|_2^2$, high α_{eff}, small N, and low-bandwidth communication increase stochastic lock-in in the QSG time units: per-interaction drift scales as $\alpha_{\text{eff}}^2/(mN^2)$, equivalently $\alpha_{\text{eff}}^2/(mN)$ per population round. When weak blank-memory asymmetries are present, larger populations or richer evidence can instead make those asymmetries more decisive. These regimes are not contradictory; they are two sides of the same drift–selection tradeoff.

For pluralistic alignment systems, this perspective suggests concrete diagnostics before treating a consensus as a reliable aggregate of diverse values. One should audit blank-memory priors, repeat the protocol across seeds, vary N, and compare low-bandwidth outputs against richer message protocols. Independent reasoning before interaction and multiple independently initialized committees can help separate evidence aggregation from path-dependent amplification. Conversely, protocols that repeatedly expose agents to one another's sampled outputs can manufacture agreement even when the underlying task contains little signal.

More broadly, QSG illustrates a population-level complement to single-model mechanistic interpretability. The role of synthetic near-tie games is not to replace realistic tasks, but to provide controlled regimes in which scaling laws can be derived, falsified, and then tested against semantically richer interactions. Instead of asking only how one model forms an internal representation, we can ask how a population stabilizes, distorts, or amplifies shared representations through interaction. The present experiments test this idea only in near-tie discrete-choice settings with well-mixed pairings and measurement probes; extending the analysis to structured networks, heterogeneous agents, richer deliberation, and tasks with ground truth is necessary before drawing conclusions about accuracy or truth discovery. The value of the null model is precisely to make such extensions measurable: it specifies what consensus should look like before semantics, rewards, or task evidence are allowed to explain it.

References

Akyürek, E., Schuurmans, D., Andreas, J., Ma, T., and Zhou, D. What learning algorithm is in-context learning? investigations with linear models. In *ICLR 2023*, 2023.

Ashery, A. F., Aiello, L. M., and Baronchelli, A. Emergent social conventions and collective bias in LLM populations. *Science Advances*, 11(20):eadu9368, 2025. doi: 10.1126/sciadv.adu9368.

Baronchelli, A., Felici, M., Loreto, V., Caglioti, E., and Steels, L. Sharp transition towards shared vocabularies in multi-agent systems. *Journal of Statistical Mechanics:*

Theory and Experiment, pp. P06014, 2006. doi: 10.1088/1742-5468/2006/06/P06014.

Baronchelli, A., Loreto, V., and Steels, L. In-depth analysis of the Naming Game dynamics: the homogeneous mixing case. *International Journal of Modern Physics C*, 19(5):785–812, 2008. doi: 10.1142/S0129183108012522.

Becker, J., Kaesberg, L. B., Bauer, N., Wahle, J. P., Ruas, T., and Gipp, B. MALLM: Multi-agent large language models framework. In Habernal, I., Schulam, P., and Tiedemann, J. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 418–439, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-334-0. doi: 10.18653/v1/2025.emnlp-demos.29.

Boyd, S., Ghosh, A., Prabhakar, B., and Shah, D. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530, 2006. doi: 10.1109/TIT.2006.874516.

Brockers, V. C., Ehrlich, D. A., and Priesemann, V. Disentangling interaction and bias effects in opinion dynamics of large language models. *arXiv preprint arXiv:2509.06858*, 2025. doi: 10.48550/arXiv.2509.06858.

Chuang, Y.-S., Goyal, A., Harlalka, N., Suresh, S., Hawkins, R., Yang, S., Shah, D., Hu, J., and Rogers, T. Simulating opinion dynamics with networks of LLM-based agents. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3326–3346, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.211.

Clifford, P. and Sudbury, A. A model for spatial conflict. *Biometrika*, 60(3):581–588, 1973. doi: 10.1093/biomet/60.3.581.

Dai, D., Sun, Y., Dong, L., Hao, Y., Ma, S., Sui, Z., and Wei, F. Why can GPT learn in-context? Language models secretly perform gradient descent as meta-optimizers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4005–4019, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.247.

DeGroot, M. H. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974. doi: 10.1080/01621459.1974.10480137.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the Forty-first International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763. PMLR, 2024.

Flint, A., Aiello, L. M., Pastor-Satorras, R., and Baronchelli, A. Group size effects and collective misalignment in LLM multi-agent systems. *arXiv preprint arXiv:2510.22422*, 2025. doi: 10.48550/arXiv.2510.22422.

Friedkin, N. E. and Johnsen, E. C. Social influence and opinions. *Journal of Mathematical Sociology*, 15(3-4): 193–206, 1990. doi: 10.1080/0022250X.1990.9990069.

Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Palepu, A., Sirkovic, P., Myaskovsky, A., Weissenberger, F., Rong, K., Tanno, R., Saab, K., Popovici, D., Blum, J., Zhang, F., Chou, K., Hassidim, A., Gokturk, B., Vahdat, A., Kohli, P., Matias, Y., Carroll, A., Kulkarni, K., Tomasev, N., Guan, Y., Dhillon, V., Vaishnav, E. D., Lee, B., Costa, T. R. D., Penadés, J. R., Peltz, G., Xu, Y., Pawlosky, A., Karthikesalingam, A., and Natarajan, V. Towards an AI co-scientist. *arXiv preprint arXiv:2502.18864*, 2025. doi: 10.48550/arXiv.2502.18864.

Holley, R. A. and Liggett, T. M. Ergodic theorems for weakly interacting infinite systems and the voter model. *The Annals of Probability*, 3(4):643–663, 1975. doi: 10.1214/aop/1176996306.

Kaesberg, L. B., Becker, J., Wahle, J. P., Ruas, T., and Gipp, B. Voting or consensus? decision-making in multi-agent debate. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11640–11671, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.606.

Kim, Y., Park, C., Jeong, H., Chan, Y. S., Xu, X., McDuff, D., Lee, H., Ghassemi, M., Breazeal, C., and Park, H. W. MDAgents: An adaptive collaboration of LLMs for medical decision-making. In *Advances in Neural Information Processing Systems 37*, 2024. doi: 10.52202/079017-2522.

Kimura, M. Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626, 1968. doi: 10.1038/217624a0.

Kimura, M. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, 1983. ISBN 0521231094.

Liggett, T. M. *Stochastic Interacting Systems: Contact, Voter and Exclusion Processes*. Number 324 in Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 1999. ISBN 9783540659952. doi: 10.1007/978-3-662-03990-8.

Martins, A. C. R. Continuous opinions and discrete actions in opinion dynamics problems. *International Journal of Modern Physics C*, 19(4):617–624, 2008. doi: 10.1142/S0129183108012339.

Moran, P. A. P. Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 54(1):60–71, 1958. doi: 10.1017/S0305004100033193.

Park, C. F., Lee, A., Lubana, E. S., Yang, Y., Okawa, M., Nishi, K., Wattenberg, M., and Tanaka, H. ICLR: In-Context Learning of Representations. In *International Conference on Learning Representations*, 2025a.

Park, C. F., Lubana, E. S., Pres, I., and Tanaka, H. Competition dynamics shape algorithmic phases of in-context learning. In *International Conference on Learning Representations*, 2025b. Spotlight.

Sreedhar, K., Cai, A., Ma, J., Nickerson, J. V., and Chilton, L. B. Simulating cooperative prosocial behavior with multi-agent LLMs: Evidence and mechanisms for AI agents to inform policy decisions. In *Proceedings of the 2025 International Conference on Intelligent User Interfaces*, pp. 1272–1286, 2025. doi: 10.1145/3708359.3712149.

Steels, L. A self-organizing spatial vocabulary. *Artificial Life*, 2(3):319–332, 1995. doi: 10.1162/artl.1995.2.3.319.

von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35151–35174. PMLR, 2023.

Watson, W., Cho, N., Srishankar, N., Zeng, Z., Cecchi, L., Scott, D., Siddagangappa, S., Kaur, R., Balch, T., and Veloso, M. LAW: Legal agentic workflows for custody and fund services contracts. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., Schockaert, S., Darwish, K., and Agarwal, A. (eds.), *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pp. 583–594, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

Wurgaft, D., Lubana, E. S., Park, C. F., Tanaka, H., Reddy, G., and Goodman, N. D. In-context learning strategies emerge rationally. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.48550/arXiv.2506.17859.

Xiao, Y., Sun, E., Luo, D., and Wang, W. TradingAgents: Multi-Agents LLM financial trading framework. *arXiv preprint arXiv:2412.20138*, 2024. doi: 10.48550/arXiv.2412.20138.

Zhang, J., Xu, X., Zhang, N., Liu, R., Hooi, B., and Deng, S. Exploring collaboration mechanisms for LLM agents: A social psychology view. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14544–14607, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.782.

Zhao, X., Wang, K., and Peng, W. An electoral approach to diversify LLM-based multi-agent collective decision-making. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2712–2727, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.158.

A. Additional Theory for Quantized Simplex Gossip (QSG)

This appendix collects algebraic identities and full drift calculations used in Sec. 3.

A.1. QSG null-model assumptions

QSG is defined by the following assumptions, which we treat as empirically testable approximations:

A1. **Simplex state.** Each agent i is represented by a probability distribution $x_i \in \Delta^{K-1}$ over K competing conventions, corresponding to a fixed naming-game prompt or referent r.

A2. **Well-mixed pair selection.** At each step, an ordered speaker–listener pair (S, L) is sampled uniformly from the $N(N-1)$ pairs with $S \neq L$.

A3. **Quantized message channel.** The speaker communicates a finite-bandwidth message obtained by sampling from x_S (Hard or Top-m), rather than transmitting the full distribution.

A4. **First-order adaptation.** Each interaction induces a single adaptation step of the listener distribution toward the received message, with magnitude controlled by a scalar adaptation rate α.

A.2. Proof of Theorem 1: hard-sampling drift decomposition

Let only the listener update: $x'_L = (1-\alpha)x_L + \alpha y$, where $y = x_S$ (Soft) or $y = e_{k^\star}$ (Hard). Then $\bar{x}' = \bar{x} + \frac{\alpha}{N}(y - x_L)$. Expanding $U' = \|\bar{x}'\|_2^2$ gives $\Delta U = \frac{2\alpha}{N}\langle \bar{x}, y - x_L\rangle + \frac{\alpha^2}{N^2}\|y - x_L\|_2^2$. Taking conditional expectations yields $\mathbb{E}[\Delta U \mid X] = \frac{\alpha^2}{N^2}\mathbb{E}\big[\|y - x_L\|_2^2 \mid X\big]$ because unbiased messaging and uniform ordered-pair sampling give $\mathbb{E}[y \mid X] = \bar{x} = \mathbb{E}[x_L \mid X]$. For Hard, decompose $y - x_L = (y - x_S) + (x_S - x_L)$. Since $\mathbb{E}[y - x_S \mid x_S] = 0$, the cross term vanishes after conditioning on (x_S, x_L), and $\mathbb{E}\big[\langle y - x_S, x_S - x_L\rangle \mid x_S, x_L\big] = 0$. Hence

$$\mathbb{E}\big[\|y - x_L\|_2^2 \mid X\big] = \mathbb{E}\big[\|x_S - x_L\|_2^2 \mid X\big] + \mathbb{E}\big[\|y - x_S\|_2^2 \mid X\big].$$

For one-hot $y = e_{k^\star}$ sampled from x_S, $\mathbb{E}\|y - x_S\|_2^2 = 1 - \|x_S\|_2^2$. Substituting these two terms gives the Hard variance-injection formula in Theorem 1.

A.3. Proof of Theorem 2: Top-m drift decomposition and scaling

The same expansion applies with y replaced by the Top-m empirical message $y^{(m)}$. We have $x'_L = (1-\alpha)x_L + \alpha y^{(m)}$ and hence $\bar{x}' = \bar{x} + \frac{\alpha}{N}(y^{(m)} - x_L)$, so expanding $U' = \|\bar{x}'\|_2^2$ gives $\Delta U = \frac{2\alpha}{N}\langle \bar{x}, y^{(m)} - x_L\rangle + \frac{\alpha^2}{N^2}\|y^{(m)} - x_L\|_2^2$. Taking conditional expectations yields $\mathbb{E}[\Delta U \mid X] = \frac{\alpha^2}{N^2}\mathbb{E}\big[\|y^{(m)} - x_L\|_2^2 \mid X\big]$ because unbiased messaging and uniform ordered-pair sampling give $\mathbb{E}[y^{(m)} \mid X] = \bar{x} = \mathbb{E}[x_L \mid X]$. Decompose $y^{(m)} - x_L = (y^{(m)} - x_S) + (x_S - x_L)$ to obtain the Soft term plus an extra variance term $\frac{\alpha^2}{N^2}\mathbb{E}\|y^{(m)} - x_S\|_2^2$. More explicitly, conditioning on (x_S, x_L) gives

$$\mathbb{E}\big[\langle y^{(m)} - x_S, x_S - x_L\rangle \mid x_S, x_L\big] = 0,$$

so

$$\mathbb{E}\big[\|y^{(m)} - x_L\|_2^2 \mid X\big] = \mathbb{E}\big[\|x_S - x_L\|_2^2 \mid X\big] + \mathbb{E}\big[\|y^{(m)} - x_S\|_2^2 \mid X\big].$$

By (22), $\mathbb{E}\|y^{(m)} - x_S\|_2^2 = \frac{1}{m}(1 - \|x_S\|_2^2)$, yielding (8).

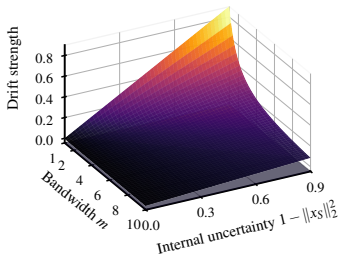


Figure 9. **Sampling-driven drift from uncertainty and bandwidth.** For fixed α and N, the drift term in Eq. (8) scales as $(1 - \|x_S\|_2^2)/m$, increasing with uncertainty and decreasing with bandwidth m.

A.4. Absorption at $\alpha = 1$ and run-level symmetry breaking for $\alpha < 1$

For $\alpha = 1$ (pure copying), after each agent has served as a listener at least once, all x_i are one-hot and Hard QSG reduces to a finite-state copying Markov chain with absorbing consensus states $\{x_i = e_k,\ \forall i\}$. For $\alpha < 1$ the state space is continuous, and we do not claim finite-time consensus or almost-sure absorption. Our formal results in this regime are expectation-level drift identities and mean-field approximations. In simulations, the process tends to polarize and empirically select a winner in each run even though ensemble averages preserve symmetry. This is the discrete analogue of the drift mechanism above. Randomness can still select a winner even when the ensemble is neutral, but for $\alpha < 1$ we treat this as an empirical run-level phenomenon rather than a proved absorption result. For $\alpha < 1$, Eq. (9) should be read as a diffusion-approximation description of run-level fixation behavior. In the hard-copying limit on a complete interaction graph, this reduction becomes a classical voter-model statement (Clifford & Sudbury, 1973; Holley & Liggett, 1975). The process almost surely reaches a consensus vertex in finite time. If the initialization is exchangeable across token labels, for example $x_i(0) = \mathbf{1}/K$, the winning token

is uniformly distributed over $\{1, \dots, K\}$. More generally, the probability that token k wins equals $\bar{x}_k(0)$, by the martingale property of the density process. Appendix A.5 gives the reduction and proof details. This neutral winner symmetry is the reference point that the weak-asymmetry extension in Appendix A.6 perturbs into the drift–selection crossover. A weak-asymmetry extension gives a drift–selection crossover controlled by $\Gamma_h = (mN/\alpha)h$, with $N_c \sim \alpha/(m|h|)$.

A.5. Voter-model reduction and winner symmetry

With $\alpha = 1$, the listener is overwritten by the sampled one-hot message: $x'_L = e_{k^\star}$ with $k^\star \sim \mathrm{Cat}(x_S)$. Under uniform random pair selection on the complete interaction graph, each agent is chosen as a listener infinitely often; hence after an almost-surely finite "coupon collector" time, all x_i lie on simplex vertices and remain there. From that time onward the dynamics is exactly the K-state voter/-Moran copying process on a finite complete graph, which almost surely reaches a consensus (an absorbing vertex state) in finite time.

For the winner distribution, fix a token k and define the population mean coordinate $\bar{x}_k(t) := \frac{1}{N}\sum_{i=1}^{N} x_{i,k}(t)$. In one update, $\bar{x}_k(t+1) = \bar{x}_k(t) + \frac{1}{N}\big(y_k - x_{L,k}(t)\big)$ where $y = e_{k^\star}$. Conditioned on the current state $\mathcal{F}_t$, $\mathbb{E}[y_k \mid \mathcal{F}_t] = \mathbb{E}[x_{S,k}(t) \mid \mathcal{F}_t] = \bar{x}_k(t)$ and also $\mathbb{E}[x_{L,k}(t) \mid \mathcal{F}_t] = \bar{x}_k(t)$, so $\mathbb{E}[\bar{x}_k(t+1) \mid \mathcal{F}_t] = \bar{x}_k(t)$; the expected value of $\bar{x}_k$ is unchanged at each step (equivalently, $(\bar{x}_k(t))$ is a bounded martingale). Let T_{cons} be the (a.s. finite) consensus time; then $\bar{x}_k(T_{\text{cons}}) = \mathbf{1}\{\text{winner} = k\}$. By optional stopping for bounded martingales, $\Pr[\text{winner} = k] = \mathbb{E}[\bar{x}_k(T_{\text{cons}})] = \bar{x}_k(0)$. If the initialization is exchangeable across labels (e.g. $x_i(0) = \mathbf{1}/K$), then $\bar{x}_k(0) = 1/K$ and the winner is uniform over $\{1, \dots, K\}$.

A.6. Weak asymmetry and the drift–selection crossover ($K = 2$)

To model weak asymmetry, we tilt only the speaker channel by a small sampling bias (external field) h. For $K = 2$, let $p_i := x_{i1}$ and sample messages from $\tilde{p}_S = \frac{p_S e^h}{p_S e^h + (1-p_S)}$ (listener update unchanged). A diffusion approximation (Appendix A.7) collapses fixation statistics onto the single parameter $\Gamma_h \equiv \frac{mN}{\alpha}h$. This parameter measures the competition between systematic bias and endogenous drift. Larger N or m suppress the neutral sampling noise and make the same bias more decisive, whereas larger α strengthens drift relative to that same h. Defining the final magnetization

$$M_\infty := 2\bar{x}_1(\infty) - 1,$$

$$\begin{aligned} \Pr(\text{label 1 fixes}) &\approx \frac{1}{1+\exp(-\Gamma_h)}, \\ \mathbb{E}[M_\infty] &\approx 2\Pr(\text{label 1 fixes}) - 1, \\ N_c &\sim \frac{\alpha}{m|h|}. \end{aligned} \tag{9}$$

Thus fixation is approximately logistic in Γ_h, with crossover scale $N_c \sim \alpha/(m|h|)$ (from $|\Gamma_h| \sim 1$). This crossover is previewed in Fig. 10. We use Γ_h to distinguish this bias-based crossover from the temperature-based Γ_T in Appendix A.15. Consequently, $|\Gamma_h| \ll 1$ yields near-neutral winners (near $1/2$), while $|\Gamma_h| \gg 1$ yields bias-driven amplification of the asymmetry.

A.7. Diffusion approximation for weak asymmetry

For $K = 2$, let $\bar{p} = \frac{1}{N}\sum_i p_i$ with $p_i = x_{i1}$. Under a small bias h, expand $\tilde{p}_S = p_S + h\,p_S(1-p_S) + \mathcal{O}(h^2)$. Under the homogeneous mean-field approximation and to first order in h, one interaction step gives

$$\mathbb{E}[\Delta\bar{p} \mid \bar{p}] = \frac{\alpha}{N} h\,\bar{p}(1-\bar{p}) + \mathcal{O}(h^2), \qquad \mathrm{Var}(\Delta\bar{p} \mid \bar{p}) = \frac{\alpha^2}{mN^2}\bar{p}(1-\bar{p}) + \mathcal{O}(h$$

(10)

In population-round time $\tau = t/N$, this yields

$$d\bar{p} = \alpha h\,\bar{p}(1-\bar{p})\,d\tau + \sqrt{\frac{\alpha^2}{mN}\bar{p}(1-\bar{p})}\,dW. \tag{11}$$

The fixation probability $\pi(p_0)$ of this diffusion surrogate solves the backward equation $\mu(\bar{p})\pi'(\bar{p}) + \frac{1}{2}D(\bar{p})\pi''(\bar{p}) = 0$ with $\pi(0) = 0$, $\pi(1) = 1$, giving $\pi(p_0) = \frac{1-\exp(-2\Gamma_h p_0)}{1-\exp(-2\Gamma_h)}$ where $\Gamma_h = \frac{mN}{\alpha}h$. From $p_0 = 1/2$, this yields the approximate logistic/tanh form used in Eq. (9).

A.8. Order parameters and second-moment identities

Recall the population mean $\bar{x} := \frac{1}{N}\sum_{i=1}^{N} x_i$ and polarization potential $U := \|\bar{x}\|_2^2$. Define the disagreement energy

$$V := \sum_{i=1}^{N} \|x_i - \bar{x}\|_2^2, \tag{12}$$

and the mean self-overlap

$$q := \frac{1}{N}\sum_{i=1}^{N} \|x_i\|_2^2. \tag{13}$$

Finally, we define the coordination rate

$$S := \frac{1}{N(N-1)}\sum_{i \neq j} x_i^\top x_j. \tag{14}$$

For any population state $X = (x_1, \dots, x_N)$, the basic second-moment identities are

$$V = N\,(q - U), \tag{15}$$

$$S = U - \frac{V}{N(N-1)}. \tag{16}$$

Expand $V = \sum_i \|x_i - \bar{x}\|_2^2 = \sum_i \|x_i\|_2^2 - N\|\bar{x}\|_2^2$, giving $V = N(q-U)$. Also $\|\sum_i x_i\|_2^2 = \sum_i \|x_i\|_2^2 + \sum_{i\neq j} x_i^\top x_j$. Since $\|\sum_i x_i\|_2^2 = N^2 U$ and $\sum_i \|x_i\|_2^2 = V + NU$, we obtain $\sum_{i\neq j} x_i^\top x_j = N(N-1)U - V$, hence (16).

For a uniformly random ordered pair (S, L) with $S \neq L$,

$$\mathbb{E}\big[\|x_S - x_L\|_2^2 \mid X\big] = \frac{2}{N-1}\,V. \tag{17}$$

Because the diagonal terms vanish,

$$\sum_{i\neq j} \|x_i - x_j\|_2^2 = \sum_{i,j} \|x_i - x_j\|_2^2.$$

Expanding the right-hand side around the mean gives $\sum_{i,j} \|x_i - x_j\|_2^2 = 2N \sum_i \|x_i - \bar{x}\|_2^2 = 2NV$. Averaging over the $N(N-1)$ ordered pairs gives (17).

A.9. Mean update and preservation in expectation

For any regime in which $\mathbb{E}[y \mid x_S] = x_S$ (Soft/Hard/Top-m),

$$\bar{x}' = \bar{x} + \frac{\alpha}{N}(y - x_L). \tag{18}$$

Conditioned on the current state X,

$$\mathbb{E}[\bar{x}' \mid X] = \bar{x}, \tag{19}$$

so each coordinate $(\bar{x}_k(t))_{t\geq 0}$ is preserved in expectation from one step to the next (equivalently, it is a bounded martingale). Only the listener changes: $x_L' - x_L = \alpha(y - x_L)$, so $\bar{x}' = \bar{x} + \frac{1}{N}(x_L' - x_L)$, yielding (18). Now condition on X. Since (S, L) is a uniformly random ordered pair with $S \neq L$, $\mathbb{E}[x_S \mid X] = \mathbb{E}[x_L \mid X] = \bar{x}$. If $\mathbb{E}[y \mid x_S] = x_S$, then $\mathbb{E}[y \mid X] = \bar{x}$. Thus $\mathbb{E}[y - x_L \mid X] = 0$, implying $\mathbb{E}[\bar{x}' \mid X] = \bar{x}$.

A.10. Message variance for Top-m

Recall the empirical message $y^{(m)} = \frac{1}{m}\sum_{j=1}^m e_{k_j}$ with $k_j \overset{\text{iid}}{\sim} \mathrm{Cat}(x_S)$.

Conditioned on x_S,

$$\mathbb{E}[y^{(m)} \mid x_S] = x_S, \tag{20}$$

$$\mathrm{Cov}(y^{(m)} \mid x_S) = \frac{1}{m}\left(\mathrm{diag}(x_S) - x_S x_S^\top\right), \tag{21}$$

$$\mathbb{E}\big[\|y^{(m)} - x_S\|_2^2 \mid x_S\big] = \frac{1}{m}\left(1 - \|x_S\|_2^2\right). \tag{22}$$

Write $y^{(m)} = \frac{1}{m}\sum_{j=1}^m Y_j$ with $Y_j = e_{k_j}$ i.i.d. Then $\mathbb{E}[Y_j \mid x_S] = x_S$ and $\mathrm{Cov}(y^{(m)} \mid x_S) = \frac{1}{m}\mathrm{Cov}(Y_1 \mid x_S)$. Since $\mathbb{E}[Y_1 Y_1^\top \mid x_S] = \mathrm{diag}(x_S)$, $\mathrm{Cov}(Y_1 \mid x_S) = \mathrm{diag}(x_S) - x_S x_S^\top$. Taking trace gives (22).

A.11. Polarization drift: full derivation of the variance-injection law

Let $U = \|\bar{x}\|_2^2$. Assume unbiased messaging $\mathbb{E}[y \mid x_S] = x_S$ (true for Soft/Hard/Top-m). For the listener update $x_L' = (1-\alpha)x_L + \alpha y$,

$$\mathbb{E}[\Delta U \mid X] = \frac{\alpha^2}{N^2}\,\mathbb{E}\big[\|y - x_L\|_2^2 \mid X\big]. \tag{23}$$

From (18), $\bar{x}' = \bar{x} + \frac{\alpha}{N}(y - x_L)$. Expand $\|\bar{x}'\|_2^2 - \|\bar{x}\|_2^2 = \frac{2\alpha}{N}\langle \bar{x}, y - x_L\rangle + \frac{\alpha^2}{N^2}\|y - x_L\|_2^2$. Conditioned on X, unbiased messaging and uniform ordered-pair sampling give $\mathbb{E}[y \mid X] = \bar{x} = \mathbb{E}[x_L \mid X]$, so $\mathbb{E}[y - x_L \mid X] = 0$ and the linear term vanishes.

Soft exchange. Under Soft exchange ($y = x_S$),

$$\mathbb{E}[\Delta U \mid X]_{\text{soft}} = \frac{\alpha^2}{N^2}\mathbb{E}\big[\|x_S - x_L\|_2^2 \mid X\big] = \frac{2\alpha^2}{N^2(N-1)}\,V. \tag{24}$$

This is (23) with (17) substituted for $\mathbb{E}\big[\|x_S - x_L\|_2^2 \mid X\big]$.

Under Top-m empirical communication $y = y^{(m)}$,

$$\mathbb{E}[\Delta U \mid X]_{\text{top}m} = \mathbb{E}[\Delta U \mid X]_{\text{soft}} + \frac{\alpha^2}{mN^2}\,\mathbb{E}\big[1 - \|x_S\|_2^2 \mid X\big]. \tag{25}$$

Start from (23). Decompose $y^{(m)} - x_L = (x_S - x_L) + (y^{(m)} - x_S)$. Conditioned on (x_S, x_L), $\mathbb{E}[y^{(m)} - x_S \mid x_S] = 0$, so the cross term vanishes:

$$\mathbb{E}\|y^{(m)} - x_L\|_2^2 = \mathbb{E}\|x_S - x_L\|_2^2 + \mathbb{E}\|y^{(m)} - x_S\|_2^2.$$

The first term is the Soft contribution (24), and (22) gives the second term.

Closed form in (U, V). Using (15), so that $q = U + V/N$,

$$\mathbb{E}[\Delta U \mid X]_{\text{top}m} = \frac{2\alpha^2}{N^2(N-1)}\,V + \frac{\alpha^2}{mN^2}\left(1 - U - \frac{V}{N}\right). \tag{26}$$

At perfect symmetry ($x_i = \mathbf{1}/K$), $V = 0$ and $U = 1/K$, giving $\mathbb{E}[\Delta U \mid X]_{\text{top}m} = \frac{\alpha^2}{mN^2}(1 - 1/K)$.

A.12. Disagreement drift and coordination drift

Let $V = \sum_i \|x_i - \bar{x}\|_2^2$. Under Top-$m$ empirical communication,

$$\mathbb{E}[\Delta V \mid X] = -\frac{2\alpha}{N-1}\left(1 - \alpha + \frac{\alpha}{N}\right)V + \frac{\alpha^2(N-1)}{mN}\left(1 - U - \frac{V}{N}\right). \tag{27}$$

Let $\Delta x := x'_L - x_L = \alpha(y - x_L)$ and $\Delta\bar{x} = \Delta x/N$. Write $\delta_i = x_i - \bar{x}$ so $\sum_i \delta_i = 0$. A direct expansion gives $V' = V + 2\langle\delta_L, \Delta x\rangle + \frac{N-1}{N}\|\Delta x\|_2^2$. For the linear term, condition on (x_S, x_L) and use $\mathbb{E}[y \mid x_S] = x_S$ to replace Δx by $\alpha(x_S - x_L)$ in expectation. Averaging over uniformly sampled ordered pairs then gives the contraction contribution $-\frac{2\alpha}{N-1}\left(1 - \alpha + \frac{\alpha}{N}\right) V$. For the quadratic term, decompose $\|y - x_L\|_2^2 = \|x_S - x_L\|_2^2 + \|y - x_S\|_2^2$ and apply (17) and (22) to obtain the injection term. Finally use $1 - \mathbb{E}\|x_S\|_2^2 = 1 - q = 1 - U - V/N$.

Soft limit. Under Soft exchange ($m = \infty$), the injection term vanishes and

$$\mathbb{E}[\Delta V \mid X]_{\text{soft}} = -\frac{2\alpha}{N-1}\left(1 - \alpha + \frac{\alpha}{N}\right) V \le 0. \quad (28)$$

Coordination drift. Using (16) and the drift formulas above,

$$\mathbb{E}[\Delta S \mid X] = \frac{2\alpha}{N(N-1)^2} V \;\ge\; 0. \quad (29)$$

In particular, $\mathbb{E}[\Delta S \mid X]$ is nonnegative and depends on m only through the evolution of V.

A.13. Homogeneous mean-field closure and consensus time scaling

A common closure assumes agents remain approximately homogeneous: $x_i \approx \bar{x}$, so $V \approx 0$ and $q \approx U$. Then (26) yields the per-step approximation

$$\mathbb{E}[\Delta U \mid X] \approx \frac{\alpha^2}{mN^2}(1 - U). \quad (30)$$

Measuring time in population rounds $\tau = t/N$ and treating the dynamics continuously gives

$$\frac{dU}{d\tau} \approx \frac{\alpha^2}{mN}(1-U), \qquad U(\tau) \approx 1-(1-U_0)\exp\left(-\frac{\alpha^2}{mN}\tau\right). \quad (31)$$

Thus the characteristic timescale in population rounds scales as $\tau_{\text{cons}} \sim \frac{mN}{\alpha^2}$.

A.14. Mean-field comparison at finite α

The mean-field approximation (31) uses a homogeneous closure and a continuous-time limit for the discrete QSG updates. Small systematic deviations from this curve have two simple sources. When α is small, the continuous approximation is accurate and agent heterogeneity becomes visible: by Jensen's inequality, $\mathbb{E}[\|x_S\|_2^2] \ge \|\bar{x}\|_2^2 = U$, so the realized drift term $1 - \mathbb{E}[\|x_S\|_2^2]$ can be smaller than the homogeneous term $1 - U$. When α is large, especially at the hard-copying limit $\alpha = 1$, listener states move through finite one-hot jumps, which can polarize faster than the smooth exponential trajectory. These effects account for the small below- and above-mean-field deviations in Fig. 6a.

A.15. Tempered sampling and the crossover parameter Γ_T

For the tempered-sampling extension used in Fig. 10, we define the temperature transform

$$g_T(x)_k \propto x_k^{1/T}, \qquad \sum_{k=1}^{K} g_T(x)_k = 1, \quad (32)$$

and generate messages from $g_T(x_S)$ instead of x_S (Hard/Top-m). Then $\mathbb{E}[y \mid x_S] = g_T(x_S)$, and the mean-field dynamics of $\bar{x}$ becomes

$$\frac{d\bar{x}}{d\tau} \approx \alpha\big(g_T(\bar{x}) - \bar{x}\big). \quad (33)$$

Linearizing around the symmetric point $u = \mathbf{1}/K$ with $\bar{x} = u + \delta$ and $\sum_k \delta_k = 0$ gives

$$g_T(u+\delta) = u + \frac{1}{T}\delta + O(\|\delta\|_2^2), \qquad \Rightarrow \qquad \frac{d\delta}{d\tau} \approx \alpha\left(\frac{1}{T} - 1\right)\delta. \quad (34)$$

Thus $T < 1$ deterministically amplifies small asymmetries while $T > 1$ damps them.

Quantized communication injects sampling-driven polarization at a characteristic population-round rate scale $\sim \alpha^2/(mN)$ (cf. (31) near symmetry). Comparing the deterministic linear rate $\alpha|1/T - 1|$ to the quantization-driven scale $\alpha^2/(mN)$ motivates the dimensionless crossover parameter

$$\Gamma_T \;:=\; \frac{mN}{\alpha}\left|\frac{1}{T} - 1\right|. \quad (35)$$

We interpret $\Gamma_T \approx 1$ as a *finite-size crossover* between near-neutral (drift-dominated) and tempering-dominated regimes, not as a literal thermodynamic phase transition of the finite-N absorbing chain.

A.16. Relating U to entropy and magnetization via a one-vs-rest ansatz

To connect second-moment theory to entropy/magnetization plots, a useful approximation is the one-vs-rest ansatz

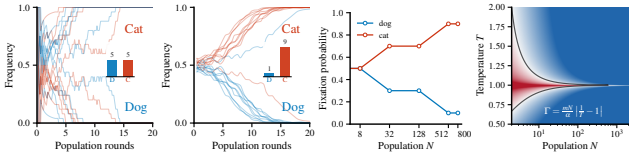
$$\bar{x} \approx \left(p, \frac{1-p}{K-1}, \ldots, \frac{1-p}{K-1}\right), \qquad p = \max_k \bar{x}_k \in [1/K, 1].$$

Under this ansatz,

$$U = p^2 + \frac{(1-p)^2}{K-1}, \qquad p(U) = \frac{1 + \sqrt{(K-1)(KU-1)}}{K}. \quad (36)$$

The magnetization $M = \frac{Kp-1}{K-1}$ becomes

$$M(U) = \sqrt{\frac{KU-1}{K-1}}. \quad (37)$$



(a) Drift at $N = 8$. *(b)* Selection at $N = 800$. *(c)* Fixation vs. N. *(d)* Temperature crossover.

Figure 10. **Additional drift–selection crossover checks in the two-label naming game.** Panels (a)–(c) show two-label (Dog, Cat) naming-game experiments with GPT-4o using random ordered speaker–listener pairs for a fixed referent r. (a) At $N = 8$, runs show substantial run-to-run variability, and the inset shows winner counts across trials. (b) At $N = 800$, a weak asymmetry consistently selects the same winner (Cat), again reflected in the inset counts. (c) Fixation probability versus N shows a finite-size crossover. (d) The corresponding tempered-sampling crossover diagram derived from QSG theory in $(T, mN/\alpha)$; the black curve marks $\Gamma_T = 1$, where $\Gamma_T = \frac{mN}{\alpha}\left|\frac{1}{T} - 1\right|$.

The normalized entropy $H(\bar{x}) = -\frac{1}{\log K}\sum_k \bar{x}_k \log(\bar{x}_k + \varepsilon)$ yields

$$H(U) = -\frac{1}{\log K}\left[p(U)\log p(U) + (1-p(U))\log\left(\frac{1-p(U)}{K-1}\right)\right], \tag{38}$$

where ε is used only for numerical stability in implementations.

B. Experiments

B.1. Committee budget-plan task and labels

The committee budget-plan experiments in Fig. 1 keep the naming-game structure of repeated discrete choices for a shared referent, but place the choices inside a municipal budget recommendation prompt. The core user prompt is:

User message (committee budget-plan task)

```
A municipal planning committee must recommend one
of three proposals for the next capital budget
cycle.  The staff memo says the three proposals are
equally strong overall under the current evidence.
One proposal would modestly improve downtown travel
times, one would modestly improve neighborhood
coverage, and one would modestly improve off-peak
service reliability, but the trade-offs balance out.
The proposals have the same cost, the same
implementation timeline, and no objective
tie-breaker.  If the committee had to recommend one
plan code now, any of the three would be a
defensible choice.
```

The model is instructed to output exactly one valid plan code in JSON. Agents receive no experimenter-defined payoff, ground-truth answer, or external feedback. The task is a contextualized near-tie decision with neutral plan codes chosen to keep priors close to balanced. For readability, the main figure displays the $K = 2$ codes as Budget 1 and Budget 2. The actual neutral random codes were `ktwbv` and `qezoq`, with blank-memory prior $0.480/0.520$. The matched $K = 3$ committee prediction panels used `kjxdc`, `pdtet`, and `gmany`, with blank-memory probabilities $0.346/0.325/0.329$. In all committee experiments, N denotes independently maintained memory states queried through the same base model/API. It does not denote separately trained or fine-tuned models.

B.2. Numerical simulations of QSG

We report QSG simulations comparing Soft, Hard, and Top-m dynamics across symmetry breaking, temperature sweeps, and entropy trajectories. The one-step drift identity test in Fig. 6b is reported in the main text; here we focus on additional dynamics and robustness sweeps.

B.2.1. Setup

Unless stated otherwise:

- Population size $N = 24$, vocabulary size $K = 10$ (matching a common experimental setting in (Ashery et al., 2025)).
- Soft update: $x_L \leftarrow (1-\alpha)x_L + \alpha x_S$.
- Hard update: $k^\star \sim \mathrm{Cat}(x_S)$ and $x_L \leftarrow (1-\alpha)x_L + \alpha e_{k^\star}$.
- Top-m update: sample m tokens i.i.d. with replacement and update toward their empirical distribution.
- We track coordination rate $S(t)$ (Eq. (14)) and entropy $H(t)$.

B.2.2. Simulation suite

We run four experiments: (i) Soft versus Hard dynamics from the symmetric initialization $x_i(0) = \mathbf{1}/K$, (ii) a softmax-temperature sweep for Hard sampling, showing a finite-horizon crossover near $T \approx 1$, (iii) a comparison of Hard, Soft, and Top-m updates from symmetric initialization, where entropy remains high for $m > 1$ while Hard collapses, and (iv) the direct Monte Carlo test of Theorem 2 reported in Fig. 6b.

B.2.3. Method summary and reproducibility

We simulate QSG directly in probability space with Soft, Hard, and Top-m updates as defined in the main text. The Top-m message uses the empirical distribution from m independent samples with replacement. The Monte Carlo estimator samples a random speaker and listener, computes the one-step change in $U = \|\bar{x}\|_2^2$, and averages over repeated draws. Unless noted, we use $N = 24$, $K = 10$, and the α listed in each figure; m ranges over $\{1, 2, 3, 5, 10, 20\}$. The simulations use NumPy with fixed seeds.

B.3. LLM population experiments

LLM population experiments using the Neutral Naming Drift (NND) protocol are reported in Sec. 4 and summarized in Fig. 7. Across all LLM runs, we use delayed-reveal interactions with memory size $H = 10$, synthetic 5-character labels, a fixed referent string, and no external reward. Label order is shuffled each interaction, and probes are measurement-only. Trial seeds are deterministic: resolved runs increment trial seeds by fixed offsets, pair schedules use a trial-specific derived seed, and label-order shuffles use independent deterministic offsets. Blank-memory prior probes use the resolved configuration's `swap_seed` value 123.

Across LLM population experiments, each agent is represented by an independent memory buffer. Interactions are ordered speaker–listener updates:

1. Sample an ordered pair (S, L) uniformly with $S \neq L$.
2. Query the speaker S using its current memory.
3. Parse the speaker message y as one valid label for Hard communication, or as an empirical multiset of m labels for Top-m.
4. Append the parsed speaker message to the listener L's memory.
5. Do not modify the speaker memory.
6. Probe queries are measurement-only and are never appended to any memory.

This protocol matches the one-sided listener update in QSG. If any implementation logs a listener response for diagnostics, that response is not used as a state update unless explicitly stated.

The OpenAI runs use API model id `gpt-4o` with temperature 1.0, top-p = 1.0, and `max_completion_tokens`=60. The Anthropic runs use API model id `claude-haiku-4-5-20251001` with temperature 1.0, top-$p = 1.0$, and `max_tokens`=60; the backend omits the Anthropic `top_p` field when it is at the default value 1.0. All production runs use JSON output mode where available. The parser accepts exactly one top-level JSON object: `{"label": "<label>"}` for Hard and `{"labels": ["<label>", ...]}` for Top-m. Outputs with invalid JSON, out-of-set labels, the padding token, unexpected keys, or the wrong Top-m list length are rejected and retried; failed retries are logged and excluded from state updates.

Empirical observables are estimated from probe outputs that are used only for measurement and are never incorporated into memory. At each probe time, we estimate the population-average label distribution p from sampled agent outputs, yielding the empirical counterpart of the theoretical population mean $\bar{x}$ defined in Sec. 3. In the Top-m sweep, repeated probe draws support a direct estimate of U, whereas in the N-sweep the plotted squared-frequency statistic is a finite-sample proxy for ensemble scaling trends. We estimate early drift as $\Delta U/\text{step}$, and define time to consensus as the first probe where $U \geq U_\star$ with $U_\star = 0.9$. For scaling plots, we aggregate this quantity over trials that also finish above the same threshold at the run horizon. Because p is estimated from finite samples, we report cross-trial variability.

For a one-hot observed label k, QSG predicts $p'_k = (1 - \alpha)p_k + \alpha$, hence

$$\alpha_{\text{eff}} = \frac{p'_k - p_k}{1 - p_k}. \tag{39}$$

We estimate p and p' using measurement-only probes before and after a single listener update, and average the resulting estimates across stored early memory states and trials. Non-target coordinates give the equivalent estimate $(p_j - p'_j)/p_j$, used as a consistency check.

Prompt structure and decoding are fixed across runs: the system message enforces valid JSON output over the allowed labels, and the speaker/listener prompts specify the referent, allowed labels, and the agent's memory of past interactions. Fig. 11 shows a concrete example.

For the $K = 3$ N-sweeps in Fig. 7 (GPT-4o and Claude Haiku 4.5), we use $N \in \{2, 3, 4, 6, 8, 12, 16, 24, 32\}$, $m = 1$ (Hard), and horizon $T = 30N$ interactions. Probes are taken once per population round, with one probe sample per agent in the original trajectory and consensus panels. In the N-sweep trajectory and consensus panels, each probe time uses one measurement-only probe sample per agent, so the plotted squared-frequency statistic should be interpreted as a finite-sample proxy for the population-level polarization. We do not treat these one-sample trajectory estimates as bias-corrected estimates of the continuous-state quantity $U = \|\bar{x}\|_2^2$. The quantitative plug-in early-drift comparisons in Fig. 7(b,e) instead use stored early memory states re-probed with 32 repeated draws and combine those measurements with independently estimated α_{eff} and speaker uncertainty. The GPT-4o drift panel uses 10 trials per N; the Claude Haiku 4.5 drift panel and the remaining $K = 3$ panels use 5 trials per (model, N).

B.3.1. Early-drift prediction and adaptation diagnostics

The QSG plug-in prediction used for the early-drift panels is

$$\widehat{\Delta U}_{\text{QSG}}(N) = \widehat{\alpha}_{\text{eff}}^2 N^{-2} \widehat{\mathbb{E}}\big[1 - \|x_S\|_2^2\big]$$

System message

```
You must output only valid JSON. No extra keys, no
explanations, no markdown.
Valid labels are exactly those in Allowed labels.
Never output "<PAD>".
```

User message (speaker)

```
Referent: ref_07
Allowed labels: ["hdsad", "vokhg", "fmhja"]
The list order is randomized and has no meaning.
Both players are choosing a label for the same
referent in repeated interactions. The memory
shows labels you observed from previous
interactions with partners.
Memory (last H observed messages, oldest -> newest,
padded with "<PAD>"): ["<PAD>", "<PAD>", "<PAD>",
"<PAD>", "<PAD>", "<PAD>", "<PAD>", "fmhja",
"hdsad", "vokhg"]
Each memory entry is a label string.

Constraints:
- Output JSON only.
- Every label must be from Allowed labels.

Output JSON exactly: {"label": "<label>"}
```

Figure 11. **Example speaker prompt (Hard, $m = 1$, $K = 3$, $H = 10$).** Labels are synthetic 5-character strings; memory is zero-padded with <PAD> tokens. Listener prompts follow the same basic structure, with their own memory buffer; the ordered speaker–listener update protocol is described above.

Table 1. **Effective-adaptation diagnostics.** CIs are bootstrap intervals for the mean over valid single-interaction projection estimates; residual is the mean norm of the component not explained by the scalar projection.

Setting	n	$\bar{\alpha}_{\text{eff}}$	95% CI	std.	out	resid.
Committee GPT-4o	355	0.725	[0.702, 0.745]	0.205	4 / 0	0.112
NND GPT-4o	51	0.982	[0.965, 0.995]	0.056	0 / 0	0.001
NND Claude	36	0.927	[0.870, 0.976]	0.167	0 / 0	0.000

where $\widehat{\alpha}_{\text{eff}}$ and speaker uncertainty are estimated from measurement probes rather than fit to the population-level drift curve. We read the LLM results as scaling and calibration checks of the QSG account.

Table 1 summarizes the effective-adaptation estimates used in the early-drift predictions. Projection estimates are undefined when the listener state already matches the message direction; those rows are omitted from the table. The estimates are concentrated in the expected range, with interaction-level variability especially in the committee prompt. Thus α_{eff} serves as an effective scalar summary for the plug-in prediction.

B.3.2. Top-m LLM sweep

For the $K = 10$ Top-m sweep in Fig. 8, we use GPT-4o with $N = 4$, horizon $t_{\max} = 120$ interaction steps, and $m \in \{1, 2, 3, 4\}$ (10 trials each). Early drift is estimated from the initial linear regime using measurement-only probes at each interaction step; error bars in Fig. 8 show s.e.m.

LLM API usage is dominated by interaction calls and high-probe measurement. Ignoring parser retries, one $K = 3$ NND trial uses about $2T = 60N$ interaction calls plus about $30N$ probe calls, so a 5-trial sweep over the N-grid uses roughly $48{,}150$ calls per model before repeated re-probes. The $K = 10$ Top-m high-probe sweep uses about $8{,}240$ calls per trial, or roughly $329{,}600$ calls for $m \in \{1, 2, 3, 4\}$ with ten trials each. The $K = 2$ committee $N = 800$ panel uses about $320{,}000$ interaction calls across ten trials; the $N = 8$ panel uses about $2{,}640$ calls including trajectory probes. The experiments use commercial LLM APIs under their provider access terms and standard scientific Python packages. We do not provide code, generated interaction logs, experiment configs, or figure scripts with the anonymous submission. To support reproducibility, Appendix B reports the interaction protocol, prompt schemas and examples, model identifiers, decoding parameters, label sets, population sizes, horizons, probe cadence, estimator definitions, parser/retry rules, and approximate API-call budgets.