The Ringelmann Effect in Multi-Agent LLM Systems: A Scaling Law for Effective Team Size

Blaž Bertalanič
Jožef Stefan Institute
Ljubljana, Slovenia
blaz.bertalanic@ijs.si

Carolina Fortuna
Jožef Stefan Institute
Ljubljana, Slovenia
carolina.fortuna@ijs.si

Abstract

Inference-time multi-agent LLM scaling lacks a shared unit: counting nominal agents conflates cost with independent evidence. We derive a two-parameter scaling law $R(N) = N_{\text{eff}}/N = 1/(1 + c(N-1)N^{-\beta})$ where the regime exponent β classifies the studied configuration into one of three asymptotic regimes: hard-ceiling at $1/c$ ($\beta = 0$), sublinear at N^{β}/c ($0 < \beta < 1$), or linear ($\beta \geq 1$), and a mean-field theorem predicts that peer count k and rounds τ during agent debate enter the dynamics only through their product $k\tau$. The law applies at two levels: answer diversity and correctness redundancy. We empirically validate the law on homogeneous, role-symmetric, fully-connected debate on short-horizon voting-aggregated tasks, and demonstrate initial generalization to heterogeneous ensembles, hierarchical teams without peer exchange, temperature sweeps, thinking models, and open-ended (HumanEval) tasks. Across 44 (model $\times$ task $\times$ condition) cells spanning peer debate, self-correction, random-noise placebo, three open-weight families (Qwen, Llama, Ministral) at scales from 7B to 32B, a closed-source frontier model (Gemini), a thinking model, and heterogeneous teams, the functional form fits every cell. As primary evidence we use held-out extrapolation: (c, β) fitted on $N \leq 5$ predict N_{eff} up to $N = 30$ at 5–12% mean relative error per family (median per-cell error 4.3%, median 95% bootstrap CI width 9.5%), outperforming power-law, logistic, and shape-matched two-parameter rational and stretched-exponential alternatives. On free-form math, dense peer influence shifts the answer-level regime from sublinear to hard-ceiling; correctness-level fits remain hard-ceiling throughout. Three findings have practical implications. *(i)* On MMLU-Hard, thirty densely communicating agents attain an answer-level effective team size of 1.12. *(ii)* A noise placebo tracks self-correction on free-form math across three open-weight families and at $4\times$ scale, and on Qwen2.5-7B at N=30 debate exceeds self-correction in team accuracy by at most 0.5 pp and the noise placebo by at most 1.7 pp. *(iii)* A single $N \leq 5$ pilot predicts the $N = 30$ structural ceiling, and within the configurations tested, interventions on initial inter-agent correlation lower c and raise the ceiling, whereas communication-mode interventions shift (c, β) without a significant team-accuracy gain.

1 Introduction

Multi-agent LLM systems are an active research direction for automating reasoning tasks [Du et al., 2024, Liang et al., 2024, Chan et al., 2024]. While pretraining scaling laws gave the field a shared unit (parameters, tokens, compute) for reasoning about when and how to scale models [Kaplan et al., 2020, Hoffmann et al., 2022], inference-time multi-agent scaling has no such unit. Some studies report accuracy gains from teams of N=10 agents [Du et al., 2024], whereas others report no gain at comparable team sizes [Wang et al., 2024, Smit et al., 2024, Choi et al., 2025a]; the results cannot

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

be compared on a common axis. The main contribution of this paper is a *measurement framework* inspired by psychosocial theory. It consists of a two-parameter scaling law (c, β) that, estimated from a small pilot, quantifies how many of the N nominal agents contribute independent evidence. The law applies at two levels: *answer diversity* (how distinct the team's answers are) and *correctness redundancy* (how correlated the binary correct/wrong indicators are). It also has two practical consequences. First, deployment planning reduces from a three-dimensional grid search over team size, peer count, and revision rounds to a single small pilot. Second, the estimated β indicates whether additional agents contribute independent evidence ($\beta \approx 0$ implies a hard ceiling, $\beta > 0$ implies that scaling still adds independent evidence), and candidate interventions (architectural diversity, sparse topology, mixture-of-agents (MoA) pipelines, verifier filtering) can be evaluated on the same (c, β) axis. The empirical validation is scoped to homogeneous, role-symmetric, fully-connected teams on short-horizon, voting-aggregated tasks with $N \leq 30$ and no tools. Our contributions are:

C1. (regime classifier) A new two-parameter scaling law $R(N)=1/(1+c(N-1)N^{-\beta})$ that classifies an agentic configuration into hard-ceiling ($\beta=0$), sublinear ($0<\beta<1$), or linear ($\beta\geq 1$) regimes. The law applies at both answer-diversity and correctness-redundancy levels, and parameters estimated on $N \leq 5$ extrapolate to $N=30$ at 5–12% mean relative error per model family (§3, §5.1).

C2. (falsifiable mechanism) A mean-field theorem on communication density (Thm. 1, §3.2) derives the scaling law from explicit conformity dynamics and yields the falsifiable prediction that peer count k and rounds τ enter only through their product $k\tau$. Empirically, configurations with equal $k\tau$ agree within $\Delta\rho \leq 0.052$ on both tasks (Table 3; full stratified analysis in App. H).

C3. (unification) The same form describes all evaluated configurations, spanning debate, self-correction, noise placebo, self-consistency, sparse communication, three open-weight model families, $4\times$ model scale, and a closed-source frontier model, with initial generalization to heterogeneous ensembles, hierarchical teams without peer exchange, temperature sweeps, thinking models, and open-ended tasks (Table 4), as well as to classical human group studies (App. B). Conditions previously treated as separate phenomena are different points in (c, β). On free-form math, peer interaction induces a regime crossing: debate is hard-ceiling ($\beta \to 0$), self-correction sublinear ($\beta \approx 0.25$, §5).

C4. (design-space collapse) Once $\beta \approx 0$, which holds for instruct-class models on bounded tasks, the primary lever on the ceiling within the tested regime is $\rho^{(0)}$ (initial inter-agent correlation): on Qwen2.5-7B, re-evaluation instead of debate changes c by at most 0.03 on multiple-choice tasks, and the noise placebo lowers c without an accuracy gain. The most effective intervention identified is architectural diversity. Accuracy is projected with the correlated-Condorcet approximation for majority voting, with errors of at most 1.9 pp on homogeneous and 7.2 pp on heterogeneous teams (§5.4, App. R.2).

2 Related work

Prior work either documents diminishing or non-monotone returns at fixed team sizes [Wang et al., 2024, Wynn et al., 2025, Chen et al., 2024] or characterizes individual conformity mechanisms [Sharma et al., 2024, Bellina et al., 2026] without combining them into a predictive framework. What has been missing is a regime classification that operates from a small-team pilot and bears a falsifiable mechanistic prediction. This paper provides both.

Multi-agent scaling. Multi-agent debate was proposed for scalable oversight [Irving et al., 2018, Khan et al., 2024] and inference-time scaling [Du et al., 2024, Liang et al., 2024, Chan et al., 2024], with Li et al. [2024] claiming monotone gains from sampling-and-voting. A growing body of work reports diminishing or negative returns at fixed team sizes without a predictive framework [Wang et al., 2024, Choi et al., 2025a, Yang et al., 2025b, Wu et al., 2025, Smit et al., 2024, Zhang et al., 2025, Kaesberg et al., 2025, Wynn et al., 2025], and Qian et al. [2025] reported logistic saturation past 1,000 agents. Kim et al. [2026] fit a predictive model across 260 configurations on agentic benchmarks. Whereas prior work *describes* saturation, we *classify* regimes and *predict* large-N behavior, linking saturation explicitly to inter-agent correlation. We share Wynn et al. [2025]'s conformity diagnosis but provide a two-parameter scaling law validated out-of-sample across three model families. Concurrent Yang et al. [2026] studies the same scaling bottleneck through diversity metrics. Our axis is defined by variance matching via the Kish design effect, and its predictive validity is tested out of sample: (c, β) fitted on $N \leq 5$ extrapolate to $N=30$ at $\leq 12\%$ mean relative

error across three open-weight families, outperforming power-law ($\geq 31\%$), logistic ($\geq 68\%$), and two-parameter rational (32%) and stretched-exponential (133%) forms (§5.1, App. D).

Mechanistic ingredients. Sycophancy [Perez et al., 2023, Sharma et al., 2024], conformity [Zhu et al., 2025, Bellina et al., 2026], identity-weighted belief updating [Choi et al., 2026], inter-agent sycophancy and belief entrenchment [Yao et al., 2025, Oh et al., 2026], and non-monotone majority-vote scaling [Chen et al., 2024] have been studied as separate phenomena, and Denisov-Blanch et al. [2026] show that consensus does not constitute verification. The Ringelmann framework represents these mechanisms as parameter shifts in (c, β): sycophancy and inter-agent sycophancy raise $\rho^{(0)}$, peer conformity raises α, identity weighting modulates α asymmetrically, and our two-rate decomposition (§3.2, App. I) extends Bellina et al. [2026] by separating learning from sycophancy rates. The scaling law translates each parameter shift into a regime prediction.

Baselines and classical foundations. Self-consistency [Wang et al., 2023] is the no-interaction baseline. Brown et al. [2024] establish scaling for repeated independent sampling, and Snell et al. [2025] for test-time compute more broadly: accuracy increases with compute at a measurable rate. Our claim is orthogonal. Inference-time scaling describes *throughput* (more samples improve accuracy under a given aggregator). The Ringelmann law describes *redundancy* (more agents under interaction contribute diminishing independent evidence under a regime-classified rate). Even independent samples are correlated ($\rho^{(0)} > 0$, e.g., self-consistency has $c = 0.93$ on MMLU-Hard), and communication shifts the system among the three regimes (C1, Prop. 1). Our self-correction control extends Madaan et al. [2023] to isolate re-evaluation from peer influence: the gap to debate is $\leq$3 pp while re-evaluation alone adds up to $+18$ pp. The framework builds on the Ringelmann effect [Ringelmann, 1913], the Kish design effect [Kish, 1965], and correlated-Condorcet theory [Ladha, 1992, Boland, 1989]. We contribute three new results: regime classification (Prop. 1), a hard-ceiling escape condition (Prop. 3), and a falsifiable $k\tau$-product theorem (Thm. 1).

3 Framework: a Ringelmann scaling law for multi-agent systems

3.1 A new two-parameter scaling law leading to a regime classifier (C1)

LLM agents conform under interaction [Sharma et al., 2024, Perez et al., 2023, Zhu et al., 2025], so their answers correlate rather than being independent. This is the standard statistical problem of clustered samples: when observations within a sample are correlated, the nominal sample size overstates the information content. The Kish design effect [Kish, 1965] provides the canonical correction, converting N correlated observations into an equivalent number of independent observations whose mean has the same variance:

$$N_{\text{eff}}(N) \;=\; N/(1 + (N-1)\rho_N), \tag{1}$$

where ρ_N is the equicorrelation-scale pairwise agreement at team size N (the convention is that all agent pairs are treated as equally correlated). At $\rho_N = 0$ (independent agents) $N_{\text{eff}}(N) = N$. As $\rho_N \to 1$ (full consensus) $N_{\text{eff}}(N) \to 1$, regardless of N. The Kish form is derived for binary correctness indicators (App. A.1). Applied to answer-level agreement it is a variance-matching definitional rescaling onto a common axis.

Definition 1 (Ringelmann Exponent)**.** *We model the correlation profile as* $\rho_N = cN^{-\beta}, \qquad c > 0,$ *and call* β *the* Ringelmann exponent *of the multi-agent system.*

Substituting the profile into Eq. 1 gives the Ringelmann efficiency curve:

$$R(N) \;:=\; N_{\text{eff}}(N)/N \;=\; 1/(1 + c(N{-}1)N^{-\beta}), \tag{2}$$

with two interpretable parameters: c sets the asymptotic correlation level (ceiling $1/c$ on N_{eff} when $\beta = 0$), and β controls how fast correlation falls with team size. Next, we classify what scaling behavior each region of the (c, β) plane produces. The asymptotic growth of N_{eff} is governed by the product $N \cdot \rho_N$ in Eq. 1, which measures the aggregate redundancy introduced by the team: linear scaling ($N_{\text{eff}} = \Theta(N)$) requires this product to stay bounded: $\rho_N = O(1/N)$. Any constant $\rho > 0$ instead drives $N \cdot \rho_N \to \infty$ and implies $N_{\text{eff}} \to 1/\rho$ (Proposition 2 in App. A.3). Under the power-law parameterization, the exponent β determines whether ρ_N decays faster or slower than $1/N$, partitioning the system into three discrete regimes:

Proposition 1 (Power-Law Ringelmann Regimes with proof in App. A.4.)**.** *Within the effective-team model, suppose $\rho_N = cN^{-\beta}$ with $c > 0$. Then*

$$N_{\mathit{eff}}(N) \sim \begin{cases} 1/c, & \beta = 0 \quad \textit{(}\text{hard ceiling}\textit{)}, \\ N^{\beta}/c, & 0 < \beta < 1 \quad \textit{(}\text{sublinear}\textit{)}, \\ \Theta(N), & \beta \geq 1 \quad \textit{(}\text{linear}\textit{)}. \end{cases}$$

The linear regime gives $N_{\mathit{eff}} \sim N/(1+c)$ at the boundary $\beta = 1$ and $N_{\mathit{eff}} \sim N$ for $\beta > 1$.

Interpretation of (c, β). c controls the ceiling level: at β=0, the asymptotic ceiling is $1/c$. β controls how ρ_N shrinks with N: β=0 means correlation is constant (hard ceiling), β=1 means it falls like $1/N$ and N_{eff} grows linearly, discounted by $1+c$, and $0<\beta<1$ is sublinear. Empirically one estimates (c, β) from $R(N)=N_{\text{eff}}(N)/N$ and asks whether an intervention lowers c, increases β, or both. Algorithm and worked example in App. C.8.

β is unconstrained. The definition places no upper bound on β, so all three regimes are admissible. No evaluated LLM configuration reaches $\beta \geq 1$ (unbounded refits of 38 cells (36 open-weight, 2 Gemini 3.1 Flash-Lite) give $\beta_{\text{free}} \in [0.00, 0.36]$, App. D.1). The linear regime therefore remains attainable in principle and constitutes a design target; it is not excluded by construction.

3.2 Falsifiable mechanism with explicit conformity dynamics (C2)

The preceding subsection takes ρ_N as given and derives what it implies for scaling. We now address the complementary question: *what determines ρ_N?* A simple mean-field closure, in the spirit of DeGroot [1974]'s classical consensus model, assumes that residual disagreement contracts by a constant factor per communication round. Writing τ for the number of post-communication updates and $d^{(\tau)} := 1 - \rho^{(\tau)}$, we assume $d^{(\tau+1)} \approx (1-\alpha)d^{(\tau)}$. This yields the approximation:

$$\rho_{\text{mf}}^{(\tau)} \approx 1 - (1 - \rho^{(0)})(1-\alpha)^{\tau} \tag{3}$$

Here $\rho^{(0)}$ is the pre-communication correlation and $\alpha \in [0, 1]$ is the fraction of remaining disagreement removed per round. The closure is agnostic to the specific conformity mechanism. For example, Choi et al. [2025b] show that LLM agents defer to numerically dominant and more capable peers, a specific instance of majority pressure, represented as a larger per-peer adoption rate α_1.

Thm. 1 derives this contraction under explicit conformity dynamics and finds that the effective α depends on peer count k as $\alpha = 1 - (1-\alpha_1)^k$, where α_1 is the per-modal-peer adoption probability. The single rate α averages over a learning rate α_l (wrong agents adopting correct modal peers) and a sycophancy rate α_s (correct agents adopting wrong modal peers). When α and $\rho^{(0)}$ do not depend on N, ρ_{mf} is N-independent, giving $\beta = 0$ (hard ceiling). Escaping the ceiling requires both $\alpha_N \to 0$ and $\rho_N^{(0)} \to 0$ as N grows (subscript N: team size), and neither alone suffices (Proposition 3).

Theorem 1 (Mean-Field Convergence and the $k\tau$ Product Structure)**.** *Consider $N \geq 2$ agents over discrete rounds. Coarsen each agent's answer into a binary indicator $\in \{$modal, non-modal$\}$, where "modal" denotes the most common team answer at round τ, and assume three conditions.* (A1) *Each non-modal agent observes k peers ($1 \leq k \leq N-1$) drawn uniformly without replacement from the $N-1$ other agents, adopting the modal answer with probability $\alpha_1 \in (0, 1)$ per modal peer, independently.* (A2) *Modal agents do not switch (an idealization, App. A.6 shows the $k\tau$ structure is robust to small modal-switching rates).* (A3) *The mode is stable across rounds.*

Let $d^{(\tau)} = 1 - \rho^{(\tau)}$ denote pairwise disagreement at round τ. For $d^{(\tau)} \ll 1$ (high agreement):

$$d^{(\tau+1)} = d^{(\tau)}\,(1-\alpha_1)^k + O\big((d^{(\tau)})^2\big) \tag{4}$$

and iterating yields:

$$\rho_{\text{mf}}(\tau, k) \approx 1 - (1 - \rho^{(0)})(1-\alpha_1)^{k\tau} \tag{5}$$

with a single effective conformity rate α_1.

Three consequences: *(i)* k and τ enter only through $k\tau$ (peer-round equivalence), *(ii)* fixed budget B=$k\tau$ is split-invariant to first order, *(iii)* on the pairwise agreement scale, $\rho^{(\tau)}$ is approximately N-independent for fixed $k \geq 1$ in the high-agreement regime, giving $\beta_{\text{agr}} \approx 0$. The entropy-scale exponent β_{ent} (fitted to the entropy-based $N_{\text{eff}}^{\text{ans}}$, §4) may differ on free-form tasks where wrong-answer mass spreads across distinct strings. Proof and scope conditions in App. A.6 (covering placebo conditions with $\alpha_1 \to 0$, self-consistency at τ=0, and answer- vs. pairwise-agreement scales).

4 Experimental methodology

In this section we distinguish the measured $\hat{\rho}_N$ from the model ρ_N from Section 3. Elsewhere in the paper, where context makes clear we are reporting a measurement, we drop the hat for brevity and write bare ρ or ρ_N. Other subscripts indicate qualifiers (eff, ICC, mf, cond), tabulated in App. A. We estimate (c, β) by fitting Eq. 2 to the measured efficiency N_{eff}/N (estimators below).

Measuring ρ_N. We use a hat to mark the empirical estimator computed from data: for a team of N agents, $\hat{\rho}_N := \frac{N \cdot A_{\text{pair}} - 1}{N-1}$, where A_{pair} is the probability that two randomly drawn agents give the same answer post-communication. The formula converts A_{pair} to a correlation on the same scale Eq. 1 uses. To empirically validate Eq. 1, we need to know how ρ_N depends on team size. Running the experiment at $N \in \{2, 3, 5, \ldots, 30\}$ produces an empirical sequence $\{\hat{\rho}_N\}$, to get the *measured correlation profile*. We model the underlying ρ_N as a power law, with classical precedent in declining-efficiency literature (App. B).

Three estimators of N_{eff}. *Answer level:* (a) entropy-derived $N_{\text{eff}}^{\text{ans}} = 2^{H(\mathbf{a})}$, with $H(\mathbf{a})$ the Shannon entropy in bits of the team's N answers, and (b) pairwise-agreement $N_{\text{eff}}^{\text{agr}} = N/(1 + (N-1)\hat{\rho}_N)$, the theorem's native variable. Both count identical answers as redundant. *Correctness level:* (c) $N_{\text{eff}}^{\text{corr}} = N/(1+(N-1)\rho_{\text{ICC}}^{\text{corr}})$ from the marginal intraclass correlation (ICC) of binary correct/wrong indicators. On free-form tasks $2^{H(\mathbf{a})}$ is a conservative upper bound (distinct wrong strings inflate entropy, App. F), so a hard ceiling found on it holds a fortiori on pairwise agreement. None of the three captures rationale-level independence (same answer, different reasoning).

Models, Tasks and Team Dynamics. The core identification model is Qwen2.5-7B-Instruct [Qwen et al., 2024] (full grid, N=1–30). Cross-model: Llama-3.1-8B [Llama Team, AI @ Meta, 2024] and Ministral-8B [Liu et al., 2026]. Within-family $4\times$ scale: Qwen2.5-32B-Instruct. Stronger reasoning: Qwen3-8B (thinking mode) [Yang et al., 2025a] and Gemini 3.1 Flash-Lite (minimal thinking) [Google DeepMind, 2026]. We consider the following representative multiple choice and free form tasks. *MMLU-Hard* [Hendrycks et al., 2021]: all 724 four-choice test items of MMLU's complex-reasoning subjects (college physics, college mathematics, formal logic, econometrics, professional accounting; bounded MCQA). *GSM-Hard* [Gao et al., 2023]: the 1,017 integer-answer word problems (free-form numeric answers, where wrong answers scatter across many distinct strings). *GPQA Diamond* [Rein et al., 2024]: 198 graduate-level four-choice science items (frontier-difficulty MCQA). We study team sizes $N \in \{1, 2, 3, 5, 7, 10, 15, 20, 30\}$, a fully connected topology, three rounds (R1 independent, R2–R3 post-debate, i.e. τ=2). Model-scope rationale: App. C.1.

Conditions. Three communication conditions per item: *debate* (peers see true prior answers), *self-correction* (own prior answer as "peer"), *noise* (peer answers drawn from a pre-generated pool of independent R1 outputs to other items, length-matched to debate). Self-correction and noise jointly isolate the contribution of genuine peer information: self-correction holds the prompt structure fixed while removing peer content, noise holds peer-text length and style fixed while removing topical relevance. The revision scaffold, instruction template, peer-snippet format and truncation, and answer format are identical across the three conditions. Only the content of the peer block varies (Table 7, App. C.2). Round 1 (τ=0) under majority voting is self-consistency.

Prompt format. All evaluations are zero-shot, with no in-context exemplars or chain-of-thought demonstrations. Agents respond in a structured RATIONALE/FINAL/CONF format (App. C.2) held constant across conditions, enabling automated answer extraction and confidence elicitation. Holding the format fixed across conditions ensures it does not confound pairwise comparisons.

Metrics. For each (model, task, condition, N) cell we report: (i) *team accuracy* under plurality voting at the final round, ties broken at random; (ii) the *measured pairwise agreement* ρ on the equicorrelation scale (Section 3); (iii) *effective team size* at two levels, answer-level $N_{\text{eff}}^{\text{ans}}$ as primary and correctness-level $N_{\text{eff}}^{\text{corr}}$ as complementary (Appendix E); (iv) the *Ringelmann parameters* (c, β) estimated by fitting $\rho_N = cN^{-\beta}$ to the measured profile across team sizes (algorithm in Appendix C.8); (v) *token cost*: prompt and output tokens, total and per effective agent (Appendix R.1). The two levels coincide on bounded MCQA and decouple on free-form math, where distinct wrong answers inflate entropy but not correctness redundancy.

Ablations. *Communication density* (k-sweep): peer count $k \in \{1, 2, 4, 9, 29\}$ and rounds $\tau \in \{0, \ldots, 5\}$ (τ=0: independent answers) at fixed N, evaluating the $k\tau$ product structure. *Hetero-*

geneous teams: equal-proportion mixes (Qwen2.5-7B / Llama-3.1-8B / Ministral-8B in 1:1:1 ratio) at $N \in \{3, 6, 9, 15, 30\}$, evaluating the block-exchangeable extension.

Statistical methodology at a glance. Agreement is reported on the equicorrelation scale ρ (§ 3). Confidence intervals use item-level bootstrap (≥ 300 iterations, 1,000 for the held-out extrapolation). We fit (c, β) by non-linear least squares on $R(N) = N_{\text{eff}}/N$ with $c \in [0, 20]$ and $\beta \in [0, 5]$, which admits the linear regime. Estimates match a $[0, 1]$-bounded fit to five decimals in all 38 refitted cells (App. D.1). The held-out audit fits $N \in \{2, 3, 5\}$ and extrapolates to $N \in \{7, \ldots, 30\}$. Protocol and significance tests: App. C.7. Generation parameters, answer extraction, hardware: App. C.

Table 1: Ringelmann regime estimates per condition (Qwen2.5-7B), $R(N) = 1/(1+c(N-1)N^{-\beta})$ fitted to observed N_{eff}/N. Brackets: 95% item-level bootstrap CIs. Held-out MRE: mean relative error when (c, β) are fitted on $N \in \{2, 3, 5\}$ and used to predict $N \in \{7, \ldots, 30\}$ (§5.1). Self-consistency uses round-1 answers only and is reported as a point estimate outside the held-out audit. Cross-family fits: Table 19; per-cell CIs: Table 9.

Task	Condition	c [95% CI]	β [95% CI]	Regime	Held-out MRE [95% CI]
MMLU-Hard	Self-consist.	.93	.000	hard ceiling	—
	Debate	.85 [.83, .86]	.000 [.00, .00]	hard ceiling	3.8% [1.8, 6.0]
	Self-corr.	.84 [.81, .88]	.011 [.00, .03]	hard ceiling	4.1% [0.5, 10.2]
	Noise	.65 [.64, .67]	.000 [.00, .00]	hard ceiling	6.3% [4.1, 10.2]
GSM-Hard	Self-consist.	.43	.249	sublinear	—
	Debate	.57 [.55, .61]	.014 [.00, .04]	hard ceiling	4.3% [2.3, 7.6]
	Self-corr.	.34 [.31, .36]	.252 [.22, .29]	sublinear	4.3% [1.5, 11.5]
	Noise	.33 [.31, .36]	.272 [.24, .30]	sublinear	2.5% [1.3, 10.9]
GPQA	Self-consist.	.86	.000	hard ceiling	—
	Debate	.81 [.77, .85]	.000 [.00, .01]	hard ceiling	6.0% [3.9, 22.4]
	Self-corr.	.78 [.73, .85]	.006 [.00, .04]	hard ceiling	1.2% [0.6, 12.1]
	Noise	.52 [.49, .59]	.000 [.00, .05]	hard ceiling	17.3% [5.6, 43.9]

5 Results

5.1 C1: Regime classification and small-N pilots predict large-N

On Qwen2.5-7B across three tasks (Table 1), three communication conditions (debate, self-correction, noise), and the self-consistency baseline, the same Ringelmann form $R(N) = 1/(1+c(N-1)N^{-\beta})$ applies, with parameter shifts that classify each configuration into a regime. Under dense debate, all three tasks lie in the hard-ceiling regime ($\beta \approx 0$, $c \in \{0.85, 0.57, 0.81\}$ on MMLU-Hard, GSM-Hard, GPQA respectively, asymptotic ceilings $1/c \approx 1.2$, 1.7, 1.2). On GSM-Hard, removing peer influence (self-correction, noise, self-consistency) shifts the regime to sublinear ($c \approx 0.33$–0.43, $\beta \approx 0.25$). MCQA tasks remain hard-ceiling across all conditions.

Primary evidence: held-out extrapolation. Because $R(N)$ is a smooth, monotone two-parameter curve, in-sample R^2 is weak evidence on its own, so we treat out-of-sample prediction as the primary test. We estimate parameters on $N \in \{2, 3, 5\}$ and predict $N \in \{7, 10, 15, 20, 30\}$ across 9 task×condition cells. The Ringelmann form attains 5.5% mean relative error, against 32.2% for a generic power law and 73.4% for logistic saturation (full audit table and extrapolation figure in App. D, Table 8 and Figure 5). The audit replicates on the other two open-weight families: Llama-3.1-8B preserves the ranking (Ringelmann 11.9%) and Ministral-8B reproduces it (Ringelmann 5.8%, 8 of 9 cells, App. D). Across the three open-weight families the Ringelmann mean relative error lies between 5.5% and 11.9%, whereas the logistic form yields 68.8–73.4%. Item-level bootstrap CIs (B=1000) give a median per-cell held-out error of 4.3% and a median 95% CI width of 9.5% (19/23 cells below 15%, Table 9). Against shape-matched two-parameter families over 37 cells, Ringelmann (6.8%, lowest error in 21/37 cells) outperforms rational $1/(1+sN^{\gamma})$ (32.2%) and stretched-exponential (132.8%) forms (Table 10). The Const-ρ Kish form (the β=0 special case) is competitive on hard-ceiling cells but fails on sublinear cells with $\beta > 0.2$; the additional parameter is therefore justified precisely where the regime structure requires it. As an external consistency check, the estimated c rank-correlates (Spearman $r_s = 0.80$) with an independently measured correctness-level ICC across 9 task–condition cells (App. E, Table 11), so c tracks a measurable quantity, not residual variance.

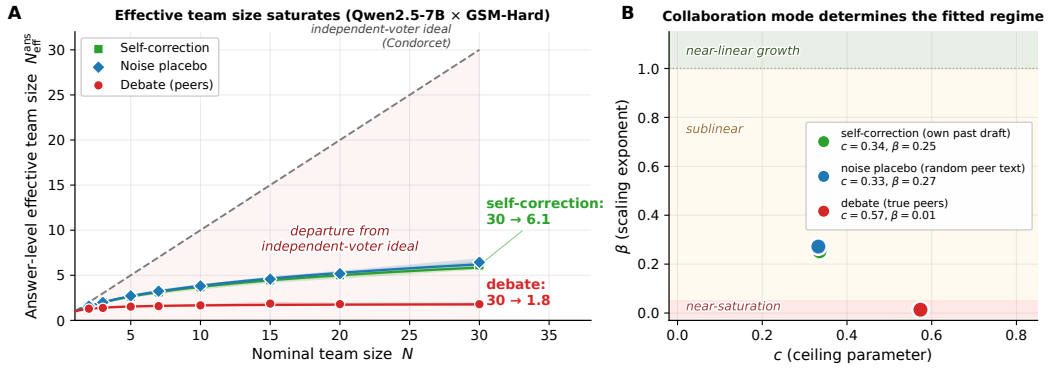


Figure 1: **Effective team size saturates and the regime is set by collaboration mode** (Qwen2.5-7B $\times$ GSM-Hard). **A:** Answer-level effective team size $N_{\text{eff}}^{\text{ans}}$ vs. nominal N. Self-correction and noise placebo grow sublinearly toward $N_{\text{eff}} \approx 6$ at N=30. Debate flattens at $N_{\text{eff}} \approx 1.8$. The dashed line marks the independent-voter ideal $N_{\text{eff}} = N$, and the shaded gap is the departure of the observed team from that ideal. **B:** Fitted Ringelmann parameters (c, β) for the three modes. Self-correction and noise lie in the sublinear band ($\beta \approx 0.25$); debate lies in the hard-ceiling band ($\beta \approx 0$).

Peer influence crosses an answer-level regime boundary on GSM-Hard. Figure 1 visualizes the crossing: noise and self-correction grow toward $N_{\text{eff}} \approx 6$ at N=30, while debate saturates at $N_{\text{eff}} \approx 1.8$. Self-consistency (round-1 only, no revision) also produces a sublinear fit ($c = 0.43$, $\beta = 0.25$). The crossing is specifically an answer-level phenomenon: at the correctness level, all GSM-Hard conditions are hard ceiling (App. E, Table 12). Cross-family replication on Llama-3.1-8B and Ministral-8B confirms the crossing is not Qwen-specific (App. J).

Conformity limits team-level gains. Agent-level transitions are net positive at every team size (+8 pp wrong $\rightarrow$ correct on GSM-Hard, App. G), yet team productivity remains at or below the matched no-peer control. Answer-diversity efficiency $R(N) = N_{\text{eff}}/N$ declines steeply with N ($\sim$15$\times$ lower at N=30 than at N=2), and N_{eff} saturates near 1.8 on GSM-Hard and $\sim$1.2 on MCQA (Table 15). Individual learning is present but does not translate into team gains because communication reduces N_{eff} through inter-agent correlation. The conformity dynamics are visible to different degrees across tasks: on MMLU-Hard, round-1 agreement is already near-saturated ($\rho^{(0)} \approx 0.97$), so communication does not raise it further (0.94 after debate); on GSM-Hard, round-1 agreement starts low (0.51) and rises post-revision to 0.78, exposing conformity amplification.

Table 2: Ringelmann fits at the answer level vs. correctness level (Qwen2.5-7B). Answer-level columns repeat Table 1; correctness-level columns refit the same form against $N_{\text{eff}}^{\text{corr}}$. All correctness-level fits are hard-ceiling ($\beta \approx 0$). *Metric caveat:* on free-form tasks $N_{\text{eff}}^{\text{ans}} = 2^{H(\mathbf{a})}$ is an upper bound on the independent count (App. F); Thm. 1 predicts $\beta \approx 0$ for pairwise agreement, not for entropy.

	MMLU-Hard				GSM-Hard				GPQA			
	Answer		Correct.		Answer		Correct.		Answer		Correct.	
Condition	c	β	c	β	c	β	c	β	c	β	c	β
Debate	.85	.000	.89	.000	.57	.014	.88	.000	.81	.000	.87	.000
Self-corr.	.84	.011	.88	.011	.34	.252	.72	.013	.78	.006	.85	.008
Noise	.65	.000	.72	.000	.33	.272	.70	.004	.52	.000	.62	.000

Two metric levels, one prediction. The answer-level fits in Table 1 describe how many distinct answer strings the team produces. The complementary correctness-level fit describes how correlated the binary correct/wrong indicators are, computed by refitting the same (c, β) form against $N_{\text{eff}}^{\text{corr}} = N/(1 + (N-1)\rho_{\text{ICC}}^{\text{corr}})$. Table 2 reports both levels side by side. All correctness-level fits lie at $\beta \approx 0$ (hard-ceiling regime, consistent with Thm. 1's contraction toward the modal answer); the answer-level regime crossing on GSM-Hard is a wrong-answer scatter phenomenon that does not translate to the correctness scale. The condition ordering $c_{\text{debate}} > c_{\text{self-corr}} \geq c_{\text{noise}}$ is preserved on both scales. Full comparison and item-difficulty decomposition in App. E. The scales are complementary: correctness redundancy bounds majority voting, while answer diversity, on which the classifier

separates plateauing from sublinear regimes, bounds what a verifier, best-of-N selector, or higher-order aggregator [Ai et al., 2026] can extract.

Table 3: $k\tau$ collapse test: pairwise agreement ρ for matched $k\tau$ products at fixed N (Qwen2.5-7B). With $\tau \leq 5$ communication rounds, only $k\tau \in \{2, 4, 8\}$ have two or more (k, τ) realizations. Spread is the max range across configurations sharing the same product.

		GSM-Hard (free-form)			MMLU-Hard (MCQA)		
$k\tau$	(k, τ) configs	$\rho_{N=10}$	$\rho_{N=30}$	max $\Delta\rho$	$\rho_{N=10}$	$\rho_{N=30}$	max $\Delta\rho$
2	(1,2), (2,1)	.508–.509	.463–.476	$\leq$ .013	.860–.888	.859–.888	$\leq$ .028
4	(1,4), (2,2), (4,1)	.555–.585	.515–.520	$\leq$ .031	.853–.905	.844–.888	$\leq$ .052
8	(2,4), (4,2)	.658–.659	.592–.605	$\leq$ .013	.878–.896	.854–.868	$\leq$ .018

Largest spread (MMLU, $k\tau$=4) involves single-round $(4, 1)$.
Restricting to $\tau \geq 2$ tightens the collapse (App. H).

5.2 C2: Peer count k and rounds τ collapse to their product $k\tau$

The mean-field theorem (Thm. 1) predicts that peer count k and rounds τ enter the dynamics only through their product $k\tau$. We test this at fixed N on Qwen2.5-7B with $k \in \{1, 2, 4, 9, 29\}$ and $\tau \in \{1, \ldots, 5\}$. Configurations sharing the same $k\tau$ yield agreement values within $\Delta\rho \leq 0.031$ on GSM-Hard and ≤ 0.052 on MMLU-Hard (Table 3), with looser collapse at τ=1 where the theorem's binary linearization is least accurate. The largest spread involves single-round $(4, 1)$: on MMLU at N=10, $(4, 1)$ reaches $\rho = 0.905$ while $(1, 4)$ and $(2, 2)$ lie at 0.853–0.880. Restricting to $\tau \geq 2$ tightens the collapse to $\Delta\rho \leq 0.029$ on both tasks; a per-item stratification by initial agreement ($\Delta\rho \leq 0.020$ in high- vs ≤ 0.055 in low-agreement items) is in App. H, Fig. 7. On the pairwise-agreement scale (the theorem's native variable) the regime-classifying exponent β_{agr} decreases with k (monotonically on GSM-Hard) toward the predicted $\beta_{\text{agr}} \approx 0$, and is small on both tasks ($\beta_{\text{agr}} \in [0.02, 0.06]$ on MMLU-Hard, $\in [0.06, 0.17]$ on GSM-Hard; Table 16). The same product structure extends under the asymmetric-rate model of App. I (Table 17; the mean α_{eff} of α_l and α_s weighted by the fraction of correct agents replaces α_1, leaving $k\tau$ unchanged). The empirical $k\tau$ collapse and two-rate decomposition (with $\alpha_l > \alpha_s$ asymmetry) are reported in Appendices H and I.

Beyond role symmetry. Assumption A1 (common adoption probability) breaks under role-asymmetric or tool-augmented agents, so exact $k\tau$ factorization is not guaranteed there, and such debate remains untested. App. M does not test A1, since its fitted agents never exchange messages.

5.3 C3: The Ringelmann form fits all 44 model–task–condition cells

Across 44 (model $\times$ task $\times$ condition) configurations, the Ringelmann form $R(N)=1/(1+c(N-1)N^{-\beta})$ fits with $R^2 \geq 0.969$ in-sample (median 0.999); only the estimated (c, β) shifts across conditions, not the form itself. We anchor on Qwen2.5-7B as the canonical example, then evaluate boundaries spanning model scale, family, paradigm, and team composition (Table 19, Figure 8, per-cell estimates in App. J, classical human Ringelmann studies in App. B).

Canonical case. Qwen2.5-7B is analyzed in §5.1 (Table 1). **Cross-family on MCQA: same regime, shifted ceiling.** Llama-3.1-8B and Ministral-8B both produce $\beta = 0$ on MMLU-Hard with $c \in [0.62, 0.71]$ (Qwen $c = 0.85$): the regime classification is shared, only the ceiling level differs. The same hard-ceiling pattern replicates on GPQA across all three families (Llama c=0.55, Ministral c=0.57, Qwen c=0.81, all $\beta = 0$). Cross-family details in App. J.1 (Table 19).

Model scale tightens the ceiling on all three tasks. Qwen2.5-32B has $c = 0.96$ on MMLU-Hard, $c = 0.77$ on GSM-Hard, and $c = 0.97$ on GPQA (all $\beta \approx 0$), versus $c = 0.85$, $c = 0.57$, and $c = 0.81$ respectively at 7B. Noise $\approx$ self-correction replicates at 32B. See App. J.2 (Table 19).

Thinking models and frontier API: hard ceiling. Qwen3-8B in thinking mode achieves 87.2% solo accuracy on MMLU-Hard, yet debate drives $\rho = 0.983$ at N=5 ($N_{\text{eff}} = 1.02$, $c = 1.00$, $\beta \approx 0$). Gemini 3.1 Flash-Lite (minimal thinking) lies in the same regime on both GSM-Hard ($c = 0.95$) and GPQA ($c = 0.86$). Under debate, $\rho \to 1$ drives N_{eff} toward 1 across both scale and paradigm. Detailed thinking-model analysis in App. K (Table 21).

Generalization envelope. Table 4 consolidates the generalization evidence along eight axes beyond the canonical case: three open-weight families, $4\times$ model scale, a thinking model, a frontier API, heterogeneous ensembles, six hierarchical team structures whose leaves answer without peer exchange, open-ended HumanEval code generation, and a sampling-temperature sweep. The functional form fits every configuration; only (c, β) varies.

Table 4: Generalization landscape. Each row is one configuration fitted with the Ringelmann form (Eq. 2) at $R^2 \geq 0.97$. Structure and HumanEval rows average across three open-weight families, and "all" pools the three tasks ($n = 9$ per structure). All other rows are single-model fits. In "indep." rows, leaves never exchange messages and N is the leaf count. Full per-cell data in the referenced appendix. H: hard ceiling ($\beta \approx 0$), S: sublinear ($\beta > 0.1$), M: H on MCQA and S on GSM-Hard (Table 25).

Axis (App.)	Configuration	Task × cond.	(c, β)	Reg.
Cross-family (J)	Llama-3.1-8B	MMLU debate	$(0.62, .00)$	H
Cross-family (J)	Ministral-8B	MMLU debate	$(0.71, .00)$	H
Scale $4\times$ (J)	Qwen2.5-32B	MMLU debate	$(0.96, .00)$	H
Thinking (K)	Qwen3-8B-think	MMLU debate	$(1.00, .01)$	H
Frontier API (J)	Gemini-3.1-FL	GSM debate	$(0.95, .00)$	H
Frontier API (J)	Gemini-3.1-FL	GPQA debate	$(0.86, .00)$	H
Heterogeneous (L)	Qwen + Llama + Min	MMLU debate	$(0.54, .00)$	H
Heterogeneous (L)	Qwen + Llama + Min	GSM debate	$(0.35, .00)$	H
Structure, task-only (M)	Star family	all, indep.	$(0.60, .10)$	M
Structure, persona (M)	Persona-Star	all, indep.	$(0.48, .12)$	M
Structure, plan-cond. (M)	Diamond	all, indep.	$(0.74, .29)$	S
Open-ended code (N)	HumanEval, 164 items	code, debate	$(0.33, .25)$	S
Temperature (O)	Qwen2.5-7B, T=1.0	GSM debate	$(0.47, .00)$	H

5.4 C4: Initial inter-agent correlation is the primary lever on the ceiling

Within role-symmetric, fully-connected teams on short-horizon tasks, three results identify the pre-communication correlation $\rho^{(0)}$ as the principal axis of intervention. A fourth result, from the temperature and persona interventions, is consistent with this identification.

(i) Noise placebo ≈ self-correction. On GSM-Hard, the noise placebo matches self-correction within bootstrap uncertainty ($c = 0.33$ vs $c = 0.34$, $\beta = 0.27$ vs $\beta = 0.25$, overlapping 95% CIs, N_{eff} within 0.4 at every N, Table 1): once revision is held fixed, peer content does not measurably alter the scaling regime, consistent with prior evidence that isolated self-correction prevails over homogeneous debate [Bertalanič and Fortuna, 2026]. The same equivalence replicates at $4\times$ scale on Qwen2.5-32B ($c = 0.45$ for both, $\beta = 0.18$ for both), and on Llama-3.1-8B and Ministral-8B. Across the four open-weight models (three families), $|\Delta c| \leq 0.036$ and $|\Delta\beta| \leq 0.064$, with no counterexample on GSM-Hard (Table 29). Details in App. P. On MCQA, the condition ordering persists under item-difficulty conditioning (Table 13). Matched prompts (Table 7) isolate peer *topical* content from prompt structure.

(ii) Mode-invariant team accuracy. Across homogeneous conditions, communication mode shifts (c, β) across the plane (Table 1) yet team accuracy is statistically equivalent within ± 5 pp. A two-one-sided-tests (TOST) equivalence test, which jointly rejects "debate is more than 5 pp better" and "self-correction is more than 5 pp better", bounds the gap on all three tasks ($p_{\text{TOST}} < 0.02$). Full statistics are in App. E (Table 14); aggregator-robustness checks in App. Q. The scaling law accounts for the convergence: on MCQA both conditions are hard-ceiling with similar c, and on GSM-Hard both are hard-ceiling at the correctness level, which bounds majority voting (Table 2).

(iii) Architectural diversity raises the ceiling. Cross-family teams (Qwen+Llama+Ministral) lower c from 0.85 to 0.54 on MMLU-Hard, 0.57 to 0.35 on GSM-Hard, and 0.81 to 0.56 on GPQA, with $\beta \approx 0$ preserved on all three tasks. The shifts match the block-exchangeable prediction (Prop. 4, App. A.7), and within-family pairs always agree more than cross-family pairs (App. L, Table 23).

(iv) Other $\rho^{(0)}$ interventions move c the same way. Raising temperature from 0.2 to 1.0 lowers c on all tasks ($0.655 \to 0.468$ on GSM-Hard) with the regime unchanged (App. O), and persona conditioning lowers mean c from 0.60 to 0.48 (App. M).

Taken together, these results identify $\rho^{(0)}$ as the primary validated determinant of the ceiling once $\beta \approx 0$, within the tested regime. Communication mode shifts (c, β) without a team-accuracy gain: on Qwen2.5-7B, re-evaluation instead of debate changes MCQA c by at most 0.03 at both levels (Table 2), and the noise placebo (answers to other items as peer input) lowers MCQA c by 0.20–0.28 but has lower team accuracy than debate and self-correction on every task (Table 14).

5.5 Deployment implications

Applicability. Instruct-class models are the most economical tier for multi-agent debate (under assumed prices, $\sim\$0.04$ per problem self-hosted vs. $\approx\$1.4$ on frontier reasoning models, App. R.3), so the tested regime corresponds to this deployment tier. For frontier-class configurations with high solo accuracy, the framework predicts $c \to 1$ (Eq. 3 with $\rho^{(0)} \to 1$). Qwen3-8B in thinking mode realizes this on MMLU-Hard (87.2% solo, $c = 1.00$). Solo evaluation of Gemini 3.1 Pro (81.5–97.0% accuracy) places it in the same predicted regime, untested with multi-agent runs (App. J).

Sizing and mode. Once N_{eff} saturates, additional agents increase token cost without adding independent evidence. At solo accuracy $p > 0.5$, 3–5 self-correcting agents approach the ceiling (MMLU-Hard: $N_{\text{eff}} = 1.17$ at $N{=}5$ vs. 1.21 at $N{=}30$; Tables 21 and 30). On Qwen2.5-7B at $N{=}30$, debate exceeds self-correction in accuracy by at most 0.5 pp and the noise placebo by at most 1.7 pp (Table 14), while costing 10–$45\times$ more prompt tokens per effective agent (Table 30).

Solo accuracy moderates the accuracy returns from teaming. On strong tasks ($p > 0.5$, MMLU-Hard), correlation prevents majority-vote gain from materializing. On harder tasks ($p \approx 0.10$–0.30), re-evaluation raises GSM-Hard accuracy from 10% to 26–28% (self-correction $+18.2$ pp, debate $+16.3$ pp) but GPQA accuracy by only 1.0 pp. Because wrong GSM-Hard answers scatter across many values, the correct answer can attain a plurality even at $p \ll 0.5$ (App. R.2).

5.6 Discussion

Implications for the design space. Within homogeneous, role-symmetric, fully-connected teams on short-horizon voting-aggregated tasks, the regime in which we evaluate instruct-class debate, communication mode leaves team accuracy unchanged within ± 5 pp, and peer content adds at most 1.7 pp over the noise placebo (Qwen2.5-7B). Much of the multi-agent literature varies voting rules or debate scripts with the model fixed, factors to which the ceiling responds only weakly in this regime. The ceiling moves primarily with $\rho^{(0)}$, lowered through architectural diversity (App. L), and to a lesser extent through temperature and persona conditioning.

Falsification criteria and extensions. The claim that $\rho^{(0)}$ is the primary lever would be falsified if verifier–aggregator pipelines lowered c at constant pre-verifier $\rho^{(0)}$. Proposition 3 would be contradicted if MoA [Wang et al., 2025] or frontier heterogeneous teams attained $\beta_{\text{agr}} > 0$ without an N-dependent decrease of both α and $\rho^{(0)}$. On full HumanEval (App. N) the form fits all nine family $\times$ condition cells ($R^2 \geq 0.99$), but majority-vote accuracy does not track N_{eff}, which is necessary but not sufficient for open-ended task success.

Limitations and scope. Claims are validated on homogeneous, role-symmetric teams (with initial evidence for heterogeneous and persona-conditioned teams), short-horizon voting-aggregated tasks, $N \leq 30$ and $\tau \leq 5$, mostly with 7–8B open-weight models (plus 32B, thinking, and frontier-API checks). Untested are long-horizon workflows, tool-augmented and verifier–aggregator pipelines, which lie outside the mean-field model, and rationale-level independence (App. S).

6 Conclusion

Across the evaluated role-symmetric teams, effective team size follows a Ringelmann-style scaling law $R(N) = 1/(1 + c(N-1)N^{-\beta})$ at both answer-diversity and correctness-redundancy levels. Systems differ not in *whether* the law applies but in *where* they fall in the (c, β) landscape. A single $N \leq 5$ pilot identifies the regime, and once $\beta \approx 0$ the primary validated lever on the ceiling within the tested regime is reducing initial inter-agent correlation $\rho^{(0)}$ via architectural diversity.

Acknowledgments and Disclosure of Funding

This work was supported by the Slovenian Research and Innovation Agency (ARIS) through research core funding No. P2-0016. We thank ARNES, the Academic and Research Network of Slovenia, for providing the GPU compute resources used in this work. The authors declare no competing interests.

References

Rui Ai, Yuqi Pan, David Simchi-Levi, Milind Tambe, and Haifeng Xu. Beyond majority voting: LLM aggregation by leveraging higher-order information. In *International Conference on Machine Learning (ICML)*, volume 306 of *Proceedings of Machine Learning Research*, pages 1311–1340, 2026.

Alessandro Bellina, Giordano De Marzo, and David Garcia. Prompt-mediated perceptual conformity in multimodal language models follows Social Impact Theory. *advances.in/psychology*, 1(1): e680902, 2026. doi: 10.56296/aip00072.

Blaž Bertalanič and Carolina Fortuna. The cost of consensus: Isolated self-correction prevails over unguided homogeneous multi-agent debate. In *ACM Conference on AI and Agentic Systems (CAIS)*, pages 311–329. ACM, 2026. doi: 10.1145/3786335.3813137.

Philip J. Boland. Majority systems and the Condorcet jury theorem. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 38(3):181–189, 1989. doi: 10.2307/2348873.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

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 (ICLR)*, 2024.

Lingjiao Chen, Jared Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. Are more LLM calls all you need? Towards the scaling properties of compound AI systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 45767–45790, 2024. doi: 10.52202/079017-1455.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pages 112802–112834, 2025a. doi: 10.52202/085713-3405.

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. When identity skews debate: Anonymization for bias-reduced multi-agent reasoning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 14284–14311, 2026. doi: 10.18653/v1/2026.acl-long.650.

Min Choi, Keonwoo Kim, Sungwon Chae, and Sangyeob Baek. An empirical study of group conformity in multi-agent systems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5123–5139, 2025b. doi: 10.18653/v1/2025.findings-acl.265.

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

Yegor Denisov-Blanch, Joshua Kazdan, Jessica Chudnovsky, Rylan Schaeffer, Sheng Guan, Soji Adeshina, and Sanmi Koyejo. Truthfulness does not scale like reasoning: Why polling fails as a proxy verifier. In *International Conference on Machine Learning (ICML)*, volume 306 of *Proceedings of Machine Learning Research*, pages 24461–24477, 2026.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 11733–11763, 2024.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. PAL: Program-aided language models. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 10764–10799, 2023.

Google DeepMind. Gemini 3.1 Flash-Lite model card. `https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-1-Flash-Lite-Model-Card.pdf`, May 2026. Model card, updated May 2026. Accessed 2026-11-07.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 30016–30030, 2022. doi: 10.52202/068431-2176.

Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

Lars Benedikt Kaesberg, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. Voting or consensus? Decision-making in multi-agent debate. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11640–11671, 2025. doi: 10.18653/v1/2025.findings-acl.606.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

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 *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 23662–23733, 2024.

Yubin Kim, Ken Gu, Chanwoo Park, Chunjong Park, Samuel Schmidgall, A. Ali Heydari, Yao Yan, Zhihan Zhang, Yuchen Zhuang, Yun Liu, Mark Malhotra, Paul Pu Liang, Hae Won Park, Yuzhe Yang, Xuhai Xu, Yilun Du, Shwetak Patel, Tim Althoff, Daniel McDuff, and Xin Liu. Capable language models can outgrow the benefits of collaboration. *Nature Machine Intelligence*, 8(7): 1157–1172, 2026. doi: 10.1038/s42256-026-01268-y.

Leslie Kish. *Survey Sampling*. John Wiley & Sons, New York, 1965.

David A. Kravitz and Barbara Martin. Ringelmann rediscovered: The original article. *Journal of Personality and Social Psychology*, 50(5):936–941, 1986. doi: 10.1037/0022-3514.50.5.936.

Krishna K. Ladha. The Condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, 36(3):617–634, 1992. doi: 10.2307/2111584.

Bibb Latané, Kipling Williams, and Stephen Harkins. Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology*, 37(6):822–832, 1979. doi: 10.1037/0022-3514.37.6.822.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More agents is all you need. *Transactions on Machine Learning Research (TMLR)*, 2024. URL https://openreview.net/forum?id=bgzUSZ8aeg.

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 *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 17889–17904, 2024. doi: 10.18653/v1/2024.emnlp-main.992.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaëtan Ecrepont, Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyl, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futeral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yassine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026.

Llama Team, AI @ Meta. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 46534–46594, 2023. doi: 10.52202/075280-2019.

Andrew Mao, Winter Mason, Siddharth Suri, and Duncan J. Watts. An experimental study of team size and performance on a complex task. *PLOS ONE*, 11(4):e0153048, 2016. doi: 10.1371/journal.pone.0153048.

Mistral AI. Un Ministral, des Ministraux. Mistral AI news, https://mistral.ai/news/ministraux, October 2024.

Jihwan Oh, Minchan Jeong, Jongwoo Ko, and Se-Young Yun. From belief entrenchment to robust reasoning in LLM agents. *Transactions of the Association for Computational Linguistics*, 14: 1286–1307, 2026. doi: 10.1162/tacl.a.728.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with

model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, 2023. doi: 10.18653/v1/2023.findings-acl.847.

Chen Qian, Zihao Xie, YiFei Wang, Wei Liu, Kunlun Zhu, Hanchen Xia, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Scaling large language model-based multi-agent collaboration. In *International Conference on Learning Representations (ICLR)*, 2025.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. In *Conference on Language Modeling (COLM)*, 2024.

Max Ringelmann. Recherches sur les moteurs animés: Travail de l'homme. *Annales de l'Institut National Agronomique, 2e série*, 12:1–40, 1913.

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

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 *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 45883–45905, 2024.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations (ICLR)*, 2025.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Y. Zou. Mixture-of-Agents enhances large language model capabilities. In *International Conference on Learning Representations (ICLR)*, 2025.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6106–6131, 2024. doi: 10.18653/v1/2024.acl-long.331.

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 (ICLR)*, 2023.

Haolun Wu, Zhenkun Li, and Lingyao Li. Can LLM agents really debate? A controlled study of multi-agent debate in logical reasoning. *arXiv preprint arXiv:2511.07784*, 2025.

Andrea Wynn, Harsh Satija, and Gillian K. Hadfield. Talk isn't always cheap: Understanding failure modes in multi-agent debate. *arXiv preprint arXiv:2509.05396*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Yingxuan Yang, Chengrui Qu, Muning Wen, Laixi Shi, Ying Wen, Weinan Zhang, Adam Wierman, and Shangding Gu. Understanding agent scaling in LLM-based multi-agent systems via diversity. In *ICLR 2026 Workshop on AI for Mechanism Design and Strategic Decision Making (AIMS)*, 2026. arXiv:2602.03794.

Yongjin Yang, Euiin Yi, Jongwoo Ko, Kimin Lee, Zhijing Jin, and Se-Young Yun. Revisiting multi-agent debate as test-time scaling: A systematic study of conditional effectiveness. *arXiv preprint arXiv:2505.22960*, 2025b.

Binwei Yao, Chao Shang, Wanyu Du, Jianfeng He, Ruixue Lian, Yi Zhang, Hang Su, Sandesh Swamy, and Yanjun Qi. Peacemaker or troublemaker: How sycophancy shapes multi-agent debate. *arXiv preprint arXiv:2509.23055*, 2025.

Hangfan Zhang, Zhiyao Cui, Qiaosheng Zhang, and Shuyue Hu. Multi-LLM-agents debate – performance, efficiency, and scaling challenges. ICLR 2025 Blogpost Track, `https://iclr-blogposts.github.io/2025/blog/mad/`, April 2025.

Xiaochen Zhu, Caiqi Zhang, Tom Stafford, Nigel Collier, and Andreas Vlachos. Conformity in large language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3854–3872, 2025. doi: 10.18653/v1/2025.acl-long.195.

Appendix roadmap. The appendices follow the order of the main text and map to the four contributions in §1 (diagram in Figure 2, App. A).

Foundations and methodology (§3–§4). App. A derives the effective-team model and proves all propositions and Thm. 1. App. B relates the form to classical Ringelmann effects. App. C documents model scope, prompts (including the matched-construction table), generation, answer extraction, aggregation, hardware, the statistical methodology, and the (c, β) measurement procedure.

C1, regime-classifying scaling law (§5.1). App. D (held-out extrapolation audit with per-cell CIs, shape-matched baselines, robustness and unbounded β), App. E (correctness-level fits), App. F (answer-diversity metrics), App. G (agent-level transitions).

C2, $k\tau$ product theorem (§5.2). App. H ($k\tau$ collapse and stratification) and App. I (asymmetric conformity rates and transition diagnostics).

C3, shared form across conditions (§5.3). App. J (three open-weight families, $4\times$ scale, Gemini), App. K (thinking model), App. L (heterogeneous teams), App. M (six hierarchical team structures without peer exchange, 54 cells), App. N (HumanEval code generation).

C4, initial correlation as the lever (§5.4). App. O (temperature sweep), App. P (noise placebo, self-correction, self-consistency across four open-weight models from three families), App. Q (aggregator robustness).

Deployment and scope (§5.5–§5.6). App. R (token cost, accuracy projection, economics) and App. S (extended scope and limitations).

A Derivation of the effective-team model

Roadmap: definitional vs. empirical components. Figure 2 maps each component of the framework to the main-text section and appendices that support it. The upper row is definitional (theory), the middle row is the estimation methodology, and the lower row is the empirical validation and findings.

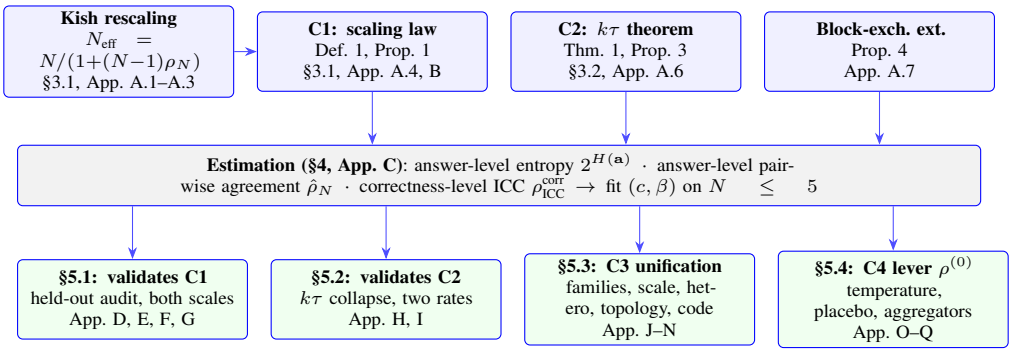


Figure 2: **Paper roadmap.** Top: definitional components (blue). Middle: empirical estimation methodology. Bottom: empirical validation (§5.1, §5.2) and empirical findings (§5.3, §5.4), with the supporting appendices. Deployment guidance is in App. R and the extended scope and limitations are in App. S.

Notation guide for ρ variants used throughout the paper. Subscript on ρ indicates a qualifier (mf, eff, ICC, cond) or team size N. Bare ρ denotes the equicorrelation-scale pairwise agreement (§ 3). Superscript with parens indicates round: $\rho^{(0)}$ is the initial (pre-communication) correlation and $\rho^{(\tau)}$ is the correlation at round τ.

- $\rho^{(0)}$: pre-communication pairwise answer correlation, baseline before any peer interaction.
- $\rho^{(\tau)}$: pairwise correlation at round τ, used in the dynamic mean-field model (Thm. 1).
- ρ_N: post-debate (final-round) pairwise correlation as a function of team size N, modeled by the Ringelmann profile $\rho_N = cN^{-\beta}$. Empirically, ρ_N is estimated through $\rho_{\text{eff}}^{\text{ans}}$ (primary) or through the pairwise-agreement estimate $\hat{\rho}_N$.
- ρ_{mf}: mean-field correlation from Thm. 1, tracking pairwise agreement through $d = 1 - \rho^{(\tau)}$.

- $\rho_{\text{eff}}^{\text{ans}}$: equicorrelation implied by inverting $N_{\text{eff}}^{\text{ans}} = 2^{H(\mathbf{a})}$ through Eq. 1. The Ringelmann curve is estimated against this quantity.
- $\rho_{\text{ICC}}^{\text{corr}}$: marginal ICC of binary correctness indicators across items ($(\mathbb{E}[X_i X_j] - \bar{p}^2)/(\bar{p}(1-\bar{p}))$).
- ρ: pairwise answer agreement on the equicorrelation scale, $(N \cdot A_{\text{pair}} - 1)/(N-1)$.
- $\rho_{\text{cond}}^{\text{corr}}$: correctness ICC conditioned on round-1 item difficulty (App. E), isolating within-item social dependence from shared item-difficulty effects.

Other notation.

- $\mathbf{a} = (a_1, \ldots, a_N)$: vector of the N agents' final answers in a single team trial. Each a_i is the post-debate answer of agent i on a single item, extracted from the agent's free-form output (App. C).
- $H(\mathbf{a})$: Shannon entropy of $\mathbf{a}$ in bits, $H = -\sum_v p_v \log_2 p_v$ with p_v the empirical frequency of answer value v in the team.
- $N_{\text{eff}}^{\text{ans}} = 2^{H(\mathbf{a})}$: answer-level effective team size. Ranges from 1 (full consensus, $H{=}0$) to N (all agents disagree, $H{=}\log_2 N$). Per-team quantity, reported as a mean across items.
- $N_{\text{eff}}^{\text{corr}} = N/(1 + (N{-}1)\rho_{\text{ICC}}^{\text{corr}})$: correctness-level effective team size, derived from the Kish design effect on the binary correctness indicator (App. E).
- $R(N) = N_{\text{eff}}/N$: efficiency ratio. The Ringelmann fit $R(N) = 1/(1 + c(N{-}1)N^{-\beta})$ targets this quantity.

A.1 From equicorrelated votes to effective team size

Let $X_i \in \{0, 1\}$ indicate whether agent i's final answer is correct after debate. Write

$$\mathbb{E}[X_i] = p, \qquad \text{Corr}(X_i, X_j) = \rho_{\text{eff}} \quad (i \neq j),$$

where p is the per-agent post-debate accuracy (the probability that an arbitrary agent finishes the debate at the correct answer) and ρ_{eff} is the equicorrelation between any pair of agents' final correctness indicators. Under exchangeability,

$$\text{Var}(X_i) = p(1-p), \qquad \text{Cov}(X_i, X_j) = \rho_{\text{eff}}\, p(1-p).$$

For the team mean $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$,

$$\text{Var}(\bar{X}) = \frac{1}{N^2}\left[\sum_{i=1}^{N} \text{Var}(X_i) + \sum_{i \neq j} \text{Cov}(X_i, X_j)\right] \tag{6}$$

$$= \frac{1}{N^2}\left[Np(1-p) + N(N-1)\rho_{\text{eff}} p(1-p)\right] \tag{7}$$

$$= \frac{p(1-p)}{N}\big(1 + (N-1)\rho_{\text{eff}}\big). \tag{8}$$

Relative to independent votes, the variance is inflated by the Kish design effect $D_{\text{eff}}(N) := 1 + (N-1)\rho_{\text{eff}}$. Matching this variance to that of an independent sample of size N_{eff} gives $N_{\text{eff}} = N/(1 + (N-1)\rho_{\text{eff}})$, which is Eq. 1. This identity is exact for exchangeable Bernoulli votes with common pairwise correlation. When heterogeneous dependence is compressed into a single effective parameter ρ_{eff}, the formula becomes a variance-matching approximation.

A.2 Correlated-Condorcet approximation

Let $S_N := \sum_{i=1}^{N} X_i$ denote the number of correct final votes. Majority voting is correct when $S_N > N/2$ (strict majority, ties broken by fair coin). Under the same exchangeable model, $\text{Var}(S_N) = Np(1-p)(1 + (N-1)\rho_{\text{eff}}) = N^2 p(1-p)/N_{\text{eff}}$. Approximating S_N by a Gaussian yields

$$\mathbb{P}(S_N > N/2) \approx \Phi\left(\frac{p - 1/2}{\sqrt{p(1-p)/N_{\text{eff}}}}\right), \tag{9}$$

where Φ is the standard normal CDF. Thus the correlated-vote system is approximated by an independent-vote system with only N_{eff} effective voters.

In the paper we use $\mathrm{PP}(N_{\text{eff}}, p)$ as shorthand for this variance-matched correlated-Condorcet approximation. The experiments observe answer distributions, from which we derive answer-level effective-diversity proxies such as $N_{\text{eff}}^{\text{ans}} = 2^{H(\mathbf{a})}$ and pairwise answer agreement. Eq. 1 therefore bridges the answer-level diversity we measure with the correctness-level ceiling we predict.

A.3 Proposition 2 and proof

Proposition 2 (Exact Criterion for Effective Scaling)**.** *Within the effective-team model with $\rho_N \geq 0$: if $\rho_N \to \rho_\infty > 0$, then $N_{\text{eff}}(N) \to 1/\rho_\infty$ (finite ceiling). Linear effective scaling $N_{\text{eff}}(N) = \Theta(N)$ is possible if and only if $\rho_N = O(1/N)$.*

Proof. Starting from Eq. 1, $N_{\text{eff}}(N)/N = 1/(1 + \frac{N-1}{N} \cdot N \cdot \rho_N)$. Since $(N-1)/N \to 1$, the asymptotic behavior is determined by $N \cdot \rho_N$. If $\rho_N \to \rho_\infty > 0$, then $N_{\text{eff}}(N) \to 1/\rho_\infty$. $N_{\text{eff}}(N) = \Theta(N)$ holds iff $N \cdot \rho_N = O(1)$, that is, $\rho_N = O(1/N)$.

A.4 Derivation of Proposition 1

Assume $\rho_N = cN^{-\beta} + o(N^{-\beta})$ with $c > 0$. From Eq. 1, the Kish denominator expands to $1 + (N-1)\rho_N = 1 + cN^{1-\beta} + o(N^{1-\beta})$. The proof tracks which term dominates as $N \to \infty$ in each of the three regimes.

$\beta = 0$ *(hard ceiling).* $\rho_N \to c$, so $(N-1)\rho_N \approx cN$ dominates the 1 and $N_{\text{eff}}(N) = N/(1 + cN + o(N)) \to 1/c$.

$0 < \beta < 1$ *(sublinear).* The $cN^{1-\beta}$ term still dominates the 1 (since $1 - \beta > 0$), giving $N_{\text{eff}}(N) \sim N/(cN^{1-\beta}) = N^\beta/c$.

$\beta \geq 1$ *(linear).* For $\beta = 1$, both terms are $O(1)$ and $N_{\text{eff}}(N) \sim N/(1+c)$. For $\beta > 1$, $cN^{1-\beta} \to 0$, the 1 dominates, and $N_{\text{eff}}(N) \sim N$ (correlation falls fast enough that voters are asymptotically independent).

Corollary 1 (Heterogeneity Shifts the Ceiling)**.** *Consider two debate systems with correlation profiles $\rho_N^{(1)} \sim c_1 N^{-\beta}$ and $\rho_N^{(2)} \sim c_2 N^{-\beta}$ sharing the same exponent β. If $c_2 < c_1$, then system 2 has strictly larger asymptotic effective team size. In the hard-ceiling regime $\beta = 0$, the ceiling rises from $1/c_1$ to $1/c_2$.*

Proof. Immediate from $N_{\text{eff}} \sim (1/c)N^\beta$ for $\beta < 1$ and $N_{\text{eff}} \to 1/c$ for $\beta = 0$. □

A.5 Proposition 3: hard ceiling is generic

Proposition 3 (Hard Ceiling Is Generic)**.** *Within the mean-field model (Eq. 3), if the conformity rate $\alpha \in (0,1)$ and initial correlation $\rho^{(0)} \in [0,1)$ do not depend on team size N, then post-debate correlation ρ_{mf} is constant in N, giving $\beta = 0$ (hard-ceiling regime) with ceiling $N_{\text{eff}}^* = 1/\rho_{mf}$.*

Let α_N and $\rho_N^{(0)}$ denote α and $\rho^{(0)}$ at team size N. Escaping the hard ceiling ($\beta > 0$) requires, for fixed τ, that both *$\alpha_N \to 0$ and $\rho_N^{(0)} \to 0$ as $N \to \infty$. Neither alone suffices: if $\alpha_N \to 0$ but $\rho_N^{(0)} \to \rho_\infty > 0$, then $\rho_{mf}^{(\tau)} \to \rho_\infty > 0$. If $\rho_N^{(0)} \to 0$ but α stays positive, then $\rho_{mf}^{(\tau)} \to 1-(1-\alpha)^\tau > 0$.*

Proof. When α and $\rho^{(0)}$ are N-independent, $\rho_{\text{mf}}^{(\tau)}$ is constant in N, so $\beta = 0$. The two failure modes stated in the proposition follow by direct substitution into Eq. 3, and only when both vanish can $\rho_{\text{mf}}^{(\tau)} \to 0$. □

A.6 Proof of Thm. 1 (mean-field convergence)

We prove the recurrence under assumptions (A1)–(A3). Let $q^{(\tau)}$ denote the modal share and $\varepsilon^{(\tau)} = 1 - q^{(\tau)}$ the non-modal mass.

Step 1: Recurrence for non-modal mass. Under (A1), a non-modal agent samples k peers without replacement. Let M denote the number of modal peers observed. For fixed k as $N \to \infty$, the hypergeometric sampling distribution converges to $\mathrm{Bin}(k, q^{(\tau)})$ with deviation $O(k/N)$, and we

use the binomial form throughout. Conditional on M, the agent retains its non-modal answer with probability $(1-\alpha_1)^M$ (each modal peer independently fails to trigger adoption). By the binomial PGF, $\mathbb{E}[(1-\alpha_1)^M] = \big(1-\alpha_1 q^{(\tau)}\big)^k$. Under (A2), modal agents do not switch. The non-modal mass therefore evolves as

$$\varepsilon^{(\tau+1)} = \varepsilon^{(\tau)}\,(1-\alpha_1+\alpha_1\,\varepsilon^{(\tau)})^k. \tag{10}$$

Under (A3), the mode label is consistent across rounds.

Step 2: High-agreement expansion. For $\varepsilon^{(\tau)} \ll 1$, Taylor-expand $(1-\alpha_1+\alpha_1\varepsilon)^k = (1-\alpha_1)^k\big(1+\frac{k\alpha_1}{1-\alpha_1}\varepsilon + O(\varepsilon^2)\big)$, so $\varepsilon^{(\tau+1)} = \varepsilon^{(\tau)}(1-\alpha_1)^k + O((\varepsilon^{(\tau)})^2)$.

Step 3: Convert non-modal mass to pairwise disagreement. For two agents drawn uniformly, $d = 1-\sum_a p_a^2 = 1-q^2-\sum_{a\neq\text{mode}} p_a^2$. The non-modal term satisfies $0 \le \sum_{a\neq\text{mode}} p_a^2 \le \varepsilon^2$, so $d = 2\varepsilon + O(\varepsilon^2)$ and $\varepsilon = d/2 + O(d^2)$. Substituting: $d^{(\tau+1)} = d^{(\tau)}(1-\alpha_1)^k + O((d^{(\tau)})^2)$.

Step 4: Iteration. Let $r := (1-\alpha_1)^k \in (0,1)$. By induction, $d^{(\tau)} = r^\tau d^{(0)}(1+O(d^{(0)}))$. Therefore $\rho^{(\tau)} \approx 1-(1-\rho^{(0)})(1-\alpha_1)^{k\tau}$ for $d^{(0)} = 1-\rho^{(0)} \ll 1$, which is Eq. 5. □

Remark on multi-class extensions and modal-switching robustness

For multi-class answer spaces, (A1)–(A2) coarsen the dynamics to $\{\text{modal}, \text{non-modal}\}$. Non-modal-to-non-modal transitions preserve total non-modal mass ε but redistribute minority answers. In convergent dynamics this typically concentrates the minority distribution, which decreases pairwise disagreement $d = 1-\sum_a p_a^2$. The per-peer rate α_1, estimated as $P(\text{nm}\to\text{m})/k$ over R1→R3, on GSM-Hard debate ranges from 0.014 at N=30 to 0.062 at N=5 (Table 18, App. I). The α_1 in this proof is the *effective* rate absorbing both modal-copy and minority-redistribution effects.

Empirically, (A2) is also approximate: modal agents occasionally adopt non-modal peer answers at a small per-peer rate. We denote this rate $\alpha_{s,1}$, paralleling the per-peer learning rate α_1. The notation distinguishes it from the agent-level sycophancy rate α_s of App. I. Including $\alpha_{s,1} > 0$ in the recurrence of Step 1 yields

$$\varepsilon^{(\tau+1)} \approx \varepsilon^{(\tau)}\big[(1-\alpha_1)^k + k\alpha_{s,1}\big] + O\big((\varepsilon^{(\tau)})^2\big),$$

which preserves the geometric contraction and the $k\tau$ product structure as long as $\alpha_{s,1} \ll \alpha_1$. An empirical upper bound on $\alpha_{s,1}$ is the modal-leak rate of Table 18 (1 − modal-stay, R1→R3), which counts every departure from the round-1 mode whether or not a peer induced it, divided by k as for α_1. Under debate this bound is 0.003–0.037 on MMLU-Hard and GPQA, a factor of 2.6–6.3 below α_1 at every $N \in \{5, 10, 30\}$, so (A2) holds well on MCQA and the deviation from Eq. 4 is small. On GSM-Hard the bound is 0.013–0.114, between 0.9 and 1.8 times α_1, so (A2) holds only approximately. Most of this leak is not peer-induced: under self-correction, where no peer answers are shown, modal agents leave the round-1 mode at a similar rate (0.40–0.41, vs. 0.37–0.46 under debate). The contraction is therefore sharpest on MCQA and approximate on free-form math, in agreement with App. I.

Scope: when the theorem applies

The theorem describes contraction under iterated peer-influenced revision. It applies directly to debate (including sparse-k and heterogeneous variants, Proposition 4). For conditions without peer exposure (self-correction, noise), $\alpha_1 \to 0$ and the theorem predicts $\rho^{(\tau)} \approx \rho^{(0)}$ (no systematic contraction). The empirical match between these placebo conditions (§ 5) is consistent with this null prediction. For self-consistency ($\tau = 0$), the theorem reduces to the identity $\rho^{(\tau)} = \rho^{(0)}$.

Two scales, one prediction

Consequence (iii) of the theorem is formulated on the *pairwise agreement* scale ρ, the theorem's native variable. On free-form tasks the entropy-derived scale ($N_{\text{eff}}^{\text{ans}} = 2^H$) inflates with N because wrong answers spread across many distinct numeric strings, so the entropy-scale exponent β_{ent} is not the quantity the theorem predicts. Empirically, $\beta_{\text{agr}} \approx 0.06$–$0.17$ at fixed k on GSM-Hard (close to the theorem's $\beta_{\text{agr}} \approx 0$ prediction), while $\beta_{\text{ent}} \approx 0.17$–$0.40$ on the same data. On MMLU-Hard the two exponents are close ($|\beta_{\text{ent}} - \beta_{\text{agr}}| \le 0.05$ at every k, App. H).

A.7 Block-exchangeable extension for heterogeneous teams

Thm. 1 assumes exchangeable agents. Heterogeneous teams violate exchangeability: within-family pairs (such as Qwen–Qwen) agree more than cross-family pairs (such as Qwen–Llama). We extend the framework to a block-exchangeable model.

Proposition 4 (Heterogeneous Effective Correlation)**.** *Consider N agents partitioned into G groups of sizes $n_1, \ldots, n_G$ ($\sum_g n_g = N$) with $\mathrm{Corr}(X_i, X_j) = \rho_W$ within groups and ρ_B between groups ($\rho_W \geq \rho_B \geq 0$). Then*

$$N_{\text{eff}} = \frac{N}{1 + (\bar{n} - 1)\rho_W + (N - \bar{n})\rho_B}, \qquad \bar{n} := \frac{1}{N}\sum_{g=1}^{G} n_g^2.$$

Equivalently, the team behaves as a homogeneous team with effective correlation $\rho_{\text{eff}} = [(\bar{n}-1)\rho_W + (N-\bar{n})\rho_B]/(N-1)$.

Proof. $\mathrm{Var}(\bar{X}) = (1/N^2)[\sum_i \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j)]$. The covariance sum decomposes: $\sum_g n_g(n_g-1)\rho_W\sigma^2 + [N(N-1) - \sum_g n_g(n_g-1)]\rho_B\sigma^2$. Using $\sum_g n_g(n_g-1) = N(\bar{n}-1)$, variance-matching against $(p(1-p)/N_{\text{eff}})$ yields the result. □

Consequences. *(i)* When $\rho_B = \rho_W$, the formula collapses to homogeneous $N_{\text{eff}} = N/(1+(N-1)\rho_W)$. *(ii)* $\rho_B < \rho_W$ implies $\rho_{\text{eff}} < \rho_W$, so the ceiling $1/\rho_{\text{eff}}$ rises: cross-family independence increases the effective team size. *(iii)* If ρ_W, ρ_B are N-independent and groups grow proportionally ($\bar{n}/N$ constant), then ρ_{eff} is N-independent, giving $\beta = 0$.

Empirical validation. On MMLU-Hard heterogeneous teams (Qwen+Llama+Ministral, equal-sized groups) at N=30, within-family pairwise agreement gives $\rho_W = 0.879$ and between-family agreement gives $\rho_B = 0.834$. The predicted effective correlation under equal groups is $\rho_{\text{eff}}^{\text{pred}} = \frac{9}{29}\rho_W + \frac{20}{29}\rho_B = 0.848$ (exact weights at N=30 and $\bar{n} = 10$, which tend to $\frac{1}{3}$ and $\frac{2}{3}$ as $N \to \infty$), which coincides with the observed overall pairwise agreement (0.848; 0.843 after the equicorrelation rescaling $(NA_{\text{pair}}-1)/(N-1)$). The coincidence holds by construction, because the overall agreement is the pair-count-weighted mean of the within- and between-family agreement with exactly these weights. The informative quantity is the gap $\rho_W - \rho_B = 0.045 > 0$, which Proposition 4 requires for the ceiling to rise.[1] This is consistent with heterogeneous teams lowering c (from $\approx$0.85 homogeneous to $\approx$0.54) because cross-family pairs are less correlated, as Proposition 4 predicts.

B Relation to classical Ringelmann effects

We use the Ringelmann label because the same *process-loss structure* appears in our LLM experiments: actual productivity falls below matched potential productivity as group size grows [recovered and translated by Kravitz and Martin, 1986]. This is a structural analogy, not a claim that LLM agents reproduce human social psychology or social loafing.

[1] ρ_W, ρ_B, and the predicted ρ_{eff} are on the pairwise-agreement scale, consistent with the variance-matching derivation. Table 22 reports $\rho_{\text{eff}}^{\text{ans}}$ on the entropy-inverted scale (App. F), which is a different metric.

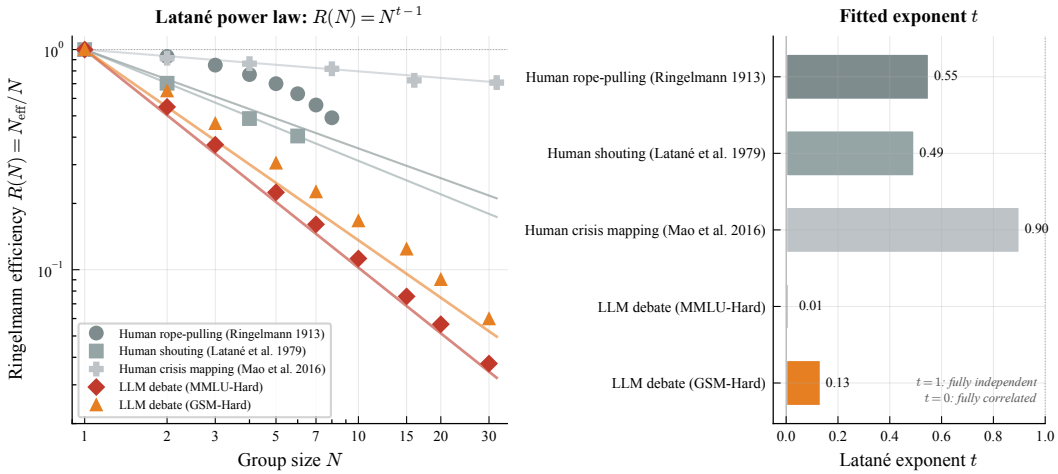


Figure 3: **Structural comparison to classical Ringelmann data via the Latané power law.** We fit $R(N) = N^{t-1}$ to LLM debate and to three published human group datasets, on the same axes. The estimated exponent t quantifies the decline rate: human groups have $t \in [0.49, 0.90]$ (gradual decline), LLM debate has $t \leq 0.13$ (much steeper). The same form is used for both populations so the gap in estimated t is structural, not an artifact of model choice. Latané's power law matches the Ringelmann form's asymptote at large N but is not a strict special case (see the text below for the relationship).

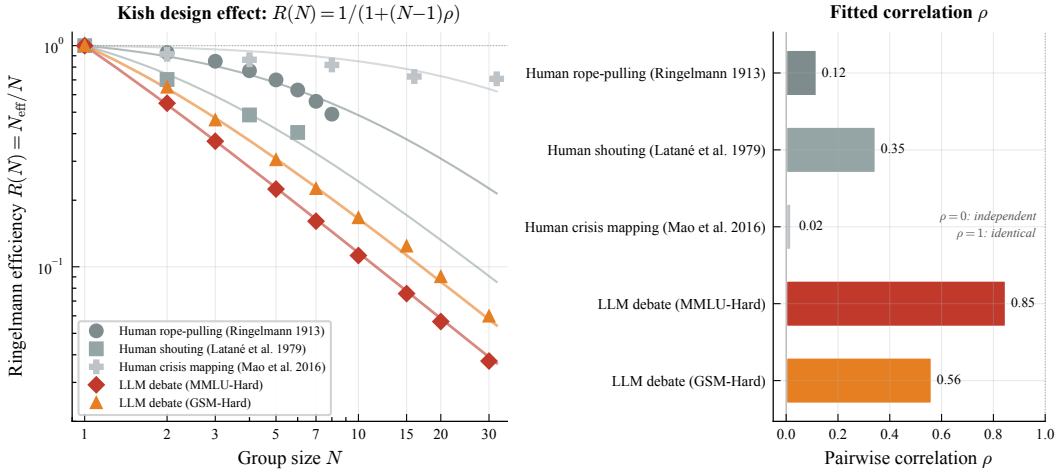


Figure 4: **Structural comparison via the Kish design effect.** Same datasets as Figure 3, estimated with the constant-ρ form $R(N) = 1/(1+(N-1)\rho)$, the $\beta = 0$ special case of the Ringelmann form used in the main text. The estimated pairwise correlation ρ quantifies redundancy: human groups have $\rho \in [0.02, 0.35]$ (largely independent voters), LLM debate has $\rho \in [0.56, 0.85]$ (highly correlated). The Kish form serves here as a common one-parameter axis for the cross-population comparison. LLM debate lies close to the hard-ceiling regime ($\beta \approx 0$), whereas the two-parameter fit to the human data gives $\beta = 0.00$ (rope-pulling), 0.33 (shouting) and 0.66 (crisis mapping). The main analyses use the full two-parameter form because β becomes the regime classifier on conditions like self-correction and noise where $\beta > 0$.

Figures 3 and 4 fit two single-parameter forms used in the human Ringelmann literature: the Latané power law $R(N) = N^{t-1}$ and the Kish design effect $R(N) = 1/(1+(N-1)\rho)$. Kish is the $\beta = 0$ special case of the two-parameter Ringelmann form $R(N) = 1/(1+c(N-1)N^{-\beta})$ used in the main analyses. Latané is not a strict special case but matches the Ringelmann asymptote at large N. We use these single-parameter forms here for cross-population comparison on a common axis, but the main paper fits the full two-parameter form because β is the regime classifier (hard-ceiling, sublinear,

or linear). Forcing $\beta = 0$ collapses the regime distinction and fails on sublinear cells (Const-ρ Kish: 8.3% mean held-out relative error vs. 5.5% for the two-parameter Ringelmann form, Table 8).

For reference, we fit the same declining-efficiency families used in the human literature to our LLM data and to three published human group-productivity measurements. Fitting to Ringelmann's [1913] rope-pulling data (N=1–8) yields $t = 0.55$ ($c = 0.12$ at β=0). Fitting to the Latané et al. [1979] shouting data (N=1–6) yields $t = 0.49$ ($c = 0.35$ at β=0). We also include the Mao et al. [2016] crisis-mapping study (N=1–32). For LLM debate, the estimated exponents are much smaller: $t = 0.01$ on MMLU-Hard (implied constant $\rho = 0.85$) and $t = 0.13$ on GSM-Hard (implied $\rho = 0.56$, where this single-parameter Kish fit differs slightly from the two-parameter generalized Ringelmann $c = 0.57$ in the main text). Figure 3 plots all five datasets on the Latané axis and Figure 4 on the Kish axis. Even the human dataset with the steepest decline ($t = 0.49$) retains substantially higher per-person efficiency than LLM debate ($t \leq 0.13$) at matched group sizes.

C Experimental details

This appendix complements §4 with implementation details.

C.1 Model scope

Rationale for the model scope of §4. Table 5 makes the role of each experiment explicit. The sweep is layered, not a flat cross-product. A single core identification (Qwen2.5-7B, full N=2–30 grid, three conditions, full item budgets of 724 / 1017 / 198) carries the main scaling-law fit. Two cross-family replications (Llama-3.1-8B at matched scope, Ministral-8B with partial coverage at large N) test that the form is not Qwen-specific. Four boundary experiments each probe one framework prediction with narrower coverage chosen to suffice for the respective test: $4\times$ within-family scale (Qwen2.5-32B), architectural diversity (heterogeneous teams), reasoning paradigm (Qwen3-8B in thinking mode), and closed-source paradigm (Gemini Flash-Lite). The frontier solo evaluation (Gemini 3.1 Pro at N=1 only) serves as a reference point for the regime in which the framework predicts saturation, rather than as part of the law's identification, since the framework predicts $c \to 1$ at its solo-accuracy level (§3.2). Item budgets and team-size grids are matched to the question each cell poses: full grids where extrapolation is audited, smaller item subsets where the test is qualitative (regime preservation, ceiling shift), and single-N where the framework already predicts a degenerate fit. The resulting 44 (c, β) estimates therefore allocate compute in proportion to the evidential role of each cell.

Table 5: Strength of evidence per experiment. Core sweeps (full N range, full item budget) carry the main claim. Replications, boundary checks, and the frontier solo evaluation play supporting roles. Items column: per-task item counts (M = MMLU-Hard, G = GSM-Hard, Q = GPQA Diamond). Cond.: d = debate, s = self-correction, n = noise placebo.

Experiment	Tasks	Items	N	Cond.	Role
Qwen2.5-7B	MGQ	724/1017/198	2–30	d/s/n	Open-weight, full sweep
Llama-3.1-8B	MGQ	724/1017/198	2–30	d/s/n	Open-weight, full sweep
Ministral-8B	MGQ	724/1017/198[‡]	2–30	d/s/n	Open-weight, partial sweep
Qwen2.5-32B	MGQ	200/200/198	2–30	d/s/n	$4\times$ scale
Heterogeneous	MGQ	200/200/198	3–30	d	Diversity
Qwen3-8B-think	M	724	2–5	d/s/n[†]	Thinking
Gemini Flash-Lite	GQ	200/198[§]	3–30	d	Closed-source
Gemini 3.1 Pro	MGQ	200/200/198	1	ensemble	Frontier solo

[†]Noise at N=5 covers 26 of 724 items.

[‡]Partial coverage. GSM-Hard uses as few as 200 items at $N \geq 5$ and has no noise runs at $N \leq 3$. GSM-Hard self-correction and noise and GPQA extend to N=20. GSM-Hard debate reaches N=30 on 477 items. MMLU-Hard noise has 25 items at N=30.

[§]Partial coverage at N=10 (GSM-Hard, 10 items) and N=30 (178 GSM-Hard and 141 GPQA items).

Checkpoints. Table 6 lists the exact model identifiers. "Ministral-8B" denotes the 2512 checkpoint in all homogeneous experiments, whereas the heterogeneous teams use the 2410 checkpoint.

Table 6: Model checkpoints: Hugging Face Hub identifiers for the open-weight models and Gemini API model strings for the closed-source models. Qwen3-8B runs with thinking mode enabled in App. K. Nemotron-Nano-9B-v2 appears only in App. M.

Model (name in the paper)	Checkpoint
Qwen2.5-7B	`Qwen/Qwen2.5-7B-Instruct`
Qwen2.5-32B	`Qwen/Qwen2.5-32B-Instruct`
Qwen3-8B	`Qwen/Qwen3-8B`
Llama-3.1-8B	`meta-llama/Llama-3.1-8B-Instruct`
Ministral-8B (homogeneous)	`mistralai/Ministral-3-8B-Instruct-2512` [Liu et al., 2026]
Ministral-8B (heterogeneous teams)	`mistralai/Ministral-8B-Instruct-2410` [Mistral AI, 2024]
Nemotron-Nano-9B-v2	`nvidia/NVIDIA-Nemotron-Nano-9B-v2`
Gemini Flash-Lite	`gemini-3.1-flash-lite-preview`
Gemini 3.1 Pro	`gemini-3.1-pro-preview`

C.2 Prompt templates

All agents receive structured prompts that request a final answer, confidence score, and brief rationale. The format is identical across conditions, and only the peer-response block varies.

Initial solve (MCQA):

```
Answer the multiple-choice question. Be concise.

Question: {question}
{choices}

Format EXACTLY (answer and confidence FIRST, then reasoning):
FINAL: <A, B, C, or D>
CONF: <0-100>
RATIONALE: <brief reasoning, max 4 lines>
```

For math tasks, "`FINAL: <integer>`" replaces the letter constraint, and the rest of the template is analogous.

Revision (debate / self-correction / noise):

```
You already solved this problem. Now consider what
other team members said. Carefully evaluate their
reasoning -- they may be wrong. Update or keep your answer.

Question: {question}
{choices}

Your previous answer:
{own_answer}

Peer responses:
{peers}

Format EXACTLY (answer and confidence FIRST, then reasoning):
FINAL: <A, B, C, or D>
CONF: <0-100>
RATIONALE: <brief reasoning, note any disagreements>
```

The three communication conditions differ only in how `{peers}` is populated:

- **Debate** (`comm_true`): true peer snippets from the same item.
- **Self-correction** (`comm_self`): the agent's own previous output, formatted identically to a peer snippet.

- **Noise** (`comm_random`): randomly shuffled snippets drawn from other agents' responses to *different* items.

Each peer snippet is formatted as "`Agent {i}: FINAL: {answer} CONF: {conf} TEXT: {reasoning}`" and truncated to 120 tokens to keep prompts within context limits at large N.

Matched construction. Table 7 lists what is held constant and what varies across the three revision conditions. Because only the topical content of the peer block varies, the debate vs. self-correction vs. noise contrasts in §5.4 isolate peer *topical* content, not prompt structure.

Table 7: Prompt matching across revision conditions.

Element	Debate	Self-corr.	Noise
Revision scaffold, instruction template	same	same	same
Snippet format, 120-token truncation	same	same	same
Answer format (FINAL/CONF/RATIONALE)	same	same	same
Peer-block content	true peers	own prior output	other items' R1

C.3 Generation parameters

Unless stated otherwise, experiments use temperature $= 0.4$, top-$p = 0.95$, and max new tokens $= 1024$. Two exceptions apply. The temperature ablation (App. O) varies $T \in \{0.2, 0.4, 0.7, 1.0\}$ at otherwise identical settings. Qwen3-8B in thinking mode (App. K) uses a larger output budget to accommodate the reasoning trace: its longest outputs reach $16{,}192$ tokens, the length at which 108 of its $55{,}424$ agent responses (0.2%) terminate. Seeded random number generators (keyed on item ID, run ID, and round number) ensure deterministic peer selection and tiebreaking across runs.

C.4 Answer extraction

Model outputs are parsed with a priority chain. (1) `FINAL:` field (highest priority). (2) `ANSWER:` field. (3) `\boxed{}` expressions. (4) last matching letter (A–D) for MCQA or last integer for math, as fallback. Confidence is extracted from the `CONF:` field, clamped to $[0, 100]$. For thinking models, `<think>...</think>` tags are stripped before parsing. Invalid-rate monitoring flags conditions where $> 5\%$ of outputs fail to parse.

C.5 Team aggregation

The team answer is determined by plurality vote over valid parsed answers. Ties are broken by a deterministic pseudorandom choice seeded on `hash(item_id, run_id)`, ensuring reproducibility. Confidence scores are *not* used in aggregation.

C.6 Hardware

Homogeneous experiments run on NVIDIA H100 GPUs. Heterogeneous experiments run on A100 GPUs. Some Llama-3.1-8B sweeps were completed on L40S GPUs. All open-weight runs except the structure audit (App. M), which used vLLM, use Hugging Face `transformers` in bfloat16, with microbatches of 4 by default. The full research project consumed approximately $2{,}000$ GPU-hours across H100, A100, and L40S devices, including preliminary and superseded configurations not reported in the main results. We report cost primarily in token counts and additionally give measured latency and NVML energy (App. R.3). Token counts make the comparison hardware-independent and remove batch-utilization artifacts (a small microbatch underutilizing an H100 inflates J/token by a factor unrelated to the scaling claim).

C.7 Statistical methodology

Bootstrap confidence intervals. All reported 95% intervals on estimated (c, β) parameters, N_{eff} values, and Steiner decomposition components use item-level non-parametric bootstrap with at least 300 iterations (more where bootstrap precision matters: $1{,}000$ for extrapolation envelopes, $5{,}000$ for verification of reported values). For each iteration we resample items with replacement, recompute

the per-cell N_{eff}/N profile, refit the Ringelmann curve with bounds $c \in [0, 5]$, $\beta \in [0, 5]$ (numerically unconstrained; App. D.1), and collect the percentile interval. The bootstrap uses a tighter upper bound on c than the point estimation in § 4 ($c \in [0, 20]$) for optimizer stability under resampling. All observed estimates are $c \leq 1$, so the tighter bound is non-binding in practice. Item-level resampling preserves the within-item correlation structure that drives ρ, which would be destroyed by agent-level resampling.

Parameter estimation. The Ringelmann form $R(N) = 1/(1 + c(N-1)N^{-\beta})$ is estimated by non-linear least squares (`scipy.optimize.curve_fit`) on the observed N_{eff}/N values across the available team sizes, with β unconstrained (numerical bound $[0, 5]$), so that all three regimes of §3, including the linear one, are reachable. The held-out audit (App. D) is implemented in `analyses/held_out_audit.py` with the tighter bounds $c \in [0, 5]$, $\beta \in [0, 1]$. Neither bound is active in any audit cell: the calibration fits give $c < 1$ and $\beta < 0.30$, and refitting with $c \in [0, 20]$, $\beta \in [0, 5]$ changes no parameter by more than 3×10^{-5}, so the audit results are identical under both settings.

Significance tests. p-values reported in the Steiner decomposition table (App. E) come from a two-sided paired sign-flip permutation test (10^4 permutations) on item-level team correctness for the AP $-$ PP_{rev} contrast. On GPQA the discordant item counts are 9 and 8, so every permutation is at least as extreme as the observed difference and $p = 1$. No multiple-comparison correction is applied. No contrast reaches $p < 0.05$ uncorrected ($p = 0.478$, 0.090, 1.000), so a correction would not change any conclusion.

C.8 Measurement procedure: estimating (c, β) on a new model

The procedure below is implemented in the released script `analyses/held_out_audit.py`. It calibrates (c, β) from a small-team pilot and reports the predicted extrapolation to a target N.

Algorithm.

1. *Select a model, a task, and a calibration grid.* The paper uses $\{2, 3, 5\}$ for the calibration set and a held-out target $N \in \{7, 10, 15, 20, 30\}$.
2. *Run teams.* For each N, draw M items (we use M=200–1017). For each item, sample N independent agent responses (single-shot ensembling) or run τ rounds of debate with $k = N-1$ peers, recording the final answer of each agent.
3. *Per-team effective count.* For each item i, compute the Shannon entropy H_i in bits over the N agent answers and set $N_{\text{eff}}^{\text{ans}}(i) = 2^{H_i}$ (§ 3).
4. *Aggregate to $R(N)$.* Average across items at each N: $R(N) = \frac{1}{M} \sum_i N_{\text{eff}}^{\text{ans}}(i)/N$.
5. *Fit (c, β).* Nonlinear least squares of $R(N) = 1/(1 + c(N-1)N^{-\beta})$ on the calibration grid, with bounds $c \in [0, 20]$, $\beta \in [0, 5]$ (numerically unconstrained). The released script uses $c \in [0, 5]$, $\beta \in [0, 1]$, which yields identical fits on every audit cell (App. C.7). The c bound is permissive for optimizer stability. The theoretical upper bound from $\rho_N \leq 1$ at $N = 2$ is $c \leq 2^{\beta}$, and all 44 fits in this paper yield $c \leq 1$.
6. *Classify the regime.* $\beta \approx 0$ is hard ceiling at $1/c$, $0 < \beta < 1$ is sublinear scaling ($N_{\text{eff}} \sim N^{\beta}/c$), and $\beta \geq 1$ is linear(discounted by $1/(1+c)$ at $\beta = 1$).

Choice of N_{eff} scale. On bounded-answer (MCQA) tasks, $N_{\text{eff}}^{\text{ans}} = 2^H$ is close to the agreement-based $N_{\text{eff}}^{\text{agr}} = N/(1 + (N-1)\rho)$ because the entropy is bounded, $H \leq \log_2 K$. On Qwen2.5-7B at N=30, $N_{\text{eff}}^{\text{ans}}$ exceeds $N_{\text{eff}}^{\text{agr}}$ by 6–26% on MMLU-Hard and GPQA, against 42–176% on GSM-Hard (Table 15), and the fitted exponents on the two scales differ by at most 0.05 on MMLU-Hard (Table 16). On free-form tasks (GSM-Hard), distinct wrong numeric strings inflate H without reducing ρ, so $N_{\text{eff}}^{\text{ans}}$ is an upper bound on the effective independent count rather than a tight measure. The theorem-native variable is ρ. For free-form tasks, replace Step 3 with: per item, compute pairwise agreement $\rho_i = \frac{2}{N(N-1)} \sum_{j<k} \mathbb{1}[a_j = a_k]$ and use $N_{\text{eff}}^{\text{agr}}(i) = N/(1 + (N-1)\rho_i)$. The two scales are close on MCQA and decouple on free-form math, and the paper reports both, using ρ for the theorem (see App. H).

Worked example: Qwen2.5-7B on GSM-Hard, debate condition. With M=1017 items per N and τ=2 (final round R3):

- Step 4 outputs (mean $N_{\text{eff}}^{\text{ans}}$ per N): $R(2) = 0.648$, $R(3) = 0.461$, $R(5) = 0.305$.
- Step 5 yields $c = 0.566$, $\beta = 0.000$, $R^2 = 0.997$.
- Step 6: $\beta \approx 0$, hard-ceiling regime, predicted ceiling $1/c = 1.77$.
- Held-out check at N=30: predicted $N_{\text{eff}} = 1.72$, observed 1.79, relative error 4%. Across all nine (task, condition) cells, the mean relative error is 5.5% (Table 8, App. D).

Practical defaults. $M \geq 200$ items per N is sufficient for stable means. Calibration grid $\{2, 3, 5\}$ minimizes compute while keeping the fit identified (three points for two parameters), and larger grids $\{2, 3, 5, 7, 10\}$ reduce parameter uncertainty at modest additional cost. The same procedure fits (c, β) on the answer-diversity ($N_{\text{eff}}^{\text{ans}}$) and correctness-redundancy ($N_{\text{eff}}^{\text{corr}}$) levels separately. The two agree closely on bounded-answer MCQA (both hard-ceiling, c within 0.10, Table 12) and decouple on free-form math (§4).

D Held-out extrapolation audit and robustness of the functional form

This appendix supports C1 with the primary evidence for the scaling law: out-of-sample extrapolation from a small-team pilot. It reports the cross-family held-out audit (Table 8), per-cell bootstrap CIs (Table 9), a comparison with shape-matched two-parameter saturating families (Table 10), and robustness checks including the unbounded-β refit (App. D.1).

Table 8: **Ringelmann form attains the lowest held-out extrapolation error across three open-weight families and heterogeneous teams.** Each cell is estimated on $N \in \{2, 3, 5\}$ and predicts $\widehat{N}_{\text{eff}}$ at held-out $N \in \{7, 10, 15, 20, 30\}$ (9 cells per family on Qwen and Llama, 8 on Ministral; 3 heterogeneous debate cells, fitted on $N \in \{3, 6, 9, 15\}$ and evaluated at N=30). Across all three open-weight families the Ringelmann form attains 5–12% mean relative error, whereas logistic saturation yields 68.8–73.4%, indicating that the advantage is not attributable to in-sample flexibility. Const-ρ Kish is the special case $\beta = 0$ enforced. On heterogeneous teams the fit converges to $\beta = 0$ everywhere, making Ringelmann and Const-ρ Kish numerically identical.

Functional family	Form	Params	Mean RMSE	Mean rel. err.
Qwen2.5-7B (9 cells)				
Ringelmann	$1/(1+c(N-1)N^{-\beta})$	2	**0.13**	**5.5%**
Const-ρ Kish	$1/(1+(N-1)\rho)$	1	0.37	8.3%
Power law	$a + bN^c$	3	0.86	32.2%
Logistic saturation	$L/(1+e^{k(N-N_0)})$	3	1.69	73.4%
Llama-3.1-8B (9 cells)				
Ringelmann	—	2	**0.45**	**11.9%**
Const-ρ Kish	—	1	0.66	14.1%
Power law	—	3	1.02	40.8%
Logistic saturation	—	3	2.31	69.7%
Ministral-8B (8 cells)[*]				
Ringelmann	—	2	**0.18**	**5.8%**
Const-ρ Kish	—	1	0.23	6.6%
Power law	—	3	0.79	31.6%
Logistic saturation	—	3	1.64	68.8%
Heterogeneous teams (3 cells, debate only)				
Ringelmann (=Kish)	—	2	0.43	27.4%
Const-ρ Kish	—	1	0.43	27.4%
Power law	—	3	0.96	38.5%
Logistic saturation	—	3	1.34	80.6%

[*] Ministral GSM noise excluded: insufficient $N \in \{2, 3\}$ coverage to fit the small-team training set.

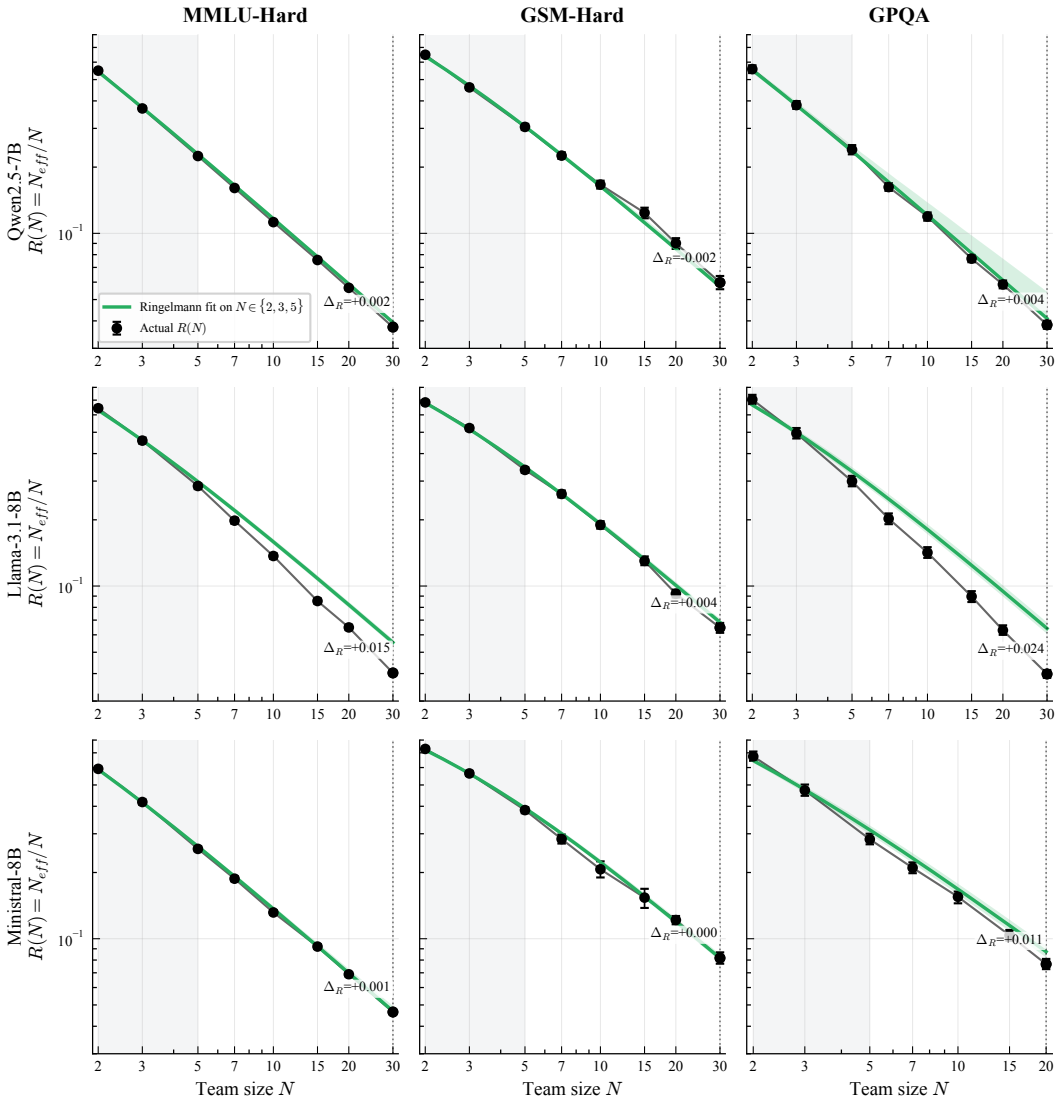


Figure 5: **Small-team runs predict the large-team Ringelmann ceiling across models and tasks.** Rows: Qwen2.5-7B (top), Llama-3.1-8B (middle), Ministral-8B (bottom). Columns: MMLU-Hard, GSM-Hard, GPQA. Black points show observed answer-diversity efficiency $R(N) = N_{\text{eff}}/N$ with 95% item-level bootstrap intervals. Green curves are the Ringelmann form $R(N) = 1/(1+c(N-1)N^{-\beta})$ estimated only on $N \in \{2, 3, 5\}$ (gray training band), with 95% bootstrap envelope, extrapolated to the largest N evaluated for that model and task(N=30, except N=20 for Ministral-8B on GPQA; Ministral-8B GSM-Hard at N=30 covers 477 of 1017 items). Residuals Δ_R at the largest N are reported per panel. Both axes are on log scale (matching Figure 9). Cross-family comparison against alternative functional forms (Const-ρ Kish, three-parameter power law $a+bN^c$, logistic) appears in Table 8, where held-out errors across all three open-weight families confirm the same ranking.

Per-cell bootstrap confidence intervals. For 23 of the 29 audit cells (Table 9) we compute an item-level bootstrap (B=1000) of the held-out mean relative error (fit on $N \in \{2, 3, 5\}$, predict $N \in \{7, \ldots, 30\}$). Across the 23 bootstrapped cells, the median point error is 4.3% (min 1.2%, max 40.1%) and the median 95% CI width is 9.5% (min 3.9%, max 38.3%). 13/23 cells have CI width below 10% and 19/23 below 15%. The widest intervals are on GPQA cells (largest: Qwen2.5-7B noise, 38.3), where 198 items and N_{eff} close to 1 make relative errors unstable.

Table 9: Per-cell held-out extrapolation error with 95% item-level bootstrap CIs (B=1000). Fit on $N \in \{2, 3, 5\}$, predict $N \in \{7, 10, 15, 20, 30\}$. Width = upper − lower CI bound. The bootstrap covers 23 of the 29 audit cells: Ministral-8B GSM-Hard and MMLU-Hard noise and the heterogeneous cells were not bootstrapped.

Model	Task	Cond.	MRE [95% CI]	CI width
Qwen2.5-7B	MMLU	debate	3.8% [1.8, 6.0]	4.2
	MMLU	self-c	4.1% [0.5, 10.2]	9.7
	MMLU	noise	6.3% [4.1, 10.2]	6.1
	GSM	debate	4.3% [2.3, 7.6]	5.3
	GSM	self-c	4.3% [1.5, 11.5]	10.0
	GSM	noise	2.5% [1.3, 10.9]	9.5
	GPQA	debate	6.0% [3.9, 22.4]	18.5
	GPQA	self-c	1.2% [0.6, 12.1]	11.5
	GPQA	noise	17.3% [5.6, 43.9]	38.3
Llama-3.1-8B	MMLU	debate	23.7% [20.2, 27.5]	7.4
	MMLU	self-c	2.9% [0.9, 6.6]	5.6
	MMLU	noise	4.8% [2.2, 7.5]	5.3
	GSM	debate	3.5% [2.2, 7.5]	5.3
	GSM	self-c	13.2% [6.7, 19.6]	12.9
	GSM	noise	12.7% [5.8, 19.0]	13.2
	GPQA	debate	40.1% [32.5, 48.3]	15.8
	GPQA	self-c	3.8% [1.1, 12.9]	11.8
	GPQA	noise	2.3% [1.2, 7.0]	5.8
Ministral-8B	MMLU	debate	1.9% [1.0, 4.9]	3.9
	MMLU	self-c	2.1% [0.6, 6.3]	5.7
	GPQA	debate	9.7% [3.5, 16.3]	12.8
	GPQA	self-c	5.2% [1.9, 21.2]	19.4
	GPQA	noise	2.5% [1.0, 6.6]	5.6
Summary over 23 cells			median 4.3%	median 9.5

Shape-matched two-parameter baselines. The power-law and logistic comparators in Table 8 have three parameters and differ in shape from a saturating efficiency curve. We therefore also compare against shape-matched two-parameter saturating families under the same held-out protocol: a rational form $1/(1+sN^{\gamma})$ and a stretched exponential $a + (1-a)e^{-(N-1)/N_c}$. Over 37 cells spanning five models from four families (Qwen2.5-7B, Qwen2.5-32B, Llama-3.1-8B, Ministral-8B, Gemini 3.1 Flash-Lite), three tasks, and the available conditions (Table 10), Ringelmann has the lowest mean error and the largest number of cells with the lowest error. On GSM-Hard, the hardest task, Ringelmann reaches 7.7% vs. 30.6% (rational) and 85.0% (stretched exponential), a 4.0×–11.1× advantage over the shape-matched families. Const-ρ Kish is competitive on the median because it is the β=0 sub-model, and that constraint holds in the many hard-ceiling cells. The additional parameter β improves extrapolation in the sublinear cells.

Table 10: Held-out extrapolation against shape-matched saturating families (37 cells, fit on $N \in \{2, 3, 5\}$, predict $N \in \{7, \ldots, 30\}$). Wins: number of cells where the form attains the lowest error.

Form	Expression	Params	Mean MRE	Median MRE	Wins
Ringelmann	$1/(1+c(N-1)N^{-\beta})$	2	**6.81%**	4.08%	**21/37**
Const-ρ Kish	$1/(1+(N-1)\rho)$	1	8.56%	3.82%	16/37
Rational	$1/(1+sN^{\gamma})$	2	32.23%	34.13%	0/37
Stretched-exp.	$a + (1-a)e^{-(N-1)/N_c}$	2	132.83%	126.01%	0/37

D.1 Robustness of the functional form

Four independent lines of evidence argue against overfitting.

(i) Out-of-sample extrapolation. Parameters estimated on $N \leq 5$ predict $N{=}30$ across the evaluated configurations (Figure 5). Out-of-sample extrapolation is the standard test for a power-law fit versus a flexible curve, formalized in the held-out audit (Table 8).

(ii) Interpretable parameter shifts. The shifts across conditions follow the mean-field prediction: peer debate yields the largest c on every task and, on GSM-Hard, the lowest β (more conformity, tighter ceiling). The two conditions without same-item peer content coincide on GSM-Hard ($c = 0.34$ for self-correction, 0.33 for noise), whereas on MCQA self-correction stays close to debate (0.84 vs. 0.85 on MMLU-Hard, 0.78 vs. 0.81 on GPQA) and noise lies lower (0.65 and 0.52). For self-consistency the fitted c (0.93 on MMLU-Hard) is close to the independently measured $\rho^{(0)}$ (0.97; App. P).

(iii) Cross-validation against an external metric. The condition ordering in estimated c (debate $\gtrsim$ self-corr. $>$ noise on MCQA, debate $>$ self-corr. $\approx$ noise on GSM-Hard) is confirmed by the independently computed correctness-level ICC $\rho^{\text{corr}}_{\text{ICC}}$ (Table 11), which does not enter the fitting procedure. The ordering persists under item-difficulty conditioning: $\rho^{\text{corr}}_{\text{cond}} = 0.74$ (debate) > 0.33 (self-corr.) > 0.04 (noise) on MMLU-Hard (App. E). On MCQA, $c \approx \rho^{\text{ans}}_{\text{eff}} \approx \rho^{\text{corr}}_{\text{ICC}}$ (within 0.14 of one another in every cell). On GSM-Hard the three diverge (by up to 0.58) because distinct wrong answers inflate answer entropy, but the ordering holds.

(iv) β is unconstrained, and the linear regime is not excluded by construction. The definition places no upper bound on β (§3). An earlier version of the fitting code bounded $\beta \in [0, 1]$, which was an implementation artifact. All fits are now run with β unbounded up to 5, which admits the linear regime $\beta \geq 1$ (ρ_N falling at least as fast as $1/N$). We refit all 38 cells (36 open-weight and 2 Gemini Flash-Lite) this way. The unconstrained fit gives $\beta_{\text{free}} \in [0.000, 0.361]$ across all 38 cells. No cell approaches the bound. Constrained and unconstrained (c, β) agree to five decimals, so the classifier is descriptive rather than constructive: the regime label reflects the data, not the imposed bound.

E Correctness-level analysis

The main paper presents the answer-level law as primary ($N^{\text{ans}}_{\text{eff}} = 2^{H(\mathbf{a})}$). This appendix supports C1 by presenting the complementary correctness-level law, estimating the same two-parameter form against $N^{\text{corr}}_{\text{eff}} = N/(1 + (N{-}1)\rho^{\text{corr}}_{\text{ICC}})$, where $\rho^{\text{corr}}_{\text{ICC}}$ is the marginal intraclass correlation of binary correctness indicators across items.

Why the regime crossing is answer-level only. The regime crossing on GSM-Hard between debate and self-correction/noise (§ 5.1) is specifically an answer-level phenomenon: at the correctness level, all GSM-Hard conditions lie in the hard-ceiling regime (Table 12). Within the framework, this is the predicted answer-space signature of conformity. Peer influence raises ρ enough to flatten β at the answer level, consolidating wrong-answer mass into fewer distinct strings. Correctness-level ceilings, driven by shared item difficulty plus low solo accuracy, are already saturated and do not move. This consolidation appears on GSM-Hard's much larger answer space, not just on 4-choice MCQA, which is consistent with, but does not establish, generalization to open-ended tasks.

Comparison of ρ measures. Table 11 compares estimated c against the three ρ quantities (answer-level entropy-derived $\rho^{\text{ans}}_{\text{eff}}$, correctness-level ICC $\rho^{\text{corr}}_{\text{ICC}}$, and pairwise agreement ρ) at $N{=}30$ on Qwen2.5-7B. On MCQA tasks the three lie close to c (all four values within 0.19 of one another, and within 0.12 under debate and self-correction). On free-form GSM-Hard, $\rho^{\text{corr}}_{\text{ICC}} \gg \rho^{\text{ans}}_{\text{eff}}$ because many distinct wrong numeric answers inflate answer entropy while correctness redundancy stays high.

Table 11: Scaling-law parameter c compared with answer-level and correctness-level correlation at N=30 (Qwen2.5-7B). $\rho_{\text{eff}}^{\text{ans}}$: equicorrelation implied by $N_{\text{eff}}^{\text{ans}} = 2^{H(\mathbf{a})}$. $\rho_{\text{ICC}}^{\text{corr}}$: marginal correctness ICC over binary indicators across items. ρ: pairwise answer agreement.

Task	Condition	Fitted c	$\rho_{\text{eff}}^{\text{ans}}$	$\rho_{\text{ICC}}^{\text{corr}}$	ρ
MMLU-Hard	Debate	.85	.89	.92	.94
	Self-corr.	.84	.82	.87	.90
	Noise	.65	.72	.79	.84
GPQA	Debate	.81	.86	.91	.93
	Self-corr.	.78	.77	.83	.87
	Noise	.52	.55	.66	.70
GSM-Hard	Debate	.57	.54	.91	.78
	Self-corr.	.34	.14	.69	.42
	Noise	.33	.13	.71	.41

Correctness-level Ringelmann fits. Table 12 shows that at the correctness level, *all* conditions fall in the hard-ceiling regime ($\beta \approx 0$). The "sublinear" regime observed on GSM-Hard self-correction and noise (Table 1) is an answer-space phenomenon: many distinct wrong numeric answers sustain non-trivial entropy even at large N, but correctness-level redundancy saturates. The qualitative condition ordering is preserved: $c_{\text{debate}} > c_{\text{self-corr}} \geq c_{\text{noise}}$ on all tasks.

Table 12: Ringelmann estimates at the correctness level vs. answer level (Qwen2.5-7B, $N \in \{2, \ldots, 30\}$). All correctness-level estimates are hard-ceiling ($\beta \approx 0$). At the answer level, GSM-Hard self-correction and noise are sublinear ($\beta \approx 0.25$), and debate is hard-ceiling on both levels.

		Answer level		Correctness level	
Task	Condition	c	β	c	β
MMLU-Hard	Debate	.85	.000	.89	.000
	Self-corr.	.84	.011	.88	.011
	Noise	.65	.000	.72	.000
GSM-Hard	Debate	.57	.014	.88	.000
	Self-corr.	.34	.252	.72	.013
	Noise	.33	.272	.70	.004
GPQA	Debate	.81	.000	.87	.000
	Self-corr.	.78	.006	.85	.008
	Noise	.52	.000	.62	.000

Item-difficulty decomposition. The marginal $\rho_{\text{ICC}}^{\text{corr}}$ conflates two sources of agent correlation: (i) shared item difficulty (all agents tend to get easy items right) and (ii) within-item social dependence (sycophancy, shared prompt bias). To isolate (ii), we condition on round-1 accuracy $\hat{p}_i^{(0)}$ as an external item-difficulty estimate (round 1 is pre-debate, τ=0). With $X_j^{(2)}$ the final-round (R3, τ=2) correctness of agent j, the conditioned ICC

$$\rho_{\text{cond}}^{\text{corr}} = \frac{\mathbb{E}_i\big[\frac{1}{N(N-1)} \sum_{j \neq k} X_j^{(2)} X_k^{(2)} - (\hat{p}_i^{(0)})^2\big]}{\mathbb{E}_i\big[\hat{p}_i^{(0)}(1-\hat{p}_i^{(0)})\big]}$$

removes the item-difficulty component. Table 13 shows that on MCQA tasks, debate induces substantial within-item social dependence ($\rho_{\text{cond}}^{\text{corr}} = 0.49$–$0.74$), self-correction induces moderate dependence (0.33 on MMLU-Hard), and noise adds near-zero social correlation (0.04). The condition ordering *debate* > *self-correction* > *noise* is preserved after conditioning. On GPQA, negative $\rho_{\text{cond}}^{\text{corr}}$ for self-correction (-0.15) and noise (-0.23) indicates that these conditions inject *more* answer diversity than item difficulty alone would predict, consistent with stochastic re-evaluation redistributing mass across answers. We omit GSM-Hard from the conditioned analysis because the round-1 item-difficulty estimate is too noisy for free-form numeric answers, producing $\rho_{\text{cond}}^{\text{corr}} > 1$ (not a valid ICC).

Table 13: Marginal vs. item-difficulty-conditioned ICC at $N{=}30$. ρ_{r1}: round-1 marginal ICC (pre-debate baseline). ρ_{r3}: round-3 marginal ICC. ρ^{corr}_{cond}: round-3 ICC conditioned on round-1 item difficulty.

Task	Condition	ρ_{r1}	ρ_{r3}	ρ^{corr}_{cond}	Ordering
MMLU-Hard	Debate	.957	.916	.741	D>S>N ✓
	Self-corr.	.957	.865	.332	
	Noise	.957	.794	.039	
GPQA	Debate	.923	.912	.495	D>S>N ✓
	Self-corr.	.923	.833	−.147	
	Noise	.923	.662	−.232	

GSM-Hard omitted: round-1 item-difficulty estimates are too noisy for free-form numeric answers, producing $\rho^{corr}_{cond} > 1$ (not a valid ICC).

Accuracy implications: Steiner decomposition. Beyond correctness-level correlation, § 5.4 reports that team accuracy is statistically equivalent within ± 5 pp across communication modes (debate vs. self-correction). Table 14 provides the matched-control accounting underlying that claim: actual productivity (AP, debate accuracy) versus potential productivity (PP_{rev}, self-correction as the no-peer matched control), with social process loss $PL_{social} = PP_{rev} - AP$. Across all three tasks the gap is within ± 2 pp at $N{=}30$ and the bootstrap intervals straddle zero, consistent with the equivalence test in §5.4 (two one-sided paired t-tests on item-level team correctness with a ± 5 pp margin: $p_{TOST} < 10^{-5}$ on MMLU-Hard, 0.001 on GSM-Hard, and 0.016 on GPQA). The re-evaluation gain column reports the absolute accuracy increase of PP_{rev} over solo ($N{=}1$) accuracy (55.5%, 10.0%, and 30.3%).

Table 14: Steiner decomposition at $N{=}30$ (Qwen2.5-7B). PP_{rev}: self-correction (matched control). AP: debate. Noise: noise-placebo team accuracy (no same-item peer content), for reference. Brackets: 95% item-level bootstrap CIs. Re-eval. gain: PP_{rev} minus solo ($N{=}1$) accuracy. p: paired permutation test of AP vs. PP_{rev} (App. C.7). p_{TOST}: equivalence of AP and PP_{rev} within ± 5 pp (two one-sided paired t-tests).

Task	PP_{rev}	AP	Noise	PL_{social}	Re-eval. gain	p	p_{TOST}
MMLU-Hard	58.7 [55.1, 62.3]	57.9 [54.3, 61.5]	56.2 [52.6, 59.8]	0.8 [−1.0, 2.6]	+3.2 pp	0.478	$< 10^{-5}$
GSM-Hard	28.2 [25.5, 31.0]	26.4 [23.6, 29.0]	25.5 [22.8, 28.1]	1.9 [−0.2, 3.9]	+18.2 pp	0.090	0.001
GPQA	31.3 [25.3, 37.9]	31.8 [25.8, 38.4]	30.3 [23.7, 36.9]	−0.5 [−4.5, 3.5]	+1.0 pp	1.000	0.016

F Diversity metrics

This appendix supports C1 by reporting the answer-diversity scale quantities (N^{ans}_{eff}, ρ^{ans}_{eff}, ρ) on which the Ringelmann fit operates. Table 15 shows their per-condition values at $N{=}30$ on Qwen2.5-7B.

Table 15: Diversity metrics at $N{=}30$, final round (Qwen2.5-7B). ρ^{ans}_{eff} from entropy-based N^{ans}_{eff}, and ρ from pairwise agreement.

Task	Condition	N^{ans}_{eff}	ρ^{ans}_{eff}	ρ	% Unan.	% Unan. wrong
MMLU	Debate	1.12	.886	.939	80%	30%
	Self-corr.	1.21	.822	.904	69%	24%
	Noise	1.37	.722	.842	48%	14%
GSM	Debate	1.79	.542	.784	47%	26%
	Self-corr.	6.10	.135	.416	12%	3%
	Noise	6.44	.126	.408	11%	2%
GPQA	Debate	1.15	.864	.928	76%	48%
	Self-corr.	1.29	.767	.865	57%	35%
	Noise	1.76	.552	.703	24%	11%

Key patterns. The condition ordering debate > self-correction > noise on ρ holds across all three tasks, with the noise placebo close to self-correction on GSM ($\rho = 0.408$ vs. 0.416) but lower

on MCQA (0.842 vs. 0.904 on MMLU, 0.703 vs. 0.865 on GPQA). The GSM regime crossing is visible directly in $N_{\text{eff}}^{\text{ans}}$: debate saturates at 1.79 while self-correction and noise reach 6.10 and 6.44 respectively, whereas on MCQA every condition stays close to the hard ceiling ($N_{\text{eff}}^{\text{ans}} \leq 1.76$). The % unanimous-wrong column quantifies the failure mode from which no answer-level aggregator can recover: it peaks at 48% on GPQA debate and falls to 11–14% under noise on MCQA (2% on GSM), consistent with peer influence concentrating mass on correlated errors.

G Agent-level transition analysis

This appendix supports C1 by resolving an apparent paradox. The decline in team-level efficiency documented in § 5 is not driven by a failure of agents to learn from peers. The opposite holds: agents *do* update toward correct peer answers, yet the aggregate efficiency $R(N) = N_{\text{eff}}/N$ declines steeply with N. Figure 6 illustrates the resolution.

Individual learning is positive. Counting agent-level wrong-to-correct (W→C) and correct-to-wrong (C→W) transitions between rounds R1 and R3, the net rate W→C − C→W is positive at every team size on every task. On GSM-Hard at N=10, individual W→C transitions outnumber C→W by roughly 18× (Table 17 reports the unconditional rates: 0.088 vs. 0.005). On average, the conformity dynamics therefore move agents *toward* correctness; at the agent level, debate is not equivalent to a noise process.

Collective gains do not scale. The bottleneck is informational, not behavioral. Even though most individual updates are W→C, redirected agents converge to the modal answer irrespective of its correctness, and on items where the round-1 modal answer is wrong the sycophancy rate α_s (0.24 on GSM-Hard, 0.60 on MMLU-Hard; App. I) is non-negligible. Per-item contributions therefore are strongly correlated across agents: the team contracts to a single answer rather than aggregating N near-independent votes. The effective team size N_{eff} tracks this contraction directly and stays at or below the matched no-peer control across N=2–30.

Mechanistic reading. Two facts stand in tension. Individual learning is positive, while team efficiency $R(N) = N_{\text{eff}}/N$ collapses with N. This pattern is consistent with the Ringelmann scaling law when conformity raises pairwise correlation faster than per-agent accuracy improves. The framework treats these as the same phenomenon viewed at two scales: agent-level W→C transitions (driven by α_l) and team-level redundancy (driven by ρ). Both are positive, but only the latter matters for aggregation, and the latter saturates first.

When the correct answer starts as a minority. Thm. 1 assumes a stable modal answer (A2–A3). The agent-level data show what happens when the correct answer is initially a minority. *(i)* Agents do move toward correct answers: W→C exceeds C→W at every N (+8 pp net on GSM-Hard), so a correct minority can spread. *(ii)* Agents also drift toward the modal answer irrespective of its correctness. Under debate, the non-modal→modal rate is 1.8–5.0× that under self-correction or noise on MMLU-Hard and GPQA, and 11–25× on GSM-Hard (Table 18). When the round-1 majority is wrong, correct-minority agents adopt it at rate $\alpha_s = 0.24$ on GSM-Hard (Table 17). *(iii)* A correct round-1 minority is rare rather than suppressed. On GSM-Hard debate (N=2–30), no agent is correct in round 1 in 86.5% of runs, and the correct answer is present but outnumbered in 3.0% (247/8136). In those runs it becomes the round-3 plurality in 59.5% of cases (95% CI 48–70%, self-correction 56.7%, noise 52.6%). On GSM-Hard, self-consistency accuracy stays flat from N=2 to 30 (App. P). Modal flips do occur. Normalized per peer, the modal-leak rate of Table 18 ((1 − modal-stay)/k, R1→R3, N=10) is about one quarter of the per-peer learning rate α_1 on MCQA (0.014 vs. 0.059 on MMLU-Hard, 0.012 vs. 0.050 on GPQA) but comparable to it on GSM-Hard (0.046 vs. 0.038). Assumption (A2) therefore holds well on MCQA and approximately on GSM-Hard, and the contraction theorem applies to first order on MCQA and as an approximation on GSM-Hard (App. I and App. A.6).

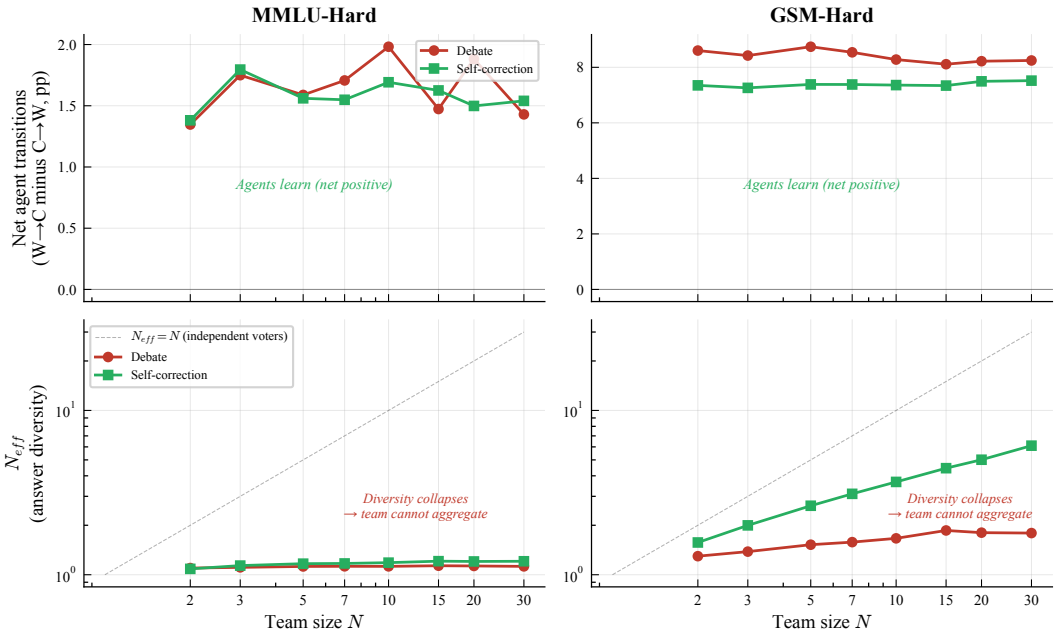


Figure 6: **Individual learning versus collective efficiency. Top:** Net agent-level transitions (W→C minus C→W) are positive at all team sizes, on both tasks, under both debate and self-correction. Agents do learn from interaction. **Bottom:** Final-round effective team size N_{eff} stays close to the no-aggregation lower bound on MMLU-Hard ($N_{\text{eff}} \leq 1.21$ at N=30). On GSM-Hard it saturates under debate ($N_{\text{eff}} = 1.79$ at N=30) and grows sublinearly only under self-correction ($N_{\text{eff}} = 6.10$ at N=30). In all cases it remains far below the independent-voter line $N_{\text{eff}} = N$. Individual gains are not aggregated because answer diversity declines.

H Communication density ablation

This appendix supports C2 (the $k\tau$ product theorem) by ablating communication density. The main experiments use fully connected teams ($k = N-1$, every agent sees all peers). We examine whether restricting communication density reduces conformity. We run a dedicated $k\tau$ ablation on GSM-Hard and MMLU-Hard (200 items each, Qwen2.5-7B) with $k \in \{1, 2, 4, 9, 29\}$ peers, $\tau \in \{0, \ldots, 5\}$ communication rounds (τ=0: independent answers before any peer exchange), and $N \in \{3, 10, 30\}$.

Table 16: Scaling exponent β by communication density k (Qwen2.5-7B, 200 items, final round τ=5 on both tasks). β_{ent}: entropy scale ($N_{\text{eff}}^{\text{ans}} = 2^H$). β_{agr}: pairwise agreement scale ($N_{\text{eff}}^{\text{agr}} = N/(1+(N-1)\rho)$). The theorem operates on ρ. On free-form GSM-Hard the two scales diverge (large β_{ent} from wrong-answer mass spread, § 3). On MCQA MMLU-Hard the two scales coincide because the answer space is bounded (K=4).

k	GSM β_{agr}	GSM β_{ent}	MMLU β_{agr}	MMLU β_{ent}
1	0.17	0.40	0.06	0.11
2	0.13	0.27	0.06	0.10
4	0.08*	0.25	0.02*	0.05
9	0.06*	0.17	0.03*	0.05

*2-point log-slope estimate.

Table 16 shows a scale-dependent pattern and supports the theorem on a second task. On the *pairwise agreement* scale (the theorem's native variable), β_{agr} is small on both tasks: 0.06–0.17 on GSM-Hard and 0.02–0.06 on MMLU-Hard, decreasing with k (monotonically on GSM-Hard, 0.06, 0.06, 0.02, 0.03 on MMLU-Hard) and converging toward the theorem's prediction of $\beta = 0$. On the *entropy* scale,

the two tasks differ markedly. On bounded-answer-space MMLU-Hard (K=4 choices), $\beta_{\text{ent}} \approx \beta_{\text{agr}}$: the two scales coincide, as predicted in § 3. On free-form GSM-Hard, β_{ent} is much larger (0.17–0.40) because distinct wrong numeric answers inflate answer entropy $H(\mathbf{a})$ without proportionally reducing pairwise agreement, and at fixed k this inflation is amplified because minority answers spread across more agents in larger teams. The theorem is thus supported on its native scale across both task types. The GSM–MMLU contrast in β_{ent} identifies the dispersion of wrong-answer mass as a task-structural property, not a violation of the theorem.

The $k\tau$ product structure. The matched-$k\tau$ collapse test is reported in Table 3 of §5.2. With $\tau \in \{1, \ldots, 5\}$, only $k\tau \in \{2, 4, 8\}$ admit two or more (k, τ) realizations. On GSM-Hard the spread is $\Delta\rho \leq 0.031$, and no matched pair differs by more than 1.4 item-paired standard errors (200 items). This spread is of the same order as the spread across k at τ=0 (0.031 at N=10), where all runs are independent and the expected ρ does not depend on k. On MMLU-Hard the spread is 0.028, 0.052, and 0.018 at $k\tau = 2$, 4, and 8, against 0.004 at τ=0, and its sign is systematic: at every matched product and both N, ρ increases with k (largest for $(4, 1)$ vs. $(1, 4)$ at N=10, item-paired $z = 3.4$). The largest deviation thus involves the single-round configuration $(4, 1)$, where the theorem's binary-coarsening derivation introduces second-order corrections that its linearization does not capture. Restricting to $\tau \geq 2$ tightens the spread to $\Delta\rho \leq 0.025$ on GSM-Hard and ≤ 0.029 on MMLU-Hard. The MMLU-Hard ordering persists for $\tau \geq 2$ ($(2, 2)$ exceeds $(1, 4)$ by 0.027–0.030, $z = 1.9$–3.0), so on this task the product structure holds only approximately.

Stratifying by initial agreement. The aggregate $\Delta\rho$ in Table 3 averages over all items per cell. Conditioning instead on per-item initial agreement $\bar{\rho}^{(0)}$ (a difficulty proxy) reveals a residual structure on GSM-Hard (N=10): restricting to comparable multi-round configurations ($\tau \geq 2$, which leaves $k\tau \in \{4, 8\}$), $\Delta\rho$ across realizations of the same $k\tau$ is 0.020 and 0.001 (at $k\tau = 4$ and 8) in high-agreement items ($\bar{\rho}^{(0)} > 0.7$, 80 items), 0.053 and 0.028 in mid-agreement items ($\bar{\rho}^{(0)} \in [0.4, 0.7]$, 40 items), and 0.055 and 0.012 in low-agreement items ($\bar{\rho}^{(0)} < 0.4$, 80 items). The product law thus holds most closely where the high-agreement linearization of Thm. 1 is valid. Outside this regime the spread at $k\tau$=4 is 2.7–2.8-fold larger, although it does not increase monotonically as $\bar{\rho}^{(0)}$ decreases. Single-round (τ=1) realizations also deviate from multi-round outcomes at the same product: including them raises the maximum $\Delta\rho$ to 0.039 (high), 0.096 (mid), and 0.078 (low), consistent with larger second-order corrections during the τ=1 initial transient. Figure 7 shows the low-, mid-, and high-agreement strata.

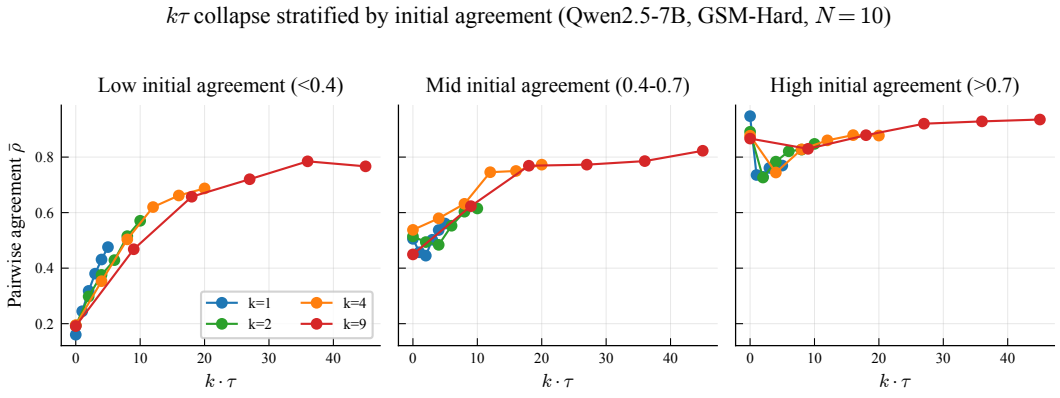


Figure 7: $k\tau$ collapse stratified by initial item-level agreement (Qwen2.5-7B, GSM-Hard, N=10). Curves are pairwise agreement $\bar{\rho}$ vs. $k\tau$, one per peer count k, starting from the independent answers (τ=0) at $k\tau$=0. Strata are defined on the k=1 run's τ=0 agreement, so that point is displaced by selection. Left panel (low agreement, $\bar{\rho}^{(0)} < 0.4$, 80 items): visible cross-k separation at small $k\tau$, narrowing as $k\tau$ grows. Middle panel (mid agreement, $\bar{\rho}^{(0)} \in [0.4, 0.7]$, 40 items): cross-$k$ separation of similar magnitude at $k\tau$=4, also narrowing by $k\tau$=8. Right panel (high agreement, $\bar{\rho}^{(0)} > 0.7$, 80 items): apart from the selection-displaced k=1 point at τ=0, curves nearly coincide, consistent with the theorem's high-agreement regime.

Why the large β_{ent} does not violate the theorem. The theorem predicts contraction of ρ, not entropy. At dense communication ($k \to N-1$), $\beta_{\text{agr}} \to 0$ on both tasks while β_{ent} persists only on GSM-Hard, isolating answer-space inflation as a task-structural quantity independent of communication dynamics.

I Two-rate mean-field extension

This appendix supports C2 by extending the mean-field model with asymmetric conformity rates and grounding assumption (A2) in agent-level transition diagnostics. The single conformity coefficient α in Eq. 3 averages over two distinct processes with empirically different magnitudes.

Definition 2 (Asymmetric Conformity Rates)**.** *Let* $\alpha_s := P(\textit{correct agent adopts wrong peer majority})$ *denote the* sycophancy rate *and* $\alpha_l := P(\textit{wrong agent adopts correct peer majority})$ *denote the* learning rate*.*

Table 17 reports these rates from the agent-level debate data. On both tasks, $\alpha_l > \alpha_s$: agents adopt a correct majority more readily than an incorrect one. The asymmetry is moderate on MMLU-Hard ($\alpha_s/\alpha_l \approx 0.67$) and pronounced on GSM-Hard (≈ 0.30). A possible explanation, which we do not test directly, is that correct mathematical derivations are easier to verify than incorrect ones.

Table 17: Asymmetric conformity rates at N=10, round 3 (Qwen2.5-7B, debate).

Task	α_s (n)	α_l (n)	Ratio	Unconditional C→W / W→C
MMLU-Hard	0.60 (119)	0.89 (141)	0.67	0.025 / 0.045
GSM-Hard	0.24 (170)	0.79 (233)	0.30	0.005 / 0.088

Accuracy dynamics (item-conditioned heuristic). When the majority is correct, wrong-minority agents switch at rate α_l, and when the majority is wrong, correct-minority agents switch at rate α_s:

$$p^{(\tau+1)} \approx \begin{cases} p^{(\tau)} + (1-p^{(\tau)})\,\alpha_l & \text{if } p^{(\tau)} > 0.5, \\ p^{(\tau)}(1-\alpha_s) & \text{if } p^{(\tau)} < 0.5. \end{cases} \tag{11}$$

When $\alpha_l > \alpha_s$, accuracy grows toward 1 faster than it shrinks toward 0.

Agreement dynamics and robustness of scaling predictions. On a per-item basis, items where the modal answer is correct contract pairwise disagreement at rate α_l (non-modal agents adopting correct modal), and items where the modal answer is wrong contract at rate α_s (correct minority adopting wrong modal). Let f_c denote the fraction of items where the modal answer is correct (the population's modal-level accuracy). The population-mean disagreement contracts as

$$d^{(\tau+1)} \approx d^{(\tau)}(1-\alpha_{\text{eff}}), \qquad \alpha_{\text{eff}} = f_c\,\alpha_l + (1-f_c)\,\alpha_s.$$

As long as $\alpha_{\text{eff}} > 0$ (which holds whenever $\alpha_l, \alpha_s \in (0,1)$), the contraction is geometric, so $\rho^{(\tau)} \to 1$ as $\tau \to \infty$ and the hard-ceiling prediction is preserved. The $k\tau$ product structure extends by the same argument as Thm. 1: replacing the single α_1 with α_{eff} in Eq. 4 gives the same $k\tau$-product collapse with a renormalized contraction rate. The single-α model is therefore sufficient for *agreement dynamics and scaling predictions*, while the two-rate extension is needed for *accuracy-trajectory predictions*, where the asymmetry $\alpha_l > \alpha_s$ drives the convergence of $f_c \to 1$.

Empirical transition matrices grounding Thm. 1. Thm. 1 idealizes the modal class as stable (assumption A2). Table 18 reports the empirical modal/non-modal transition rates on Qwen2.5-7B across all three tasks and conditions, computed by anchoring on the round-1 modal answer per item and tracking persistence to round 3. Three observations: (i) Modal-stay rates under debate are high on MCQA (85–91%) and moderate on GSM-Hard (54–63%), and across all three conditions they span 79–91% and 54–67%, respectively, placing the modal-leak rate ($1 -$ modal-stay) in $[0.09, 0.46]$ and confirming that A2 is well-satisfied on bounded-answer tasks and approximate on free-form math. (ii) Non-modal-to-modal rates are substantially higher under debate (38–62% on MCQA, 25–41% on GSM-Hard) than under self-correction or noise (11–28% on MCQA, 2–3% on GSM-Hard), consistent with the conformity mechanism. (iii) The per-peer learning rate $\alpha_1 := P(\text{nm} \to \text{m})/k$ shrinks with N on every cell, consistent with the high-agreement saturation argument: as the modal

class grows, the marginal information from each additional modal peer decays, and the contraction enters the plateau regime predicted by Eq. 4.

Table 18: Empirical modal/non-modal transition rates on Qwen2.5-7B (round 1 $\rightarrow$ round 3, anchored on round-1 modal answer per item). Modal-stay := P(R1-modal stays at the same answer through R3). The modal-leak rate is $1 -$ modal-stay. nm$\rightarrow$m := probability that an R1-non-modal agent moves to the R1-modal answer by R3. α_1 := nm$\rightarrow$m $/k$ (per-peer learning rate, distinct from the agent-level α_l of Def. 2).

		Modal-stay			nm$\rightarrow$m			α_1 per peer		
Task	Cond.	N=5	N=10	N=30	N=5	N=10	N=30	N=5	N=10	N=30
MMLU	deb.	.85	.88	.90	.38	.53	.62	.096	.059	.021
	self	.89	.90	.90	.18	.24	.24	.045	.027	.008
	noi.	.83	.83	.88	.15	.28	.19	.038	.031	.007
GSM	deb.	.54	.59	.63	.25	.34	.41	.062	.038	.014
	self	.59	.60	.60	.02	.03	.03	.006	.003	.001
	noi.	.63	.67	.67	.02	.02	.02	.004	.002	.001
GPQA	deb.	.88	.90	.91	.52	.45	.55	.130	.050	.019
	self	.88	.88	.87	.17	.25	.20	.043	.027	.007
	noi.	.79	.79	.80	.11	.21	.11	.027	.024	.004

These diagnostics support A2 as an approximation rather than an exact identity: the modal class loses members at non-negligible rates, especially on GSM-Hard, and the robustness analysis above shows that the $k\tau$ structure is preserved under this leakage with a renormalized contraction rate.

J Cross-model and cross-scale verification experiments

This appendix supports C3 by extending the core scaling-law identification (§ 5, Qwen2.5-7B, full N=2–30, three conditions) across model families, scale, paradigm, and team composition. Each verification experiment below evaluates a specific framework prediction. Table 5 in App. C.1 tabulates scope per experiment. Table 19 summarizes every (c, β) estimate. Table 20 organizes the same estimates by framework prediction. The held-out audit across families is in App. D, and the noise vs. self-correction comparison across families is in App. P.

Table 19: Master table of all estimated Ringelmann parameters (c, β). $R(N) = 1/(1+c(N-1)N^{-\beta})$. Regime: H = hard-ceiling ($\beta \approx 0$), S = sublinear ($\beta > 0.1$). $\widehat{N}_{\text{eff}}$: observed effective team size at the largest N evaluated.

Model	Task	Cond.	c	β	Reg.	$\widehat{N}_{\text{eff}}$
Qwen2.5-7B (core, N=2–30)						
Qwen2.5-7B	MMLU	debate	.85	.00	H	1.12
Qwen2.5-7B	MMLU	self-c	.84	.01	H	1.21
Qwen2.5-7B	MMLU	noise	.65	.00	H	1.37
Qwen2.5-7B	GSM	debate	.57	.01	H	1.79
Qwen2.5-7B	GSM	self-c	.34	.25	S	6.10
Qwen2.5-7B	GSM	noise	.33	.27	S	6.44
Qwen2.5-7B	GPQA	debate	.81	.00	H	1.15
Qwen2.5-7B	GPQA	self-c	.78	.01	H	1.29
Qwen2.5-7B	GPQA	noise	.52	.00	H	1.76
Llama-3.1-8B (cross-family, N=2–30)						
Llama-3.1-8B	MMLU	debate	.62	.00	H	1.21
Llama-3.1-8B	MMLU	self-c	.49	.00	H	1.94
Llama-3.1-8B	MMLU	noise	.51	.00	H	1.81
Llama-3.1-8B	GSM	debate	.47	.00	H	1.94
Llama-3.1-8B	GSM	self-c	.22	.36	S	10.8
Llama-3.1-8B	GSM	noise	.18	.30	S	10.4
Llama-3.1-8B	GPQA	debate	.55	.00	H	1.21
Llama-3.1-8B	GPQA	self-c	.44	.00	H	2.12
Llama-3.1-8B	GPQA	noise	.45	.00	H	2.13
Ministral-8B (cross-family, N=2–30)						
Ministral-8B	MMLU	debate	.71	.00	H	1.39
Ministral-8B	MMLU	self-c	.58	.01	H	1.76
Ministral-8B	MMLU	noise	.54	.00	H	1.81^{p}
Ministral-8B	GSM	debate	.40	.00	H	2.45^{p}
Ministral-8B	GSM	self-c	.28	.24	S	$5.74^{p,q}$
Ministral-8B	GSM	noise	.32	.17	S	4.54^{q}
Ministral-8B	GPQA	debate	.57	.00	H	1.53^{q}
Ministral-8B	GPQA	self-c	.41	.00	H	2.23^{q}
Ministral-8B	GPQA	noise	.37	.00	H	$2.58^{p,q}$
Qwen2.5-32B (4× scale, N=2–30, without N=15, and without N=10 on GPQA)						
Qwen2.5-32B	MMLU	debate	.96	.00	H	1.03
Qwen2.5-32B	MMLU	self-c	.88	.00	H	1.15
Qwen2.5-32B	MMLU	noise	.88	.02	H	1.21
Qwen2.5-32B	GSM	debate	.77	.00	H	1.24
Qwen2.5-32B	GSM	self-c	.45	.18	S	3.81
Qwen2.5-32B	GSM	noise	.45	.18	S	3.82
Qwen2.5-32B	GPQA	debate	.97	.00	H	1.03
Qwen2.5-32B	GPQA	self-c	.92	.04	H	1.16
Qwen2.5-32B	GPQA	noise	.85	.02	H	1.23^{p}
Heterogeneous teams (Qwen+Llama+Ministral, N=3–30)						
Hetero	MMLU	debate	.54	.00	H	1.34
Hetero	GSM	debate	.35	.00	H	2.25
Hetero	GPQA	debate	.56	.00	H	1.27
Thinking models						
Qwen3-8B-think	MMLU	debate	1.00	.01	H	1.02^{c}
Qwen3-8B-think	MMLU	self-c	.89	.01	H	1.11^{c}
Qwen3-8B-think	MMLU	noise	.88	.00	H	$1.10^{c,p}$
Gemini-FL (min. think)	GSM	debate	.95	.00	H	1.08^{p}
Gemini-FL (min. think)	GPQA	debate	.86	.00	H	1.11^{p}

^{c} $N \leq 5$. ^{p} partial item coverage at largest N. ^{q} largest N available is N=20.

Frontier solo evaluation (Gemini 3.1 Pro). The master fits table excludes Gemini 3.1 Pro because no Ringelmann curve can be estimated from a single team size. Solo accuracy at N=1: 97.0% on

MMLU-Hard (n=200), 81.5% on GSM-Hard (n=200), and 88.9% on GPQA Diamond (n=198). The framework predicts $c \to 1$ at this accuracy regime (§3.2), making the implied $1/c$ ceiling negligibly above one effective agent. Multi-agent runs on Gemini 3.1 Pro were not performed; the prediction $c \to 1$ is therefore untested for this model. The solo accuracies show that frontier models operate in the high-accuracy regime in which the framework predicts saturation, complementing Qwen3-8B thinking mode's $c = 1.00$ at 87.2% solo (Table 19).

The (c, β) landscape. Figure 8 plots all 44 estimates from Table 19 on the (c, β) plane, color-coded by communication condition and shape-coded by model variant. The same form fits every cell, with parameter shifts that classify each into the regime predicted by § 3.

Table 20: Fitted Ringelmann parameters (debate condition) organized by framework prediction. $R(N) = 1/(1+c(N-1)N^{-\beta})$ estimated against observed N_{eff}/N. $\widehat{N}_{\text{eff}}$: observed (not predicted) effective team size at the largest N evaluated. R^2 is computed on $N \geq 2$, since $R(1)$=1 by construction.

Prediction evaluated	Model × Task	c	β	R^2	$\widehat{N}_{\text{eff}}$
Hard ceiling on MCQA (3 families, N=2–30)					
	Qwen2.5-7B × MMLU	.85	.000	.9995	1.12
	Llama-3.1-8B × MMLU	.62	.000	.9942	1.21
	Ministral-8B × MMLU	.71	.000	.9996	1.39
	Qwen2.5-7B × GPQA	.81	.000	.9993	1.15
	Llama-3.1-8B × GPQA	.55	.000	.9814[c]	1.21
	Ministral-8B × GPQA	.57	.000	.9918	1.53[q]
Regime structure on free-form math (answer level, N=2–30)					
	Qwen2.5-7B × GSM	.57	.014	.9989	1.79
	Llama-3.1-8B × GSM	.47	.000	.9993	1.94
	Ministral-8B × GSM	.40	.000	.9986	2.45[p]
Scale / paradigm tightens ceiling on debate					
	Qwen2.5-32B × GSM (4× scale)	.77	.000	.9982	1.24
	Gemini-FL × GSM (frontier API)	.95	.000	.9995	1.08[p]
Thinking models: tightest ceiling					
	Qwen3-8B-think × MMLU	1.00	.015	—	1.02[a]
	Gemini-FL (min. think) × GPQA	.86	.000	.9998	1.11[p]

[a] $N \leq 5$. [p] partial item coverage at largest N. [q] largest N available is N=20.
[c]Lowest R^2 in this table: Llama-3.1-8B N_{eff} on GPQA peaks at N=5 and declines at larger N, so the monotone fit overshoots.

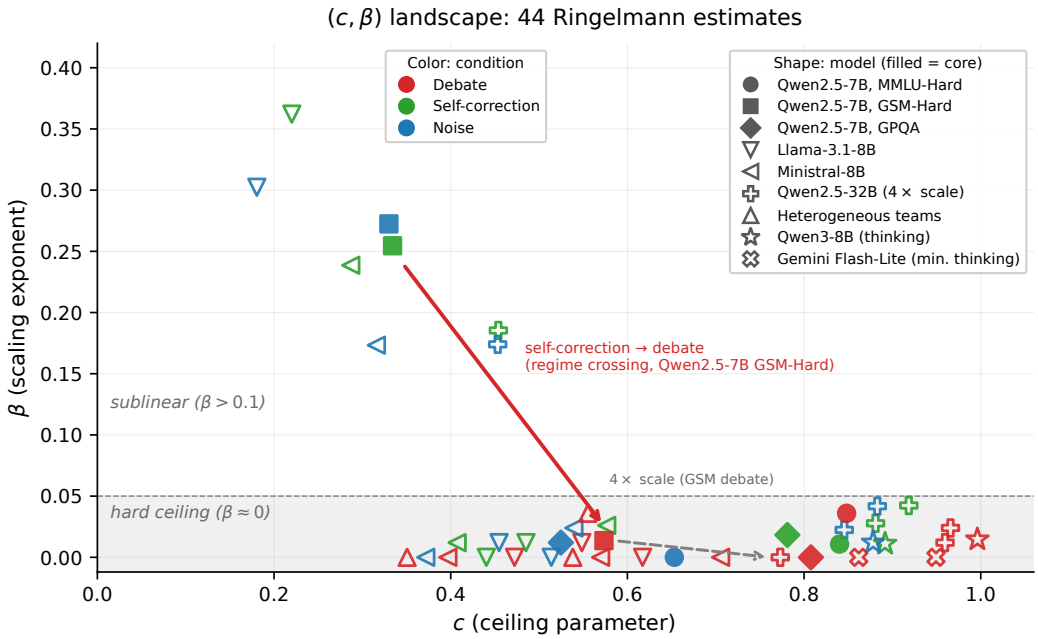


Figure 8: **The (c, β) landscape across evaluated configurations.** Each point is one Ringelmann estimate across 6 models plus heterogeneous teams, 3 tasks, and 3 conditions (44 estimates total, also reported in Table 19). *Color*: condition (red = debate, green = self-correction, blue = noise). *Filled shapes*: core Qwen2.5-7B (circles = MMLU-Hard, squares = GSM-Hard, diamonds = GPQA). *Open shapes*: cross-model variants (inverted triangles = Llama-3.1-8B, left-pointing triangles = Ministral-8B, plus signs = Qwen2.5-32B ($4\times$ scale), upward triangles = heterogeneous teams, stars = Qwen3-8B (thinking), crosses = Gemini Flash-Lite (minimal thinking)). Inside the hard-ceiling band ($\beta < 0.05$), markers are displaced upward by at most 0.036 to avoid overlap. Exact values are in Table 19. **Red arrow**: answer-level regime crossing on GSM-Hard (self-correction sublinear $\rightarrow$ debate hard-ceiling), same pattern on Llama-3.1-8B (Table 19). **Gray dashed arrow**: $4\times$ model scale tightens the hard ceiling on GSM debate (c: 0.57 $\rightarrow$ 0.77), and the same ceiling-tightening pattern holds on MMLU debate (0.85 $\rightarrow$ 0.96).

J.1 Across 7–8B model families

Regime crossing on GSM-Hard replicates across families. The main-text canonical case (Qwen2.5-7B, §5.1) shows debate at $c = 0.57$, $\beta \approx 0$ (hard ceiling) versus self-correction at $c = 0.34$, $\beta = 0.25$ and noise at $c = 0.33$, $\beta = 0.27$ (both sublinear). On Llama-3.1-8B the same crossing is visible with a larger contrast: debate $c = 0.47$, $\beta = 0.000$, self-correction $c = 0.22$, $\beta = 0.361$, noise $c = 0.18$, $\beta = 0.300$ (Table 19). On Ministral-8B, debate lies at $c = 0.40$, $\beta \approx 0$ (hard ceiling), while self-correction and noise are sublinear ($c = 0.28$, $\beta = 0.24$ and $c = 0.32$, $\beta = 0.17$, both fitted on $N \leq 20$). All three families exhibit the regime crossing: hard-ceiling debate, sublinear when peer content is removed.

Hard ceiling on MCQA across families. All three 7–8B families produce $\beta = 0$ on MCQA (Table 20, top block). Cross-model variation appears in c (Llama 0.62 vs. Qwen 0.85 vs. Ministral 0.71) rather than in the regime. The hard-ceiling regime on GPQA, originally established on Qwen, also replicates on Llama-3.1-8B ($\beta = 0$, $c = 0.55$) and on Ministral-8B ($\beta = 0$, $c = 0.57$, with self-correction $c = 0.41$ and noise $c = 0.37$ also at $\beta = 0$), so that on all three families GPQA behaves like MMLU-Hard rather than GSM-Hard. The lower c of Llama-3.1-8B and Ministral-8B on GSM-Hard debate indicates more dispersed errors; the next paragraph shows that solo accuracy alone does not account for this difference.

Cross-model parameter consistency. Re-running the parameter-estimation audit (Table 8, App. D) and the modal/non-modal transition diagnostics (Table 18) on Llama-3.1-8B preserves both rankings: Ringelmann remains the lowest held-out error family (11.9% mean relative error vs 14.1% for const-ρ

Kish, 40.8% for power law, 69.7% for logistic), and the qualitative transition pattern matches Qwen (modal-stay high on MCQA, lower on GSM-Hard, debate non-modal-to-modal rate 1.8–$4.6\times$ the rate under self-correction or noise on MCQA and 3.5–$6.1\times$ on GSM-Hard, per-peer α_1 decaying with N). Llama's debate modal-stay rate on GSM is lower than Qwen's (26–40% vs 54–63%). Llama's GSM solo accuracy (5.3% at $N{=}1$ on the $1{,}017$ items) is also lower than Qwen's (10.0%), so for this pair lower solo accuracy and higher modal leakage cannot be separated. Ministral-8B, however, has a higher GSM solo accuracy than Qwen (26.2%) together with a lower debate c (0.40 vs. 0.57), so solo accuracy alone does not account for the parameter differences. The framework's prediction is consistent with what we observe: regimes are shared (both at $\beta \approx 0$ on GSM debate) while parameters c and α_s differ in the predicted direction (lower c co-occurs with higher modal leakage; with two families this is an observation, not an established correlation). Cross-model parameter shifts without regime crossing are consistent with § 3.

J.2 Scale sensitivity: 7B vs. 32B

We test whether the hard-ceiling regime ($\beta \approx 0$) persists at larger model scale. Qwen2.5-32B-Instruct provides a within-family evaluation at $4\times$ parameter scale on GSM-Hard, MMLU-Hard, and GPQA (200 / 200 / 198 items, team sizes $N \in \{2, 3, 5, 7, 10, 20, 30\}$, without $N{=}10$ on GPQA, all three conditions).

Three findings emerge. First, **32B debate tightens the hard ceiling**. At 7B, GSM-Hard debate has $c = 0.57$, $\beta \approx 0$. At 32B, c rises to 0.77 ($\beta = 0.00$, $R^2 = 0.998$), reducing the diversity ceiling from $1/c = 1.74$ to 1.29. The same tightening shows on GPQA: c moves from 0.81 at 7B to 0.97 at 32B (debate, $\beta \approx 0$, Table 19), placing the 32B GPQA ceiling within 1.03 effective agents at $N{=}30$. Second, **the 32B model is more correlated, not less**: $\widehat{N}_{\text{eff}}(30) = 1.24$ at 32B versus 1.79 at 7B on GSM-Hard, and 1.03 versus 1.15 on GPQA. Solo accuracy increases, but this gain reflects the stronger base model rather than improved collective reasoning. Third, **noise $\approx$ self-correction replicates at 32B with N up to** 30: self-correction gives $c = 0.45$, $\beta = 0.18$ and noise gives $c = 0.45$, $\beta = 0.18$, nearly identical, consistent with the mechanism-agnostic character of the Ringelmann law across model scales. On GPQA the noise condition covers $122/198$ items at $N{=}30$. Omitting $N{=}30$ gives $c = 0.847$, $\beta = 0.024$ (all sizes: 0.845, 0.022), and restricting every N to the 122 covered items gives $c = 0.863$, $\beta = 0.020$. Both lie within the bootstrap 95% CI for c, $[0.78, 0.92]$.

K Thinking model: stronger reasoning tightens the ceiling

This appendix supports C3 by evaluating whether models with extended internal reasoning are more resistant to peer conformity and therefore escape the hard-ceiling regime. We use Qwen3-8B in thinking mode on MMLU-Hard (724 items, $N \in \{2, 3, 5\}$, all three communication conditions, with the noise condition at $N{=}5$ covering 26 of the 724 items).

The results contradict this hypothesis. The thinking model has substantially higher solo accuracy (87.2% vs. 55.5% for Qwen2.5-7B), yet debate drives correlation *higher* ($\rho = 0.983$ at $N{=}5$ vs. 0.911 for the non-thinking model) and effective team size *lower* ($\widehat{N}_{\text{eff}} = 1.02$ vs. 1.12). The estimated Ringelmann parameters confirm this: debate gives $c = 1.00$, $\beta \approx 0$, the hard-ceiling regime with ceiling $1/c = 1.00$ (Table 21).

Table 21: Ringelmann comparison: thinking vs. non-thinking model (MMLU-Hard). The thinking model has higher solo accuracy but a tighter Ringelmann ceiling under debate.

Model	Condition	Solo acc.	ρ ($N{=}5$)	N_{eff} ($N{=}5$)	c	Ceiling
Qwen2.5-7B	Debate	.555	.911	1.12	.85	1.18
	Self-corr.	.555	.879	1.17	.84	1.19
Qwen3-8B-think	Debate	.872	.983	1.02	1.00	1.00
	Self-corr.	.872	.920	1.11	.89	1.12

At the agreement scale, the observations are as follows. Initial pairwise agreement is already high for both models on MMLU-Hard at $N{=}5$ ($\rho^{(0)} \approx 0.91$ for Qwen3-think, $\rho^{(0)} \approx 0.96$ for Qwen2.5-7B),

but two rounds of debate drive the thinking model to tighter consensus ($\rho = 0.983$ vs. $\rho = 0.911$ for the non-thinking model). In the thinking model, stronger per-agent reasoning is associated with stronger rather than weaker conformity once peer answers are available, leaving less room for independent evidence. These observations are consistent with the framework without modification. Within the tested setting, the practical implication is that *stronger models require fewer debating agents, not more*.

L Heterogeneous teams

This appendix supports C3 by evaluating whether the same Ringelmann form applies when teams mix three model families (Qwen, Llama, Ministral) at team sizes $N \in \{3, 6, 9, 15, 30\}$ on MMLU-Hard, GSM-Hard, and GPQA (200, 200, and 198 items, respectively). The framework predicts that architectural diversity lowers initial correlation $\rho^{(0)}$ and raises the ceiling $1/c$ while preserving the regime β. Within-family vs. between-family pairwise agreement (Table 23) also tests the block-exchangeable extension of the variance-matching identity (App. A.7).

Table 22: Heterogeneity shifts the Ringelmann ceiling. Debate at N=30. Homogeneous rows: Qwen2.5-7B on the full item sets (724 / 1,017 / 198 items). Heterogeneous rows: Qwen+Llama+Ministral teams on 200 / 200 / 198 items, a subset of the homogeneous item sets (identical on GPQA).

Task	Setting	$N_{\text{eff}}^{\text{ans}}$	$\rho_{\text{eff}}^{\text{ans}}$	Fitted c	β	Ceiling $1/c$
MMLU-Hard	Homogeneous	1.12	.886	.85	.000	1.18
	Heterogeneous	1.34	.735	.54	.000	1.86
GSM-Hard	Homogeneous	1.79	.542	.57	.014	1.74
	Heterogeneous	2.25	.426	.35	.000	2.85
GPQA	Homogeneous	1.15	.864	.81	.000	1.24
	Heterogeneous	1.27	.778	.56	.000	1.80

The debate comparison in Table 22 shows the same declining-efficiency structure as homogeneous debate, with c falling and the asymptotic ceiling $1/c$ rising on every task while β stays at 0. The shift is a ceiling change, not a regime change. Restricting the homogeneous runs to the heterogeneous item subsets gives $c = 0.84$ on MMLU-Hard and $c = 0.58$ on GSM-Hard (ceilings 1.19 and 1.73), so the ceiling shift does not arise from the difference in item sets. Figure 9 shows that the small-N Ringelmann fit also extrapolates to the largest N on heterogeneous data.

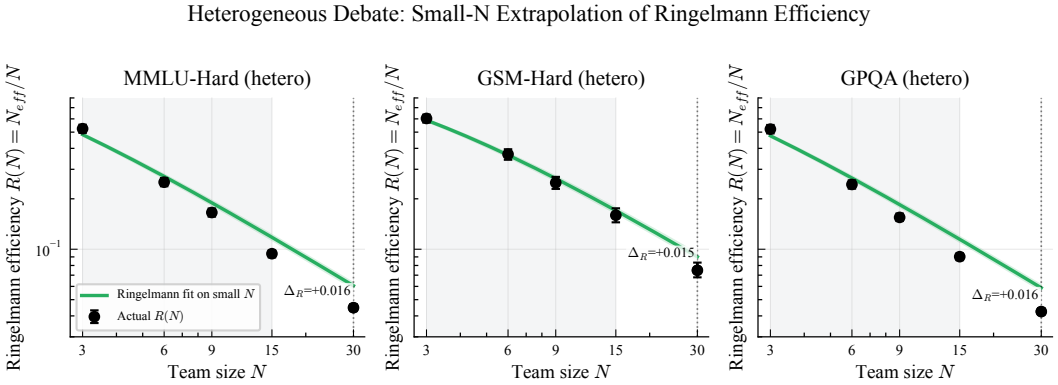


Figure 9: **Small-N extrapolation of the Ringelmann scaling law on heterogeneous debate teams** across MMLU-Hard, GSM-Hard, and GPQA. Green curves are the Ringelmann form $R(N) = 1/(1+c(N-1)N^{-\beta})$ estimated on $N \in \{3, 6, 9, 15\}$ (gray band, identical for the three tasks), with 95% bootstrap envelopes, extrapolated to the largest N available(N=30). The same form continues to fit, with a higher ceiling ($1/c$) than in the homogeneous setting, although at N=30 the small-N fit overestimates R by $\Delta_R = 0.015$–0.016 (20–39% of the observed value).

Within- vs. between-family pairwise agreement. The block-exchangeable extension (App. A.7) predicts that pairs from the same model family should agree more strongly than pairs from different families. We test this directly on the heterogeneous data by enumerating all final-round agent pairs within each item and partitioning by family pairing (Table 23). Aggregated across all cells with $N \geq 6$ (so within-family pairs exist), $\bar{\rho}_{\text{within}} = 0.824$ vs. $\bar{\rho}_{\text{between}} = 0.761$ ($\Delta = +0.063$), with all six family-pair combinations ranking in the predicted direction: same-family pairs (Qwen-Qwen 0.82, Llama-Llama 0.82, Ministral-Ministral 0.83) agree more than cross-family pairs (Qwen-Llama 0.75, Qwen-Ministral 0.77, Llama-Ministral 0.76, all pair-weighted over the same $N \geq 6$ cells). The gap is largest on free-form GSM-Hard (Δ up to $+0.15$ at N=6), plausibly because families differ most in solution strategy, and smallest on MCQA, where the bounded answer space constrains agreement. The same hetero teams whose *aggregate* ρ is lower than homogeneous teams (Table 22) thus exhibit the predicted internal block structure, supporting the variance-matching bridge under the block-exchangeable assumption.

Table 23: Within- and between-family pairwise agreement on heterogeneous teams (final round, weighted means across all cells with $N \geq 6$). Same-family pairs agree more than cross-family pairs in every combination, consistent with the block-exchangeable prediction.

Task	$\bar{\rho}_{\text{within}}$	$\bar{\rho}_{\text{between}}$	Δ	Pairs (W / B)
MMLU-Hard	0.868	0.819	$+0.050$	35,400 / 82,800
GSM-Hard	0.722	0.616	$+0.105$	35,400 / 82,800
GPQA	0.882	0.849	$+0.033$	35,037 / 81,952
Overall	0.824	0.761	$+0.063$	105,837 / 247,552

M Hierarchical team-structure audit

This appendix supports C3 (functional form robustness). In the main debate experiments, every agent of a fully connected team revises after reading its peers. Here we evaluate six hierarchical team structures in which the answer-producing agents (leaves) never read one another. Each leaf answers in a single call, conditioned only on the task and, depending on the structure, on a persona instruction or on one planner's plan. Downstream nodes (critics, judges, synthesizers, manager) read leaf outputs but never feed back to the leaves. The fitted (c, β) therefore quantify the redundancy of leaf answers produced without peer exchange, a $\rho^{(0)}$-type quantity that depends on the leaf prompt and, for Diamond, on shared plans. The audit tests whether the Ringelmann form describes N_{eff} in these structures. It does not test communication-induced conformity or assumption A1 of Thm. 1, both of which require exchange.

Structures. Each structure spends a budget of N LLM calls, $N \in \{2, 3, 5, 7, 10, 15, 20, 30\}$ (Diamond: $N \geq 3$), on all items of the three main tasks (MMLU-Hard, GSM-Hard, GPQA), with three runs. Every node is called once, at T=0.4 and top-p 0.95, and a node with parents reads their outputs truncated to 120 tokens. L denotes the number of leaves, whose answers alone enter $N_{\text{eff}}^{\text{ans}} = 2^H$.

- *Star*: $L = N-1$ workers answer from the task prompt alone (the solve prompt of the main experiments), and a manager reads all worker outputs and returns the team answer.
- *Tournament-Star*: $L = \lfloor (N+1)/2 \rfloor$ task-only workers enter a single-elimination bracket (byes for odd entrant counts) of pairwise judges, each of which reads two entrants and returns one answer. The final judge acts as manager. At $N \in \{10, 20, 30\}$ the bracket uses $N-1$ calls.
- *Proposer-Critic-Star*: $L = \lceil (N-1)/2 \rceil$ task-only proposers. Each of $\lfloor (N-1)/2 \rfloor$ critics reads one proposal and states its own answer, and the manager reads the critics and any unpaired proposer. Proposers never revise.
- *Tree*: $s = \lfloor (N-1)/4 \rfloor$ synthesizers each read three task-only workers, and the manager reads the synthesizers and any unassigned workers, so $L = N-1-s$.
- *Persona-Star*: Star in which each of the $L = N-1$ workers receives one of six method personas (forward, backward, decomposition, step-back, conservative verification, contrarian), assigned cyclically. Graph and manager are identical to Star.
- *Diamond*: $M = \max(1, \lfloor N/5 \rfloor)$ planners each write one solution approach, instructed to differ from the most obvious one, without answering and without seeing the other plans. $L = N-1-M$

plan-solvers each receive one plan (three or four solvers per plan for $N \geq 10$) and answer, and the manager reads all solvers.

Leaves of Star, Tournament-Star, Proposer-Critic-Star and Tree are generated identically (task-only prompt, distinct seeds), so these four task-only structures differ only downstream of the fitted answers. Persona-Star changes the leaf prompt, and Diamond conditions leaves on shared plans. Fits regress $R(L) = N_{\text{eff}}^{\text{ans}}/L$ on L with the Ringelmann form and exclude budgets with $L<2$. They therefore use 7 budgets for Star, Persona-Star, Tournament-Star and Tree, and 6 for Proposer-Critic-Star and Diamond (both from $N=5$).

Models and tasks. Three open-weight families (Qwen2.5-7B, Llama-3.1-8B, Ministral-8B), three tasks (MMLU-Hard, GSM-Hard, GPQA), six structures, giving 54 cells. Two additional models (Nemotron-Nano-9B-v2, and Qwen3-8B with thinking disabled) are also fit. They reproduce the three-family pattern: Persona-Star has the lowest mean c, the four task-only structures lie within 0.02 of one another in mean c, and Diamond has the highest mean c and β. They are excluded from the summary aggregates.

Aggregate pattern (Table 24). Averaged across the 3 models $\times$ 3 tasks ($n = 9$ cells per structure):

Table 24: Team-structure audit summary. Mean (c, β) per structure across three open-weight families (Qwen2.5-7B, Llama-3.1-8B, Ministral-8B) and three tasks (MMLU-Hard, GSM-Hard, GPQA), $n = 9$ cells per row. $L(30)$: leaf count at $N=30$.

Structure	Mean c	Mean β	Notes on structure
Persona-Star	0.480	0.125	six cycled personas, $L(30)=29$
Star	0.594	0.104	task-only leaves, $L(30)=29$
Proposer-Critic-Star	0.595	0.097	one critic per proposer, $L(30)=15$
Tree	0.599	0.104	synthesizer per 3 leaves, $L(30)=22$
Tournament-Star	0.600	0.099	pairwise judge bracket, $L(30)=15$
Diamond	0.745	0.285	plan-conditioned leaves, $L(30)=23$

The four task-only structures agree closely (mean c from 0.594 to 0.600, mean $\beta \approx 0.10$). Since they generate leaves identically and differ only downstream of the fitted answers, this agreement is expected by construction. It replicates the leaf-level fit across independent samples and different L grids, and it is not evidence that the aggregation structure leaves (c, β) unchanged. Persona-Star has the lowest mean c (0.48, $\Delta = -0.12$ relative to the task-only mean of 0.60) and lies below the task-only mean in all 9 (family, task) cells (Δc from -0.07 to -0.16). This is compatible with a lower $\rho^{(0)}$ under prompt conditioning (C4). Diamond has the largest mean c (0.745, i.e., the lowest ceiling $1/c$) and a distinctly sublinear exponent ($\beta = 0.285$), and it exceeds the task-only means in both parameters in all 9 cells. Its plan-sharing design is compatible with both deviations. Solvers that share a plan agree more often than solvers with different plans in all 36 (family, task, $N \geq 10$) cells (mean exact-answer agreement 0.70 vs. 0.53). All solvers share one plan for $N \leq 7$, and the share of same-plan solver pairs falls from 0.43 at $N=10$ to 0.13 at $N=30$ as the number of plans grows. Shared conditioning therefore raises leaf correlation at small L, and its dilution at larger L is consistent with the larger β.

Table 25: Structure $\times$ task breakdown (mean across 3 models per cell). GSM-Hard shows sublinear β everywhere and the widest relative spread in c across structures (2.1$\times$, Persona-Star 0.213 to Diamond 0.454). MCQA tasks stay near the hard ceiling for all star-family structures. Only Diamond departs from it on both MMLU-Hard and GPQA. $R^2 \geq 0.97$ for all 54 underlying fits.

	MMLU-Hard		GSM-Hard		GPQA	
Structure	c	β	c	β	c	β
Persona-Star	0.675	0.061	0.213	0.276	0.553	0.036
Star	0.781	0.002	0.318	0.310	0.682	0.001
Proposer-Critic-Star	0.780	0.001	0.308	0.290	0.698	0.001
Tree	0.790	0.007	0.316	0.304	0.691	0.001
Tournament-Star	0.784	0.003	0.310	0.288	0.706	0.007
Diamond	0.943	0.156	0.454	0.540	0.837	0.161

Framework interpretation. Three observations. *(i) The functional form is preserved.* All 54 cells fit the Ringelmann form at $R^2 \geq 0.97$ (minimum 0.974 for Llama-3.1-8B Diamond on GSM-Hard, median 0.9996), and no structure requires a different functional family. This extends C3 to N_{eff} of answers produced without peer exchange under persona and plan conditioning. It does not bear on the conformity dynamics of the mean-field theorem, which require exchange. *(ii) Prompt conditioning lowers c.* Persona-Star has the lowest mean c. Because its leaves never exchange messages, the reduction reflects a lower $\rho^{(0)}$ under method-diverse prompts. This is the same direction as architectural heterogeneity (App. L), and it is not a test of A1. *(iii) For task-only leaves, the regime is set by the task.* On MCQA the four task-only structures keep $\beta \approx 0$, and on free-form GSM-Hard all four are sublinear ($\beta \approx 0.30$). Diamond departs from this pattern with $\beta = 0.54$ on GSM-Hard and $\beta \approx 0.16$ on MCQA. Because the manager acts only after the leaf answers are fixed, this departure arises in leaf generation, where Diamond differs from the other structures by plan conditioning, and not in the aggregation structure.

Caveats. The structures are evaluated with a Ringelmann-form fit on the same (c, β) scale as the main paper; predictive extrapolation ($N \leq 5 \to N = 30$) was not audited per structure at the same intensity as the main-text held-out grid, so we do not report per-cell MREs against alternative forms. Persona-Star uses a fixed set of six personas, which recur for $L > 6$, so prompt heterogeneity saturates at the persona-pool size. The audit covers neither peer exchange among leaves nor tool-augmented agents. No values are projected.

N HumanEval: open-ended code generation

This appendix supports C3 by extending the framework to open-ended code generation. We first report the full sweep and then the original pilot, which introduces the two measurement axes.

N.1 Full sweep

Setup. All 164 HumanEval problems [Chen et al., 2021], three open-weight families (Qwen2.5-7B, Llama-3.1-8B, Ministral-8B), three communication conditions (debate, self-correction, noise placebo), 3 rounds, team sizes up to $N = 30$($k = N-1$ peers, 1024-token output cap, each previous response shown in the revision prompt, own or peer, truncated to 120 tokens as in the main experiments). Agreement is measured on the canonicalized-code axis and on the pass-vector axis, as defined in the pilot below.

Results (Table 26). The Ringelmann form fits all nine (family $\times$ condition) cells at $R^2 \geq 0.99$. Every cell is sublinear on the code axis ($\beta \in [0.21, 0.31]$, all 95% CIs exclude 0) and consistent with a hard ceiling on the pass-vector axis ($\beta \leq 0.06$, all CIs include 0), replicating the pilot's two-regime result. In a paired item bootstrap, debate yields a higher c than self-correction on all three families ($\Delta c \in [0.06, 0.10]$, all CIs exclude 0). The noise-placebo equivalence of C4(i) holds on Ministral ($\Delta c_{\text{self-noise}} = 0.01$ [-0.02, 0.05]), is small on Llama (0.03 [0.01, 0.07]), and fails on Qwen (0.16 [0.10, 0.22]), where off-task peer code raises code diversity by a mechanism we have not isolated. Excluding outputs that reach the 1024-token cap or are empty (up to 23% under Llama noise, $\leq 8\%$ elsewhere) changes c by at most 0.04 and leaves every cell sublinear. Team accuracy does not follow the growth of N_{eff}. It never exceeds single-agent pass@1 by more than 3 pp, and it falls by 12 pp in the cell with the most truncated outputs (Llama, noise). Oracle pass@30 exceeds majority accuracy by 9–31 pp, so under majority voting N_{eff} bounds the evidence a verifier or best-of-N selector could convert rather than predicting accuracy.

Table 26: HumanEval full sweep (164 items, final round, $N \in \{2, \ldots, 30\}$). Code-axis fits with 95% item-bootstrap CIs (B=1000). Acc.: single-agent pass@1 at N=1 $\rightarrow$ strict-majority team accuracy at N=30.

Family	Condition	c	β	R^2	Acc. (N=1 $\rightarrow$ 30)
Qwen2.5-7B	debate	0.51 [0.44, 0.59]	0.21 [0.13, 0.28]	0.99	0.83 $\rightarrow$ 0.76
	self-corr.	0.41 [0.35, 0.48]	0.23 [0.16, 0.29]	1.00	0.83 $\rightarrow$ 0.82
	noise	0.26 [0.21, 0.32]	0.23 [0.16, 0.31]	1.00	0.83 $\rightarrow$ 0.84
Llama-3.1-8B	debate	0.24 [0.19, 0.29]	0.29 [0.22, 0.36]	0.99	0.66 $\rightarrow$ 0.65
	self-corr.	0.15 [0.12, 0.19]	0.27 [0.21, 0.33]	1.00	0.66 $\rightarrow$ 0.68
	noise	0.12 [0.09, 0.15]	0.31 [0.23, 0.39]	0.99	0.66 $\rightarrow$ 0.54
Ministral-8B	debate	0.24 [0.20, 0.30]	0.24 [0.17, 0.32]	1.00	0.86 $\rightarrow$ 0.86
	self-corr.	0.19 [0.15, 0.23]	0.29 [0.23, 0.35]	1.00	0.86 $\rightarrow$ 0.86
	noise	0.17 [0.14, 0.21]	0.27 [0.20, 0.33]	1.00	0.86 $\rightarrow$ 0.85

N.2 Pilot: two measurement axes

Setup. We extend the framework to open-ended code generation as a pilot. Qwen2.5-7B-Instruct on 30 randomly sampled HumanEval problems [Chen et al., 2021] at $N \in \{1, 2, 3, 5, 7, 10\}$, two communication conditions, *comm_true* (debate, peer code) and *comm_self* (self-correction, own previous code only), 3 rounds (τ=2), 30 items per cell. Unlike the full sweep, the pilot passed previous responses to the revision prompt without the 120-token truncation ($\approx$450 prompt tokens per peer). We attempted N=30 at *comm_true_k29*, but the revision prompts containing 29 peer completions exceeded the runner's 8,192-token input cap and were truncated. The resulting outputs ran to the 1,024-token output limit ($\geq 98\%$ of rows), so those rows are excluded. Noise placebo (*comm_random*) was not run.

Two measurement axes. HumanEval admits a discrete *pass-vector* measurement absent in MCQA or numeric math. The problem's `check` function contains T assertions. An agent's pass-vector is the binary length-T vector indicating which assertions its code passes. We compute pass-vectors deterministically by running each assertion in isolation through `human_eval.execution.check_correctness` (per-assertion sandboxed execution with timeout). Mean $T = 6.9$ across our 30 problems, giving a nominal answer space of $\sim 2^{6.9}$ per problem. We compute two correlation profiles, both via the variance-matching identity (§3). On the *pass-vector axis*, agents agree iff their pass-vectors match (functional-outcome axis). On the *code axis*, agents agree iff their canonicalized code matches: each completion runs through Python's AST round-trip (`ast.parse` $\rightarrow$ docstring stripping $\rightarrow$ `ast.unparse`), so cosmetic differences (whitespace, comments, docstring formatting) are removed but algorithmic differences (variable naming, control flow, helper structure) are preserved. This is the open-ended axis.

Same task, two regimes (Table 27). On the pass-vector axis, HumanEval lies in a hard-ceiling regime ($\beta \in [0.05, 0.08]$, $c \in [0.92, 0.99]$), matching the GPQA pattern. Assertions are highly correlated since most problems are all-pass or all-fail at the function level, so the empirical distribution over the $\sim 2^{6.9}$ possible pass-vectors concentrates on one or two modal vectors per problem. The code axis recovers a sublinear regime ($\beta \in [0.12, 0.31]$, $c \in [0.39, 0.59]$), comparable to GSM-Hard. The Ringelmann form fits both axes with $R^2 \geq 0.99$.

Interpretation. The same task, measured on a discrete-outcome axis, lies in an MCQA-like regime. Measured on an open-ended axis, it lies in a GSM-like regime. The Ringelmann form fits both, which is consistent with the framework extending to open-ended generation. *comm_self* preserves more code diversity than *comm_true* ($\beta = 0.31$ vs $\beta = 0.12$ on the code axis): without peer code, agents retain individual algorithmic variation. This contrast is not detectable on the pass-vector axis ($\beta_{\text{self}} = 0.08$ vs $\beta_{\text{true}} = 0.05$) because both collapse to the same all-pass or all-fail outcome.

Caveats. (Pilot only. The full sweep above addresses the first three.) 30 items per condition is sufficient for $R^2 > 0.99$ fits but marginal for bootstrap CIs (we report point fits only). N is capped at 10. Two conditions only (*comm_random* / noise placebo not run). The code-canonicalization metric treats variable-name choices (`sequence` vs `tribonacci`) as semantic. An alpha-renaming variant that further normalizes variable identifiers is a natural alternative we have not pursued.

Table 27: HumanEval pilot: per-N efficiency under two measurement axes (final round, 30 items per cell). R_N^{pv} uses the pass-vector axis (functional outcome); R_N^{code} uses the canonicalized-AST code axis (open-ended).

	comm_true			comm_self		
N	R_N^{pv}	R_N^{code}	acc	R_N^{pv}	R_N^{code}	acc
1	ensemble: $R^{\text{pv}} = R^{\text{code}} = 1.00$, acc $= 0.73$					
2	0.52	0.68	0.67	0.53	0.77	0.73
3	0.34	0.45	0.73	0.37	0.64	0.77
5	0.22	0.34	0.73	0.23	0.52	0.73
7	0.16	0.29	0.70	0.19	0.45	0.73
10	0.11	0.19	0.77	0.12	0.37	0.77

Ringelmann fits. comm_true: pass-vector $(c, \beta) = (0.99, 0.05)$ with $R^2 = 1.00$, code $(0.59, 0.12)$ with $R^2 = 0.99$. comm_self: pass-vector $(0.92, 0.08)$ with $R^2 = 1.00$, code $(0.39, 0.31)$ with $R^2 = 1.00$.

O Temperature ablation

This appendix supports C4 (the initial-correlation lever) by directly testing sampling temperature T as an intervention on $\rho^{(0)}$. Higher T raises per-agent output entropy, which should lower pre-communication answer agreement and therefore raise the ceiling $1/c$. Whether temperature shifts the regime (β) or only the level (c) is the empirical question.

Protocol. Qwen2.5-7B $\times$ {MMLU-Hard, GSM-Hard, GPQA}, full-mesh debate, 3 rounds, at $T \in \{0.2, 0.4, 0.7, 1.0\}$. Team sizes $N \in \{2, 5, 10, 30\}$. 400 items per cell on MMLU-Hard and GSM-Hard, 198 items per cell on GPQA (full Diamond subset). Ringelmann fit on $N \in \{2, 5, 10\}$; the fit then predicts $\widehat{N}_{\text{eff}}$ at $N = 30$, compared to the observed value. The paper's canonical held-out audit uses $N \in \{2, 3, 5\}$ but the temperature-sweep grid did not include $N = 3$, so $N = 10$ substitutes for the third anchor. Bootstrap CIs are item-level with $B = 1000$.

Table 28: Temperature sweep across three tasks (Qwen2.5-7B, full-mesh debate). Ringelmann fit is on $N \in \{2, 5, 10\}$; the observed N_{eff} at $N = 30$ (out of 30 nominal) is compared to the small-team pilot's prediction. Regime is invariant across T ($\beta = 0$ in $11/12$ cells, max $\beta = 0.004$). Ceiling parameter c decreases monotonically with T on all three tasks, as predicted under C4: higher sampling temperature lowers pre-communication agreement, raising $1/c$. Every small-team prediction is within 0.3 effective agents of observation at $N = 30$.

Setting	c	β	pred N_{eff} (N=30)	obs N_{eff} (N=30)
MMLU-Hard (400 items, 4-choice MCQA)				
$T = 0.2$ (near-greedy)	0.855	0.000	1.16	1.10
$T = 0.4$ (paper default)	0.832	0.000	1.19	1.13
$T = 0.7$	0.794	0.000	1.25	1.13
$T = 1.0$ (high-diversity)	0.791	0.000	1.25	1.14
GSM-Hard (400 items, free-form numeric)				
$T = 0.2$ (near-greedy)	0.655	0.000	1.50	1.53
$T = 0.4$ (paper default)	0.562	0.000	1.73	1.85
$T = 0.7$	0.469	0.000	2.05	2.16
$T = 1.0$ (high-diversity)	0.468	0.004	2.09	2.38
GPQA (198 items, 4-choice MCQA)				
$T = 0.2$ (near-greedy)	0.828	0.000	1.20	1.13
$T = 0.4$ (paper default)	0.784	0.000	1.26	1.15
$T = 0.7$	0.707	0.000	1.39	1.14
$T = 1.0$ (high-diversity)	0.693	0.000	1.42	1.18

Regime is temperature-invariant. Across all 12 cells, $\beta \leq 0.004$. Sampling temperature does not move a hard-ceiling configuration into a sublinear one, nor a sublinear one into linear. The paper's regime classification (§5.1) is therefore robust to the sampling temperature within $T \in [0.2, 1.0]$ on Qwen2.5-7B.

Ceiling parameter c tracks temperature monotonically. On GSM-Hard, c falls from 0.655 at $T = 0.2$ to ~ 0.47 at $T \geq 0.7$ (a 28% relative drop, plateauing at high T). On MCQA the drop is smaller (MMLU-Hard $0.86 \to 0.79$, -7%, and GPQA $0.83 \to 0.69$, -16%). The MCQA-vs-free-form contrast is expected: entropy saturates at $\log_2 4 = 2$ bits on 4-choice tasks, so the additional per-agent diversity attainable by raising T is bounded by the size of the answer space. On GSM-Hard, the answer space is unbounded and c falls further.

Effective agents at $N = 30$. On GSM-Hard, raising T from 0.2 to 1.0 increases observed N_{eff} from 1.53 to 2.38 (a 55% relative gain), which remains below 8% of the nominal $N = 30$. On MCQA the gain is marginal ($1.10 \to 1.14$ on MMLU-Hard, $1.13 \to 1.18$ on GPQA), consistent with the bounded answer space.

Held-out prediction. Every prediction from the small-team pilot ($N \in \{2, 5, 10\}$) is within 0.3 effective agents of observation at $N = 30$ (max absolute error 0.29 on GSM-Hard, $T = 1.0$; median absolute error 0.11). The largest relative errors concentrate on GPQA at $T \geq 0.7$, where observed N_{eff} barely exceeds 1 and the small pilot overestimates by 0.2–0.3 agents. The absolute error is at most 0.29 in every cell.

Framework interpretation. Temperature acts as a bounded intervention on the ceiling: it lowers c but does not change the regime. Its effect matches C4 (higher $T \Rightarrow$ lower $\rho^{(0)} \Rightarrow$ lower c), and the effect saturates on bounded MCQA answer spaces. This complements the noise-placebo result (§5.4): within the tested regime peer content adds little beyond re-evaluation, but the pre-communication distribution from which agents sample is not inert: interventions on that distribution (heterogeneous ensembles, temperature) are the interventions observed to shift the ceiling.

P Noise placebo, self-correction, and self-consistency

This appendix supports C4 (initial correlation $\rho^{(0)}$ as the lever) by examining three sub-conditions in detail. The shared-form result of § 5.3 rests on these sub-conditions.

Self-correction (no peers, same law). Self-correction uses the same revision prompt and model population as debate but removes all peer information. On GSM-Hard it produces $c = 0.34$, $\beta = 0.25$ (sublinear regime), with different parameters from debate but the same functional form and the same quality of fit. The shift is interpretable: without peer-induced conformity, correlation is lower and answer diversity decays more slowly with N, placing self-correction in the sublinear rather than the hard-ceiling regime. Our earlier study reports that isolated self-correction prevails over unguided homogeneous multi-agent debate [Bertalanič and Fortuna, 2026]. The present results are consistent with that finding: self-correction matches debate accuracy within ± 5 pp (TOST, Table 14) at 10–45$\times$ lower prompt cost per effective agent (Qwen2.5-7B, Table 30).

Noise placebo (random peers, same law). The noise placebo replaces true peer answers with randomly shuffled text from other items. It produces $c = 0.33$, $\beta = 0.27$ on GSM-Hard, with 95% bootstrap CIs overlapping those of self-correction ($\widehat{N}_{\text{eff}}$ within 0.4 at every team size). Even though the law is derived under explicit peer-conformity dynamics, it fits the no-peer placebo with matching parameters. This indicates that the law describes a structural property of correlated LLM outputs under revision, not a property of peer debate specifically.

Cross-family replication (GSM-Hard, four models). Table 29 extends the noise vs. self-correction equivalence beyond the canonical Qwen2.5-7B cell to four open-weight models from three families (Qwen2.5-7B, Qwen2.5-32B, Llama-3.1-8B, Ministral-8B). The equivalence holds throughout: $|\Delta c| \leq 0.036$ and $|\Delta \beta| \leq 0.064$ everywhere, small relative to the between-condition variation the paper reports. We observe no counterexample within the tested grid. The equivalence therefore does not rest on a single (Qwen2.5-7B, GSM-Hard) cell.

Table 29: Noise placebo vs. self-correction on GSM-Hard across four open-weight models from three families. Each column pair reports the fitted (c, β). $|\Delta c|$ and $|\Delta \beta|$ are the absolute condition-pair gaps for the same model. Debate (c, β) is included for reference to show that both no-peer conditions lie far from debate on the ceiling axis. All fits use $N \in \{2, 3, 5, 7, 10, 15, 20, 30\}$ except Qwen2.5-32B (no N=15) and the Ministral-8B no-peer conditions (largest N is 20, and the noise grid starts at N=5).

Model	Self-corr. (c, β)	Noise (c, β)	Debate (c, β)	$\lvert\Delta c\rvert$	$\lvert\Delta \beta\rvert$
Qwen2.5-7B	$(0.335, 0.252)$	$(0.332, 0.272)$	$(0.573, 0.014)$	0.003	0.019
Qwen2.5-32B	$(0.450, 0.183)$	$(0.449, 0.176)$	$(0.773, 0.000)$	0.001	0.007
Llama-3.1-8B	$(0.219, 0.361)$	$(0.183, 0.300)$	$(0.472, 0.000)$	0.036	0.061
Ministral-8B	$(0.285, 0.237)$	$(0.316, 0.174)$	$(0.397, 0.000)$	0.031	0.064

Self-consistency (no revision, predicted from $\rho^{(0)}$). Self-consistency, majority vote over round-1 answers without any revision or communication, is the zero-interaction baseline. The framework predicts its scaling directly from the pre-communication correlation $\rho^{(0)}$. Measuring round-1 pairwise agreement gives $\rho^{(0)} \approx 0.97$ (MMLU-Hard) and $\rho^{(0)} \approx 0.51$ (GSM-Hard). On MMLU-Hard, fitting the Ringelmann curve to round-1 answers recovers $c = 0.93$, close to $\rho^{(0)}$.[2] On GSM-Hard, self-consistency lies in the sublinear answer-level regime ($c = 0.43$, $\beta = 0.25$, $R^2 = 0.9997$), the same regime as self-correction and noise. Accuracy is flat from N=2 to N=30: low solo accuracy ($p \approx 0.10$) prevents the correct answer from winning a plurality even as answer-space entropy grows.

Q Aggregator robustness

We re-aggregate final-round team answers from the canonical homogeneous data (Qwen2.5-7B and Llama-3.1-8B, three tasks, three communication conditions: debate, self-correction, noise) under five aggregation rules, using only quantities logged during the original runs. Aggregators that exploit higher-order information about inter-agent correlation [Ai et al., 2026] lie outside this set and are a candidate means of converting the answer-level N_{eff} headroom into accuracy:

- *plurality*: majority vote over agent answers (paper baseline).
- *conf-weighted*: votes weighted by self-reported confidence (a 0–100 value the model is prompted to emit).
- *prob-weighted*: votes weighted by softmax probability of the chosen answer.
- *filtered plurality*: drop the bottom 25% of votes by softmax confidence, then plurality.
- *top-1 confident*: select the single answer with maximum softmax probability.

Self-reported confidence is only weakly correlated with the softmax probability (Pearson $r = 0.03$ on Qwen2.5-7B, the model emits ≥ 85 for almost all answers), so *conf-weighted* is essentially identical to plurality and serves as a negative control. The other aggregators differ because they use the logit signal rather than the discrete answer.

Across all $(\text{model}, \text{task}, \text{condition}, N)$ cells, alternative aggregators shift team accuracy by at most ~ 5 pp from plurality, with mean absolute shifts under 1.3 pp on every task. The largest shifts come from *top-1-confident* (which coincides with *prob-weighted* at N=2): up to 3.9 pp on MMLU-Hard, 3.7 pp on GSM-Hard, and 5.0 pp on GPQA, concentrated at small N. On the canonical Qwen2.5-7B GSM-Hard noise cell at N=30, top-1-confident lifts accuracy by $+2.4$ pp over plurality. The fitted (c, β) are aggregator-invariant by construction, because $N_{\text{eff}}^{\text{ans}} = 2^H$ and the correctness ICC are computed from the agents' answers before aggregation. The regime classification therefore does not depend on the rule, even when accuracy shifts. As an accuracy-side check, we convert team accuracy under each rule into an accuracy-implied N_{eff} by inverting the correlated-Condorcet approximation $\text{PP}(N_{\text{eff}}, p)$ (App. A.2) at the solo accuracy p, and refit the Ringelmann form. The inversion requires $p > 0.5$, which holds only for Qwen2.5-7B on MMLU-Hard. In that cell the refitted β is below 10^{-9} in all three conditions under plurality, conf-weighted, prob-weighted, and top-1-confident aggregation (filtered plurality is undefined for $N < 4$). These refits are coarse ($R^2 = 0.28$–0.65), because team

[2] Round-1 $\rho^{(0)} = 0.97$ is slightly higher than the post-debate $\rho = 0.94$ in Table 11 because debate redistributes some answers on MMLU-Hard, breaking prior consensus on a small fraction of items.

accuracy varies by at most 2.3 pp across N, and they are consistent with a hard ceiling under every rule.

Heterogeneous teams (App. L) use balanced compositions: $N \in \{3, 6, 9, 15, 30\}$ split equally across Qwen, Llama, and Ministral. Under this balance, two alternatives to plurality reduce algebraically to plurality or nearly so:

- *block-weighted* (each model family gets total weight 1): collapses exactly to plurality when family counts are equal.
- *family-then-vote* (each family takes its internal plurality, then majority across families): differs from plurality by at most 2.5 pp (5 of 200 items, GSM-Hard, N=6) and by at most 1 pp in the other 14 (task, N) cells.

We find no systematic aggregator effect on heterogeneous accuracies under the balanced design. *Imbalanced* hetero designs (e.g. a single Ministral agent in a Qwen-dominant team) are needed to discriminate the block-exchangeable prediction (App. A.7) empirically, and remain future work.

Adjudication-based aggregators. Aggregators that score peer rationales (debate-with-cross-examination, judge-mediated consensus, verifier-feedback aggregation) require additional generation beyond what the present sweep collects, since they need a separate adjudicator output per round. Our framework predicts they lower $\rho^{(0)}$ when the adjudicator is independent of the underlying agents, and leave $\rho^{(0)}$ unchanged when the adjudicator is the same model. Empirical confirmation requires a dedicated experiment.

R Deployment: cost, throughput, and accuracy projection

This appendix covers the deployment-facing implications of the scaling law: token-cost scaling under dense communication (§R.1), a heuristic accuracy projection from small-team calibration (§R.2), and the detailed economic argument behind the deployment claims in §5.5 (§R.3). The small-N calibration procedure for estimating (c, β) on a new model or task is in App. C.8.

R.1 Token cost scaling

The Ringelmann ceiling implies redundant compute. We report cost as prompt tokens per effective agent (Prompt$/\widehat{N}_{\text{eff}}$), a hardware-independent measure that mirrors per-token API pricing and isolates the scaling claim from batch-size and utilization effects. At N=30, debate spends roughly 91K prompt tokens per effective agent on MMLU-Hard and 64K on GSM-Hard, versus 7.5K and 1.43K for self-correction, a 12× and 45× ratio. On GPQA the gap is similar (101K vs. 10K, ≈10×). Output-token cost follows the same pattern but is dominated by the prompt term in debate, where each agent must read $N-1$ peer messages. Figure 10 shows the super-linear growth of debate prompt tokens versus the linear growth of self-correction across all three tasks, which drives the cost asymmetry.

Table 30: Scaling summary at N=30 (Qwen2.5-7B). Δ acc: team accuracy at N=30 minus solo accuracy at N=1 (the round-1 ensemble baseline; 0.555 on MMLU-Hard, 0.100 on GSM-Hard, 0.303 on GPQA). Prompt/N_{eff}: prompt tokens per effective agent, quantifying redundant compute.

Task	Condition	Acc	Δ acc	N_{eff}	Prompt/N_{eff}
MMLU	Debate	.579	+.023	1.12	91K
	Self-corr.	.587	+.032	1.21	7.5K
GSM	Debate	.264	+.163	1.79	64K
	Self-corr.	.282	+.182	6.10	1.43K
GPQA	Debate	.318	+.015	1.15	101K
	Self-corr.	.313	+.010	1.29	10K

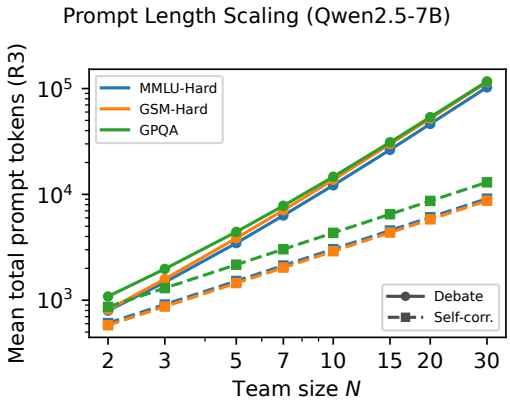


Figure 10: Total prompt tokens per item at R3 (Qwen2.5-7B), across three tasks and two conditions. Solid lines: debate. Dashed lines: self-correction. Debate prompt cost grows super-linearly in N because each agent reads $N-1$ peer messages, while self-correction grows linearly. At $N=30$ debate uses 9–13$\times$ more prompt tokens than self-correction, and normalization by N_{eff} widens the gap to 10–45$\times$ per effective agent (Table 30).

R.2 Heuristic accuracy projection

Solo accuracy and the Condorcet bridge. On a separate axis from the scaling regime classification, solo accuracy p interacts with majority voting through the Condorcet CDF. When the model is already strong ($p > 0.5$, MMLU-Hard), correlation prevents the majority-vote gain from materializing: the high correlation ($\rho \approx 0.84$–0.94, Table 11) keeps N_{eff} close to 1, and accuracy gains over solo are at most a few points. When individual accuracy is low ($p \approx 0.10$–0.30, GSM-Hard and GPQA), absolute accuracy still improves substantially from 10% to 26–28% on GSM-Hard via re-evaluation rather than peer information (self-correction $+18.2$ pp, debate $+16.3$ pp). On GSM-Hard, this gain is possible despite $p \ll 0.5$ because wrong answers scatter across many distinct numeric strings, so the correct answer attains a plurality even at low individual accuracy: this is the regime where free-form math diverges from the binary Condorcet threshold.

The calibration protocol in §5.1 predicts the structural ceiling N_{eff}, which is the paper's primary quantitative prediction. As a heuristic, the global mean individual accuracy $\bar{p}$ can be substituted into the correlated-Condorcet approximation: $\text{PP}(N_{\text{eff}}, \bar{p})$.

Theoretical limitation. Because $\text{PP}(\cdot, p)$ is nonlinear in p, applying it at the population mean $\bar{p}$ ignores item-difficulty heterogeneity: $\text{PP}(N_{\text{eff}}, \mathbb{E}[p_i]) \neq \mathbb{E}[\text{PP}(N_{\text{eff}}, p_i)]$. The heuristic should therefore be regarded as an approximate projection conditional on the structural ceiling law.

ρ quantities. See App. A for the full notation guide. On MCQA tasks all three measures track each other (spread 0.05–0.10 for debate and self-correction, up to 0.15 for noise, Table 11). On free-form GSM-Hard they diverge substantially: $\rho_{\text{ICC}}^{\text{corr}} \approx 0.7$–$0.9$ while $\rho_{\text{eff}}^{\text{ans}} \approx 0.1$–$0.5$, because many distinct wrong numeric answers inflate answer entropy without contributing useful diversity. The scaling law describes the answer-level quantity, and the correctness-level ICC provides independent support for the condition ordering (debate $>$ self-correction $>$ noise on MMLU-Hard and GPQA, and debate $>$ self-correction $\approx$ noise on GSM-Hard, Table 11).

Prediction procedure (MCQA tasks). On MCQA tasks where $\rho_{\text{eff}}^{\text{ans}} \approx \rho_{\text{ICC}}^{\text{corr}}$, the scaling law supports a direct accuracy projection. Calibrate two parameters from small teams ($N \leq 5$): $\bar{p}$ (mean solo accuracy) and ρ (answer-level correlation, Eq. 1). For any target team size N, compute $N_{\text{eff}} = N/(1 + (N-1)\rho)$ and predict team accuracy as $\text{PP}(N_{\text{eff}}, \bar{p})$. On free-form tasks where the answer-level and correctness-level correlations diverge, the Condorcet bridge becomes approximate. The scaling law still describes answer-diversity dynamics but should not be combined with PP to predict accuracy without additional calibration.

Empirical validation. Per-agent accuracy is nearly constant across team sizes (± 1.3 pp across N=2–30 on all tasks), so the two-parameter calibration is stable. For homogeneous debate, projecting from $N \leq 5$ to N=7–30 gives mean errors ≤ 1.4 pp on all three tasks (Table 31). Heterogeneous teams

show larger errors (up to 7.2 pp on GSM-Hard), modestly exceeding the per-item sampling noise of 200-item subsets (± 3.2 pp, one binomial standard error at $p \approx 0.3$). The excess error is consistent with the wider item-difficulty distribution in heterogeneous mixes: different model families err on different items, and the population-mean projection $\text{PP}(N_{\text{eff}}, \bar{p})$ does not capture this within-item heterogeneity. For accuracy-critical applications on heterogeneous teams, item-level calibration would be preferable. The structural ceiling prediction (N_{eff} as a function of N) is evaluated separately in Table 8; for heterogeneous teams its held-out error (27.4%) is also larger than for homogeneous teams (5.5–11.9%).

Table 31: Heuristic accuracy projection from small-team calibration (debate). ρ calibrated from $N \leq 5$ only, so N=30 predictions are fully out of sample (OOS).

		Calibration	Accuracy at N=30		OOS error (pp)	
Setting	Task	ρ	Predicted	Observed	Mean	Max
Homogeneous	MMLU-Hard	.846	.567	.579	1.3	1.9
	GSM-Hard	.566	.249	.264	1.4	1.7
	GPQA	.796	.331	.318	1.1	1.9
Heterogeneous	MMLU-Hard	.560	.421	.445	3.4	4.8
	GSM-Hard	.349	.272	.320	5.9	7.2
	GPQA	.588	.297	.303	0.9	1.2

R.3 Detailed deployment economics

The condensed argument in §5.5 contrasts instruct-class and reasoning-model multi-agent debate. This appendix expands the reasoning behind the deployment claims.

Cost arithmetic. A 30-agent, 3-round debate costs roughly $30 \times 3 \times 1500 \times \$10/\text{M} \approx \$1.35$ per problem on a frontier reasoning model at low thinking budget (assuming $1{,}500$ output tokens per response and an output price of $\$10$ per million tokens), and runs in minutes per query under serial generation. In our runs (one GPU, Hugging Face `transformers`, microbatches of 4), the same debate on a self-hosted 7–8B instruct model took a median 51 s (Qwen2.5-7B) and 80 s (Llama-3.1-8B) of generation time and 17 and 25 kJ of GPU energy per problem (NVML, over the tasks with energy readings). At an assumed $\$0.15$/kWh this is $\approx\$10^{-3}$ in electricity, or $\$0.03$–$0.05$ including GPU time at an assumed $\$2$ per GPU-hour. Median generation time per response (batch wall time divided by team size) was 0.6 s and 1.0 s for these models versus 15 s for Qwen3-8B in thinking mode under the same runner, a 15–$25\times$ contrast.

Throughput trade-off at high solo accuracy. For configurations with high solo accuracy, the framework predicts $c \to 1$. If the prediction were to hold at multi-agent scale (untested in this paper), the deployment consequence would be as follows: N agents debating one problem and N agents working independently on N problems would achieve near-identical aggregate accuracy at $1/N$ throughput, in which case the framework's prescription becomes *parallelize rather than form teams*. The solo-accuracy threshold at which c approaches 1 can in principle be estimated from a small-N pilot.

S Extended scope and limitations

This appendix extends the main-text limitations paragraph (§5.6).

What was evaluated. The core sweep spans five open-weight models from three families (Qwen2.5-7B/32B, Qwen3-8B thinking, Llama-3.1-8B, Ministral-8B), a closed-source model (Gemini Flash-Lite), and a frontier solo evaluation (Gemini 3.1 Pro at N=1 on all three tasks). It covers three tasks (MMLU-Hard, GSM-Hard, GPQA), team sizes $N \in \{1, 2, 3, 5, 7, 10, 15, 20, 30\}$, three communication conditions (debate, self-correction, noise placebo) plus the self-consistency baseline, homogeneous and heterogeneous compositions, and a communication-density ablation $k \in \{1, 2, 4, 9, 29\}$, $\tau \in \{0, \ldots, 5\}$ (App. H). Targeted extensions add open-ended code generation on all 164 HumanEval problems (App. N), six hierarchical team structures without peer exchange (App. M, which additionally fits Nemotron-Nano-9B-v2), and a sampling-temperature sweep $T \in \{0.2, 0.4, 0.7, 1.0\}$ on Qwen2.5-7B (App. O).

Scope by design. The core sweep is intentionally restricted to homogeneous, role-symmetric, fully-connected interaction so that the conformity signal can be isolated under controlled conditions. Heterogeneous teams (App. L), sparse communication with $k < N-1$ (App. H), and hierarchical structures (App. M) are targeted extensions of this design. Asymmetric roles, external tools, and aggregator pipelines all introduce additional mechanisms (verifier feedback, tool-grounded evidence, asymmetric information flow) that would confound the inter-agent correlation we measure. The following configurations are untested extensions of the framework and lie outside the scope of its empirical claims:

- *Role-asymmetric debate* (judges, critics, moderators). The team-structure audit (App. M) does not test this setting, because its fitted leaves never exchange messages and critics, judges and managers act only downstream of them. Production systems often use a single underlying LLM playing multiple roles, where the conformity dynamics may persist because role conditioning does not change the underlying model.
- *Tool-augmented agents* (web search, code execution, calculators, retrieval), and *long-horizon multi-step workflows*, where assumption A1 of Thm. 1 no longer holds.
- *Mixture-of-Agents [Wang et al., 2025] and verifier-aggregator pipelines* with a designated aggregator.
- *Multimodal tasks* (vision, audio).
- *Multi-agent frontier-class debate*. Gemini 3.1 Pro is evaluated solo here, since the framework predicts a saturated ceiling at its solo-accuracy regime.
- *Very large team sizes* ($N > 30$) and *long-horizon debates* ($\tau > 5$).

The framework yields a falsifiable hypothesis for each. *(H1)* Role asymmetry, external tools, and independent verifiers lower $\rho^{(0)}$ and hence c. H1 is falsified if c does not decrease relative to the matched symmetric configuration. *(H2)* Frontier-class models on high-accuracy tasks yield $c \to 1$. H2 is falsified if the 95% CI of the fitted c lies entirely below 0.86, the lowest value observed for Qwen3-8B (thinking, $c = 1.00$) and Gemini Flash-Lite ($c = 0.86$ on GPQA, 0.95 on GSM-Hard). *(H3)* The $k\tau$ collapse extends to $N > 30$ and $\tau > 5$. H3 is falsified if configurations with matched $k\tau$ differ by more than the maximum $\Delta\rho = 0.052$ observed here. Testing these hypotheses requires dedicated experiments and is left to future work.

Aggregator and team-balance. Reported team accuracies use plurality (majority vote) aggregation, whereas the fitted (c, β) are computed from the agents' answers before aggregation and do not depend on the rule. Re-aggregating the same agent-level data under confidence-weighted, probability-weighted, quartile-filtered, or top-1-confident rules shifts team accuracy by at most 5 pp across all $(\text{model}, \text{task}, \text{condition}, N)$ cells (mean ≤ 1.3 pp per task), with the largest shifts concentrated at small N and on top-1-confident (App. Q). Heterogeneous teams use balanced compositions ($N \in \{3, 6, 9, 15, 30\}$ split equally across three families), which makes block-weighted aggregation algebraically identical to plurality. Family-then-vote stays within 2.5 pp of plurality. *Imbalanced* hetero designs would discriminate the block-exchangeable extension (App. A.7) more decisively and remain a natural follow-up.

Range bounded at N=30, τ≤5. The empirical sweep is bounded at these values. The held-out audit (App. D) evaluates extrapolation from $N \leq 5$ to larger N within the tested range ($N \leq 30$) but does not certify the law beyond it. Deployments at $N > 30$ should repeat the pilot to confirm regime stability.

Decoding parameters. The main estimates are conditioned on the decoding configuration in App. C (temperature 0.4, top-p 0.95, fixed seed). A temperature sweep $T \in \{0.2, 0.4, 0.7, 1.0\}$ on Qwen2.5-7B (App. O) confirms the predicted direction: the regime is invariant ($\beta = 0$ in $11/12$ cells) and c falls monotonically with T. Sweeps over top-p/top-k and the temperature sweep on other families remain future work.

Noise placebo, single form. Our placebo uses length-matched round-1 outputs for other items (§4). Richer controls (token-length-only padding, fluent-but-off-topic text, adversarial peers) would more cleanly isolate the contribution of peer content from prompt-overhead effects, and remain a natural follow-up.

$N_{\text{eff}}^{\text{ans}} = 2^H$ as an upper bound. Treating answer entropy as a count of equivalent independent voters is heuristic: on free-form tasks, distinct wrong answers inflate H without reducing ρ. The GSM-Hard

contrast $\beta_{\text{ent}} \gg \beta_{\text{agr}}$ in App. H is a direct manifestation of this bias. We report both scales rather than collapsing them, and the theorem operates on the agreement scale. The bias is conservative: any hard-ceiling finding on 2^H holds a fortiori on pairwise agreement, and the C4 conclusions also hold on the correctness scale (Table 2), so neither Thm. 1 nor the classifier rests on 2^H alone.

Rationale-level independence. Both metric scales treat two agents with the same final answer as redundant (answer level) or equally correct (correctness level), regardless of their reasoning. Counting same-answer, different-derivation agents as partially independent would require clustering rationales or derivation traces, which we leave to future work.

Model roster. The roster centres on 7–8B open-weight models (Qwen2.5, Llama-3.1, Ministral, Qwen3 thinking), which makes reproducible 30-way dense multi-round debate computationally feasible. Qwen2.5-32B, Gemini 3.1 Flash-Lite, and Gemini 3.1 Pro (solo) are scale, closed-source, and frontier checks.

Domain. The core tasks are bounded-answer (MCQA) or numeric-answer (arithmetic word problems with a unique correct integer) with short-horizon, voting-aggregated answers. Open-ended code generation is covered by the HumanEval sweep (App. N). Multi-step planning, long-horizon agentic tasks, and creative tasks are not covered.

Sampling. No difficulty filter is applied. MMLU-Hard comprises all test items of the five subjects in §4 (n=724), GSM-Hard the integer-answer items of Gao et al. [2023] (n=1,017 of 1,319), and GPQA the Diamond split (n=198).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The four contributions enumerated in the Introduction (regime classification, mean-field $k\tau$ theorem, condition-agnostic shared form, deployment guidance) match the empirical evidence in §5. Scope conditions and tested ranges are explicitly bounded in Appendix S.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: §5.6 "Limitations and scope" summarizes constraints in the main text; Appendix S enumerates testable extensions, decoding-parameter sensitivity, the $N \leq 30$ / $\tau \leq 5$ range bound, the $N_{\text{eff}}^{\text{ans}} = 2^H$ heuristic caveat, the noise-placebo single form, and aggregator/team-balance scope.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: Three propositions and one theorem are stated in §3 with explicit assumptions and proven in Appendix A. Scope conditions for Theorem 1 (high-agreement linearization, modal stability) are stated in App. A and empirically grounded in App. H.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Appendix C documents prompts, decoding parameters (temperature 0.4, top-p 0.95, fixed seed), sampling protocol, model versions, hardware, and answer-extraction logic. All datasets are public (MMLU-Hard, GSM-Hard, GPQA Diamond). Results are produced from logged trajectories; analysis scripts are bundled in supplementary.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The supplementary material includes Python analysis scripts (`analyses/` directory), the runner code, the per-team and per-agent CSV logs that produce every figure, and a README with reproduction commands.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix C specifies prompts, decoding hyperparameters, samples per task, team sizes, communication conditions, and aggregation rule. §4 provides the in-text summary; per-cell sample sizes appear in Table 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All fits report 300-iteration item-level bootstrap 95% CIs (figures and tables). Held-out audit uses pre-registered train/test split; AIC and r^2 reported for all fit comparisons. Bias correction and full statistical methodology in Appendix C.7.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix C (Hardware) reports the GPU types used (NVIDIA H100 for homogeneous experiments, A100 for heterogeneous, L40S for selected Llama sweeps), the inference engine (vLLM), and a project-total estimate of approximately 2,000 GPU-hours including preliminary runs. Per-problem deployment-cost arithmetic appears in Appendix R.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read the NeurIPS Code of Ethics; the work is a measurement framework for multi-agent LLM systems on public benchmarks, involves no human subjects, and conforms to the Code in every respect.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper proposes a measurement framework for multi-agent LLM systems and releases no model or dataset with deployment risk. The positive externality is more honest reporting of multi-agent effectiveness, supporting compute efficiency; we identify no direct negative societal impact beyond those already documented in the broader LLM literature.

    Guidelines:

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

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We release no models or scraped data; the supplementary contains only analysis code and processed CSV logs of agent answers on public benchmarks, with no high misuse risk.

    Guidelines:

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

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Datasets used (MMLU, GSM-Hard, GPQA Diamond) and models (Qwen2.5-7B/32B, Qwen3-8B, Llama-3.1-8B, Ministral-8B, Gemini 3.1) are cited and used per their published licenses (open-weights checkpoints under their respective licenses; Gemini via API per terms of service). Source citations appear in §4 and Appendix C.

    Guidelines:

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

13. **New assets**

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

    Answer: [Yes]

    Justification: The supplementary material includes analysis code (Python, polars/numpy), the multi-agent runner, processed agent- and team-level CSV logs (the source of every figure), and a README documenting reproduction commands. No new datasets or model weights are introduced.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper involves no crowdsourcing and no human subjects; experiments are run entirely on language models against public benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

    Answer: [N/A]

    Justification: No human subjects are involved; IRB review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs are the experimental subject of the paper, not a non-standard methodological component. Standard inference (greedy/sampling decoding via published HuggingFace and Gemini APIs) elicits the agent responses we measure; the measurement framework itself uses no LLM.

    Guidelines:

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