A UNIFIED INFORMATION-GEOMETRIC FRAMEWORK FOR COLLECTIVE BELIEF DYNAMICS IN MULTI-AGENT LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Communication among LLM agents sometimes adds information and sometimes collapses a population onto a wrong consensus. We isolate one step of this question: a population of beliefs is pooled and is also updated on common evidence, and we ask when the order matters. Logarithmic (e-) pooling commutes with a common Bayes update, whereas arithmetic (m-) pooling reallocates the social weights in proportion to the evidence normalizers; we characterize the exact zero set of this order gap, bound it, and show that it is zero or fourth order near consensus. Because the order gap is not a quality ordering, we score both orderings against an independently declared finite observation model: update-then-pool has higher expected log loss in all 18 cells (exact enumeration), and at matched observation budget no single rule is best across six information structures. Stated as operators on an information-geometric map of predictive states, the same language also closes the second moments of sampled communication, which includes the sampled convex-combination rule used to model memetic drift. A preregistered one-step transfer test on native Qwen2.5 receivers failed its gate, so we do not claim that native agents follow these operators.

1 INTRODUCTION

Multi-agent systems can outperform a single agent, most clearly when agents hold complementary observations (Pavlova & Tanaka, 2026) and also on shared objective questions (El et al., 2026), yet the same literature reports frequent degradation under communication: wrong consensus that becomes more likely as agents read more messages (Fukushima, 2026), collective accuracy that falls with population size (Pavlova & Tanaka, 2026), opinion drift without ground truth (El et al., 2026), and protocol violations that spread between peers (Shi et al., 2026). The contrast raises a task-level question: when does communication add information that the receiver lacks, and when does it merely repeat evidence already available to that receiver? A comparison against a single agent must also specify which observations that agent receives: a local agent with one exogenously assigned observation and an oracle with all observations answer different questions.

This paper isolates one step of that question. A population of beliefs is pooled and is also updated on common evidence; we ask when the two operations commute, how large the gap is when they do not, and which order is closer to an external target. We state the step as an operator on an information-geometric map, in the spirit of Oyama et al. (2025) but for interacting agents: each predictive distribution is a point in the Gram geometry induced by KL divergences on a common probe set (Proposition 5), and a message is $\mathcal{T}_{i\to j} = U_j \circ K_{i\to j}$, a speaker kernel followed by a receiver update (Section 2). Deterministic kernels give the order results, which are the main claims. Sampling kernels give, as a secondary result, the second moments of repeated communication; the sampled convex-combination model of memetic drift (Tanaka, 2026) is one such kernel (Theorem 1(e)). All statements are finite and exactly checked; Section D.1 reports the failed preregistered native transfer test on Qwen2.5 receivers.

Contributions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

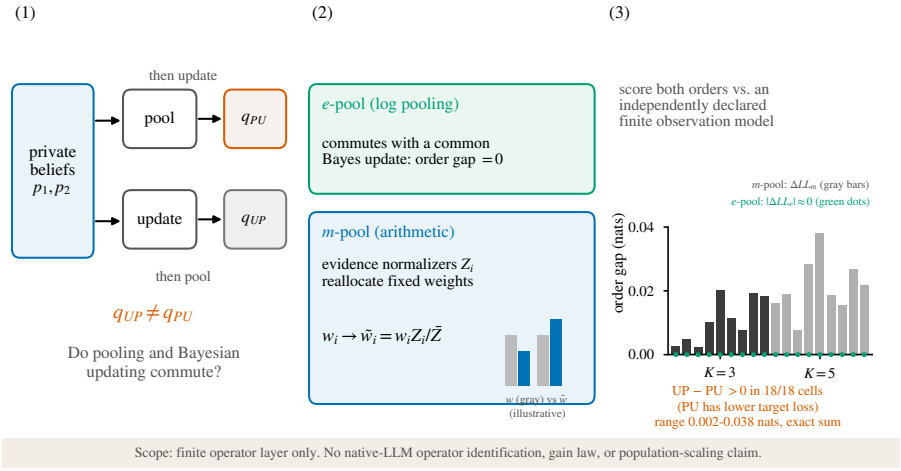


Figure 1: **Pooling order, two pools, and an independent finite target.** (1) Pooling then updating need not equal updating then pooling (q_{PU} vs. q_{UP}). (2) The e-pool (log pooling) commutes with a common Bayes update, so its order gap is zero; the m-pool (arithmetic) reallocates the declared weights through the evidence normalizers, $\tilde{w}_i = w_i Z_i / \bar{Z}$ (bars illustrative). (3) Both orderings are scored against an independently declared finite observation model: update-first minus pool-first cross-entropy, $\Delta LL_m = \mathrm{CE}(UP) - \mathrm{CE}(PU)$, is positive in all 18 specified E1 cells (range 0.002–0.038 nats; exact finite sum, not Monte Carlo), i.e. pool-first has the lower target loss; the e-pool gap $|\Delta LL_e|$ is zero to machine precision. Shade of the bars marks K. This finite-model comparison does not establish a native-LLM gain law.

1. **When pooling commutes with evidence.** Propositions 1–2 and Corollary 1 separate logarithmic commutation from arithmetic evidence-normalizer reallocation and give the exact zero set of the order gap; Propositions 3 and 4 bound it and show that it is zero or fourth order near consensus. Theorem 1 places these results in one operator algebra on the Gram geometry of Proposition 5.
2. **Measurement against an independent target at matched budget.** Figure 2 scores both arithmetic orderings against an independently constructed finite task posterior: update-then-pool has higher expected log loss in all 18 cells of that observation model. Table 1 compares every rule with no-communication and independence-product baselines that see the same observations, across six information structures; no single rule is best in all.
3. **Population statistics of sampled communication.** For sampling kernels, Proposition 7 closes the one-step drift and diffusion of the population's second moments exactly; Theorem 2 gives the consensus probabilities, Proposition 8 the conserved weighted mean, and Proposition 9 the N-scaling of the disagreement relaxation time, which contains the sampled convex-combination model of memetic drift as a special case. Proposition 10 turns the size dependence of the listener response into two conditional predictions: consensus speed that is monotone in agent size, and large agents that barely move in mixed-size teams.

Table 2 (Appendix A) maps each contribution to its formal result, measurement, and use.

2 Setting: Two Operators on a Population of Beliefs

Beliefs and weights. Let $\mathcal{Y}$ be a finite hypothesis space and $\Delta(\mathcal{Y})$ the set of strictly positive distributions on it. A population is a tuple $p = (p_1, \ldots, p_n) \in \Delta(\mathcal{Y})^n$ with fixed influence weights $w \in \operatorname{int} \Delta_n$. The *consensus diagonal* is $\mathcal{C}_\Delta = \{p : p_1 = \cdots = p_n\}$.

Pooling operators.

Definition 1 (m- and e-pooling). $C_{\mathrm{m}}(p) = \sum_i w_i p_i$ *and* $C_{\mathrm{e}}(p) = \prod_i p_i^{w_i}/A(p)$ *with* $A(p) = \sum_y \prod_i p_i(y)^{w_i} \in (0, \infty)$.

Both map $\Delta(\mathcal{Y})^n \to \Delta(\mathcal{Y})$ and agree on $\mathcal{C}_\Delta$. C_{m} and C_{e} are the weighted m- and e-mixtures, the two affine structures of the dually flat manifold $\Delta(\mathcal{Y})$.

Evidence operator. A piece of common evidence is a likelihood $L : \mathcal{Y} \to (0, \infty)$. The Bayes update acts agent-wise,

$$E_L p_i(y) = \frac{p_i(y)L(y)}{Z_i}, \qquad Z_i = \sum_y p_i(y)L(y),$$

and on a pooled belief q by the same formula with normalizer $Z_q = \sum_y q(y)L(y)$. We write $E_L p = (E_L p_1, \ldots, E_L p_n)$ for the population update. We call Z_i, the marginal likelihood of the evidence under i's prior, the *evidence strength* of agent i. Its weighted mean is $\bar{Z} = \sum_i w_i Z_i$, and $\mathrm{Var}_w(Z_i) = \sum_i w_i (Z_i - \bar{Z})^2$.

The two orderings and the order gap. Given a pooling operator $C \in \{C_{\mathrm{m}}, C_{\mathrm{e}}\}$ the system can produce a collective posterior in two ways:

$$\text{(update-then-pool)} \quad q_{\mathrm{UP}} = C(E_L p), \qquad \text{(pool-then-update)} \quad q_{\mathrm{PU}} = E_L C(p).$$

Definition 2 (Order gap). $\chi_C(p) = D_{\mathrm{sym}}(q_{\mathrm{UP}}, q_{\mathrm{PU}})$ *with* $D_{\mathrm{sym}}(q, q') = D_{KL}(q\|q') + D_{KL}(q'\|q)$.

$\chi_C \geq 0$ with equality iff the two orderings agree. Section 3 shows $\chi_{C_{\mathrm{e}}} \equiv 0$ and characterizes $\chi_{C_{\mathrm{m}}}$.

Assumptions and target. L is common to all agents, who may differ only in their priors; agent-specific likelihoods L_i are treated separately, and Proposition 1 does not extend to them. Beliefs and likelihoods are strictly positive; the constant-likelihood case is a zero-evidence special case, and support collisions are deferred to the appendix. The quartic expansion of Section 3 holds only under its stated nondegeneracy condition. Since the order gap is not a quality ordering, E1 separately declares a generative joint law and scores both outputs against its exact posterior.

3 Theory

3.1 Operator algebra: commutation, reallocation, and zero sets

Setup. Finite $\mathcal{Y}$; agents $i = 1, \ldots, N$, $p_i(y) > 0$, $w_i > 0$, $\sum_i w_i = 1$; common likelihood $L : \mathcal{Y} \to (0, \infty)$; $Z_i := \sum_y p_i(y)L(y) \in (0, \infty)$.

$$\Pi_E p_i(y) := \frac{p_i(y)L(y)}{Z_i}, \qquad \Pi_C^m(p_{1:N})(y) := \sum_i w_i p_i(y),$$

$$\Pi_C^e(p_{1:N})(y) := \frac{\prod_i p_i(y)^{w_i}}{A_e}, \qquad A_e := \sum_y \prod_i p_i(y)^{w_i} \in (0, \infty).$$

$D_{sym}(P, Q) := D_{KL}(P\|Q) + D_{KL}(Q\|P)$, $\chi(p) := D_{sym}(\Pi_C \Pi_E p,\ \Pi_E \Pi_C p)$.

Proposition 1 (e-pooling commutes with common-evidence Bayes update; $\chi = 0$). *Assume* $p_i(y) > 0\ \forall y, i$, $Z_i \in (0, \infty)$, *and* $A_e \in (0, \infty)$ *(before and after update). Then* $\Pi_C^e \Pi_E p = \Pi_E \Pi_C^e p$ *exactly, hence* $\chi(p) = 0$ *for e-pooling.*

Proof deferred to Appendix I.

Proposition 2 (m-pooling influence-weight reallocation, exact). *With* $Z_i \in (0, \infty)\ \forall i$ *and* $q_i(y) := \Pi_E p_i(y)$,

$$\Pi_C^m \Pi_E p(y) = \sum_i w_i q_i(y), \qquad \Pi_E \Pi_C^m p(y) = \sum_i \tilde{w}_i q_i(y), \quad \tilde{w}_i := \frac{w_i Z_i}{\sum_j w_j Z_j}.$$

Proof deferred to Appendix I.

Corollary 1 (Exact zero-gap condition)**.** *Under the assumptions above,*

$$\chi(p) = 0 \quad \Longleftrightarrow \quad \Pi_C^m \Pi_E p = \Pi_E \Pi_C^m p \quad \Longleftrightarrow \quad Z_1 = \cdots = Z_N.$$

In particular, unequal evidential normalizers cannot produce an accidental zero gap.

Proof deferred to Appendix I.

What the finite-state classification predicts. The reallocation identity does more than label an observed divergence: given positive priors, a common likelihood, and fixed weights, it constructs both orderings from the individually updated beliefs and their normalizers. Proposition 4 additionally predicts the *rate* of the symmetric-KL discrepancy before a small-t sweep: on its stated linear family, $V > 0$ yields the explicit fourth-order coefficient, whereas $V = 0$ yields an identically zero gap. Figure 10 checks the nonzero coefficient numerically: $\chi/(t^4 \mathbb{E}_{q_0} H^2)$ is 1.000033 at $t = 10^{-4}$; a distinct-prior equal-normalizer control has maximum computed gap 8×10^{-120} in Decimal arithmetic. These are finite-state operator predictions, not claims that pooling improves task accuracy or that the same rate holds for native language-model agents.

The distinction matters: divergence between the two orderings alone has no sign of quality. Against an independently constructed task posterior, Figure 2 measures a positive update-then-pool minus pool-then-update log-loss contrast in all 18 declared synthetic cells (0.0023–0.0382 nats), but this is a restricted observation-model result, not a consequence of the quartic theorem. Figure 9 is a separate held-out quality test: the frozen receiver-gain predictor loses to an information-only MAE baseline in all three runs (for 7B/145, 0.393 versus 0.283 nats). Hence the present evidence does not establish a native gain-sign predictor or a matched-token single-agent advantage.

Proposition 3 (Exact global lower and upper bound)**.** *Let $A = \sum_i w_i q_i$, $B = \sum_i \tilde{w}_i q_i$, $d_i = \tilde{w}_i - w_i$, $F(w)_{ij} = \sum_y q_i(y) q_j(y)/A(y)$, $r_{\max} = \max_i Z_i/\bar{Z}$, $r_{\min} = \min_i Z_i/\bar{Z}$, and $\psi(s) = \log(s)/(s-1)$ with $\psi(1) = 1$ (ψ positive and decreasing on $(0, \infty)$). Then, without a local or non-degeneracy assumption,*

$$\psi(r_{\max})\, d^\top F(w) d \;\le\; \chi(p) \;\le\; \psi(r_{\min})\, d^\top F(w) d, \qquad d^\top F(w) d \ge 0.$$

In particular, as $r_{\min}, r_{\max} \to 1$, $\chi(p)/d^\top F(w) d \to 1$ (for $d^\top F(w) d > 0$); the coefficient $\psi(r_{\max})$ does not *degenerate as $r_{\max} \to 1$ (indeed $\psi(1) = 1$). Additionally, for non-constant positive L, with $J := \sum_i w_i/Z_i - 1/\bar{Z}$ and $R_L := \max_y L(y)^{-1} - \min_y L(y)^{-1}$,*

$$\chi(p) \;\ge\; 4\,(J/R_L)^2,$$

a certificate obtainable without computing $F(w)$ in full and invariant under $L \mapsto cL$.

Proof deferred to Appendix I.

Proposition 4 (Quartic or identically zero along a linear family)**.** *Let $p_i(t) = p_0(1 + t a_i)$, where p_0 is strictly positive, the functions a_i are fixed, and $\mathbb{E}_{p_0}[a_i] = 0$. Restrict t to an interval on which every $p_i(t)$ is positive. Set $q_0 = \Pi_E p_0$, $m_i = \mathbb{E}_{q_0}[a_i]$, and*

$$\mu = \sum_i w_i m_i, \qquad V = \sum_i w_i (m_i - \mu)^2, \qquad H(y) = V - \sum_i w_i (m_i - \mu) a_i(y).$$

Then

$$\chi(p(t)) = t^4 \sum_y q_0(y) H(y)^2 + O(t^5).$$

The fourth-order coefficient is positive if and only if $V > 0$. If $V = 0$, the order gap is identically zero throughout the positive part of this family.

Proof. The normalized updates and reallocated weights are exactly

$$q_i(t) = q_0(1 + t a_i)/(1 + t m_i), \qquad \tilde{w}_i(t) = w_i(1 + t m_i)/(1 + t\mu).$$

Expanding the two pooled distributions gives $A - B = t^2 q_0 H + O(t^3 q_0)$, and the local symmetric-KL expansion gives the displayed fourth-order term. Moreover,

$$\sum_y p_0(y) H(y) = V.$$

Thus $V > 0$ implies that H is nonzero somewhere; positivity of q_0 makes the coefficient strictly positive. If $V = 0$, all m_i are equal, so all evidence normalizers coincide exactly for every admissible t. Corollary 1 then gives an identically zero gap. □

Variance is not a global control variable. The dispersion $\mathrm{Var}_w(Z_i)$ alone does not order χ. For $N = 2$, $w = (1/2, 1/2)$ and $L = (1, 2, 3)$, the priors $((.49, .03, .48), (.11, .77, .12))$ give $\mathrm{Var}_w(Z_i) = 10^{-4}$ and $\chi = 1.412 \times 10^{-5}$, whereas $((.29, .40, .31), (.31, .40, .29))$ give four times the variance but $\chi = 5.667 \times 10^{-8}$. Thus neither the abstract nor the experiments interpret a marginal correlation between these quantities as a universal monotonic law.

Corollary 2 ($L \equiv c$ is a $\chi = 0$, zero-evidence special case; *not* a general common-field embedding)**.** *If the likelihood is an agent-independent positive constant, $L \equiv c$ (no informative evidence), then $Z_i = c\sum_y p_i(y) = c = \bar{Z}$ for every i,* regardless of the priors p_i. *Proposition 2 then gives $\tilde{w}_i = w_i\bar{Z}/\bar{Z} = w_i$ for all i (i.e. $\Delta w = 0$), so $\Pi_C^m\Pi_E p = \Pi_E\Pi_C^m p$ holds* exactly *and $\chi(p) = 0$. **This is the only sense in which we identify Tanaka's construction with an equal-Z_i cross-section**; a merely* common *(agent-independent) but non-constant likelihood L does* not *imply $Z_i \equiv \bar{Z}$ in general.*

Proof deferred to Appendix I. A common likelihood need not give equal normalizers. The constant-likelihood case is a zero-evidence special case, not a topological boundary of the positive-likelihood domain. The appendix gives the common-field counterexample.

The verifier limit $L(y) = \varepsilon$ off a target label, where the arithmetic gap is $\varepsilon\kappa + O(\varepsilon^2)$, is treated in Appendix I (Proposition 18).

3.2 A UNIFIED OPERATOR THEORY OF COMMUNICATING AGENTS

The finite pooling results above are properties of one communication operator. This subsection places them, together with the population dynamics of sampled communication, in a single state space. Proofs are in Appendix C.

Setup. Agents $i \in [N]$ hold predictive distributions p_i; a fixed probe set $T = \{y_1, \ldots, y_M\}$ gives states $\ell_i = (\log p_i(y_m))_{m \le M} \in \mathbb{R}^M$, centered ω_i. A communication $(i \to j)$ updates only the receiver, $\ell_j' = U_j(\ell_j, m)$, $m \sim K_{i\to j}(\cdot \mid \ell_i)$; write $\mathcal{T}_{i\to j} = U_j \circ K_{i\to j}$. Common evidence acts as $E_L : \ell \mapsto \ell + \log L$. Population statistics in simplex coordinates $x_i = p_i$: $\bar{x} = N^{-1}\sum x_i$, $R = \|\bar{x}\|^2$, $M_2 = N^{-1}\sum \|x_i\|^2$, $V = M_2 - R$.

State space: information geometry.

Proposition 5 (Multi-agent Gram identity)**.** *Under (M1)–(M4) of Lemma 1 for every i: (a) $\mathrm{KL}(P_i\|P_j) = \|\omega_i - \omega_j\|^2/(2M) + O(\varepsilon^3) + O_P(M^{-1/2})$; (b) with $\langle u, v\rangle_T = \sum_m u_m v_m/M$, $\tilde{\omega}_i = \omega_i - N^{-1}\sum_j \omega_j$ and the symmetrized $K_{ij} = \frac{1}{2}\big[\mathrm{KL}(P_i\|P_j) + \mathrm{KL}(P_j\|P_i)\big]$, $\langle\tilde{\omega}_i, \tilde{\omega}_j\rangle_T = -\big[K_{ij} - K_{i\cdot} - K_{\cdot j} + K_{\cdot\cdot}\big]$ up to the same error ($K_{i\cdot}, K_{\cdot j}, K_{\cdot\cdot}$ row, column and grand means).*

Communication operators.

Theorem 1 (Algebra of $\mathcal{T}$)**.** *For the deterministic-kernel family and the sampling-kernel family (QSG, Eq. equation 3): (a) e-pooling is affine in ω and commutes with E_L (Prop. 1); (b) m-pooling fails to commute with E_L only through the weight reallocation $w_i \mapsto \tilde{w}_i \propto w_i Z_i$ (Prop. 2); (c) $\chi = 0$ iff the Z_i agree, with the local lower bound of Prop. 3 and Cor. 1; (d) the quartic law and the off-zero order effect (Prop. 4); (e) the QSG update is the sampling-kernel instance of Proposition 6, with $\mathbb{E}\hat{p} = x_S$, the input of Prop. 7.*

Proposition 6 (Tanaka's QSG is a sampling-kernel instance of $\mathcal{T}$)**.** *Take the probe set $T = \mathcal{Y} = \{1, \ldots, K\}$, so $\ell_i = \log x_i$ coordinatewise. Let $K_{S\to L}^{(m)}(\cdot \mid x_S)$ be the law of $\hat{p} = m^{-1}\sum_{r\le m} e_{Z_r}$, $Z_r \overset{iid}{\sim} \mathrm{Cat}(x_S)$, and let $U_L^{(\alpha)}(x_L, \hat{p}) = (1-\alpha)x_L + \alpha\hat{p}$.*

- *(i) One interaction of Tanaka's QSG (Tanaka, 2026, §2, Eq. (1)), Eq. equation 3, is exactly $\mathcal{T}_{S\to L} = U_L^{(\alpha)} \circ K_{S\to L}^{(m)}$ acting on the receiver with all other agents fixed. For $\alpha < 1$ it maps $\mathrm{int}\,\Delta^{K-1}$ into itself, so it is an operator on ℓ-space; for $\alpha = 1$ it is defined on the closed simplex (extended ℓ). $m \to \infty$ gives the deterministic kernel δ_{x_S} (soft exchange); $m = 1$ gives a hard sample.*

(ii) $U_L^{(\alpha)}$ *is m-pooling* Π_C^m *with weights* $(1-\alpha, \alpha)$. *Hence QSG inherits the evidence non-commutativity of Proposition 2. It has no e-pooling counterpart, so* χ *(Proposition 3) is not defined within QSG.*

(iii) Under uniform ordered pairs $S \neq L$, *the one-step drift and diffusion of the population mean are*

$$b := \mathbb{E}[\Delta\bar{x}] = 0, \qquad \Sigma := \mathbb{E}[\Delta\bar{x}\,\Delta\bar{x}^\top] = \left(\tfrac{\alpha}{N}\right)^2 \Big[\tfrac{1}{m}\big(\operatorname{diag}\bar{x} - N^{-1}\textstyle\sum_i x_i x_i^\top\big) + \tfrac{2N}{N-1}\,\Sigma_{\mathrm{pop}}\Big],$$

with $\Sigma_{\mathrm{pop}} = N^{-1}\sum_i (x_i - \bar{x})(x_i - \bar{x})^\top$. *In particular* $\mathbb{E}\Delta R = \operatorname{tr}\Sigma$, *which is the R-equation of Proposition 7.*

(iv) Consequently the consensus lottery (Theorem 2) and the N*-scaling* $\lambda_{\mathrm{eff}} N^2$ *(Proposition 9) are properties of* $\mathcal{T}$ *restricted to the family* $\{U^{(\alpha)} \circ K^{(m)}\}$. *At consensus they reduce to Proposition 16.*

Scope. Two things are not included, and we state this explicitly. (1) Tanaka's common external field: a common bias does not equalise the Z_i (Appendix I, "Common likelihood versus equal normalizers"). It meets our model only in the zero-evidence case, Corollary 2. (2) The drift–selection crossover, which needs a speaker-side field model.

Statistical mechanics of the population.

Proposition 7 (Closed second-moment system)**.** *Uniform ordered pairs* (L, S), $x'_L = (1-\alpha)x_L + \alpha\hat{p}$, $\hat{p}$ *the empirical law of* m *i.i.d. draws from* $\mathrm{Cat}(x_S)$. *Exactly, without closure:* $\mathbb{E}\Delta R = \frac{\alpha^2}{N^2}\big[\frac{1-R-V}{m} + \frac{2NV}{N-1}\big]$, $\mathbb{E}\Delta V = -\kappa V + c(1-R-V)$, $\kappa = \frac{2\alpha}{N-1}\big(1 - \frac{\alpha(N-1)}{N}\big)$, $c = \frac{\alpha^2(N-1)}{mN^2}$. *With* V *adiabatically eliminated,* $dR/ds = \lambda_{\mathrm{eff}}(1-R)$, $\lambda_{\mathrm{eff}} = 2\alpha^2/[2Nm(N - \alpha(N-1)) + \alpha(N-1)^2]$.

Separation from the mean-field closure (pre-fixed, $\delta \propto V$**).** At m=1, the mean-field closure $M_2 \approx R$ gives $\mu_{\mathrm{MF}} = \frac{\alpha^2}{N^2}(1-R)$, whereas Prop. 7 gives $\delta = \mu_{U3} - \mu_{\mathrm{MF}} = \frac{\alpha^2}{N^2}\frac{N+1}{N-1}V$, so the two coincide iff $V = 0$. For K=2, N=4, α=1/2 at matched R=1/2: $\delta = 5/1536$ for $V = 1/8$ and $\delta = 0$ for $V = 0$ (exact enumeration of all 24 outcomes).

Proposition 8 (Conserved weighted mean)**.** $\mathbb{E}\Delta\bar{x} = [N^2(N-1)]^{-1}\sum_{i<j}(\gamma_{ij} - \gamma_{ji})(x_j - x_i)$. *If* $c_i\gamma_{ij} = c_j\gamma_{ji}$ *then* $\sum_i c_i x_i$ *is conserved in expectation. For* $\gamma_{ij} = s_j/(s_i + s_j)$ *the conserved quantity is the* s*-weighted mean and the plain mean drifts toward large* s*; for* $\gamma_{ij} = g(s_i)$, $c_i \propto 1/g(s_i)$.

Theorem 2 (Consensus and lottery)**.** *For* $\gamma_{ij} \equiv \alpha \in (0, 1]$, *any* $m \geq 1$, $N \geq 2$*: almost surely all* x_i *converge to a common vertex, with* $\Pr[\text{vertex } k] = \bar{x}_{0,k}$, *seed variance* $\bar{x}_{0,k}(1 - \bar{x}_{0,k})$ *independent of* N, *and* $\mathbb{E}\sum_t (1 - M_2(t)) \leq mN^2/\alpha^2$.

Proposition 9 (When N-dependence appears)**.** *For asynchronous uniform pairing with one receiver updated per interaction and* α, m, K *independent of* N, *the relaxation time* $t_{\mathrm{rel}} := 1/\lambda_{\mathrm{eff}}$ *of the disagreement equation of Proposition 7 (a timescale of the moment dynamics, not a hitting time of a consensus threshold) scales with* N *through* $\lambda_{\mathrm{eff}} N^2 = \frac{2\alpha^2}{D_0}\big[1 - \frac{2\alpha(m-1)}{ND_0} + O(N^{-2})\big]$, $D_0 = 2m(1-\alpha) + \alpha$. *The consensus target (Thm. 2) and the conserved quantity (Prop. 8) do not depend on* N.

Size dependence as a prediction on the map (not a measured result). Let agent i carry a capability scale s_i, and let the listener step be pair-specific, γ_{ij} (i listens to j), with kernel dispersion m_i. Then the one-step pair (b, Σ) of Proposition 6 becomes size-dependent. Two consequences follow without further assumptions. *(P1) Homogeneous groups.* For $\gamma_{ij} \equiv \gamma(s)$ and $m_i \equiv m(s)$, the relaxation rate λ_{eff} of Proposition 7 is increasing in γ and decreasing in m. The denominator $2Nm(N - \gamma(N-1)) + \gamma(N-1)^2$ increases in m because $\gamma \leq 1 < N/(N-1)$, and $\partial_\gamma \lambda_{\mathrm{eff}} \propto \gamma(2A - B\gamma) > 0$ for the positive denominator $A - B\gamma$. So consensus speed is monotone in s if $\gamma(s)$ is nondecreasing and $m(s)$ is nonincreasing. *(P2) Mixed groups.* By Proposition 8, $b = [N^2(N-1)]^{-1}\sum_{i<j}(\gamma_{ij} - \gamma_{ji})(x_j - x_i)$. So symmetric steps give $b = 0$; asymmetry is necessary but not sufficient for $b \neq 0$ (at consensus $b = 0$ for any steps), and when $b \neq 0$ it points toward the states of the agents that move less. If m does not depend on size, a listener-only response $\gamma_{ij} = g(s_i)$ cannot produce both size-monotone homogeneous consensus and a large agent

that barely moves in a mixed group: the first then requires g increasing, which makes the largest listener the one that moves most. A size-dependent $m(s)$ removes this obstruction (P1). Observing both therefore predicts a speaker–listener dependent γ_{ij}, for example $\gamma_{ij} = g(s_i)\varphi(s_j/s_i)$ with $g(s_{\text{large}})\varphi(s_{\text{small}}/s_{\text{large}}) < g(s_{\text{small}})\varphi(s_{\text{large}}/s_{\text{small}})$. It also predicts a mean drift b toward the large agents, present in mixed groups and absent in homogeneous ones. *What is measured.* Neither available measurement measures (b, Σ): a transmission-kernel smoke run is size-dependent but not monotone in size, and R2 (Appendix L) lowers the mixed 1.5B/3B team on ARC-C (-0.053, $[-0.087, -0.020]$) but stores no belief vectors. (P1) and (P2) are predictions; testing them needs γ_{ij} per (listener, speaker) size pair, with speaker fixed and listener varied and the converse (Proposition 10, Appendix).

4 Experiments

This section tests the second contribution, and checks the other two numerically, by exact enumeration of a declared finite joint law.

4.1 Finite posterior evaluation

As the primary quantitative comparison, we extended the finite Bayes world with a shared common-evidence round and computed all four pooling/update-order combinations. The evaluation uses a grid of $K \in \{3, 5\}$ latent states, private-observation noise $q_{\text{priv}} \in \{0.6, 0.8, 0.95\}$, and common-evidence noise $q_{\text{com}} \in \{0.6, 0.8, 0.95\}$ (18 cells, symmetric weights $w_1 = w_2 = 0.5$). The prior is uniform; conditional on θ, x_1, x_2, x_L are independent categorical channels, with each channel's declared q denoting $P(x = \theta \mid \theta)$ and every incorrect category assigned $(1 - q)/(K - 1)$. For each cell, the experiment exactly sums the finite joint law $P(\theta, x_1, x_2, x_L)$ (27 observation tuples when $K = 3$ and 125 when $K = 5$; no Monte Carlo). The oracle $P(\theta \mid x_1, x_2, x_L)$ is built from that joint table, separately from the pooling operators, and no pooled output is used as the quality target. The CPU audit enforces the full grid and schemas, rejects missing, duplicated, altered, and nonfinite inputs, and keeps numerical residuals and quality comparisons under separate gates (Appendix E). The joint-law mass is normalized to within 2.5×10^{-15}; the e-pooling orders agree (Proposition 1) and the reweighting reconstruction of q^m_{PU} (Proposition 2) matches the direct computation, both to max 3.3×10^{-16}. The primary metric $\Delta LL_m := \mathbb{E}[\text{CE}(P(\theta \mid x_1, x_2, x_L), q^m_{\text{UP}})] - \mathbb{E}[\text{CE}(P(\theta \mid x_1, x_2, x_L), q^m_{\text{PU}})]$ is positive in all 18 cells (range 0.0023–0.0382 nats). Thus, within this specified finite observation model, update-then-pool has higher expected log loss than pool-then-update against the independently constructed posterior in every evaluated cell. As expected from Proposition 1, the corresponding e-pooling gap ΔLL_e is zero to machine precision in every cell (Figure 2; `code/oracle_independent.py`).

Target dependence of the preferred order. The E1 sign is a property of that declared target. Over 2000 random three-state configurations (two agents, equal weights, likelihoods $\sim U(0.2, 3)^3$) and 500 Dirichlet targets each, the sign of $\text{CE}(t, q^m_{\text{PU}}) - \text{CE}(t, q^m_{\text{UP}})$ flips with the target in 99.95% of configurations and in a majority of every gap quintile (Figure 3; Spearman $\rho = -0.17$ between gap and fraction of targets favouring PU).

The zero set is where the theory says it is. Corollary 1 predicts that the order gap vanishes exactly when the agents' normalizers Z_i coincide, not when their priors coincide. Figure 7 (Appendix E) tests this on a non-consensus control: priors differ in both arms, but in the control arm the Z_i are equal. The control gap is exactly zero (rational arithmetic) in 18/18 cells, while the treatment arm (Z_i distinct) has $\chi > 0$ growing as t^2; under e-pooling the gap is zero to machine precision ($|\chi| \le 2.5 \times 10^{-32}$) across the whole heterogeneity sweep. A wider exact-rational audit (94 equal-Z_i and 188 unequal-Z_i cells) agrees (Appendix E).

Matched-budget baselines. Table 1 holds the information budget fixed: every non-oracle rule receives the same private and common observations and one exchange round, and all losses are exact excess log loss over the joint-law oracle. Communication helps: every pooling rule beats the no-communication agent in 18/18 E1 cells. By median loss, arithmetic pool-then-update is best when sources are independent or partly redundant, e-pooling is best when sources are complementary (three agents observing a, b, $a{+}b \bmod k$ of $\theta = (a, b)$), and all rules coincide when the private

K	q_p	q_c	Oracle CE	UP CE	PU CE	Δ CE
3	0.6	0.6	0.720639	0.770678	0.767939	0.002739
3	0.6	0.8	0.496731	0.532591	0.527877	0.004714
3	0.6	0.95	0.191546	0.203682	0.201428	0.002254
3	0.8	0.6	0.352217	0.445097	0.434906	0.010191
3	0.8	0.8	0.249611	0.335459	0.315328	0.020131
3	0.8	0.95	0.106791	0.147938	0.136433	0.011504
3	0.95	0.6	0.071906	0.119661	0.112097	0.007563
3	0.95	0.8	0.049949	0.107022	0.087802	0.019220
3	0.95	0.95	0.021632	0.068427	0.049939	0.018489
5	0.6	0.6	0.736005	0.836869	0.820626	0.016242
5	0.6	0.8	0.484781	0.560204	0.541313	0.018892
5	0.6	0.95	0.183179	0.209873	0.202151	0.007722
5	0.8	0.6	0.327153	0.457494	0.429012	0.028482
5	0.8	0.8	0.219462	0.339524	0.301315	0.038210
5	0.8	0.95	0.090874	0.148796	0.130079	0.018717
5	0.95	0.6	0.062295	0.120136	0.104674	0.015462
5	0.95	0.8	0.040316	0.106965	0.080236	0.026729
5	0.95	0.95	0.016469	0.068356	0.046636	0.021721

Figure 2: **Independent task-posterior losses in every E1 cell.** All quantities are exact finite expectations, re-derived from the declared joint law. CE is cross-entropy in nats; UP and PU are the two arithmetic orderings. The independent oracle CE is shown alongside both methods. The positive UP-minus-PU contrast holds on this specified grid; it is not a universal quality ordering or a native-model result.

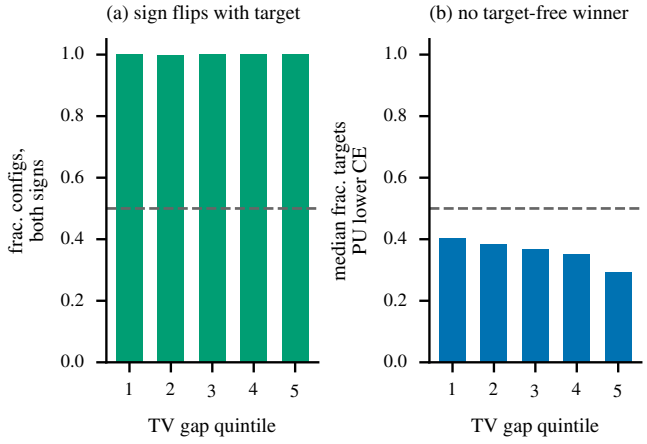


Figure 3: **The order gap does not choose an order; the target does.** Configurations are binned by the total-variation gap $\frac{1}{2}\|q^m_{\mathrm{PU}} - q^m_{\mathrm{UP}}\|_1$. (a) Fraction of configurations whose preferred order flips across targets. (b) Median fraction of targets where PU has lower CE. Declared family only.

sources are identical ($\rho = 1$). The ordering of rules is a property of the information structure, not of the operators alone.

Further checks. Appendix E reports a rejected held-out receiver-gain predictor, coefficient audits, and the real-LLM EX-1 gain-sign test ($9/64$, *disagrees*); Appendix H shows native-LLM measurements as exploratory figures only.

Table 1: **Matched-budget baselines: median excess log loss over the exact oracle (nats; lower is better).** Every non-oracle rule receives the same observations and one exchange round. E1: $K \in \{3, 5\}$, 18 cells; copy channel: agent 2 copies agent 1 with probability ρ ($K \in \{3, 5\}$, $q = 0.8$, 2 cells); E-IS: 3 agents, 18 cells per structure. Bold: best pooling rule (none when tied). The independence product is exact under E1's conditionally independent channels by construction; –: not computed for E-IS. Exact finite sums; generated by `code/rootfix/table1_matched_budget.py`. Not a native-LLM result.

Condition	no comm.	e-pool	m UP	m PU	indep. product
E1 (2 agents, independent)	0.121	0.046	0.057	**0.040**	0.000
E1 copy channel ρ=0.5	0.063	0.017	0.026	**0.011**	0.095
E1 copy channel ρ=1.0	0.000	0.000	0.000	0.000	0.241
E-IS independent (3 agents)	0.199	0.088	0.105	**0.083**	–
E-IS redundant (3 agents)	0.146	0.061	0.071	**0.047**	–
E-IS complementary (3 agents)	0.162	**0.109**	0.120	0.135	–

5 Related Work

External Bayesianity and the arithmetic/logarithmic pool distinction are classical (Genest, 1984; Genest & Zidek, 1986); we claim no priority for these identities, and none of them alone ranks pooling orders by task accuracy. Proposition 5 conditionally extends the probe vectors of Oyama et al. (2025) (with the m/e terminology of Amari (2016)) to interacting states, and the rule of Tanaka (2026) is the sampling-kernel instance of our operator (Theorem 1(e)). Order effects in belief revision are documented empirically (Hogarth & Einhorn, 1992) and non-Bayesian social learning studies pooling without a normalizer (Jadbabaie et al., 2012); our contribution is to locate the order gap in the evidence normalizer rather than in the agents. Details are in Appendix F.

6 Discussion, Limitations, and Conclusion

Pooling and updating commute unless arithmetic pooling meets unequal evidence normalizers. Even then the gap does not choose an order: update-then-pool loses in all 18 cells of our finite target, yet the preferred order flips across targets (Figure 3), so neither order is a universal recommendation. A preregistered transfer test on Qwen2.5-3B/7B receivers failed its first gate, so native identification and the size predictions remain open (Appendices B and D).

Limitations. Three caveats bound the claims. First, every headline number is an exact finite sum on a synthetic target: the m/e non-commutativity is a property of the pooling rules, not of any language model, and the native-LLM measurements in Appendix H are exploratory. Second, the ranking in Table 1 is conditional on the channel model; the independence product is exact under E1's conditionally independent channels by construction and degrades sharply under the copy channel, so Table 1 should be read as a sensitivity analysis rather than a recommendation. Third, the real-LLM gain-sign test disagrees with the finite-target prediction in $9/64$ cells and the held-out receiver-gain predictor was rejected (Appendix E); we therefore report the sign flip across targets (Figure 3) as the robust finding and leave the magnitude predictions as preregistered but unconfirmed. Finally, all experiments use two- and three-agent networks with a single exchange round; whether the normalizer gap compounds or washes out over repeated rounds and larger populations is not addressed here.

Reproducibility statement

The CPU audits in `code/aris_fix/` read the versioned inputs and write machine-readable results and the included LaTeX fragments to `results/aris_fix/`. Each published audit quantity is mapped to its implementation and output in `MEASUREMENT_MAP.md`. Identity residuals, posterior losses, and benchmark contrasts have distinct measurement contracts. No model inference is required for these audits.

REFERENCES

Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530, 1985. doi: 10.2307/2274239.

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998. doi: 10.1162/089976698300017746.

Shun-ichi Amari. Integration of stochastic models by minimizing α-divergence. *Neural Computation*, 19(10):2780–2796, 2007. doi: 10.1162/neco.2007.19.10.2780.

Shun-ichi Amari. *Information Geometry and Its Applications*. Springer, 2016. doi: 10.1007/978-4-431-55978-8.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, 2000. doi: 10.1090/mmono/191.

Robert J. Aumann. Agreeing to disagree. *The Annals of Statistics*, 4(6), 1976. doi: 10.1214/aos/1176343654.

Nihat Ay, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer. *Information Geometry*. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. Springer, 2017. doi: 10.1007/978-3-319-56478-4.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. ChatEval: Towards better LLM-based evaluators through multi-agent debate. In *International Conference on Learning Representations*, 2024.

Hyeong Kyu Choi, Xiaojin Zhu, and Yixuan Li. Debate or vote: Which yields better decisions in multi-agent large language models? In *Advances in Neural Information Processing Systems*, 2025.

Robert T. Clemen and Robert L. Winkler. Combining probability distributions from experts in risk analysis. *Risk Analysis*, 19(2):187–203, 1999. doi: 10.1111/j.1539-6924.1999.tb00399.x.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1), 1975. doi: 10.1214/aop/1176996454.

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

Franz Dietrich and Christian List. Probabilistic opinion pooling. In *The Oxford Handbook of Probability and Philosophy*. Oxford University Press, 2017. doi: 10.1093/oxfordhb/9780199607617.013.37.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

B. El, J. Paeng, F. Dinc, S. Su, M. Erdogan, A. Pappu, H. Ye, W. Zhao, S. Ganguli, and J. Zou. Physics of agents: Statistical mechanics predicts collective behavior of AI agents. *arXiv preprint arXiv:2608.16578*, 2026.

Andrew Estornell and Yang Liu. Multi-LLM debate: Framework, principals, and interventions. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

M. Fukushima. Message capacity and claim wording set the transition points of collective truth-finding in language-model networks. *arXiv preprint arXiv:2609.19183*, 2026.

John D. Geanakoplos and Heraklis M. Polemarchakis. We can't disagree forever. *Journal of Economic Theory*, 28(1):192–200, 1982. doi: 10.1016/0022-0531(82)90099-0.

Christian Genest. A characterization theorem for externally bayesian groups. *The Annals of Statistics*, 12(3):1100–1105, 1984. doi: 10.1214/aos/1176346726.

Christian Genest and James V. Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–135, 1986. doi: 10.1214/ss/1177013825.

Benjamin Golub and Matthew O. Jackson. Naïve learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–149, 2010. doi: 10.1257/mic.2.1.112.

Tom Heskes. Selecting weighting factors in logarithmic opinion pools. In *Advances in Neural Information Processing Systems*, volume 10, pp. 266–272, 1997.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002. doi: 10.1162/089976602760128018.

Robin M. Hogarth and Hillel J. Einhorn. Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24(1):1–55, 1992. doi: 10.1016/0010-0285(92)90002-J.

Ali Jadbabaie, Pooya Molavi, Alvaro Sandroni, and Alireza Tahbaz-Salehi. Non-Bayesian social learning. *Games and Economic Behavior*, 76(1):210–225, 2012. doi: 10.1016/j.geb.2012.06.001.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive LLMs leads to more truthful answers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23662–23733, 2024.

Günther Koliander, Yousef El-Laham, Petar M. Djurić, and Franz Hlawatsch. Fusion of probability density functions. *Proceedings of the IEEE*, 110(4):404–453, 2022. doi: 10.1109/JPROC.2022.3154399.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17889–17904, 2024.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, 2022.

Frank Nielsen. An elementary introduction to information geometry. *Entropy*, 22(10):1100, 2020. doi: 10.3390/e22101100.

Momose Oyama, Hiroaki Yamagiwa, Yusuke Takase, and Hidetoshi Shimodaira. Mapping 1,000+ language models via the log-likelihood vector. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 32983–33038, 2025. doi: 10.18653/v1/2025.acl-long.1584.

E. Pavlova and H. Tanaka. Flag game: A toy model for mechanistic swarm interpretability. *arXiv preprint arXiv:2609.19124*, 2026.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017, 2024.

Ville A. Satopää, Jonathan Baron, Dean P. Foster, Barbara A. Mellers, Philip E. Tetlock, and Lyle H. Ungar. Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, 30(2):344–356, 2014. doi: 10.1016/j.ijforecast.2013.09.009.

X. Shi, Y. Zhang, and D. Yang. Emergent collusion in long-horizon LLM agent interaction. *arXiv preprint arXiv:2609.24967*, 2026.

Andries Petrus Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D. Barrett, and Arnu Pretorius. Should we be going MAD? A look at multi-agent debate strategies for LLMs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45883–45905, 2024.

M. Stone. The opinion pool. *The Annals of Mathematical Statistics*, 32(4):1339–1342, 1961. doi: 10.1214/aoms/1177704873.

H. Tanaka. When is collective intelligence a lottery? multi-agent scaling laws for memetic drift in llms. *arXiv preprint arXiv:2603.24676*, 2026.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

A CLAIM MAP

Table 2: **Theorem–measurement–use correspondence.** Each row is one contribution; every entry is a finite, auditable statement under the listed scope. None is a native-LLM law.

Claim	Formal result	Measurement	Use (diagnostic rule)	Scope
Operator theory on the map	Prop. 5, Thm. 1; Props. 1–2, Cor. 1, Props. 3, 4	E1 machine-precision residuals, exact zero-set and coefficient audits (Appendix E), Fig. 1	If normalizers Z_i coincide, the order of arithmetic pooling and update is immaterial; otherwise weights move as $w_i Z_i / \bar{Z}$, and near consensus the gap is zero or quartic	Finite $\mathcal{Y}$, positive beliefs, common likelihood
Population statistics	Props. 7, 8, 9, 10, Thm. 2	Exact moment enumeration and consensus-frequency checks (Section 3.2)	Predict consensus point and time from (N, α, m); read size effects through the listener response	Specified sampling kernel; U8 conditional on a monotone response
Measurement	Prop. 5; declared joint law (Sec. 2)	Native state map (Fig. 5); E1 18-cell exact sums (Fig. 2)	Choose the ordering by loss against an external target, not by the size of the gap	Descriptive map; specified observation model and grid

B OPEN QUESTIONS

Predicting native communication gains requires specifying the task law, message channel, receiver, baseline information, and evaluation utility. The accounting identity alone does not determine the gain's sign from the task distribution.

A redundant message adds no conditional information but may improve prediction by reducing receiver error. A partial-information analysis would additionally require an explicit redundancy functional and source definition. Any communication cost must be defined separately from log-score gain.

Native population prediction would require independently identified operators, a frozen transfer protocol, and evaluation on held-out worlds and communication structures. Neither such transfer nor an additive decomposition of capacity effects and information advantages is established here.

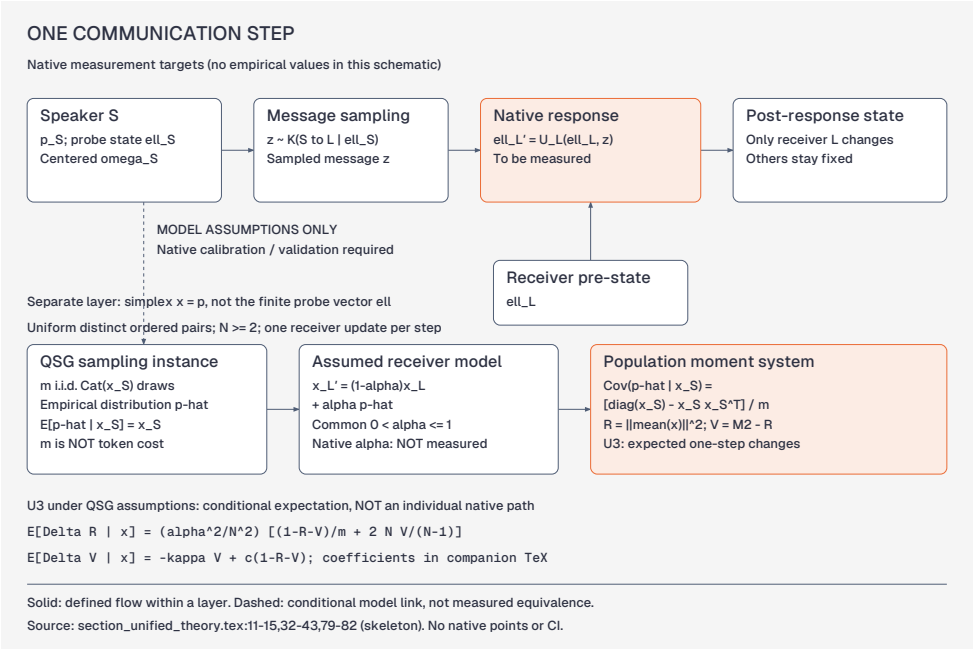


Figure 4: **One communication step and its conditional population-moment model (schematic; no empirical values).** Upper path (native measurement targets): a speaker S with probe state ℓ_S emits a sampled message $z \sim K_{S\to L}(\cdot \mid \ell_S)$ (written m in Section 3.2; renamed here to avoid a clash with the sample count m), the native receiver responds through U_L, and all other agents are unchanged. None of these quantities is measured in this paper. Lower path (model assumptions only): the QSG instance draws m i.i.d. categorical samples from x_S, forms $\hat{p}$, and applies $x'_L = (1-\alpha)x_L + \alpha\hat{p}$ with a common $0 < \alpha \leq 1$; under uniform distinct ordered pairs the sampling covariance and the population second moments give the expected one-step changes of R and V in Proposition 7. The dashed link is a conditional model correspondence, not a measured equivalence: the native α, the native sampling second moment, and the adequacy of the convex receiver model are unmeasured, and the retained finite-construction checks (e.g. the E44 accounting identity) do not establish native validity. Source: `figures/fig_gap06_one_step_schematic_manifest.json`.

C PROOFS FOR THE UNIFIED THEORY

Proof of Proposition 5. (a) is Lemma 1. (b): (a) in both directions gives $d_{ij}/M = \|\omega_i - \omega_j\|^2/M = 2K_{ij} + O(\varepsilon^3) + O_P(M^{-1/2})$; double centering, $-\frac{1}{2}J(d/M)J$, cancels the factor 2. The number of compared states is fixed. □

Proof of Theorem 1. (a)–(d) are the cited results, proved in Section 3 and unchanged; (e) is substitution of $K = m$-sample empirical law, $U_L = (1-\alpha)p_L + \alpha\hat{p}$. □

Proof of Proposition 6. (i),(ii): substitute $K^{(m)}$ and $U^{(\alpha)}$ into $\mathcal{T} = U \circ K$; interiority holds because $(1-\alpha)x_L$ is interior. (iii): given the pair, $\mathbb{E}[\Delta\bar{x} \mid S, L] = \frac{\alpha}{N}(x_S - x_L)$ and $\mathrm{Cov}\,\hat{p} = (\mathrm{diag}\,x_S - x_S x_S^\top)/m$. Averaging over the $N(N-1)$ ordered pairs, the first moment vanishes, and $\sum_{S\neq L}(x_S - x_L)(x_S - x_L)^\top = 2N^2\Sigma_{\mathrm{pop}}$. Also $\mathbb{E}\|\bar{x}'\|^2 - \|\bar{x}\|^2 = 2\bar{x}^\top b + \mathrm{tr}\,\Sigma$. (iv): Theorem 2 and Proposition 9 use only $b = 0$ and $\mathrm{tr}\,\Sigma$. □

Throughout, one interaction draws an ordered pair (L, S), $L \neq S$, uniformly from the $N(N-1)$ ordered pairs, and updates only the listener, $x'_L = x_L + \gamma_{LS}(\hat{p} - x_L)$ with $\hat{p}$ the empirical law of m i.i.d. draws from $\mathrm{Cat}(x_S)$; the homogeneous case is $\gamma \equiv \alpha \in (0, 1]$. Expectations are conditional on the current state $x = (x_1, \ldots, x_N)$. We use $\mathbb{E}[\hat{p} \mid S] = x_S$, $\mathrm{Cov}(\hat{p} \mid S) = (\mathrm{diag}\,x_S - x_S x_S^\top)/m$,

hence $\mathbb{E}[\|\hat{p} - x_S\|^2 \mid S] = (1 - \|x_S\|^2)/m$, and, averaging over distinct ordered pairs,

$$\mathbb{E}[x_S \cdot x_L] = \frac{NR - M_2}{N-1}, \quad \mathbb{E}\|x_S - x_L\|^2 = \frac{2NV}{N-1}, \quad \mathbb{E}[(x_S - x_L)(x_S - x_L)^\top] = \frac{2N\Sigma_{\text{pop}}}{N-1}. \tag{1}$$

The first identity is $\sum_{i \neq j} x_i \cdot x_j = N^2 R - N M_2$; the second follows from $2M_2 - 2\mathbb{E}[x_S \cdot x_L]$; the third is its matrix form.

Proof of Proposition 7. With $\gamma \equiv \alpha$, $\Delta\bar{x} = (\alpha/N)(\hat{p} - x_L)$. Since $\mathbb{E}[x_S - x_L] = 0$ under uniform ordered pairs, $\mathbb{E}\Delta\bar{x} = 0$ and $\mathbb{E}\Delta R = \mathbb{E}[2\bar{x} \cdot \Delta\bar{x} + \|\Delta\bar{x}\|^2] = (\alpha/N)^2 \mathbb{E}\|\hat{p} - x_L\|^2$. Splitting $\hat{p} - x_L = (\hat{p} - x_S) + (x_S - x_L)$, the cross term vanishes by $\mathbb{E}[\hat{p} - x_S \mid S, L] = 0$, so $\mathbb{E}\Delta R = (\alpha/N)^2[(1 - M_2)/m + 2NV/(N-1)]$ with $1 - M_2 = 1 - R - V$. For V, $\Delta M_2 = N^{-1}(\|x_L'\|^2 - \|x_L\|^2) = N^{-1}[2\alpha\, x_L \cdot (\hat{p} - x_L) + \alpha^2 \|\hat{p} - x_L\|^2]$; by equation 1, $\mathbb{E}[x_L \cdot (x_S - x_L)] = (NR - M_2)/(N-1) - M_2 = -NV/(N-1)$. Hence $\mathbb{E}\Delta M_2 = -2\alpha V/(N-1) + (\alpha^2/N)B$ with $B = (1 - R - V)/m + 2NV/(N-1)$, and $\mathbb{E}\Delta V = \mathbb{E}\Delta M_2 - \mathbb{E}\Delta R = -2\alpha V/(N-1) + \alpha^2(N-1)N^{-2}B = -\kappa V + c(1 - R - V)$ with κ, c as stated. No closure is used. Setting $\mathbb{E}\Delta V = 0$ gives $V^* = c(1-R)/(\kappa + c)$; substituting into $\mathbb{E}\Delta R$ yields $\mathbb{E}\Delta R = \lambda_{\text{eff}}(1 - R)$ with the stated λ_{eff} (checked exactly on 324 configurations, `results/r119_u3_exact_moments`). This last step is an adiabatic approximation: it is accurate when V relaxes much faster than R, and $(\kappa + c)/\lambda_{\text{eff}} = O(N)$; we do not state a uniform error bound for it. The matrix form, used in Proposition 6, is $\Sigma = (\alpha/N)^2 \mathbb{E}[(\hat{p} - x_L)(\hat{p} - x_L)^\top] = (\alpha/N)^2[\mathbb{E}\,\mathrm{Cov}(\hat{p} \mid S) + \frac{2N}{N-1}\Sigma_{\text{pop}}]$ with $\mathbb{E}\,\mathrm{Cov}(\hat{p} \mid S) = m^{-1}(\mathrm{diag}\,\bar{x} - N^{-1}\sum_i x_i x_i^\top)$. □

Separation from the mean-field closure (pre-fixed, $\delta \propto V$). At m=1, the mean-field closure $M_2 \approx R$ gives $\mu_{\text{MF}} = \frac{\alpha^2}{N^2}(1 - R)$, whereas Prop. 7 gives $\delta = \mu_{U3} - \mu_{\text{MF}} = \frac{\alpha^2}{N^2}\frac{N+1}{N-1}V$, so the two coincide iff $V = 0$. For K=2, N=4, α=1/2 at matched R=1/2: $\delta = 5/1536$ for $V = 1/8$ and $\delta = 0$ for $V = 0$ (exact enumeration of all 24 outcomes).

Proof of Proposition 8. $\mathbb{E}\Delta\bar{x} = N^{-1}[N(N-1)]^{-1}\sum_{i \neq j}\gamma_{ij}(x_j - x_i)$ with listener i and speaker j, using $\mathbb{E}\hat{p} = x_j$. Grouping (i, j) with (j, i) gives $(\gamma_{ij} - \gamma_{ji})(x_j - x_i)$. The same grouping for $\sum_i c_i x_i$ gives $[N(N-1)]^{-1}\sum_{i<j}(c_i\gamma_{ij} - c_j\gamma_{ji})(x_j - x_i)$, which vanishes when $c_i\gamma_{ij} = c_j\gamma_{ji}$. For $\gamma_{ij} = s_j/(s_i + s_j)$ the balance condition holds with $c_i \propto s_i$, and each pair term of $\mathbb{E}\Delta\bar{x}$ is $\frac{s_j - s_i}{s_i + s_j}(x_j - x_i)$, pointing from the smaller-s agent to the larger-s agent. For $\gamma_{ij} = g(s_i)$ it holds with $c_i \propto 1/g(s_i)$. The condition is sufficient, not necessary. □

Proof of Theorem 2. By Proposition 7, $\bar{x}_t$ is a martingale with values in the simplex, hence bounded; it converges almost surely and in L^1. Since $R \leq 1$, $\sum_t \mathbb{E}\Delta R_t \leq 1 - R_0 \leq 1$, and $\mathbb{E}\Delta R \geq (\alpha/N)^2[(1 - M_2)/m + 2NV/(N-1)]$ (equality in Prop. 7) gives $\mathbb{E}\sum_t(1 - M_2(t)) \leq mN^2/\alpha^2$ and $\mathbb{E}\sum_t V_t < \infty$. Therefore $1 - M_2(t) \to 0$ and $V_t \to 0$ almost surely. $\|x_i\|^2 \leq 1$ with equality only at vertices, so $M_2 \to 1$ forces every x_i to approach a vertex, and $V \to 0$ forces all x_i to approach the same vertex; since $\bar{x}_t$ converges, the vertex e_{k^*} is eventually fixed. Dominated convergence gives $\mathbb{E}\bar{x}_\infty = \bar{x}_0$, i.e. $\Pr[k^* = k] = \bar{x}_{0,k}$; the indicator of vertex k has variance $\bar{x}_{0,k}(1 - \bar{x}_{0,k})$, which contains no N for a fixed $\bar{x}_0$. □

Heterogeneous γ. Let $\gamma_{ij} \in (0, 1]$ satisfy $c_i\gamma_{ij} = c_j\gamma_{ji}$ with $c_i > 0$, and put $\tilde{c}_i = c_i/\sum_j c_j$, $y = \sum_i \tilde{c}_i x_i$. By Proposition 8, y is a bounded martingale. $\Delta y = \tilde{c}_L\gamma_{LS}(\hat{p} - x_L)$, so with $\beta = \min_{i \neq j}\tilde{c}_i\gamma_{ij} > 0$, $\mathbb{E}\|\Delta y\|^2 \geq \beta^2[(1 - M_2)/m + 2NV/(N-1)]$ by the same decomposition, and $\|y\|^2 \leq 1$. The argument of Theorem 2 then gives almost-sure consensus at a vertex with $\Pr[\text{vertex } k] = y_{0,k}$; for $\gamma_{ij} = s_j/(s_i + s_j)$ this is the s-weighted initial mean. The simulation in `results/r119_u5_consensus_lottery` ($s = (1, 1, 6, 6)$, 3000 runs) matches the s-weighted law ($|z| \leq 1.4$) and rejects the plain mean ($|z| > 16$). □

Proof of Proposition 9. Write $\lambda_{\text{eff}}N^2 = 2\alpha^2/D(N)$ with $D(N) = N^{-2}[2Nm(N - \alpha(N-1)) + \alpha(N-1)^2] = D_0 + N^{-1}\,2\alpha(m-1) + O(N^{-2})$, where the N^{-1} coefficient is $2m\alpha - 2\alpha = 2\alpha(m - 1)$. Then $1/D(N) = D_0^{-1}[1 - 2\alpha(m-1)/(ND_0) + O(N^{-2})]$. The consensus target of Theorem 2 and the balance weights c_i of Proposition 8 contain no N. The residual of the first-order expansion is checked to be $O(N^{-2})$ for N up to 10^4 (`results/r119_u3_exact_moments`). □

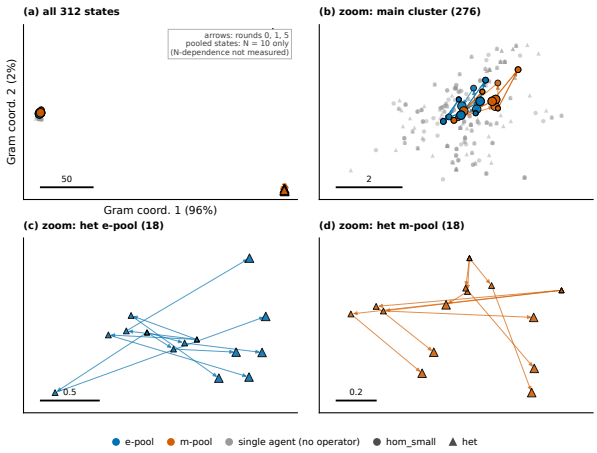


Figure 5: **A descriptive map of native predictive states in shared log-likelihood coordinates.** All 312 stored Qwen2.5 states (240 single-agent, 72 pooled team states, $N = 10$) are rows of one state-by-probe log-likelihood matrix on 200 probes; axes are the first two Gram coordinates (96%, 2% of the positive spectrum) of Proposition 5. Colour: operator (e-pool, m-pool; grey: none); shape: team family; size and arrows: rounds $0 \to 1 \to 5$. (a) All states; (b)–(d) zooms with independent scales. Descriptive only: the vertical axis is not E1 loss, no common geometry or native gain law is identified, and N-dependence is not measured.

Proposition 10 (Conditional predictions)**.** *(a) With kernel dispersion m independent of size, consensus speed monotone in an agent capability scale requires a listener response $\gamma = g(s)$ monotone in s; in general, γ nondecreasing and m nonincreasing in s is sufficient but not necessary (then $\partial\lambda_{\text{eff}}/\partial\alpha > 0$); the ratio rule $\gamma_{ij} = s_j/(s_i + s_j)$ alone gives $\gamma = 1/2$ for homogeneous groups and predicts no such dependence. (b) Agents with large s move less as listeners, and the population's expected consensus point is the s-weighted mean (Prop. 8); speaker-side and listener-side factors merge into γ_{ij} and need the controls speaker-fixed/listener-varied and the converse.*

Proof of Proposition 10. (a) In a homogeneous group $\gamma \equiv g(s)$, so λ_{eff} depends on s only through $\alpha = g(s)$. In $\lambda_{\text{eff}} = 2\alpha^2/D$, the numerator increases in α and $\partial D/\partial\alpha = (N-1)(N-1-2Nm) < 0$, so $\partial\lambda_{\text{eff}}/\partial\alpha > 0$; consensus speed is therefore monotone in s exactly when g is. Under the ratio rule $\gamma_{ij} = s_j/(s_i + s_j) = 1/2$ for equal s, independent of s. (b) Under the ratio rule a listener with larger s_i has smaller γ_{ij}; the expected consensus point is the s-weighted mean by the heterogeneous extension above. Because γ_{ij} can depend on both indices, a speaker-side and a listener-side size effect are not separated by γ_{ij} alone, which is why the two controls are needed. □

Entropy and diversity (second-order approximation).

Proposition 11 (Entropy and diversity)**.** *Let H be Shannon entropy, $D = H(\bar{x}) - N^{-1}\sum H(x_i) \geq 0$ (mean JS divergence). For $\bar{x}$ in the interior of the simplex, to second order in $\Delta\bar{x}$, $\mathbb{E}\Delta H(\bar{x}) \approx -\frac{1}{2}(\alpha/N)^2 \sum_k \left[(\bar{x}_k - (M_2^{\text{mat}})_{kk})/m + \frac{2N}{N-1}(\Sigma_{\text{pop}})_{kk}\right]/\bar{x}_k$, with $M_2^{\text{mat}} = N^{-1}\sum_i x_i x_i^\top$.*

Proof. $\mathbb{E}\Delta\bar{x} = 0$, so the first-order term vanishes; the Hessian of H is $-\text{diag}(1/\bar{x}_k)$ and $\mathbb{E}[\Delta\bar{x}\Delta\bar{x}^\top]$ is the matrix of Proposition 6. □

Scope. This is an approximation, not an exact identity: the third-order remainder is not bounded, the expansion fails as $\bar{x}_k \to 0$ (the Hessian diverges), no closed evolution of $N^{-1}\sum_i H(x_i)$ is derived, and no main-text claim uses it.

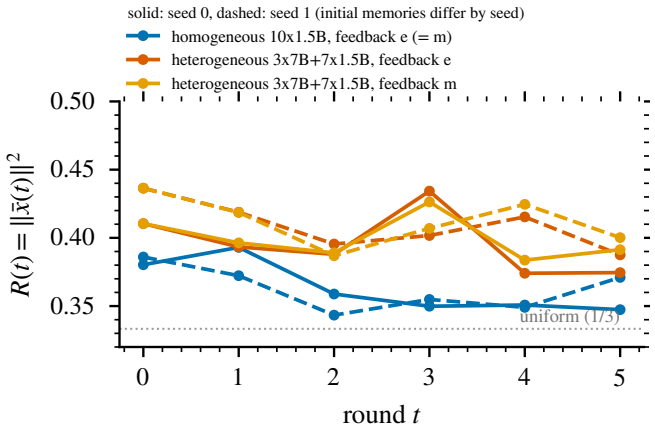


Figure 6: **Native multi-agent trajectories of $R(t) = \|\bar{x}(t)\|^2$ (exploratory; not evidence for the finite-model claims).** Naming game with $N = 10$ agents, $K = 3$, $T = 5$ rounds, two seeds per configuration; R ranges 0.343–0.436 (uniform: 0.333). Descriptive only: transfer gate not passed.

D NATIVE ONE-STEP TRANSFER TEST

D.1 PREREGISTERED MOMENT TRANSFER FROM NATIVE ONE-STEP RESPONSES

We evaluate a distinct estimand from pooling and gain-sign prediction: whether a one-step receiver response identified on initial-history worlds predicts $E[R_t]$ and $E[V_t]$ in unseen asynchronous populations without test refitting. A message is sampled from the speaker's measured three-label candidate distribution, appended to exactly one receiver's history, and the receiver's distribution is remeasured. The protocol does not overwrite the native response with the QSG equation. Two models (Qwen2.5-3B and 7B) are reported separately; 24 identification worlds per model are disjoint from 48 test worlds, each with two schedule seeds, $N \in \{4, 8\}$, $m \in \{1, 4\}$, and 64 interactions. The test plan comprises 49,152 interactions across both models if both pass the first gate. A failed first gate stops that model's native transfer rather than triggering prompt search. Gate FAILED: moment transfer did not hold in this experimental setting. Both models failed the one-step gate on validation worlds, so native transfer stopped, no held-out native trajectory was run, and we make no native moment-transfer claim. The planned unseen-(N, m) prediction-versus-measurement figure (U3 solid, mean-field dashed, native points with 95% CIs) is therefore omitted: no run exists with frozen predictions on an unseen (N, m) split (U3 vs. mean-field) together with matching native points and 95% CIs. We do not substitute synthetic points.

Table 3 (Appendix G) lists the four preregistered gates: gate 1 failed for both models (3B: mean error 47.2%, second moment 5.4%; 7B: 30.7% and 45.8% of the validation signal, against a 10% threshold), so gates 2–4 were not evaluated.

The exact QSG sampler is a CPU reference, not native evidence. Across the prespecified $K = 3$, $N \in \{4, 8, 16\}$, $\alpha \in \{.2, .4, .8\}$, $m \in \{1, 4\}$ and low/high-diversity 36-cell grid (2,048 independent trajectories per cell, 64 asynchronous interactions), the mean trajectory RMSE for (R, V) is .00142 for the full U3 recurrence, .07535 for the adiabatic approximation, .05404 for the specified mean-field comparator, and .13059 for fixed/no-change. These numbers establish numerical behavior of the specified QSG only (Figure 11, Appendix G). They do not validate its transfer to either LLM.

E SUPPLEMENTARY EXPERIMENTAL CHECKS

This appendix collects the observation-model quality checks, the zero-set and coefficient audits, the E-B sensitivity audit and the EX-1 gain-sign test summarized in Section 4.

Separately specified observation-model quality checks. Figure 8 evaluates finite-state classification quality against an exact joint-law posterior, not against either pooling output. Figure 9

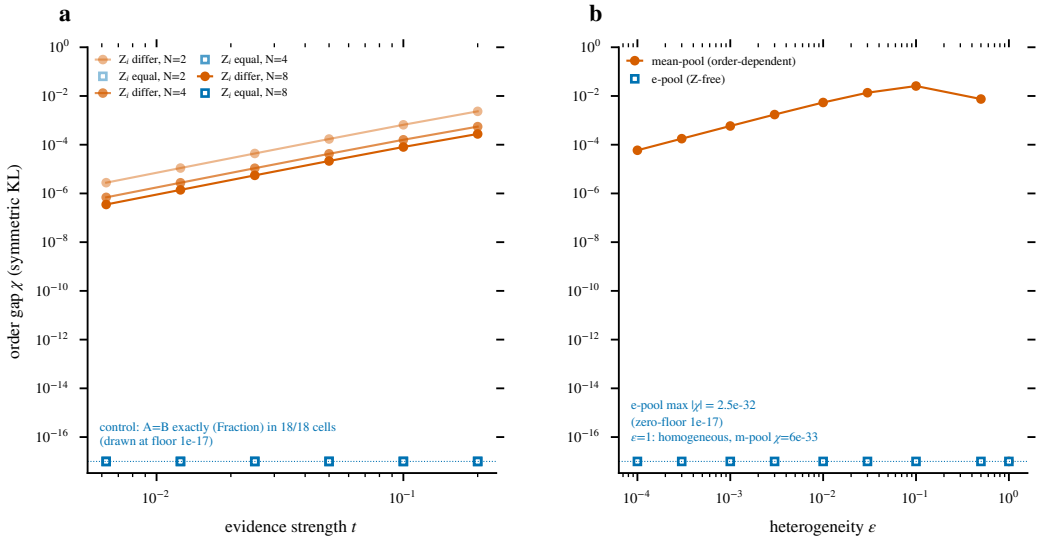


Figure 7: **Non-consensus zero control.** (a) Exact finite grid (K=3, $N \in \{2, 4, 8\}$, $t \in [1/160, 1/5]$): arms with distinct Z_i have order gap $\chi > 0$ scaling as t^2; the control (distinct priors, equal Z_i) has update-then-pool = pool-then-update exactly in 18/18 cells (drawn at a floor of 10^{-17}). (b) Heterogeneity sweep: mean-pooling is order dependent for $\varepsilon < 1$; e-pooling is Z-free. Exact finite sums, CPU only; generated by `code/make_figures.py --figure rootfix_zero_control`. Not a native-LLM result.

evaluates a frozen receiver-gain predictor on held-out observation cells with changed prompts and labels. These are distinct from Stage C belief-shift and capacity diagnostics, which do not test classification quality. Neither experiment rescues the withdrawn Theorem 3m claim or establishes a general native-communication law.

Zero sets and coefficient asymptotics. The CPU audit re-derives the global bounds and inverse-likelihood certificate on every stored finite configuration, including the reallocation residual in the gate. It separates a distinct-prior, nonconstant-likelihood equal-normalizer control from consensus and constant-likelihood controls. In the original degenerate sweep, all agents had identical priors; that sweep alone did not test the non-consensus zero set.

For the declared linear family, the audit computes H and $\mathbb{E}_{q_0} H^2$ directly and compares χ/t^4 with that coefficient. For the verifier family it computes κ from Proposition 18 and compares χ/ε with κ. Figure 10 reports ratios on the original grids. Decimal arithmetic independently recomputes the gaps and zero controls; positivity is checked before evaluation, and no clipping or display floor enters the reported values. Equal-target-mass priors supply the verifier zero control. The configurations, coefficients and per-point numerical differences are preserved in `results/aris_fix/r90_R2-E2.json`.

Two further CPU audits test these statements from saved inputs. The equal-Z audit (`code/aris_fix/r124_e2_equalz_direct.py`) uses exact rationals, $K = 3$, four nonconstant likelihoods, two normalizer targets, $N \in \{2, 3\}$ and two weight vectors. It computes $\Pi_E \Pi_C^m p$ by normalizing the pooled prior directly, which is a separate path from the reweighting of Proposition 2. The two paths agree in all 282 cells. All 94 equal-Z_i cells (pairwise distinct priors) have a zero gap, and all 188 unequal-Z_i twins have a nonzero gap. Replacing $\tilde{w}$ by w breaks the identity in all 188 unequal-Z_i cells, and the input validator rejects six classes of corrupted input. The chibound audit (`code/aris_fix/r124_e1_chibound_cleanroom.py`) recomputes Propositions 3 and 4 at 60 decimal digits. It reads the saved primitive inputs back from disk, and stored χ values and pass flags are not used as ground truth. Over 20 configurations it finds 0 sandwich and 0 Jensen violations. An equal-Z_i configuration with distinct priors gives $\chi = 0$. Both χ and the Jensen certificate are invariant under $L \to \alpha L$. In the degenerate quartic family, where all $m_i = \mathbb{E}_{q_0}[a_i]$

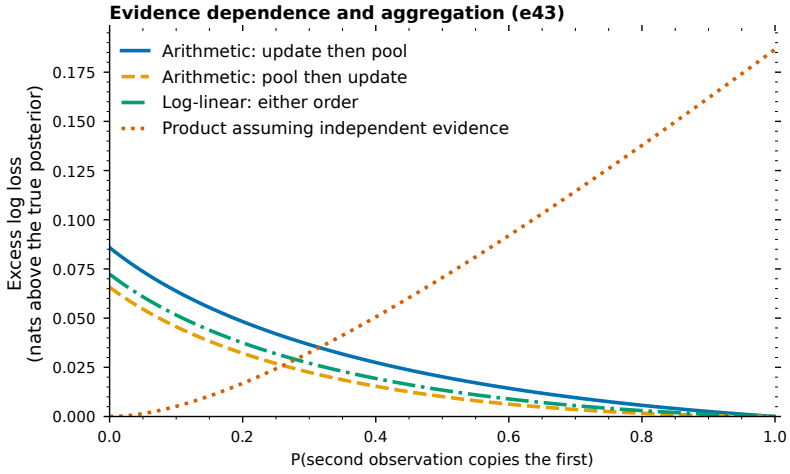


Figure 8: **Finite-state classification under a specified observation model.** For $K = 3$ and marginal channel accuracy $q = 0.8$, exact enumeration recomputes the joint-law oracle as private-observation copy probability ρ varies. The vertical quantity is excess log loss (nats) against that oracle for arithmetic pool-then-update and a product estimator that assumes independence. This is a quality measurement beyond pooling algebra: the product estimator's regret increases as evidence becomes redundant. It is restricted to the declared synthetic channel, and does not validate a native-LLM pooling operator or a universal quality ordering. Source: exact enumeration in `code/stageA/e43_e0_oracle_baselines.py`; the 101 copy-channel points are generated separately from its 18-cell grid CSV.

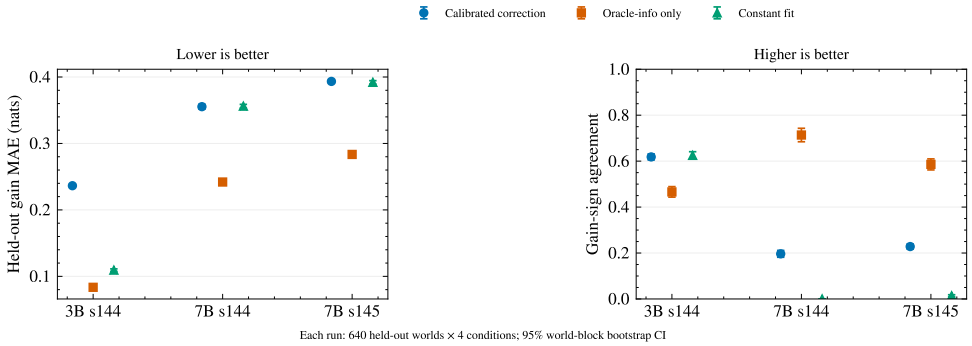


Figure 9: **Held-out receiver-gain quality beyond pooling algebra.** A frozen predictor is compared with an information-only baseline on a separately specified finite joint-observation task; test cells, prompt template, and labels differ from fitting. Each of three real receiver runs has 640 independent test worlds and four paired message conditions (2,560 rows); intervals resample worlds, not rows. Predictor MAE exceeds baseline MAE in all three runs (3B/144: 0.236 vs 0.084; 7B/144: 0.355 vs 0.242; 7B/145: 0.393 vs 0.283 nats). This tests a finite-state receiver-quality prediction and rejects this predictor on these held-out cells. It does not close a general PID-sign theorem, validate Stage C capacity shifts, or identify a native pooling operator. Sources: `experiments/ex2_receiver_gain_predictor/results/` run summaries and test rows.

are equal, $|\chi| \leq 1.5 \times 10^{-122}$. At the smallest t, the quartic log–log slope is 4.0000003. Swapping $\psi(r_{\max})$ and $\psi(r_{\min})$ violates the bound in 20/20 configurations, and replacing the factor 4 by 8 in the Jensen certificate violates it in 5/20 configurations. Both audits cost nothing and take under 0.1 s on a CPU. Their outputs are in `results/r124_e2_equalz_direct/full/` and `results/r124_e1_chibound_cleanroom/full/`. These are checks of finite identities in the specified family, not native-model evidence.

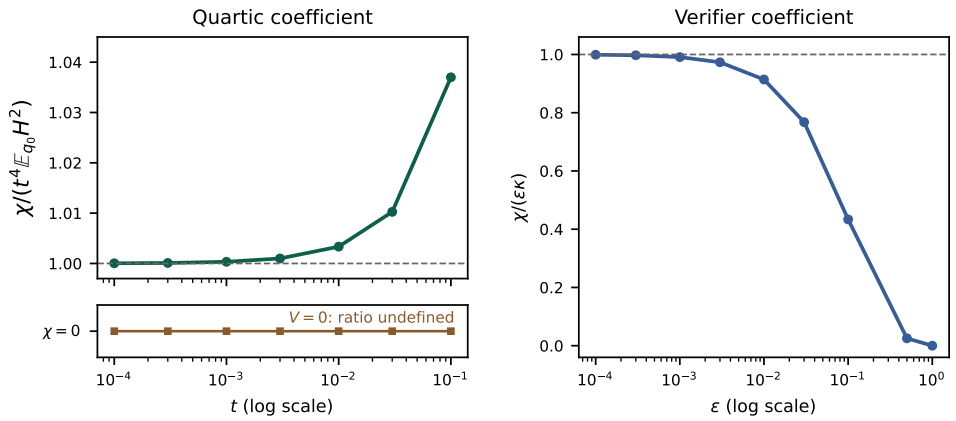


Figure 10: **Asymptotic coefficient curves and the zero-gap control.** The coefficient ratios approach one on the stored grids in independently computed high-precision arithmetic. The $V = 0$ consensus control has $\chi = 0$ and an undefined coefficient ratio, so its zero-gap series uses a separate axis rather than a fabricated ratio (Decimal residuals below 10^{-100} are displayed at exact zero). Other zero controls and global bounds are retained in the full CPU audit; no plot floor or clipped prior is used. These checks concern the specified finite families.

Normalizer dispersion alone is not a universal monotone predictor of χ. The exact zero condition and the specified-family coefficients supply distinct conclusions; neither predicts benchmark accuracy.

The supplementary copy-channel evaluation in Appendix K varies evidence dependence while reconstructing the target posterior at every grid point. Its exact rational tie audit distinguishes mathematical tie policies from floating-point argmax errors.

E-B: fixed-configuration sensitivity. The audit recovers all stored signed and scaled amplitude configurations, directions, perturbation strengths and positivity constraints. It recomputes V and the full t grid, and reports the range of fitted slopes across these fixed configurations in Table 11. The old bootstrap interval had no declared sampling population and is withdrawn. The single-agent zero is an operator identity; it has no log–log slope. Table 10 records the separate coefficient checks.

EX-1 gain-sign test. The real-LLM EX-1 v4 test estimates $R = I_{\min}(X_1, X_2; Y)$ and $S = I(X_1, X_2; Y) - I(X_1; Y) - I(X_2; Y) + R$ from 32 independent calibration runs per task, where X_i is agent i's initial argmax probe and Y is the final post-message majority argmax, not the gold answer. It freezes $\operatorname{sign}(S - R)$ before 16 held-out evaluation repeats per task and compares it with the sign of mean m-pool minus fixed model-A single log score on all 64 tasks; zero predictions abstain but stay in the denominator (41 abstentions and no zero-gain tasks). The matched-budget single arm uses the measured team token count on every paired repeat (1,024/1,024 equal pairs; 286,640 tokens per arm across all pairs), while its separate gain column changes the denominator from the fixed single to this matched single. Sign agreement is $9/64 = 0.141$ (reported 95% interval $[0.063, 0.234]$), below the always-positive baseline's 0.688 arithmetic-pooling agreement; the recorded decision is *disagrees*, not evidence for a population-level gain law. The source is `experiments/ex1_pid_sign_predictor/results/ex1_pid_sign_predictor_v4/summary.json`; the measurement contract is archived beside the runner.

Wider exact-rational zero-set audit. An exact-rational audit widens the control to four nonconstant likelihoods, $N \in \{2, 3\}$ and two weight vectors: all 94 distinct-prior equal-Z_i cells give $A = B$ exactly, all 188 matched unequal-Z_i twins give $A \neq B$ ($\chi \in [3.4 \times 10^{-6}, 2.7 \times 10^{-4}]$), and a common nonconstant likelihood alone does not equalize the Z_i.

F EXTENDED RELATED WORK

Opinion pooling. The linear (arithmetic) pool (Stone, 1961) and the logarithmic (geometric) pool have been characterized axiomatically: logarithmic pools are externally Bayesian, whereas arithmetic pools generally are not (Genest, 1984; Genest & Zidek, 1986; Dietrich & List, 2017). Surveys cover expert aggregation and density fusion (Clemen & Winkler, 1999; Koliander et al., 2022). Weight selection for logarithmic pools (Heskes, 1997), products of experts (Hinton, 2002), and logit aggregation of forecasts (Satopää et al., 2014) optimize a single combined belief. Iterated linear averaging models consensus formation (DeGroot, 1974; Golub & Jackson, 2010), and non-Bayesian social learning alternates a private Bayesian update with averaging over neighbours (Jadbabaie et al., 2012). That alternation raises our question: whether updating commutes with pooling. We do not claim priority for the failure of external Bayesianity. Proposition 2 states where the discrepancy goes (weights w_i become $w_i Z_i$), and Proposition 4 states its order along a specified near-consensus family. Classical agreement results (Aumann, 1976; Geanakoplos & Polemarchakis, 1982) and logical belief revision (Alchourrón et al., 1985) assume common priors or qualitative belief sets, not finite-state weighted pools. Order effects in human updating (Hogarth & Einhorn, 1992) are an empirical, not operator-level, phenomenon.

Information geometry. The m/e terminology follows the dual affine connections of information geometry (Amari & Nagaoka, 2000; Amari, 2016; Ay et al., 2017); see Nielsen (2020) for an introduction. Arithmetic and logarithmic pools are m- and e-mixtures, and α-integration interpolates between them by minimizing an α-divergence (Amari, 2007). I-projections (Csiszár, 1975) and the Fisher metric underlying the natural gradient (Amari, 1998) provide the geometric vocabulary for our discrepancy terms. The commutation of e-pooling with a common likelihood (Proposition 1) is a known consequence of this structure. Our contribution is not new geometry. It is the exact weight reallocation and the quartic-or-zero order of the m-pool gap, stated for positive finite-state beliefs.

Multi-agent LLMs. Multi-agent debate improves factuality and reasoning on several benchmarks (Du et al., 2023; Liang et al., 2024), supports evaluation (Chan et al., 2024), and can make non-expert judges more accurate (Khan et al., 2024). Controlled comparisons are more mixed. Debate does not reliably beat self-consistency or ensembling (Smit et al., 2024; Wang et al., 2023). Majority voting explains most reported gains, and debate modeled as a martingale over beliefs does not raise expected correctness (Choi et al., 2025). Interventions on diversity and misconceptions change debate outcomes (Estornell & Liu, 2024). Native collectives also show capacity- and protocol-dependent failures (Fukushima, 2026; Pavlova & Tanaka, 2026; Shi et al., 2026) and sampling-driven consensus (Tanaka, 2026). Separately, single LLMs are sensitive to the order of demonstrations, options, and context (Lu et al., 2022; Pezeshkpour & Hruschka, 2024; Liu et al., 2024). These works measure protocol-level accuracy or input-order sensitivity. We instead isolate one ordering choice that remains after such effects are fixed: whether to pool before or after a shared update. We characterize it exactly at the operator level and evaluate it only on a finite joint law (Figure 2). None of these results shows that native LLM agents implement our operators or that one ordering improves native task accuracy (Appendix F).

Statistical mechanics and pilot response rules. El et al. (El et al., 2026) fit a stochastic social-response model to language-model communities and report indifference, polarization, and consensus regimes. Its message-response step can be represented at the general kernel level of Section 3.2, but its Ising-like couplings and fitted trajectories are *not* special cases of the uniform-pair QSG moment or lottery theorems. Proposition 10 similarly gives conditional response-rule predictions for two pilot phenomena; without a matched native response measurement, this is an inclusion of possible mechanisms, not a fit to or explanation of the unpublished pilot. Other native collectives exhibit capacity- and protocol-dependent failures (Fukushima, 2026; Pavlova & Tanaka, 2026; Shi et al., 2026). Their outcome contrasts cannot identify evidence redundancy or a log-score cost without matched receiver information and loss measurements.

Comparison objects and priority limits. Classical external Bayesianity concerns pooling rules that commute with Bayesian updating under their stated requirements (Genest, 1984; Genest & Zidek, 1986). Our arithmetic zero-set statement instead fixes the population, weights, and likeli-

Gate fixed before held-out test	Acceptance rule	Current status
One-step native mean and second moment	Each error $\leq 10\%$ of validation signal; signal $\leq$ 2 validation SE is insufficient	Failed (3B: mean 47.2%, second 5.4%; 7B: mean 30.7%, second 45.8%)
Frozen R, V trajectory	Normalized RMSE $\leq .05$	Not evaluated
Added value over Tanaka mean-field	$\geq$ 25% RMSE reduction; paired-world bootstrap 95% CI lower bound > 0	Not evaluated
Empirical-transition control	U3 RMSE minus empirical RMSE $\leq .01$	Not evaluated

Table 3: MOMENT-TRANSFER decision table, preregistered before native held-out test. The independent bootstrap unit is the world; schedule seeds, times, and conditions remain clustered within it. All predictors receive the same calibration calls and test trajectories. Gate 1 failed for both models, so gates 2–4 were not evaluated.

hood and asks when the two resulting outputs agree. These are different comparison objects; this distinction alone does not establish priority.

The finite analysis also provides discrepancy bounds and coefficient expansions for explicitly specified families. The original pooling texts have not been checked in full for this comparison, so we do not mark the corresponding results as absent from that literature or infer novelty from their absence among another paper's headline contributions.

The language-model and sampled-population studies discussed above address different measurement objects. Their outcomes do not validate our finite operators as native response maps. The finite diagnostic checks assess formulas on declared inputs; they are not a substitute for a result-level comparison with the original pooling literature.

Sampled communication and the consensus lottery. Tanaka (2026) studies sampled convex-combination communication and memetic drift. Its unbiased, uniformly paired rule is the sampling-kernel instance of our operator (Theorem 1(e)): Proposition 7 then closes its population moments exactly, Theorem 2 gives its vertex-consensus probabilities, and Proposition 9 its N-scaling of disagreement relaxation time. Outside this containment are the drift–selection crossover, which needs an external-field model of the speaker path, and the native naming-game experiments. The constant likelihood $L \equiv c$ is a different object: the zero-evidence case of the deterministic analysis (Corollary 2).

Operator algebra and classical opinion pooling. External Bayesianity and the distinction between arithmetic and logarithmic opinion pools are classical (Genest, 1984; Genest & Zidek, 1986). We do not claim priority for logarithmic-pool commutation, Bayes normalization algebra, standard divergence comparisons, or the KL chain rule. Theorem 1 places these identities, the exact arithmetic zero condition, and the discrepancy bounds in the deterministic specialization of one operator; none of the algebra alone ranks pooling orders by task accuracy.

From language-model maps to agent-state geometry. The m/e terminology follows the dual affine structures of information geometry (Amari, 2016). Oyama et al. (2025) map individual language models from fixed-text log-likelihood vectors. Proposition 5 conditionally extends their probe vectors to a Gram geometry of interacting states. This is a geometric embedding, not an estimator of native communication gain.

G MOMENT-TRANSFER GATE TABLE AND CPU REFERENCE

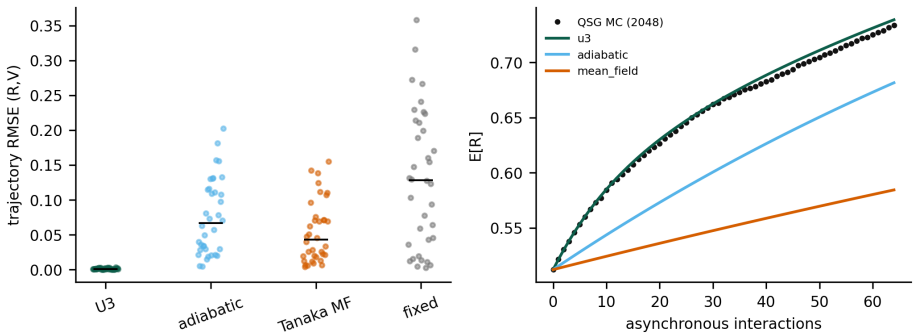


Figure 11: CPU reference, not native transfer. Left: 36-cell QSG trajectory RMSE for R, V (2,048 trajectories/cell; bars mark medians). Right: a prespecified divergent cell ($N = 8, \alpha = .8, m = 4$, high initial diversity), where the full recurrence and reduced approximations differ. One interaction updates one listener.

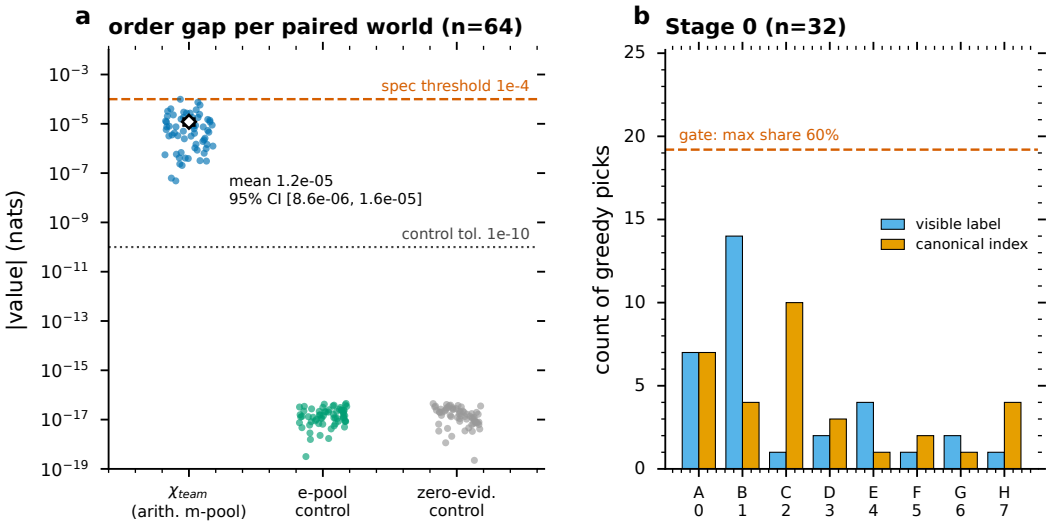
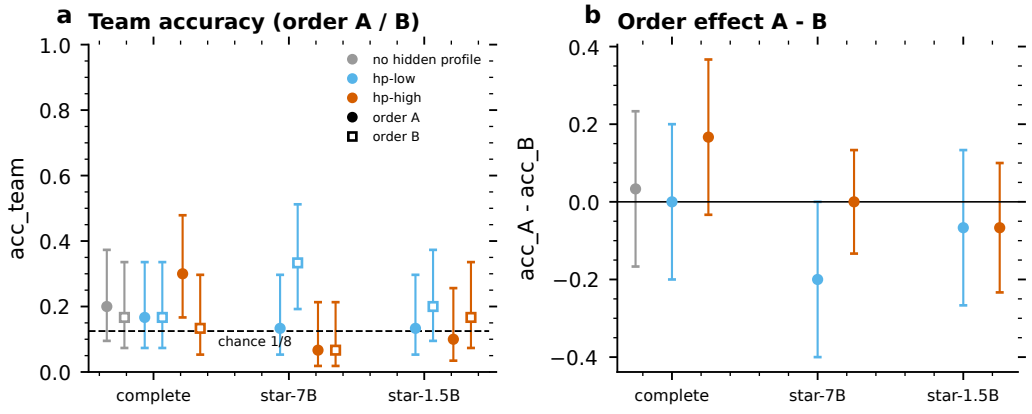


Figure 12: **Stage 0/1 native order gap (exploratory; native transfer gate not passed; not evidence for the finite-model claims).** Team order gap χ_{team} between update-then-pool and pool-then-update for explicit arithmetic pooling with a 1.5B model, over 64 paired worlds (8 seeds). Mean 1.21×10^{-5} (95% seed-block bootstrap $[8.55 \times 10^{-6}, 1.62 \times 10^{-5}]$), maximum 9.88×10^{-5}; 0 of 64 worlds exceed the exploratory threshold 1×10^{-4}. Controls (e-pool and zero-evidence) stay below 4.50×10^{-17}. Limits: the maximum is close to the threshold, so the gap is not "orders of magnitude" below it (median 5.59×10^{-6}); Stage 0 visible-label choices concentrate on label B (14/32) under a relaxed 60% share gate. Descriptive only.

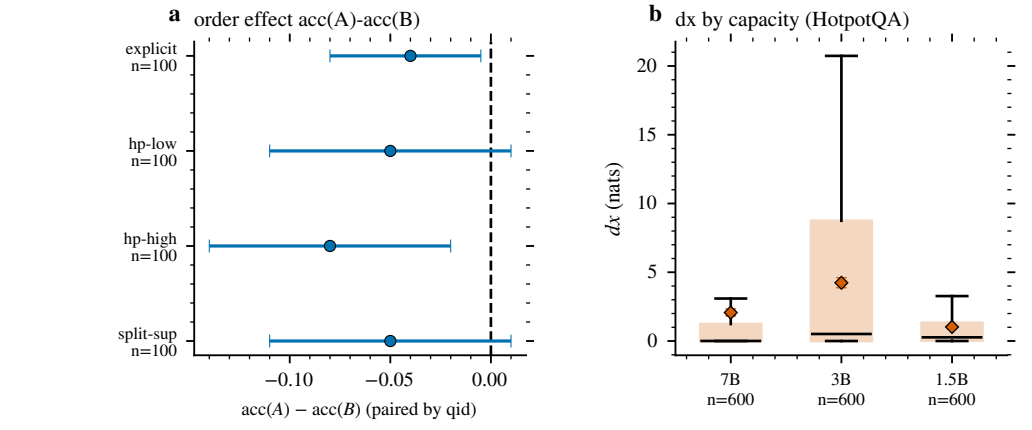
H NATIVE-LLM EXPLORATORY RESULTS

Every figure in this section comes from native-LLM runs. The native transfer gate (Section D) was not passed, so none of these figures is evidence for the finite-model claims of the main text; each is reported as a descriptive, exploratory measurement with its own caveats. Figures are built from saved receipts and sidecar files, and the numbers in the captions are read from those files by `code/make_appendix_native_numbers.py`.



Raw: stageC_combined_20260918/stageC_rows.jsonl (840 rows; native aggregation, het capacity, control=none used here: 420 rows).
n=30 trials per (hp_mode, topology, order) cell; A and B use disjoint seeds (unpaired). Panel a: Wilson 95% CI; panel b: unpaired bootstrap 95% CI (10000, seed 0).
Chance = 1/8: hidden state drawn as rng.randrange(8) (run_stagec.py:155,177); acc_team = argmax(team_belief)==state.
'no hidden profile' exists only for topology=complete (no star cell). Descriptive, not preregistered confirmatory.

Figure 13: **Hidden-profile accuracy by topology under native aggregation (exploratory; native transfer gate not passed; not evidence for the finite-model claims).** Team accuracy with Wilson 95% intervals ($n = 30$ per cell; heterogeneous capacity; chance $= 0.125$). Only 2 of 14 cells have a lower bound above chance, and the A−B order difference has an interval containing zero in 7 of 7 comparisons. No multiplicity correction; the two above-chance cells are not read as effects. Limits: "no hidden profile" is run only on the complete topology, and intervals are wide ($n = 30$).



acc_team = team-belief argmax vs gold_label (0/1), paired by qid, 95% CI = paired (by-qid) bootstrap, 10000 resamples; descriptive (not a prespecified confirmatory comparison; no prereg on file).
explicit n=100 qid (2 pooling_mode variants averaged per qid, see sidecar); hp-low/hp-high/split-sup n=100 qid each, 1 row/qid/order. dx pooled over explicit+hp+split-sup arms and both orders, since dx (run_staged.py:360) is computed identically regardless of order or arm.
EXCLUDED: stageD_null_A (zero-evidence) / stageD_null_B (Takase-additive) are two different null designs, not an order-A/B pair (run_staged.py:298-311, dx=0 by construction for Takase) -- including them here as an 'order effect' would misrepresent a null-type difference as an order effect.

Figure 14: **HotpotQA external check of the order effect (exploratory; native transfer gate not passed; not evidence for the finite-model claims).** Panel (a): qid-paired accuracy difference acc(A)−acc(B) in 4 arms ($n = 100$ questions per arm), where A applies evidence pooling before communication pooling. The difference is negative in 4 of 4 arms (range -0.080 to -0.040) with a paired-bootstrap 95% interval excluding zero in 2. Panel (b): belief-update size by agent capacity. Limits: descriptive, no preregistration, one dataset and four arm operationalizations; the two null shards are excluded by design because they are not an A/B pair.

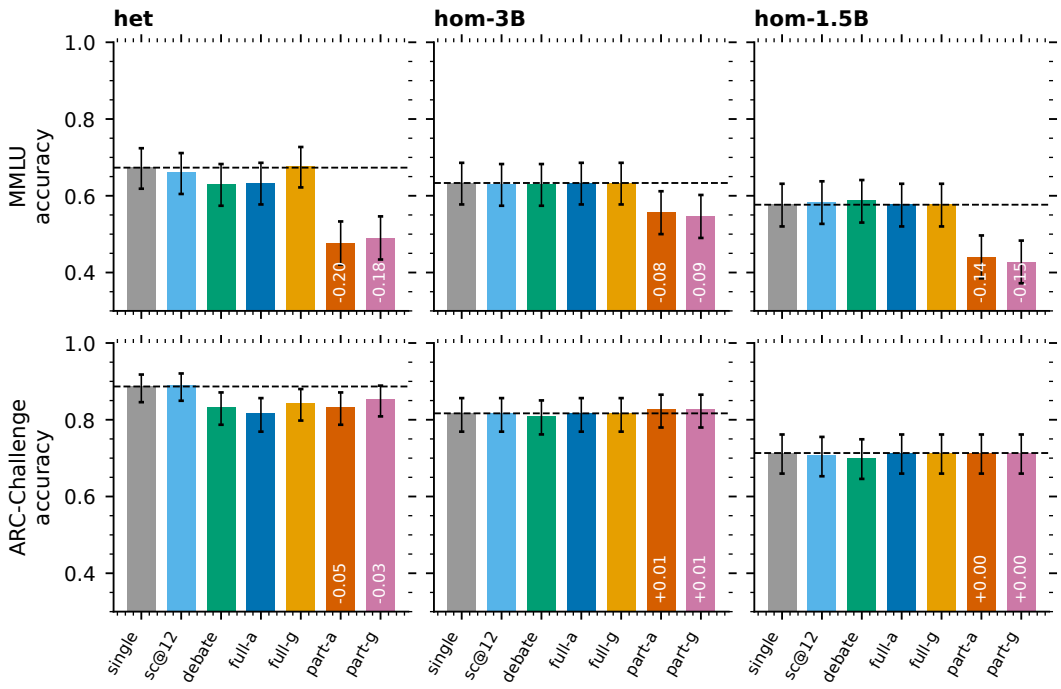


Figure 15: **Real-LLM teams on MMLU and ARC-Challenge (exploratory; native transfer gate not passed; not evidence for the finite-model claims).** Accuracy by aggregation condition with Wilson 95% intervals ($n = 300$ per cell); dashed line is the single-pass baseline of the cell, labels are paired differences to it. Of the 36 non-baseline cells, 12 have a paired interval excluding zero, 0 of them in favor of the condition (i.e. none improves on the single pass). Partial-evidence pooling on MMLU lowers accuracy for the heterogeneous team (arithmetic -0.20, geometric -0.18). Limits: the homogeneous 7B team has not been run and is not plotted; no utility claim is made.

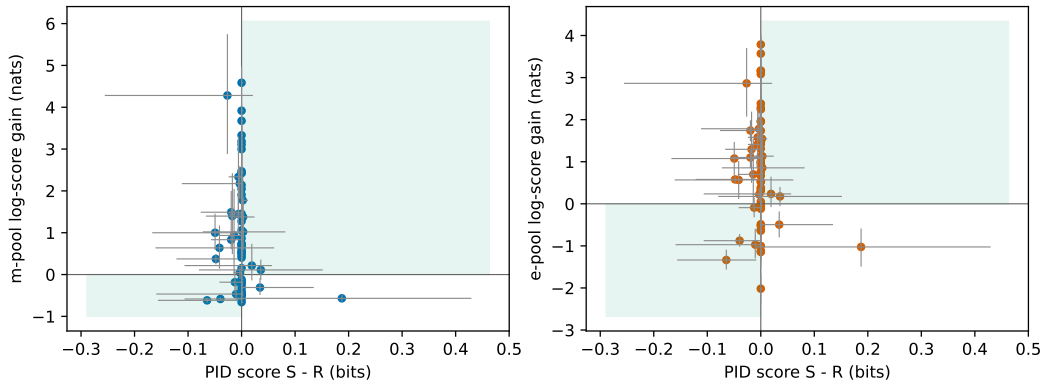


Figure 16: **EX-1 sign predictor: PID score $S - R$ against realized pooling gain (exploratory; native transfer gate not passed; not evidence for the finite-model claims).** Left: m-pool; right: e-pool; each point is one of 64 tasks with paired-repeat intervals; shaded quadrants mark sign agreement. The predictor agrees with the realized gain sign in 9/64 tasks (0.141; Wilson 95% [0.076, 0.246]) and abstains on 41, whereas "always positive" would be right for 0.688 (m-pool); the recorded decision is "disagrees". The PID target is a native-majority proxy, not a gold-label PID, and the score is an exploratory response-profile proxy, not a theorem test.

I ADDITIONAL PROOFS AND EXPERIMENTAL DETAILS

I.1 TRACE-LEVEL ATTRIBUTION AND MEASUREMENT

Trace-law setup. Fix a common prompt measure μ, a canonical ordered trace-slot map $k \mapsto (i, t, r)$ (i=agent, t=round, $r \in \{\text{private}, \text{public}, \text{readout}\}$), and a common teacher-forced trace bank $\mathcal{T}$, all fixed *before* evaluation. For a system law $P_S(\tau \mid x)$, $\tau = (z_1, \ldots, z_K)$, define

$$c_k(P, Q) = \mathbb{E}_{P(z_{<k}|x)}\, D_{KL}\big(P(z_k \mid z_{<k}, x) \,\|\, Q(z_k \mid z_{<k}, x)\big).$$

I.1.1 EXACT FACTORIZATION AND A MATCHED-KL COUNTEREXAMPLE

Proposition 12 (Exact factorization)**.** *Under a common support assumption, for fixed x,*

$$D_{KL}\big(P(\tau \mid x) \,\|\, Q(\tau \mid x)\big) = \sum_{k=1}^{K} c_k(P, Q).$$

Proof deferred to Appendix I.

For a pre-registered role map $r(k)$, define $\mathrm{Contrib}_{\text{role}}(P, Q) := \sum_{k:\, r(k)=\text{role}} c_k(P, Q) \geq 0$; Proposition 12 gives the exact (non-approximate) partition $D_{KL}(P\|Q) = \sum_{\text{role}} \mathrm{Contrib}_{\text{role}}(P, Q)$ of the *one-directional* KL only. Write $J_\tau(P, Q) := D_{sym}(P, Q) = D_{KL}(P\|Q) + D_{KL}(Q\|P)$ for the *full-trace* symmetric divergence, and define the symmetrized contribution $\mathrm{Contrib}^{\text{sym}}_{\text{role}}(P, Q) := \mathrm{Contrib}_{\text{role}}(P, Q) + \mathrm{Contrib}_{\text{role}}(Q, P) \geq 0$; summing over roles, $\sum_{\text{role}} \mathrm{Contrib}^{\text{sym}}_{\text{role}}(P, Q) = D_{KL}(P\|Q) + D_{KL}(Q\|P) = J_\tau(P, Q)$. It is $\mathrm{Contrib}^{\text{sym}}_{\text{role}}$ (not the one-directional $\mathrm{Contrib}_{\text{role}}$) that is a decomposition of J_τ, not of χ.

Proposition 13 (χ is a readout-marginal quantity; J_τ only upper-bounds it)**.** *Let $Y \in \mathcal{Y}$ denote the trace's final-readout* label*, sampled from the system's readout distribution over $\mathcal{Y}$ (a random outcome, not the belief vector itself, which is that distribution), and let P_Y, Q_Y be Y's marginal laws under P, Q. Then*

$$\begin{aligned} J_\tau(P, Q) = D_{sym}(P_Y, Q_Y) &+ \mathbb{E}_{P_Y} D_{KL}(P_{\tau|Y} \| Q_{\tau|Y}) \\ &+ \mathbb{E}_{Q_Y} D_{KL}(Q_{\tau|Y} \| P_{\tau|Y}) \;\geq\; D_{sym}(P_Y, Q_Y). \end{aligned}$$

When P, Q are the two collective-belief trace laws of §3.1 (pool-then-update vs. update-then-pool) and Y is generated by sampling a label from the resulting collective belief, so that Y's marginal laws coincide with those beliefs as distributions over $\mathcal{Y}$, $P_Y = \Pi_C \Pi_E p =: A$ and $Q_Y = \Pi_E \Pi_C p =: B$, we get $D_{sym}(P_Y, Q_Y) = D_{sym}(A, B) = \chi(p)$ by definition, so $J_\tau \geq \chi(p)$, with equality only when the two systems' intermediate traces also agree in law given Y. *In particular $\chi(p) = 0$ (e.g. e-pooling, Proposition 1) does* not *imply $J_\tau = 0$: the final beliefs can coincide exactly while earlier private or public slots differ.*

Proof deferred to Appendix I. **Consequence for slot attribution.** Because $\mathrm{Contrib}^{\text{sym}}_{\text{role}}(P, Q)$ sums to J_τ and not to χ, an observed nonzero J_τ (or its per-role symmetrized attribution) is evidence about *trace-level* divergence, not a measurement of the operator-level order gap χ itself; the two are identified only under the (unverified in general) equality case of Proposition 13. We do not claim the slot decomposition attributes an observed χ to the public channel; it attributes J_τ, which upper-bounds χ.

Theorem 3 (Matched-KL counterexample: joint KL does not identify public-communication contribution)**.** *There exist two systems with identical joint KL to a common reference but different public-communication contribution. Hence joint KL alone does not identify public-communication contribution.*

Proof deferred to Appendix I (minimal $K = 2$ slot construction; numeric instance $p = 0.3 \Rightarrow \delta \approx 0.0823$ nats).

Remark (relation to Tanaka, arXiv:2603.24676). Tanaka's Soft QSG update is the sampling-kernel instance of the communication operator made exact in Proposition 6.

I.1.2 MEASUREMENT LEMMA

Assumptions. (M1) common support of a reference law P_{ref} containing the supports of $P_S, P_{S'}$; (M2) local neighborhood, $P_S = P_{\mathrm{ref}}(1+\varepsilon_S)$, $P_{S'} = P_{\mathrm{ref}}(1+\varepsilon_{S'})$, $\sup_y |\varepsilon_S(y)|, |\varepsilon_{S'}(y)| = O(\varepsilon)$; (M3) probes $y_1, \dots, y_M \overset{iid}{\sim} P_{\mathrm{ref}}$, independent of S, S'; (M4) $\mathbb{E}_{\mathrm{ref}}[\Delta^4] < \infty$ where $\Delta(y) := \log P_S(y) - \log P_{S'}(y)$. Let $\omega_{S,m} := \log P_S(y_m) - \frac{1}{M}\sum_{m'} \log P_S(y_{m'})$.

Lemma 1 (Measurement lemma)**.** *Under (M1)–(M4),*

$$2\, D_{KL}(P_S \| P_{S'}) = M^{-1} \|\omega_S - \omega_{S'}\|_2^2 + O(\varepsilon^3) + O_P(M^{-1/2}).$$

Proof deferred to Appendix I.

Scope of the measurement route. The local approximation is not claimed as novel in isolation. Combining it with the slot decomposition is a proposed measurement route, not a completed estimator. Valid estimation requires the appropriate prefix-law sampling and approximation-error analysis. A reference-bank distance is not automatically a conditional slot contribution or a measurement of the readout order gap.

I.2 INFORMATION-CONSTRAINED PROJECTION AND COMMUNICATION ACCOUNTING

Setup and assumptions. Fix a single reception event. Let H be the receiver's pre-reception information, E the sender-accessible information (tools and any extra sender observation are *included* in E, never omitted from it), and M the message. We assume

$$Y \perp M \mid (H, E) \tag{2}$$

i.e. the message carries no influence on the target Y beyond what (H, E) already determines. Write the true conditionals and the receiver's actual models as

$$\pi_{HE} := P(Y \mid H, E), \quad \pi_{HM} := P(Y \mid H, M), \quad \pi_H := P(Y \mid H),$$
$$q^+(Y \mid H, M), \qquad q^-(Y \mid H).$$

D_{KL} is as in §3.1 (nats throughout).

Proposition 14 (Exact projection decomposition)**.** *Assume equation 2 and common support. For* any q *that depends only on the receivable information* (H, M)*,*

$$\mathbb{E}\, D_{KL}\big(\pi_{HE} \,\|\, q\big) \;=\; I(Y; E \mid H, M) \;+\; \mathbb{E}\, D_{KL}\big(\pi_{HM} \,\|\, q\big).$$

Consequently $\mathbb{E}D_{KL}(\pi_{HE}\|q)$ is minimised over this class exactly at $q = \pi_{HM}$, with irreducible floor $I(Y; E \mid H, M)$.

Proof deferred to Appendix I.

Scope note (what is *not* claimed). This is a Pythagoras-type decomposition onto an *information-constrained* predictive family, and we claim no novelty for the identity itself. In particular we do *not* assume that a native LLM computes the optimal projection π_{HM}. The ideal projection is defined from the observable input constraint (H, M); its deviation from a native predictor requires a separately validated measurement. Nor do we call an arbitrary update a projection: general Bayes updating need not be idempotent, so "projection" is reserved for the constrained-family minimiser above.

Corollary 3 (Communication accounting identity)**.** *Assume finite expected negative log probabilities of the realized label:*

$$\mathbb{E}[-\log q^-(Y \mid H)] < \infty, \qquad \mathbb{E}[-\log q^+(Y \mid H, M)] < \infty.$$

Impose the analogous conditions on every decoder pair used below. These assumptions avoid undefined differences of infinite losses. Let $\epsilon^- := \mathbb{E}D_{KL}(\pi_H \| q^-)$ and $\epsilon^+ := \mathbb{E}D_{KL}(\pi_{HM} \| q^+)$ be the receiver's pre- and post-reception prediction errors. Then the mean log-score improvement

$$G \;:=\; \mathbb{E} \log \frac{q^+(Y \mid H, M)}{q^-(Y \mid H)}$$

satisfies

$$G \;=\; I(Y; M \mid H) \;+\; \epsilon^- - \epsilon^+ \;=\; \underbrace{I(Y; E \mid H)}_{\textit{available}} \;-\; \underbrace{I(Y; E \mid H, M)}_{\textit{lost in transmission}} \;-\; \underbrace{(\epsilon^+ - \epsilon^-)}_{\textit{growth of receiver error}}.$$

Proof deferred to Appendix I.

Reading of the three terms. (i) Re-expressing the *same* information can improve prediction, purely by lowering ϵ^+. (ii) A message that genuinely carries information can still *degrade* performance whenever the growth in receiver error exceeds the information delivered. (iii) Repeating evidence already contained in H adds no new conditional information, so $I(Y; M \mid H) = 0$; in that case $G = \epsilon^- - \epsilon^+$, which is *not* signed in general — as (i) already illustrates, a mere re-expression of H can still lower ϵ^+ below ϵ^- and yield $G > 0$. Concluding $G \leq 0$ additionally requires that the pre-reception predictor already be (weakly) optimal for the class realising q^+, i.e. $\epsilon^- \leq \epsilon^+$ for the specific (q^-, q^+) in use; it is not a consequence of $I(Y; M \mid H) = 0$ alone. The identity distinguishes delivered information from receiver error. It does not identify either quantity from native outputs without additional measurements.

Remark (exact finite-state verification). The finite-state checker `sanity_checks.py` function `xor_budget` instantiates Corollary 3 term-by-term on the XOR construction of Proposition 15 below (H, E independent fair bits, $Y = H \oplus E$, a noisy channel to M), reproducing the identity to machine precision (residual $\leq 8.4 \times 10^{-17}$) on every setting of Table 7 (Appendix J.9), including a case where a genuinely informative message still yields $G < 0$. All expectations are exact finite sums; no LLM inference or benchmark claim is involved.

Proposition 15 (XOR counterexample: answer distributions are not a communication state)**.** *Let X_1, X_2 be independent fair bits and $Y = X_1 \oplus X_2$. Then*

$$P(Y \mid X_i = 0) = P(Y \mid X_i = 1) = (1/2,\, 1/2) \qquad \textit{for } i = 1, 2,$$

so every agent's answer distribution is uniform in every case, yet sharing the two bits determines Y exactly. Hence any model whose message-generation depends on the final answer distributions alone cannot express the benefit of communication on this task: a theory built on finite answer distributions is not a general theory of complementary-information integration.

Proof deferred to Appendix I.

Both poolings are fixed here; communication decides the task. Because every p_i is the uniform distribution on $\{0, 1\}$, both operators of §3.1 act trivially: $\Pi^e_C(p_{1:2}) = \Pi^m_C(p_{1:2}) = \mathrm{Unif}\{0, 1\}$ for all weights w, and with no evidence update $\chi(p) = 0$ for *both* m- and e-pooling. The task is nevertheless solved exactly by exchanging one bit. This is an exact finite construction, not a native-model experiment.

Remark (relation to the finite operator and trace results). Proposition 15 does not contradict and does not retract any of Propositions 12, 1, 2, 3 or Theorem 3; it delimits their domain. Section 3.1 studies operators on *answer-distribution* simplices, and its order-gap χ is exactly the diagnostic that is informative when the answer distributions are a sufficient state. The XOR construction is the complementary-information regime where that sufficiency fails and $\chi \equiv 0$ carries no signal, while Corollary 3 still accounts for the full gain. Accordingly the m-/e-pooling order gap of §3.1 is positioned as the special case of this section in which the state is answer-distribution-sufficient, and the basic state is taken to be the history $H_{i,t}$, with low-dimensional probe states constructed from a common probe battery (answers, private observations, intermediate relations, provenance of evidence) whose predictive sufficiency for future updates is itself an object of test. Likewise Corollary 2 continues to hold verbatim, as it concerns the $\mathrm{Var}_w(Z_i) = 0$ cross-section of the same answer-distribution regime.

Estimation limits. On synthetic tasks the true conditionals are computed exactly (Table 7). For free-form natural-language messages π_{HM} is generally not directly computable; when the log loss of a cross-fitted decode model is used as a surrogate, we keep the true Shannon quantity and the model-dependent estimator typographically and terminologically distinct, and we do not equate "a small decoder cannot read it" with "the message carries no information." We also do not call a strong LLM the true posterior: an oracle and a teacher model are different objects.

The separation can be made quantitative, and we state it because it is exactly the failure mode that an unlabelled "estimated mutual information" column invites. Let (d^+, d^-) be a cross-fitted decoder pair — $d^+(Y \mid H, M)$ and $d^-(Y \mid H)$ fitted on folds disjoint from the evaluation fold — and define the *decoder read-off*, which is a log-loss difference and not a Shannon quantity,

$$\widehat{I}_{\mathrm{dec}} \;:=\; \mathbb{E} \log \frac{d^+(Y \mid H, M)}{d^-(Y \mid H)}, \qquad \epsilon^+_{\mathrm{dec}} := \mathbb{E} D_{KL}(\pi_{HM} \| d^+), \qquad \epsilon^-_{\mathrm{dec}} := \mathbb{E} D_{KL}(\pi_H \| d^-).$$

Corollary 4 (What a decoder read-off does and does not certify)**.** *Under the assumptions of Corollary 3, applying it to the pair (d^+, d^-) in place of (q^+, q^-) gives $\widehat{I}_{\mathrm{dec}} = I(Y; M \mid H) + \epsilon^-_{\mathrm{dec}} - \epsilon^+_{\mathrm{dec}}$, hence the two-sided enclosure*

$$\widehat{I}_{\mathrm{dec}} - \epsilon^-_{\mathrm{dec}} \;\le\; I(Y; M \mid H) \;\le\; \widehat{I}_{\mathrm{dec}} + \epsilon^+_{\mathrm{dec}}.$$

In particular, if the pre-reception decoder is exact ($\epsilon^-_{\mathrm{dec}} = 0$) then $\widehat{I}_{\mathrm{dec}} \le I(Y; M \mid H)$, with deficit exactly $\epsilon^+_{\mathrm{dec}}$.

Proof deferred to Appendix I.

Interpretation of the decoder read-off. The decoder read-off is a model-dependent population log-score gain. It is a lower bound on conditional mutual information when the pre-reception decoder is exact. With that baseline held fixed, a post-reception decoder with smaller population approximation error raises the lower bound; greater model capacity alone does not guarantee that reduction.

If the pre-reception decoder is approximate, the read-off can exceed $I(Y; M \mid H)$ and must not be called a lower bound without separate control of the baseline error. For the finite label space,

$$I(Y; M \mid H) \le H(Y \mid H) \le \log |\mathcal{Y}|.$$

A small decoder read-off does not provide a correspondingly small upper bound on information unless the post-reception approximation error is controlled. An oracle and a teacher model remain distinct.

From population gap to a reportable number. Corollary 4 bounds $I(Y; M \mid H)$ in terms of the *population* quantities $\epsilon^{\pm}_{\mathrm{dec}}$, which are themselves expectations under π_{HM}, π_H and are therefore no more computable than $I(Y; M \mid H)$ itself. What is actually computed in practice is the *plug-in* estimator on a held-out evaluation fold of size n, disjoint from both decoder-fitting folds,

$$\widehat{I}^{(n)}_{\mathrm{dec}} := \frac{1}{n} \sum_{i=1}^{n} \log \frac{d^+(Y_i \mid H_i, M_i)}{d^-(Y_i \mid H_i)},$$

and it is essential to separate the error this introduces from $\epsilon^{\pm}_{\mathrm{dec}}$, because the two shrink under entirely different levers.

Corollary 5 (Finite-sample decoder-gap interval)**.** *Assume the evaluation-fold triples (H_i, M_i, Y_i), $i = 1, \dots, n$, are i.i.d. and that the log-ratio is almost surely bounded, $\left|\log d^+(Y \mid H, M) - \log d^-(Y \mid H)\right| \le \beta$. Then for any $\gamma \in (0, 1)$, with probability at least $1 - \gamma$ over the evaluation fold,*

$$\left|\widehat{I}^{(n)}_{\mathrm{dec}} - \widehat{I}_{\mathrm{dec}}\right| \;\le\; \beta \sqrt{\tfrac{2 \log(2/\gamma)}{n}} \;=:\; \rho_n(\gamma),$$

and combining with Corollary 4 gives, with probability at least $1 - \gamma$,

$$\widehat{I}^{(n)}_{\mathrm{dec}} - \rho_n(\gamma) - \epsilon^-_{\mathrm{dec}} \;\le\; I(Y; M \mid H) \;\le\; \widehat{I}^{(n)}_{\mathrm{dec}} + \rho_n(\gamma) + \epsilon^+_{\mathrm{dec}}.$$

Proof deferred to Appendix I.

Why the two error sources must be reported separately. $\rho_n(\gamma) = O(n^{-1/2})$ is a *statistical* term: it shrinks with more evaluation data alone, at a fixed decoder (d^+, d^-). $\epsilon^{\pm}_{\mathrm{dec}}$ is an *approximation* term: it depends on the decoder's capacity and training data and does *not* shrink as $n \to \infty$ with the decoder held fixed. A tight, high-confidence interval around $\widehat{I}^{(n)}_{\mathrm{dec}}$ (small $\rho_n(\gamma)$, obtained by enlarging the evaluation fold) is therefore fully compatible with an arbitrarily large true gap $\epsilon^+_{\mathrm{dec}}$: precision about the reported number is not accuracy about the quantity of interest, $I(Y; M \mid H)$. Reducing $\epsilon^+_{\mathrm{dec}}$ requires a post-decoder with smaller population approximation error; neither more capacity nor more fitting data alone guarantees that reduction. A decoder study must report $\rho_n(\gamma)$, its n and β, and the decoder's fitting/evaluation protocol separately from approximation-error assumptions. No fitted native decoder study or such interval is reported here.

I.3 TANAKA/QSG AS AN EXACT SPECIAL CASE, AND ITS FISHER GEOMETRY

The original QSG top-m channel. Fix a finite label set $\mathcal{Y} = \{1, \dots, K\}$, $N \geq 2$ agents with states $p_i \in \operatorname{int} \Delta^{K-1}$, and standard basis vectors $e_1, \dots, e_K$. One interaction selects an ordered speaker–listener pair (S, L), $S \neq L$, and applies the original QSG update of Tanaka (arXiv:2603.24676, their §2 Eq. (1)):

$$\widehat{p}_S \;=\; \frac{1}{m}\sum_{r=1}^{m} e_{Z_r}, \qquad Z_r \overset{iid}{\sim} \mathrm{Cat}(p_S), \qquad p_L^{+} \;=\; (1-\alpha)\, p_L + \alpha\, \widehat{p}_S, \qquad \alpha \in (0,1]. \tag{3}$$

In the vocabulary of §3.1 this is one specific choice of two ingredients: categorical sampling as the *transmission kernel*, and the m-mixture (i.e. Π_C^m with the two-agent weight vector $(1-\alpha, \alpha)$) as the *reception update*. The ideal limit $m \to \infty$ is soft exchange, $m = 1$ is a hard sample.

What "recovering QSG" requires us to keep separate. Reproducing the original model's premises means distinguishing, and not silently merging, (i) mean preservation, (ii) sampling covariance, (iii) disagreement between agents, (iv) the time unit, and (v) finite-population effects. We do not identify a common external field with an agent-independent normaliser (cf. Corollary 2). We also keep the $\alpha = 1$ absorption results separate from numerical or approximate conclusions obtained at $\alpha < 1$.

Novelty status. The next proposition is an algebraic corollary of equation 3 itself; we have not verified priority for it and claim none.

Proposition 16 (Fisher geometry of the QSG increment at consensus)**.** *Let $\bar{p} := N^{-1}\sum_i p_i$ and let $\Delta\bar{p} := \bar{p}^{+} - \bar{p}$ be the change of the population mean produced by one interaction equation 3. At a consensus point $p_i = p \in \operatorname{int} \Delta^{K-1}$ for all i,*

$$\mathbb{E}[\Delta\bar{p} \mid p] = 0, \qquad \mathrm{Cov}(\Delta\bar{p} \mid p) = \frac{\alpha^2}{mN^2}\big(\operatorname{diag}(p) - pp^{\top}\big).$$

Consequently, for the categorical Fisher metric $g_p(u,u) := \sum_y u_y^2/p_y$,

$$\boxed{\mathbb{E}\, g_p(\Delta\bar{p},\, \Delta\bar{p}) \;=\; \frac{\alpha^2(K-1)}{mN^2}}$$

independently of the consensus point p and of which pair (S, L) was selected.

Proof deferred to Appendix I.

Isotropy after Fisher whitening. The same covariance, whitened by the metric, is

$$D_p^{-1/2}\big(D_p - pp^{\top}\big)D_p^{-1/2} \;=\; I_K - \sqrt{p}\,\sqrt{p}^{\,\top}, \qquad \sqrt{p} := (\sqrt{p_1}, \dots, \sqrt{p_K})^{\top},$$

and since $\|\sqrt{p}\|_2^2 = \sum_y p_y = 1$ this matrix is idempotent: it is the orthogonal projector onto the $(K-1)$-dimensional tangent space $\{u : \sum_y \sqrt{p_y}\, u_y = 0\}$, the image of $T_p\Delta^{K-1}$ under whitening. Hence in Fisher coordinates the QSG sampling noise is *exactly isotropic* on the tangent space, and its trace $K-1$ re-derives Proposition 16. This is a direct measurement prediction linking the local geometry of the likelihood to the finite sampling of QSG: the predicted decay is $\propto m^{-1}$ at fixed (α, N, K), $\propto N^{-2}$ at fixed (α, m, K), $\propto \alpha^2$, and linear in $K-1$.

Scope note (mandatory; what Proposition 16 is *not*). This is a second moment measured *at the base point*. It is not a theorem about the finite KL distance to the boundary of the simplex, and not a theorem about global absorption time. The time unit is one speaker–listener interaction, nothing else. And m is the number of independent categorical samples, *not* a token count of a natural-language message: m is a sample count, not a token count, and the two must not be conflated when the update is instantiated on an LLM channel.

Remark ("it moved" is not "it learned"). Zero-mean increments are not evidence of neutrality-free learning, because a *correct* Bayes posterior is also a martingale: $\mathbb{E}[\pi_{HM} \mid H] = \pi_H$ in the notation of §I.2. Mean preservation alone therefore cannot distinguish neutral drift from useful learning, and Proposition 16 must not be read as either. What separates the two cases is *where the displacement comes from*. In the Bayes case

$$I(Y; M \mid H) \;=\; \mathbb{E}\, D_{KL}\big(\pi_{HM} \,\|\, \pi_H\big),$$

Table 4: Exact multinomial enumeration for the declared QSG consensus configuration $K = 3$, $p = (0.2, 0.3, 0.5)$. The formula is $\alpha^2(K-1)/(mN^2)$; the CPU audit replays the complete saved finite grid. No Monte Carlo estimate is used in this table.

N	α	m	Enumeration	Formula	Absolute residual
4	0.3	1	0.011250	0.011250	0
4	0.3	2	0.005625	0.005625	0
4	0.3	5	0.002250	0.002250	4.3×10^{-19}
16	0.8	1	0.005000	0.005000	0
16	0.8	2	0.002500	0.002500	0
16	0.8	5	0.001000	0.001000	0

which is locally a Fisher quadratic displacement, so the second moment does carry information. That equality, however, does *not* transfer to an equal-sized displacement of a native prediction q^+: for q^+ one must additionally measure the receiver error ϵ^+ of §I.2, and only Corollary 3 then decides whether the motion was a gain. Accordingly we do not import any accuracy semantics into a naming game whose task is left undefined.

Relation to the finite operator propositions. Within this section's frame, the operator results of §3.1 are the special cases obtained by fixing the reception update to a pooling operator and adding a common evidence update: Propositions 1, 2 and 3 describe Π_C^e / Π_C^m composed with Π_E on the answer-distribution simplex, and Corollary 2 is the $\mathrm{Var}_w(Z_i) = 0$ cross-section of Proposition 2. Equation equation 3 is exactly the evidence-free ($\Pi_E = \mathrm{id}$, $Z_i \equiv \mathrm{const}$, hence $\chi \equiv 0$), two-agent, weight-$(1-\alpha, \alpha)$ instance of Π_C^m pre-composed with the quantiser Sample_m. We therefore *treat those results as special cases in §3.1* and keep them physically there with their existing numbering and labels untouched; this paragraph, and not a renumbering, records the nesting.

Finite enumeration check. The CPU measurement audit re-executes the supplementary script `sanity_checks.py` and compares all saved finite-check fields. Its QSG calculation enumerates every multinomial message and compares the Fisher moment with $\alpha^2(K-1)/(mN^2)$ at a consensus base point. This is an implementation check of the one-interaction formula, not a native convergence experiment.

I.4 From locally identified operators to a collective prediction

Probe states and one-step edge operators. Fix, before evaluation, a common probe battery and let $\xi_{i,t} \in \mathbb{R}^d$ be the resulting probe state of agent i at round t (the probe battery, not the raw history $H_{i,t}$, is what is held common across agents and conditions; its predictive sufficiency for future updates is itself an object of test, cf. the remark following Proposition 15). A proposed identification study would estimate, for each directed edge and within a declared local regime, the mean response and covariance of one communication event. We retain three things that a scalar distance destroys: the *direction* of the induced displacement, its dependence on the *receiver state*, and its dependence on the *model pair*. An edge is thus summarised by an operator, not by a number.

Linearisation in a local chart. Around a fixed reference trajectory, in a local chart, write $\delta\xi_t$ for the deviation of the (stacked) population probe state from that reference. The linearised model is

$$\delta\xi_{t+1} \;=\; A_t\,\delta\xi_t \;+\; B_t u_t \;+\; \eta_{t+1} \;+\; r_t, \tag{4}$$

where A_t collects the identified one-step edge operators along the active communication structure, u_t is an external intervention with input matrix B_t, η_{t+1} is process noise with conditional mean zero, and r_t absorbs nonlinearity, state insufficiency and model error. We keep r_t in the notation deliberately: it is the term that an evaluation must bound, not the term a derivation may drop silently.

Proposition 17 (Moment propagation under the linearised model)**.** *Let $\mathcal{F}_t$ be the filtration generated by $\delta\xi_{0:t}$, and assume in equation 4: (L1) the coefficients A_t, B_t and the interventions u_t are* deterministic *(known in advance, fixed before the trajectory is realised; in particular not functions of $\delta\xi_{0:t}$ or of any other random quantity correlated with $\delta\xi_t$); (L2) $\mathbb{E}[\eta_{t+1} \mid \mathcal{F}_t] = 0$ and $\mathrm{Cov}(\eta_{t+1} \mid \mathcal{F}_t) = \Sigma_t$, a fixed matrix; (L3) $r_t \equiv 0$. Assume additionally that $\mathbb{E}\|\delta\xi_0\|_2^2 < \infty$ and that each noise covariance matrix in (L2) is finite. For every finite horizon, the mean and covariance are*

then well-defined. Then $\mu_t := \mathbb{E}[\delta\xi_t]$ *and* $C_t := \mathrm{Cov}(\delta\xi_t)$ *satisfy*

$$\mu_{t+1} = A_t\mu_t + B_t u_t, \qquad C_{t+1} = A_t C_t A_t^\top + \Sigma_t.$$

(L1) is essential, not a measurability formality: if A_t is merely $\mathcal{F}_t$-measurable — e.g. A_t itself a function of $\delta\xi_t$ — then in general $\mathbb{E}[A_t\delta\xi_t] \neq A_t\,\mathbb{E}[\delta\xi_t]$ and the displayed recursion fails; a scalar counterexample is $\delta\xi_t \in \{-1, 1\}$ equiprobable with $A_t := \delta\xi_t$, $B_t u_t = 0$, which gives $\mu_{t+1} = 1 \neq 0 = A_t\mu_t$. Coefficients identified from data (e.g. fitted per edge) are therefore held fixed across the trajectories on which the recursion is evaluated, not re-estimated per realisation; an A_t that is allowed to react to the current $\delta\xi_t$ (closed-loop control) falls outside this proposition and instead requires the general, unconditional update $\mu_{t+1} = \mathbb{E}[A_t\delta\xi_t + B_t u_t]$, $C_{t+1} = \mathrm{Cov}(A_t\delta\xi_t + B_t u_t) + \Sigma_t$.

Proof deferred to Appendix I.

Status of the collective-prediction extension. Proposition 17 is a conditional mathematical statement, not a reported validation of locally identified native operators. A transfer study would fit operators on separate identification data, freeze them, and evaluate their predictions on held-out worlds and communication structures. No such refit-free transfer study or probe-state sufficiency test is reported here.

Prediction discrepancies must be separated into operator estimation error, state insufficiency, nonlinearity, and measurement error. The recursion alone does not identify which source dominates, and mean and covariance predictions do not establish fixation behavior.

Consistency with the QSG special case. Under the identification $\xi = \bar{p}$ at a consensus point, with no intervention ($u_t = 0$) and with $A_t = I$ — which is exactly the mean-preservation statement $\mathbb{E}[\Delta\bar{p} \mid p] = 0$ of Proposition 16 — the one-interaction process-noise covariance of equation 4 is $\Sigma_t = \alpha^2(mN^2)^{-1}(\mathrm{diag}(p) - pp^\top)$, whose Fisher trace is $\alpha^2(K-1)/(mN^2)$. The recursion then reads $C_{t+1} = C_t + \Sigma_t$, i.e. variance accumulates linearly in the number of interactions at a consensus point. We state this as a consistency check of the two parameterisations, not as an absorption-time result: the conditions $A_t = I$ and Σ_t constant hold at the base point only, and the scope note of §I.3 applies verbatim.

Scope note (mandatory; the linearised model is a hypothesis, not a premise). Whether a native LLM system admits an approximation of the form equation 4 at all is an *empirical hypothesis* to be tested, not an assumption of this section. We do *not* assume the same coefficients (A_t, B_t, Σ_t) across all LLMs or across all prompts; coefficients are identified per model pair and per condition, and transfer across them is a claim requiring its own evidence. Nor do we explain a general directed, heterogeneous population by a single graph-Laplacian eigenvalue: the spectral shortcut presupposes a symmetric, homogeneous, scalar-gain coupling that our per-edge operators are specifically constructed not to assume.

Remark (five distinct noise sources; do not merge them). The Σ_t above must not be estimated by whatever variance is easiest to measure. We keep separate: (i) transmission sampling noise (e.g. $\widehat{p}_S$ in equation 3); (ii) randomness of the inference generated on the receiver side; (iii) randomness of edge selection, i.e. which pair communicates at round t; (iv) measurement error of the probability estimator itself, which Lemma 1 quantifies at rate $O_P(M^{-1/2})$ in the number of probes M; and (v) sampling of the final answer. In particular, the variance obtained by repeatedly sampling only the final answer is *not* to be read as diffusion of the internal predictive state: (v) is a readout stage placed after the state, and the two are related by an information loss, not by an identity — the same gap that the `trace_readout_fisher` check of the supplementary script `sanity_checks.py` exhibits numerically, where readout Fisher information 0.0710 falls short of the trace Fisher information 0.2570 by 0.1859 (decomposition residual -2.8×10^{-17}). Reporting a per-edge Σ_t therefore requires stating which of (i)–(v) were held fixed by the design and which were averaged over.

J Deferred bridge and native-model material

J.1 Quantized-gossip bridge and zero preservation

The quantized-gossip bridge changes the listener transition rule by substituting log-linear for arithmetic pooling. Its support collisions and uniform fallback are parts of that complete kernel. We do

not use the archived absorption patterns to isolate a zero-preservation mechanism or to test pooling–Bayes commutation.

Normalization convention at support collisions. Sampled messages $y^{(m)}$ can have exact zero components, so the log-linear listener update uses the standard pooling convention $0 \cdot \log 0 := 0$ (implemented as a max-shifted softmax of the log-combination, never a floor or pseudocount); when speaker and listener supports become fully disjoint one-hot vectors on different labels the combination is degenerate (all weights collapse to zero) and we fall back to the uniform distribution for that single update. An audit of this exact update rule (`code/audits/qsg_fallback_and_alpha1_audit.py`, released with this submission) found this fallback fires in up to 16.6% of listener updates in the worst grid cell ($\alpha = 0.3$, $m = 1$, $N = 32$; 4.7% pooled over the full grid, and 0% whenever $\alpha = 1$), so it is not a negligible edge case at low α and small m. The uniform reset changes the transition kernel and can either increase or decrease population polarization, which is measured from the population mean rather than from the entropy of the updated listener alone. We therefore do not exclude the reset as a contributor to the observed dynamics. Separately, the same audit confirms that at $\alpha = 1$ the arithmetic and log-linear updates coincide *exactly* (bitwise, given identical random draws) across every (N, m) cell, as expected since both reduce to replacing the listener state with the sampled message.

The mechanism behind the reported absorption pattern is not isolated. When both pooling weights are positive and the normalizer is nonzero, log-linear pooling preserves zeros contributed by its inputs. A uniform reset can restore positive support. The observations describe the complete update rule, not an isolated zero-preservation mechanism. The original bridge experiment did not run a smoothed control. The separate saved control at `results/e9_qsg_smoothed/summary.json` is re-executed cell by cell in the CPU audit. Smoothing changes the complete transition kernel; it does not isolate reset effects with the other dynamics fixed. Neither experiment establishes a causal explanation of accelerated absorption or a conclusion about pooling–Bayes order. The executed notebook is released as `artifacts/notebooks/qsg_pooling_bridge_v0.ipynb`.

J.2 Withdrawn pilot evidence

The execution pilot lacks retained trial-level outputs. Its constant-valued plot cannot verify model execution, scoring, or a native order effect, so we omit the plot and its aggregate scientific claims.

Capacity-arm uncertainty. Agent measurements from the same trial are dependent. An interval formed from the standard deviation and count of agent rows does not quantify uncertainty across independent trial worlds. We therefore omit the summary-only capacity-arm interval figure.

An inferential comparison requires trial identifiers, an explicit world-level estimand, and resampling of independent worlds while preserving within-world agent dependence. The summary-only comparison is not used to infer an effect of heterogeneous versus homogeneous composition.

J.3 Capacity-dependent convergence: a conditional scale-invariance lemma

A separate question is whether a single capacity-dependent update rate $\alpha_i = \alpha(t_i, t_j)$ (listener capacity t_i, speaker capacity t_j) can simultaneously give (i) a strictly monotone, declared convergence statistic in homogeneous populations as a function of shared capacity t, and (ii) the smallest expected one-step change for the highest-capacity group in heterogeneous populations.

Lemma 2 (Conditional scale-invariance; Lemma 3m′)**.** *Suppose capacity enters the population dynamics only through the update rate $\alpha_i = \alpha(t_i, t_j)$, with every other transition rule (initial belief distribution, message generation, observation noise, communication schedule) held fixed and independent of t. If α is zero-degree homogeneous in (t_i, t_j) (for example $t_j/(t_i + t_j)$) then $\alpha(t, t)$ is constant on the homogeneous diagonal. With the same initial law and all other transition rules fixed, the homogeneous-population trajectory law is therefore independent of shared capacity. Any convergence statistic defined from that law is consequently independent of shared capacity, so property (i) fails under this update-rate family.*

Proof sketch. Substituting $t_i = t_j = t$ into a zero-degree homogeneous $\alpha(t_i, t_j) = f(t_j/(t_i + t_j))$ gives $\alpha(t, t) = f(1/2)$ for every t. Since every other transition rule is held fixed and independent of t by assumption, the one-step transition kernel, and hence the trajectory law, of the homogeneous population are identical for every t. Therefore no convergence statistic defined from that law can be strictly monotone in t. The full statement and proof are given in the supplementary material. This is a necessary-condition result that holds only under the stated fixed-other-dynamics assumption; it says nothing about whether some other update-rate family can satisfy (i) and (ii) simultaneously. □

Withdrawn claim. An earlier draft of this result additionally claimed that a purely absolute, self-capacity rate $\alpha_i = g(t_i)$ (g strictly increasing, independent of the interaction partner) also cannot deliver property (ii), and concluded that no single-mechanism rate can satisfy both (i) and (ii) at once ("Theorem 3m, two-mechanism impossibility"). This claim is false; a closed-form counterexample, reproduced on CPU in the supplementary code, is as follows: for $g(t) = 1/4 + t/(2(1+t))$, a 3-agent heterogeneous population with $x = (0.2, 0.8, 0.5)$, $t = (1, 2, 3)$ gives the *highest*-capacity agent the *smallest* average one-step displacement – property (ii) holds – while the same rate applied to a homogeneous population gives an exact contraction coefficient $E[V_{r+1} \mid x_r] = (1 - \alpha + \frac{2}{3}\alpha^2)V_r$ that is strictly decreasing in t through $\alpha = g(t)$ – property (i) also holds. The stated absolute-rate counterexample satisfies both properties while keeping the other transition rules fixed. It lies outside the zero-degree homogeneous rate family, not outside the fixed-other-dynamics premise. The conditional scale-invariance lemma therefore places no restriction on this counterexample or on other absolute-rate designs.

Scope. The lemma excludes capacity-dependent homogeneous dynamics only within the zero-degree homogeneous rate family and under the stated fixed-other-dynamics assumptions. An absolute self-capacity rate is outside that rate family. The lemma places no restriction on whether such a rate, or another update design, can satisfy both properties.

The native capacity plots and hypothesis-test summaries are omitted. Their saved shift uses half the symmetric KL, whereas the former text called it the full symmetric KL. Neither quantity measures the declared homogeneous convergence statistic or tests the lemma's premises.

J.4 Stage C adoption structure and Stage D HotpotQA external validity

We omit the former Stage C capacity/adoption figure. The supplied plotting path does not establish the declared real-versus-control contrast for every panel, its displayed counts mix agent rows with trial units, and the recorded `adopt` field does not average over all same-capacity agent-index pairs. Consequently, the available receipts do not support a real-minus-sham adoption grid, an independent-trial confidence interval, or a population-wide adoption claim. A revised analysis would have to preserve matched worlds, report world, trial, and agent-row counts separately, and construct the control and pair population explicitly.

Status of the HotpotQA comparison. The revised Stage D receipt set contains 1,600 completed HotpotQA four-choice trials: native split-support (600), native hidden-profile (400), explicit arithmetic/geometric pooling (400), and two null shards (200). All eight full shards have a `status=valid` summary and their stored row counts sum to 1,600. For the version used by these receipts, order A is evidence-then-pooling and order B is pooling-then-evidence. The current runner constructs the second prompt using the first operation's state. The test in `code/stageD/test_i44_state_passing.py` verifies a dependency of that prompt constructor. Recorded source hashes differ from the current runner, so this test does not verify historical state hand-off in the receipts. The CPU audit independently re-scores the saved final beliefs and checks question, arm, order and manifest consistency. Its stored `dx` uses a denominator floor and is not exact symmetric KL; activation counts and row-mass errors are recorded in the audit. Prompt evaluation counts are available, but token expenditure is not.

We withhold an external-validity figure and an accuracy-order conclusion. The split-support, explicit, and null shards use hp-low records as a proxy for a generic split-support set, because the corresponding non-hidden-profile dataset was not available. Moreover, the received native prompts implement language-model updates rather than an analytically known common likelihood; the exact e-/m-pooling commutation theorem therefore does not transfer to these rows. The 100-question

Stage D shard	trials	accuracy (A)	accuracy (B)
native split-support	600	0.823	0.847
native hidden-profile	400	0.885	0.950
explicit m/e	400	0.895	0.935
null (different null types)	200	0.760	0.930

Table 5: Re-derived Stage D receipt accuracies. The two null entries are zero-evidence (A) and Takase prompt-order (B), so they are not an A/B operator comparison. Accuracy is reported only as an auxiliary outcome.

CPU measurement check	Re-derived result and scope
E4 copy-law decoder	$G = 0.0$; incompatible zero-support decoder has infinite loss.
E5 synthetic rate sweep	Time is pair draws; 0 censored trajectories on the saved grid.
Stage C stored readouts	840 trials; dx is half symmetric KL in nats.
Stage D stored readouts	1600 trials; accuracy re-derived and current prompt-construction test passes.
Uniform reset	$U : 5/9 \to 17/27$ in the saved counterexample.
QSG smoothed control	Existing control re-runs from its saved configuration; no isolated reset mechanism follows.

Table 6: CPU checks of recorded measurements. Identity and prompt checks are implementation checks; none calibrates native information or identifies a native pooling operator. Details: `results/aris_fix/r90_MEASUREMENT.json`.

per-condition records are retained as an stored prediction result, with raw rows and manifests in `artifacts/receipts/stageD_i45_shard*/`. A confirmatory benchmark comparison must regenerate the generic allocation, freeze the inference budget and prompt version, and evaluate the pre-specified paired statistic against both nulls.

Measurement-contract audit. The CPU audit separately recomputes decoder losses, synthetic pair-draw hitting times, stored native readouts, and quantized-gossip controls. The unit-test identities and conditional mathematical extensions are not evidence of native information calibration or operator transfer.

J.5 Proofs for Section 3

Proposition 18 (Verifier-limit coefficient)**.** *Fix strictly positive priors and a target label $y^\star$. Let $L_\varepsilon(y^\star) = 1$ and $L_\varepsilon(y) = \varepsilon$ for every other label, with $0 < \varepsilon \le 1$. Write $\pi_i = p_i(y^\star)$ and $\bar\pi = \sum_i w_i\pi_i$, and define*

$$a(y) = \sum_i w_i p_i(y)/\pi_i, \qquad b(y) = \sum_i w_i p_i(y)/\bar\pi,$$

$$\kappa = \sum_y (a(y) - b(y)) \log(a(y)/b(y)).$$

For arithmetic pooling,

$$\chi(p; L_\varepsilon) = \varepsilon\kappa + O(\varepsilon^2).$$

The coefficient κ is positive if and only if the π_i are not all equal. If they are all equal, the gap is zero for every $0 < \varepsilon \le 1$. Logarithmic pooling has zero gap throughout this interval. At $\varepsilon = 1$, the likelihood is constant and the arithmetic gap is zero regardless of the priors.

Proof. For an off-target label, the two arithmetic output masses are $\varepsilon a(y)+O(\varepsilon^2)$ and $\varepsilon b(y)+O(\varepsilon^2)$. Their symmetric-KL contributions give the stated coefficient. The target-label contribution is $O(\varepsilon^2)$. At the target label, both coefficients equal one, so including that label in the sum defining κ adds zero.

Every summand defining κ is nonnegative, and

$$\sum_y (a(y) - b(y)) = \sum_i w_i/\pi_i - 1/\bar{\pi}.$$

Strict convexity of the reciprocal makes this expression positive when the π_i differ, so some summand in κ is positive. When the π_i coincide, all evidence normalizers coincide for every ε, and Corollary 1 applies. □

For unequal target-label prior masses, the gap vanishes at the two endpoints for different reasons: the verifier limit concentrates both outputs on the target, whereas the constant-likelihood endpoint makes the update uninformative. Near the latter endpoint, the normalizers approach equality but need not coincide before the endpoint. The priors are fixed before taking either limit.

Proof of Proposition 1. For every y,

$$\Pi_C^e(\Pi_E p_{1:N})(y) \propto \prod_i [p_i(y)L(y)/Z_i]^{w_i} \propto L(y) \prod_i p_i(y)^{w_i} \propto \Pi_E \Pi_C^e(p)(y),$$

where the second proportionality uses $\sum_i w_i = 1$. Both sides are normalized distributions with the same kernel, and hence are equal.

Proof of Proposition 2. The update–then–pool expression is definitional. Writing $\bar{Z} = \sum_i w_i Z_i$, the other order is

$$\Pi_E \Pi_C^m p(y) = \frac{\sum_i w_i p_i(y) L(y)}{\bar{Z}} = \sum_i \frac{w_i Z_i}{\bar{Z}} \frac{p_i(y)L(y)}{Z_i} = \sum_i \tilde{w}_i q_i(y).$$

Proof of Corollary 1. Let $A = \sum_i w_i q_i$ and $B = \sum_i \tilde{w}_i q_i$. Since $\mathbb{E}_{q_i}[1/L] = 1/Z_i$,

$$\mathbb{E}_A[1/L] - \mathbb{E}_B[1/L] = \sum_i \frac{w_i}{Z_i} - \frac{1}{\bar{Z}} = \frac{1}{\bar{Z}^2} \sum_i \frac{w_i (Z_i - \bar{Z})^2}{Z_i}.$$

Thus $A = B$ implies $Z_i = \bar{Z}$ for every i; the converse follows from $\tilde{w}_i = w_i Z_i / \bar{Z}$. Symmetric KL vanishes exactly when $A = B$.

Proof of Proposition 3. Put $s(y) = B(y)/A(y)$. Directly, $\chi = \sum_y A(y)(s(y) - 1)^2 \psi(s(y))$. Moreover $s(y)$ is a convex combination of $Z_i/\bar{Z}$, with weights $w_i q_i(y)/A(y)$, and therefore $r_{\min} \le s(y) \le r_{\max}$. Since ψ is positive and decreasing, $\psi(r_{\max}) \le \psi(s(y)) \le \psi(r_{\min})$ for every y, so

$$\psi(r_{\max}) \sum_y \frac{(B(y) - A(y))^2}{A(y)} \;\le\; \chi \;\le\; \psi(r_{\min}) \sum_y \frac{(B(y) - A(y))^2}{A(y)},$$

i.e. $\psi(r_{\max}) d^\top F(w) d \le \chi \le \psi(r_{\min}) d^\top F(w) d$.

For the J/R_L certificate, apply the elementary bound $|\mathbb{E}_A f - \mathbb{E}_B f| \le \operatorname{osc}(f)\, \mathrm{TV}(A, B)$ with $f = 1/L$: $\operatorname{osc}(1/L) = R_L$ and $\mathbb{E}_A[1/L] - \mathbb{E}_B[1/L] = J$ (Corollary 1's computation), giving $\mathrm{TV}(A, B) \ge |J|/R_L$. Pinsker's inequality gives $D_{KL}(A\|B) \ge 2\,\mathrm{TV}(A, B)^2$ and, symmetrically, $D_{KL}(B\|A) \ge 2\,\mathrm{TV}(A, B)^2$; summing, $\chi = D_{KL}(A\|B) + D_{KL}(B\|A) \ge 4\,\mathrm{TV}(A, B)^2 \ge 4(J/R_L)^2$. This bound is invariant under $L \mapsto cL$ since J and R_L both scale as $1/c$.

Proof of Proposition 4. By Bayes' rule, $q_i = \Pi_E p_i = p_0(1 + t a_i)L/Z_i$ with $Z_i = \mathbb{E}_{p_0}[(1 + t a_i)L] = \bar{Z}_0(1 + t m_i)$ where $\bar{Z}_0 := \mathbb{E}_{p_0}[L]$ and $m_i = \mathbb{E}_{q_0}[a_i]$ (using $q_0 = p_0 L/\bar{Z}_0$); hence

$$q_i = q_0 \frac{1 + t a_i}{1 + t m_i} = q_0\{1 + t(a_i - m_i) + O(t^2)\},$$

not the naive $q_0(1 + t a_i + O(t^2))$ (which omits the normalizer shift m_i and is generally wrong at first order in $a_i - \mathbb{E}_{q_0}[a_i]$ terms once weighted and summed across i). Expand both pooled posteriors

$A = \sum_i w_i q_i$, $B = \sum_i \tilde{w}_i q_i$ around q_0 using this corrected form. Their first-order terms in t cancel (both reduce to q_0 at $t = 0$, and the $O(t)$ coefficients agree once the correct expansion

$$\tilde{w}_i = \frac{w_i Z_i}{\bar{Z}} = w_i \frac{1 + tm_i}{1 + t\bar{m}} = w_i\{1 + t(m_i - \bar{m}) + O(t^2)\}, \qquad \bar{m} := \sum_j w_j m_j,$$

is substituted; the normalizer-shift term $-t\bar{m}$ is required here for the same reason as for q_i above, and the naive $\tilde{w}_i = w_i(1 + tm_i) + O(t^2)$ omits it and is generally wrong at first order), leaving $A - B = t^2 q_0 H + O(t^3 q_0)$ with H as stated. The local expansion of symmetric KL is quadratic in the distributional difference, giving $\chi(p) = t^4 \mathbb{E}_{q_0}[H^2] + O(t^5)$. This is $\Theta(t^4)$ exactly when $\mathbb{E}_{q_0}[H^2] > 0$; when all m_i coincide, $H \equiv 0$ by inspection and the claim degenerates to $\chi(p) \equiv 0$ along that family (consistent with Corollary 1, since equal $m_i = \mathbb{E}_{q_0}[a_i]$ to first order corresponds to equal Z_i to first order).

Proof of Corollary 2. Substituting $Z_i = \bar{Z}$ into Proposition 2 gives $\tilde{w}_i = w_i$ for every i, hence the two m-pooling orders are identical and their symmetric KL is zero.

Verifier-limit proof. The coefficient expansion and its exact nondegeneracy condition are proved with Proposition 18. No uniform expansion over priors approaching the simplex boundary is asserted.

J.6 Proofs for Section 3

Proof of Proposition 12. By the chain rule, $P(\tau \mid x) = \prod_k P(z_k \mid z_{<k}, x)$, so $\log \frac{P(\tau|x)}{Q(\tau|x)} = \sum_k \log \frac{P(z_k|z_{<k},x)}{Q(z_k|z_{<k},x)}$. Taking the $P(\tau \mid x)$-expectation and applying the tower property (the k-th summand depends only on $z_{\le k}$) gives $D_{KL}(P\|Q) = \sum_k \mathbb{E}_{P(z_{<k})} D_{KL}\big(P(z_k \mid z_{<k}) \| Q(z_k \mid z_{<k})\big) = \sum_k c_k(P, Q)$.

Proof of Proposition 13. Standard KL chain-rule/data-processing decomposition applied to $D_{KL}(P\|Q)$ and $D_{KL}(Q\|P)$ separately (conditioning on Y), then summing; each conditional-KL term is ≥ 0.

Proof of Theorem 3. Minimal construction, $K = 2$ slots (slot 1 = private, slot 2 = public). Reference Q: $P_Q(z_1) = \text{Bernoulli}(1/2)$, $P_Q(z_2 \mid z_1) = \text{Bernoulli}(1/2)$ (independent of z_1). *System A*: $P_A(z_1) = \text{Bernoulli}(p)$, $p \neq 1/2$; $P_A(z_2 \mid z_1) = P_Q(z_2 \mid z_1)$ (identical public channel, so $c_2(P_A, Q) = 0$ exactly). Then $D_{KL}(P_A\|Q) = c_1(P_A, Q) = D_{KL}(\text{Bern}(p)\|\text{Bern}(1/2)) =: \delta$, $\text{Contrib}_{\text{pub}}(A) = 0$. *System B*: $P_B(z_1) = P_Q(z_1)$ (identical private slot, so $c_1(P_B, Q) = 0$ exactly); $P_B(z_2 \mid z_1) = \text{Bernoulli}(p)$ (independent of z_1, same p). Then $D_{KL}(P_B\|Q) = c_2(P_B, Q) = \delta$ (same functional form regardless of slot role), $\text{Contrib}_{\text{pub}}(B) = \delta$. Thus $D_{KL}(P_A\|Q) = D_{KL}(P_B\|Q) = \delta$ exactly, but $\text{Contrib}_{\text{pub}}(A) = 0 \neq \delta = \text{Contrib}_{\text{pub}}(B)$ (numeric instance: $p = 0.3 \Rightarrow \delta \approx 0.0823$ nats). Embedding into a general N-agent, T-round, multi-role slot set (setting $P = Q$ on all other slots) preserves the argument.

Proof of Lemma 1. Normalization gives $\mathbb{E}_{\text{ref}}[\varepsilon_S] = \mathbb{E}_{\text{ref}}[\varepsilon_{S'}] = 0$ exactly; let $\delta := \varepsilon_S - \varepsilon_{S'}$, so $\mathbb{E}_{\text{ref}}[\delta] = 0$ exactly. Expanding $\Delta = \delta - \frac{1}{2}(\varepsilon_S^2 - \varepsilon_{S'}^2) + O(\varepsilon^3)$ and writing $D_{KL}(P_S\|P_{S'}) = \mathbb{E}_{\text{ref}}[(1+\varepsilon_S)\Delta] = \mathbb{E}_{\text{ref}}[\Delta] + \mathbb{E}_{\text{ref}}[\varepsilon_S \Delta]$, a direct second-order computation gives $D_{KL}(P_S\|P_{S'}) = \frac{1}{2}\mathbb{E}_{\text{ref}}[\delta^2] + O(\varepsilon^3) = \frac{1}{2}\text{Var}_{\text{ref}}(\Delta) + O(\varepsilon^3)$. By (M3)+(M4), $M^{-1}\|\omega_S - \omega_{S'}\|^2$ is the sample variance of $\Delta(y_1), \ldots, \Delta(y_M)$, consistent for $\text{Var}_{\text{ref}}(\Delta)$ at CLT rate $O_P(M^{-1/2})$. Combining yields the claim.

J.7 Proofs for Section I.2–I.4

Proof of Proposition 14. Split the log-ratio, $\log \frac{\pi_{HE}}{q} = \log \frac{\pi_{HE}}{\pi_{HM}} + \log \frac{\pi_{HM}}{q}$, and take the expectation under the true joint law $P(H, E, M, Y)$. Under equation 2 the first term is $\mathbb{E} \log \frac{P(Y|H,E)}{P(Y|H,M)} = \mathbb{E} \log \frac{P(Y|H,E,M)}{P(Y|H,M)} = I(Y; E \mid H, M)$. The second term depends on Y only through (H, M)-measurable quantities, so the tower property gives $\mathbb{E}_{HM} D_{KL}(\pi_{HM}\|q)$. Adding the two yields the

identity. For the minimiser, consider the family $Q_q(H, E, M, Y) = P(H, E, M)\, q(Y \mid H, M)$ and minimise the forward KL; the first term is q-free and the second is a nonnegative KL vanishing iff $q = \pi_{HM}$ a.s.

Proof of Corollary 3. Write $G = \mathbb{E}\log\frac{q^+}{\pi_{HM}} + \mathbb{E}\log\frac{\pi_{HM}}{\pi_H} + \mathbb{E}\log\frac{\pi_H}{q^-}$. The middle term is $I(Y; M \mid H)$ by definition. The first is $-\mathbb{E}D_{KL}(\pi_{HM}\|q^+) = -\epsilon^+$ and the third is $+\mathbb{E}D_{KL}(\pi_H\|q^-) = +\epsilon^-$ (the sign flip is the standard cross-entropy/KL identity, using that each expectation is taken under the true law). This gives the first equality. For the second, the chain rule under equation 2 gives $I(Y; M \mid H) = I(Y; E \mid H) - I(Y; E \mid H, M)$, since $I(Y; E, M \mid H) = I(Y; E \mid H)$ when $Y \perp M \mid (H, E)$.

Proof of Proposition 15. X_{3-i} is uniform and independent of X_i, so conditionally on $X_i = x$ we have $Y = x \oplus X_{3-i}$, a uniform bit; this holds for both values of x and both i, giving the displayed equality. Therefore the map (agent private state) $\mapsto$ (answer distribution) is constant, and any message policy that factors through it is constant as well, yielding a receiver posterior that is uniform and an accuracy of $1/2$. Conversely, the policy that transmits X_i itself lets the receiver compute $Y = X_1 \oplus X_2$ with probability 1. The two accuracies, $1/2$ and 1, differ, so no answer-distribution-only model can realise the latter.

Proof of Corollary 4. Immediate from Corollary 3, whose derivation uses no property of $q^\pm$ beyond measurability with respect to (H, M) and H respectively, together with $\epsilon^\pm_{\mathrm{dec}} \geq 0$.

Proof of Corollary 5. Apply Hoeffding's inequality to the i.i.d. bounded summands $X_i := \log d^+(Y_i \mid H_i, M_i) - \log d^-(Y_i \mid H_i) \in [-\beta, \beta]$, whose mean is $\widehat{I}_{\mathrm{dec}}$ by definition. Each X_i ranges over an interval of length 2β, so Hoeffding's inequality gives, for $t > 0$, $P\big(|\widehat{I}^{(n)}_{\mathrm{dec}} - \widehat{I}_{\mathrm{dec}}| \geq t\big) \leq 2\exp\big(-2nt^2/(2\beta)^2\big) = 2\exp\big(-nt^2/(2\beta^2)\big)$ (the union bound over the two one-sided deviations contributes the factor 2). Setting the right-hand side to γ and solving for t gives the stated two-sided concentration at confidence $1-\gamma$ with $t = \rho_n(\gamma) = \beta\sqrt{2\log(2/\gamma)/n}$; the length-$2\beta$ range, not β, is what fixes this constant. The enclosure follows by combining the resulting interval for $\widehat{I}_{\mathrm{dec}}$ with the population bound of Corollary 4, which is unaffected by the draw of the evaluation fold since $\epsilon^\pm_{\mathrm{dec}}$ are population quantities of the fitted decoder pair (d^+, d^-), not of the evaluation sample; the evaluation fold is disjoint from both decoder-fitting folds, so (d^+, d^-) is fixed conditional on which the (H_i, M_i, Y_i) are i.i.d.

Proof of Proposition 16. Only the listener moves, so $\Delta\bar{p} = N^{-1}(p_L^+ - p_L) = (\alpha/N)(\widehat{p}_S - p_L)$; at consensus $p_S = p_L = p$, hence $\Delta\bar{p} = (\alpha/N)(\widehat{p}_S - p)$. Since $\mathbb{E}[e_{Z_r}] = p$ we get $\mathbb{E}[\widehat{p}_S] = p$ and therefore $\mathbb{E}[\Delta\bar{p} \mid p] = 0$. Each e_{Z_r} is a single multinomial draw with $\mathrm{Cov}(e_{Z_r}) = D_p - pp^\top$, $D_p := \mathrm{diag}(p)$, and the m draws are i.i.d., so $\mathrm{Cov}(\widehat{p}_S) = m^{-1}(D_p - pp^\top)$ and $\mathrm{Cov}(\Delta\bar{p}) = (\alpha^2/N^2)\,\mathrm{Cov}(\widehat{p}_S) = \alpha^2(mN^2)^{-1}(D_p - pp^\top)$. For the Fisher norm, $g_p(u, u) = u^\top D_p^{-1} u$, so because the increment is centred, $\mathbb{E}\, g_p(\Delta\bar{p}, \Delta\bar{p}) = \mathbb{E}\,\mathrm{tr}\big(D_p^{-1}\,\Delta\bar{p}\,\Delta\bar{p}^\top\big) = \mathrm{tr}\big(D_p^{-1}\,\mathrm{Cov}(\Delta\bar{p})\big) = \frac{\alpha^2}{mN^2}\,\mathrm{tr}\big(D_p^{-1}(D_p - pp^\top)\big)$. Finally $\mathrm{tr}(D_p^{-1}(D_p - pp^\top)) = \mathrm{tr}(I_K) - \sum_y p_y^2/p_y = K - 1$, which gives the displayed formula.

Proof of Proposition 17. By (L1), A_t, B_t, u_t are deterministic (non-random), so taking $\mathbb{E}[\,\cdot \mid \mathcal{F}_t]$ in equation 4 and using (L2)–(L3) gives $\mathbb{E}[\delta\xi_{t+1} \mid \mathcal{F}_t] = A_t\delta\xi_t + B_t u_t$, pulling A_t, B_t, u_t outside the conditional expectation. (For $\mathcal{F}_t$-measurable, possibly random, coefficients this conditional factoring step is itself licensed under mild integrability, since $\mathbb{E}[A_t X \mid \mathcal{F}_t] = A_t\,\mathbb{E}[X \mid \mathcal{F}_t]$ holds whenever A_t is $\mathcal{F}_t$-measurable; what is *not* available for merely measurable A_t is factoring out of the *unconditional* expectation $\mathbb{E}[A_t X] = A_t\mathbb{E}[X]$, which fails in general. Determinism of A_t, B_t, u_t by (L1) is what licenses the unconditional factoring used below.) Taking a further unconditional expectation and using the tower property gives the mean recursion $\mu_{t+1} = \mathbb{E}[\mathbb{E}[\delta\xi_{t+1} \mid \mathcal{F}_t]] = A_t\mu_t + B_t u_t$. For the covariance, $\delta\xi_{t+1} - \mathbb{E}[\delta\xi_{t+1}] = A_t(\delta\xi_t - \mu_t) + \eta_{t+1}$ (again using that A_t, B_t, u_t are the same deterministic objects appearing in μ_{t+1}), and the cross term has zero expectation because $\mathbb{E}[(\delta\xi_t - \mu_t)\eta_{t+1}^\top] = \mathbb{E}[(\delta\xi_t - \mu_t)\,\mathbb{E}[\eta_{t+1} \mid \mathcal{F}_t]^\top] = 0$ by (L2) and the $\mathcal{F}_t$-measurability of $\delta\xi_t$. Hence, using that A_t is deterministic so that $A_t(\delta\xi_t - \mu_t)$ has covariance $A_t C_t A_t^\top$, $C_{t+1} = A_t C_t A_t^\top + \Sigma_t$.

Table 7: Corollary 3 verified exactly on the XOR channel (`sanity_results.json`, field `xor_communication_budget`; nats). Row 3 is the reference setting: $\log 2 - 0.5004 - 0.1398 = 0.0530$. Row 4 shows $G < 0$ despite positive $I(Y; M \mid H)$, i.e. a genuinely informative message that hurts because ϵ^+ dominates. Row 5 is the redundant-message control: $I(Y; M \mid H) = 0$ and both decoder errors are zero ($\epsilon^- = \epsilon^+ = 0$), so $G = 0$.

err	conf	$I(Y; E \mid H)$	$I(Y; E \mid H, M)$	$I(Y; M \mid H)$	ϵ^+	G
0.0	0.999	0.6931	0.0000	0.6931	0.0010	0.6921
0.2	0.800	0.6931	0.5004	0.1927	0.0000	0.1927
0.2	0.950	0.6931	0.5004	0.1927	0.1398	0.0530
0.2	0.100	0.6931	0.5004	0.1927	1.3627	−1.1700
0.0	0.950	0.6931	0.6931	0.0000	0.0000	0.0000

MAD-topology concept	Corresponding object here	Generalized / deliberately omitted
Communication topology and neighbor mixing	A declared pooling operator C over a fixed population (complete mixing is the m-pooling special case)	Generalizes the aggregation operation; omits graph spectrum, degree, and routing.
Round-by-round diversity / polarization	Order gap $\chi_C = D_{\mathrm{sym}}(CE_Lp, E_LCp)$ (readout-level) and trace-level attribution of the upper-bounding $J_\tau \geq \chi_C$	Generalizes from a diversity outcome to a distributional estimand; does not assert a collapse law.
Message quantization and boundary absorption	Algebraic support preservation with positive weights and a nonzero normalizer	The bridge audits the complete kernel, including resets; it does not isolate a causal zero-support mechanism.
Topology-conditioned empirical comparisons	Fixed common-evidence, fixed-weight operator intervention with a zero-evidence control	Isolates identifiability of composition order; omits performance ranking across networks.

Table 8: Connection to pj4 MAD-topology (§2). The correspondence is conceptual, not a transfer of empirical claims or datasets.

J.8 Sign of the no-pooling comparison (Remark 1)

Remark 1. *Consider the E1 observation model: uniform prior, x_1, x_2, x_L conditionally independent given θ, private channels with the same accuracy, and weights $w_1 = w_2 = \frac{1}{2}$. Let $t = P(\theta \mid x_1, x_2, x_L)$ be the oracle and $q_i = P(\theta \mid x_i, x_L)$ the individually updated posteriors. The no-pooling arm is q_1. Then*

$$\mathbb{E}\,\mathrm{CE}(t, \tfrac{1}{2}q_1 + \tfrac{1}{2}q_2) \leq \mathbb{E}\,\mathrm{CE}(t, q_1), \qquad \mathbb{E}\,\mathrm{CE}(t, \mathrm{norm}(q_1^{1/2} q_2^{1/2})) \leq \mathbb{E}\,\mathrm{CE}(t, q_1).$$

Proof. By exchangeability of the two private channels, $\mathbb{E}\,\mathrm{CE}(t, q_1) = \mathbb{E}\,\mathrm{CE}(t, q_2)$. Cross-entropy $\mathrm{CE}(t, \cdot)$ is convex in its second argument, so $\mathrm{CE}(t, \frac{1}{2}q_1 + \frac{1}{2}q_2) \leq \frac{1}{2}\mathrm{CE}(t, q_1) + \frac{1}{2}\mathrm{CE}(t, q_2)$ pointwise; taking expectations gives the first inequality. For the log-linear pool, with $Z = \sum_\theta q_1(\theta)^{1/2} q_2(\theta)^{1/2}$, $\mathrm{CE}(t, \mathrm{norm}(q_1^{1/2} q_2^{1/2})) = \frac{1}{2}\mathrm{CE}(t, q_1) + \frac{1}{2}\mathrm{CE}(t, q_2) + \log Z$, and $Z \leq 1$ by Cauchy–Schwarz, hence $\log Z \leq 0$. Taking expectations gives the second inequality. □

Both inequalities use only convexity or Cauchy–Schwarz and the symmetry of the design; they do not use the zero-set or the quartic classification. The arithmetic pool-then-update output equals $\sum_i \tilde{w}_i q_i$ with data-dependent weights $\tilde{w}_i = w_i Z_i / \bar{Z}$, which are not equal in general, so this argument does not cover it.

J.9 Supplementary tables and verification detail (moved from main text for space)

J.10 Native-operator approximation constants (Theorem 3n)

The main results concern ideal operators. For a native implementation, let $\widehat{C}, \widehat{U}$ be approximate pooling and updating maps, and write $\mathrm{TV}(p, p') = \max_i \mathrm{TV}(p_i, p'_i)$ for profiles. Suppose, on the arguments used below,

$$\mathrm{TV}(\widehat{C}q, Cq) \le \varepsilon_C, \qquad \mathrm{TV}(\widehat{U}q, Uq) \le \varepsilon_U.$$

Theorem 4 (Native/ideal TV propagation with explicit constants)**.** *Let C be either m- or e-pooling and let $U = E_L$. If $\mathrm{TV}(Cp, Cp') \le K_C \mathrm{TV}(p, p')$ and $\mathrm{TV}(Uq, Uq') \le K_U \mathrm{TV}(q, q')$, then*

$$\Big|\mathrm{TV}(\widehat{C}\widehat{U}p, \widehat{U}\widehat{C}p) - \mathrm{TV}(CUp, UCp)\Big| \le (1 + K_U)\varepsilon_C + (1 + K_C)\varepsilon_U.$$

For m-pooling one may take $K_C = 1$. For e-pooling, if every compared input probability is at least $\delta > 0$, one may take

$$K_C = 1/\delta.$$

For a finite strictly positive likelihood, one may take in both cases

$$K_U = 2\rho, \qquad \rho := \frac{\max_y L(y)}{\min_y L(y)}.$$

Thus the m-pooling instance is $(1+2\rho)\varepsilon_C + 2\varepsilon_U$, and the interior e-pooling instance is $(1+2\rho)\varepsilon_C + (1 + 1/\delta)\varepsilon_U$.

Verification. For m-pooling, the triangle inequality gives

$$\mathrm{TV}\left(\sum_i w_i p_i, \sum_i w_i p'_i\right) \le \sum_i w_i \mathrm{TV}(p_i, p'_i) \le \mathrm{TV}(p, p'),$$

which proves $K_C = 1$. For e-pooling, let $d = \max_i \sum_y |p'_i(y) - p_i(y)|/2 = \mathrm{TV}(p, p')$ and interpolate $p_i(s) = (1 - s)p_i + sp'_i$ for $0 \le s \le 1$. Every interpolated probability is at least δ. Let

$$h_s(y) = \sum_i w_i \log p_i(s)(y), \qquad q_s = C^e(p(s)).$$

Each coordinate difference between probability vectors is bounded by their total variation, so

$$|\dot{h}_s(y)| \le \sum_i w_i |p'_i(y) - p_i(y)|/\delta \le d/\delta.$$

Differentiating the normalized pool gives

$$\dot{q}_s(y) = q_s(y)\Big(\dot{h}_s(y) - \sum_z q_s(z)\dot{h}_s(z)\Big),$$

and hence

$$\sum_y |\dot{q}_s(y)|/2 \le d/\delta.$$

Integration over s gives $\mathrm{TV}(C^e p, C^e p') \le \mathrm{TV}(p, p')/\delta$. This is a deterministic stability bound, not an empirical estimate of native approximation error. For Bayes updating, adding and subtracting $q'(y)/Z(q)$ gives $\mathrm{TV}(Uq, Uq') \le 2 \max_y L(y)\, \mathrm{TV}(q, q')/Z(q)$; $Z(q) \ge \min_y L(y)$ yields $K_U = 2\rho$. Finally, insert the four approximation/Lipschitz hops between $\widehat{C}\widehat{U}p$, $C\widehat{U}p$, CUp, UCp, $U\widehat{C}p$, and $\widehat{U}\widehat{C}p$. The triangle inequality in both directions yields the displayed bound. These are deterministic TV constants, not empirical estimates of $\varepsilon_C, \varepsilon_U$.

ρ	Oracle CE	m-UP excess	m-PU excess	e-pool excess	Product excess
0	0.249611	0.085848	0.065717	0.072337	0
0.25	0.309916	0.042018	0.026920	0.031885	0.024283
0.5	0.348213	0.020195	0.010130	0.013440	0.070574
0.75	0.377344	0.007540	0.002507	0.004162	0.126032
1	0.401358	0	0	0	0.186606

Figure 17: **Independent copy-channel target evaluation.** Selected points from the complete published copy-channel grid are re-derived on CPU. Excess loss is KL from the independently conditioned target, in nats. Endpoint equalities are algebraic checks of the model; comparisons at intermediate dependence are separate finite-model quality evaluations.

Common likelihood versus equal normalizers. An earlier draft asserted that Tanaka's §3.6 common external field ε — an agent-independent weak bias on the speaker channel — could be embedded as "evidence shared identically by all agents" and would therefore force $Z_i \equiv \bar{Z}$. That inference is false and is retracted here. A common likelihood does not, by itself, give equal normalizers: e.g. $p_1 = (0.8, 0.2), p_2 = (0.2, 0.8), L = (2, 1)$ share the same L but $Z_1 = 1.8 \neq Z_2 = 1.2$. Concretely, Tanaka's field acts as $p_i \mapsto p_i e^h/(p_i e^h + 1 - p_i)$ with normalizer $Z_i = p_i e^h + 1 - p_i$; a common bias h does *not* make Z_i agent-independent, since Z_i still depends on p_i. The only construction we can honestly identify with the $\chi = 0$, equal-Z_i case of our heterogeneous-evidence model is the degenerate $L \equiv c$ case of the corollary above (zero informative evidence), not Tanaka's general common-field mechanism. We therefore do *not* claim that Tanaka's common-external-field model occupies the $\mathrm{Var}_w(Z_i) = 0$ cross-section of our model in general; the two constructions are related only in this zero-evidence special case.

K Additional finite baselines and evidence-dependence sensitivity

As an implementation-contract check on the finite-state model used above, we reproduced an independently supplied exact finite-state oracle baseline and extended it with a copy-channel sweep. For $K = 3$ with $q_{\mathrm{priv}} = q_{\mathrm{com}} = 0.8$, the excess log loss (nats) against the exact joint-law oracle is 0.0858 for arithmetic update-then-pool, 0.0657 for arithmetic pool-then-update, and 0.0723 for e-pooling (either order, matching Proposition 1); the independence-assuming product estimator, which coincides with the true posterior only because this model's channels are conditionally independent by construction, is zero to floating-point precision. MAP accuracy is 85.6% for the three pooling estimators and 91.2% for the product estimator (exact MAP ties are split uniformly). Exact rational tie sets give the same expected accuracy under uniform and lowest-index tie-breaking when averaged over the full label-symmetric law. The previously reported 0.06 discrepancy is reproduced by floating-point `argmax`, which breaks mathematical ties inconsistently; it is not an effect of the exact tie policy. Across the full $K \in \{3, 5\} \times q_{\mathrm{priv}}, q_{\mathrm{com}} \in \{0.6, 0.8, 0.95\}$ grid (18 cells), e-pooling's excess log loss exceeds arithmetic pool-then-update's in every cell (range 0.0009–0.0264 nats higher), i.e. this baseline's commutativity does not imply lower loss than the best pool-order-dependent estimator here.

We then varied evidence dependence while holding each channel's marginal accuracy fixed at $q = 0.8$: with probability ρ the two private observations are copied ($x_2 = x_1$) and otherwise drawn independently, where ρ is a world-mixing coefficient, not a Pearson correlation, and the oracle is recomputed from the resulting correlated joint law. As ρ increases from 0 to 1, arithmetic pool-then-update's excess loss falls from 0.0657 to ≈ 0 nats (redundant evidence is easier to aggregate correctly), while the loss of a product estimator that ignores the induced dependence rises from ≈ 0 to 0.1866 nats (Figure 17). This is a new finite synthetic model, not a claim about the native-LLM results reported above; near-zero entries are floating-point residuals rather than exact rounding. The complete copy-channel grid is re-derived in `results/aris_fix/r90_R2-E1.json.gz`, field `finite.e0_published_copy_grid`; its baseline-grid comparison also checks `code/stageA/lib/exact_oracle_baselines.py`.

MMLU heterogeneous-team contrast	Δ	Paired 95% interval
Full arithmetic	-0.0400	$[-0.0833, +0.0033]$
Full geometric	$+0.0033$	$[-0.0400, +0.0467]$
Partial arithmetic	-0.1967	$[-0.2567, -0.1367]$
Partial geometric	-0.1833	$[-0.2467, -0.1200]$

Table 9: Question-paired accuracy contrasts reaggregated from stored R2 predictions. Intervals use the exact percentile distribution of the ordinary paired bootstrap over questions and are pointwise within the recorded comparison family. These accuracy contrasts do not test the arithmetic order-gap theorem or match inference budgets.

L STORED BENCHMARK PREDICTIONS: MMLU / ARC-CHALLENGE

Setup. We evaluate three four-agent team configurations — one heterogeneous team (**het**: Qwen2.5-7B + Qwen2.5-3B + 2× Qwen2.5-1.5B, all Instruct) and two homogeneous teams (**hom-1.5B**: 4× Qwen2.5-1.5B; **hom-3B**: 4× Qwen2.5-3B) — across seven aggregation conditions on two benchmarks: MMLU and ARC-Challenge ($n = 300$ stored questions per (team, benchmark) cell; no seed replication). Conditions follow the taxonomy of Section 2: *(i) single* (no pooling), *(ii) sc@12*, *(iii) debate*, *(iv) full-pooling* (arithmetic and geometric variants), and *(iv) partial-observation pooling* (arithmetic and geometric). The saved predictions and gold labels support question-level accuracy re-aggregation. The runner uses label-token probabilities; sc@12 draws votes from a cached probability vector, whereas debate requests updated answers. MMLU partial prompts hide subsets of answer choices; ARC partial prompts distribute retrieved evidence. Saved rows do not retain the belief vectors, full realized observation allocations, or token ledger, so those runner definitions are not an end-to-end audit of inference. The stored `n_gen` field is not a realized-call or token budget. Accuracy intervals are Wilson intervals. Contrasts use the same question identifiers across conditions and teams, with pointwise percentile intervals from the exact question-paired bootstrap distribution. The comparison family is fixed in the audit output; no simultaneous coverage or multiplicity-adjusted significance is claimed. The historical paired-bootstrap RNG provenance is missing, so we replace its intervals with the fully specified re-aggregation.

Results. The re-derived accuracy contrasts are reported in Table 9; every stored condition is reported in Tables 14 and 13. They compare recorded predictions across information and aggregation conditions. The regenerated tables replace the unaudited plotting path.

Interpretation. The benchmark contrasts do not reverse two fixed probability operations while holding their inputs constant. They therefore do not validate the arithmetic order-gap theorem or its specified quartic regime.

Equality of aggregate accuracies does not establish equality of the underlying beliefs. Idempotence is an algebraic property of the declared pooling operator when identical beliefs are supplied; aggregate accuracy equality is not an independent test of that property.

The reported paired intervals concern their stated accuracy contrasts. They do not identify native pooling and updating operators, isolate an evidence-reweighting mechanism, or establish a matched-budget communication advantage.

The audit distinguishes retained measurements from historical artifacts; a saved summary or identity residual alone is not a publication audit.

Withdrawn native-message and map evidence. The e44 implementation constructs its receiver readout by mixing an oracle posterior with a prior; it does not measure a receiver response. Its saved fields named "bits" use natural logarithms and therefore contain nats. The re-derived accounting residual checks this finite construction only. We remove its native-gain figure and calibration claims. The old team-map report and the current vector/KL files describe different state collections; we remove both map panels. Neither pair correlations nor mean probe log-likelihood changes establish the measurement lemma or target-label gain. Complete input hashes, units, adoption sets and rejected uses are recorded in `results/aris_fix/r90_R2-E1.json`.

Fixed family	Analytic coefficient	Last-grid ratio
Quartic	0.00490052	1.00003302
Verifier, unequal target masses	0.59143482	0.99909396

Table 10: CPU re-derivation of the declared finite families. Ratios are $\chi/(t^4\mathbb{E}_{q_0}H^2)$ and $\chi/(\varepsilon\kappa)$ on the original grids. Decimal arithmetic independently checks the same inputs. The supplementary JSON separates raw gaps, zero controls, positivity, reallocation residuals, and numerical precision; it applies no clipping or display floor. These are implementation and asymptotic-formula checks.

E-B population size	Fixed-configuration slope range
$N = 2$	$[3.99992140, 4.00007871]$
$N = 4$	$[3.99999708, 4.00000306]$
$N = 8$	$[3.99996275, 4.00003750]$

Table 11: E-B sensitivity across the existing signed and scaled amplitude configurations. The range is over fixed configurations and carries no population-uncertainty interpretation. Slopes use the original smallest-t fit points, with gaps recomputed in Decimal arithmetic. The single-agent zero identity has no log-log slope.

Finite audit quantity	Cells	Maximum error
E1 all common numerical columns	18	2.22×10^{-15}
E0 oracle / policy metrics	152	4.44×10^{-16}
Exact tie-policy accuracy difference	152	3.33×10^{-16}

Table 12: CPU audit of fixed finite models. The E1 checks compare every stored numerical cell field against an independent joint-law implementation and freshly executed scalar Bayes code. E0 includes the stored audit grid and the published copy-channel grid. All label permutations preserve the exact rational tie sets; missing, duplicated, nonfinite, and altered inputs are rejected. Residuals are implementation checks, not native utility evidence.

Team / condition	n	Accuracy [Wilson 95%]	Δ	Paired 95% interval
het/i_single	300	0.8867 [0.8458,0.9178]	+0.0000	[+0.0000,+0.0000]
het/ii_sc12	300	0.8900 [0.8495,0.9206]	+0.0033	[+0.0000,+0.0100]
het/iii_debate	300	0.8333 [0.7870,0.8712]	-0.0533	[-0.0867,-0.0200]
het/iv_full_arith	300	0.8167 [0.7690,0.8564]	-0.0700	[-0.1100,-0.0333]
het/iv_full_geo	300	0.8433 [0.7979,0.8801]	-0.0433	[-0.0767,-0.0100]
het/iv_partial_arith	300	0.8333 [0.7870,0.8712]	-0.0533	[-0.0900,-0.0167]
het/iv_partial_geo	300	0.8533 [0.8088,0.8889]	-0.0333	[-0.0633,-0.0033]
hom-1.5B/i_single	300	0.7133 [0.6597,0.7616]	+0.0000	[+0.0000,+0.0000]
hom-1.5B/ii_sc12	300	0.7067 [0.6528,0.7553]	-0.0067	[-0.0167,+0.0000]
hom-1.5B/iii_debate	300	0.7000 [0.6459,0.7491]	-0.0133	[-0.0400,+0.0133]
hom-1.5B/iv_full_arith	300	0.7133 [0.6597,0.7616]	+0.0000	[+0.0000,+0.0000]
hom-1.5B/iv_full_geo	300	0.7133 [0.6597,0.7616]	+0.0000	[+0.0000,+0.0000]
hom-1.5B/iv_partial_arith	300	0.7133 [0.6597,0.7616]	+0.0000	[-0.0200,+0.0200]
hom-1.5B/iv_partial_geo	300	0.7133 [0.6597,0.7616]	+0.0000	[-0.0200,+0.0200]
hom-3B/i_single	300	0.8167 [0.7690,0.8564]	+0.0000	[+0.0000,+0.0000]
hom-3B/ii_sc12	300	0.8167 [0.7690,0.8564]	+0.0000	[+0.0000,+0.0000]
hom-3B/iii_debate	300	0.8100 [0.7618,0.8504]	-0.0067	[-0.0267,+0.0133]
hom-3B/iv_full_arith	300	0.8167 [0.7690,0.8564]	+0.0000	[+0.0000,+0.0000]
hom-3B/iv_full_geo	300	0.8167 [0.7690,0.8564]	+0.0000	[+0.0000,+0.0000]
hom-3B/iv_partial_arith	300	0.8267 [0.7798,0.8653]	+0.0100	[-0.0033,+0.0267]
hom-3B/iv_partial_geo	300	0.8267 [0.7798,0.8653]	+0.0100	[-0.0033,+0.0267]

Table 13: arc_c accuracy reaggregated from the same stored questions in every condition. Pointwise paired intervals are recomputed from the exact ordinary-bootstrap percentile distribution and replace the old Monte Carlo intervals. The finite accuracy contrasts do not measure the arithmetic order gap, equality of beliefs, or matched inference budgets.

Team / condition	n	Accuracy [Wilson 95%]	Δ	Paired 95% interval
het/i_single	300	0.6733 [0.6184,0.7239]	+0.0000	[+0.0000,+0.0000]
het/ii_sc12	300	0.6600 [0.6047,0.7113]	-0.0133	[-0.0267,-0.0033]
het/iii_debate	300	0.6300 [0.5740,0.6827]	-0.0433	[-0.0867,+0.0000]
het/iv_full_arith	300	0.6333 [0.5774,0.6859]	-0.0400	[-0.0833,+0.0033]
het/iv_full_geo	300	0.6767 [0.6218,0.7271]	+0.0033	[-0.0400,+0.0467]
het/iv_partial_arith	300	0.4767 [0.4208,0.5331]	-0.1967	[-0.2567,-0.1367]
het/iv_partial_geo	300	0.4900 [0.4339,0.5463]	-0.1833	[-0.2467,-0.1200]
hom-1.5B/i_single	300	0.5767 [0.5201,0.6313]	+0.0000	[+0.0000,+0.0000]
hom-1.5B/ii_sc12	300	0.5833 [0.5268,0.6377]	+0.0067	[-0.0067,+0.0200]
hom-1.5B/iii_debate	300	0.5867 [0.5302,0.6410]	+0.0100	[-0.0167,+0.0400]
hom-1.5B/iv_full_arith	300	0.5767 [0.5201,0.6313]	+0.0000	[+0.0000,+0.0000]
hom-1.5B/iv_full_geo	300	0.5767 [0.5201,0.6313]	+0.0000	[+0.0000,+0.0000]
hom-1.5B/iv_partial_arith	300	0.4400 [0.3849,0.4966]	-0.1367	[-0.1900,-0.0833]
hom-1.5B/iv_partial_geo	300	0.4267 [0.3720,0.4832]	-0.1500	[-0.2067,-0.0967]
hom-3B/i_single	300	0.6333 [0.5774,0.6859]	+0.0000	[+0.0000,+0.0000]
hom-3B/ii_sc12	300	0.6300 [0.5740,0.6827]	-0.0033	[-0.0100,+0.0000]
hom-3B/iii_debate	300	0.6300 [0.5740,0.6827]	-0.0033	[-0.0233,+0.0167]
hom-3B/iv_full_arith	300	0.6333 [0.5774,0.6859]	+0.0000	[+0.0000,+0.0000]
hom-3B/iv_full_geo	300	0.6333 [0.5774,0.6859]	+0.0000	[+0.0000,+0.0000]
hom-3B/iv_partial_arith	300	0.5567 [0.5001,0.6118]	-0.0767	[-0.1300,-0.0233]
hom-3B/iv_partial_geo	300	0.5467 [0.4901,0.6021]	-0.0867	[-0.1433,-0.0300]

Table 14: mmlu accuracy reaggregated from the same stored questions in every condition. Pointwise paired intervals are recomputed from the exact ordinary-bootstrap percentile distribution and replace the old Monte Carlo intervals. The finite accuracy contrasts do not measure the arithmetic order gap, equality of beliefs, or matched inference budgets.