UNDERSTANDING LLM MULTI-AGENT SCALING: AGENT COMPOSITION AND EVIDENCE COVERAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Collective intelligence depends not only on how many agents participate, but also on what each contributes. We examine how model and persona composition shapes the returns from scaling LLM-based multi-agent systems. An evidence-coverage framework shows that equal individual coverage need not yield equal collective recovery: observations can overlap in what they miss. Across seven benchmarks and groups of two to sixteen agents, combining model and persona diversity achieves the highest seven-task macro-average accuracy among the tested configurations at every evaluated agent count in both voting and debate. Persona diversity provides larger average gains in mixed-model than in single-model groups across these settings, although benefits vary across tasks. We further analyze multi-round response trajectories using semantic effective rank to characterize how output diversity varies across agent configurations. Together, these findings highlight agent composition as a key dimension of LLM multi-agent scaling, motivating a focus on the complementary contributions of additional agents rather than their number alone.

1 INTRODUCTION

Large language models (LLMs) have achieved strong performance across diverse tasks, including reasoning, coding, and open-domain question answering (Wei et al., 2023; Achiam et al., 2023). However, single LLMs still struggle with problems that require multi-step reasoning, diverse perspectives, or complementary expertise (Huang et al., 2024).

To address these limitations, LLM-based multi-agent systems (MAS) orchestrate multiple agents through communication, coordination, or aggregation, enabling them to tackle challenges that are difficult for a single model (Wu et al., 2024; Hong et al., 2024; Du et al., 2023). Recent studies show that multi-agent collaboration can substantially improve over single-agent baselines on tasks ranging from software engineering (Qian et al., 2024) to scientific reasoning (Guo et al., 2024).

Given the effectiveness of MAS, a natural question is: *can we improve performance simply by scaling the number of agents?* Aggregating multiple outputs can provide ensemble-style gains (Li et al., 2024a; Wang et al., 2023). However, recent work (Kim et al., 2025) and our experiments show that these gains do not necessarily persist as systems grow. Figure 1 shows diminishing or unstable returns from homogeneous scaling across the evaluated tasks and models. These observations motivate examining whether additional calls contribute complementary evidence or repeat evidence already available to the system.

Our experiments (Figure 2) show that model-diverse configurations can outperform the mean single-backbone baseline. More broadly, differences in backbone models and prompts can expose a system to different responses and reasoning approaches (Wang et al., 2024a; Zhang et al., 2024; Qian et al., 2025). The benefits of persona variation may themselves depend on model composition. A role prompt that contributes useful responses within a single-backbone group may have a different effect in a mixed-model group, where the other agents already provide different reasoning approaches. In this study, we vary backbone models and persona prompts to examine how these design choices affect scaling. Figure 3 illustrates the evidence-coverage model motivating our analysis. We ask: **what limits MAS scaling, and how can diversity improve the value of additional agents?**

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



Figure 1: Scaling behavior of homogeneous multi-agent voting. This separate, extended homogeneous sweep plots success rate against agent count N on seven tasks for three base models. Many settings show limited gains at larger agent counts, while some are flat or deteriorate.

Prior work has documented diminishing returns from homogeneous scaling (Wang et al., 2024b; Kim et al., 2025; Chen et al., 2024; Choi et al., 2025) and improvements from heterogeneous designs (Wang et al., 2024a; Zhang et al., 2024; Wu and Ito, 2025; Yuen et al., 2025; Tang et al., 2025). We connect these observations by asking what additional agents contribute and how diversity changes their value. Our coverage model separates individual marginal coverage from joint omission to explain how overlapping observations can limit evidence recovery. We then compare model and persona diversity across agent counts and workflows, examining both aggregate gains and task-level differences. Finally, we analyze semantic output diversity to characterize the response differences associated with these performance patterns. Together, these analyses link an explicit account of evidence complementarity to a systematic study of practical agent compositions.

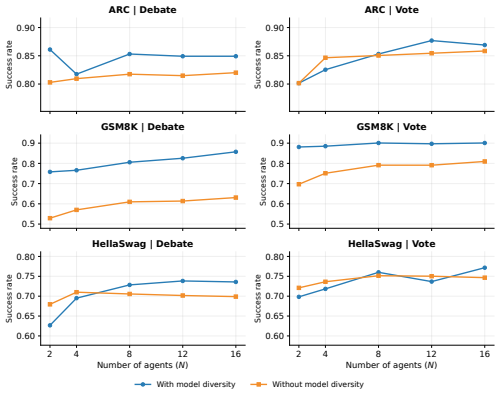


Figure 2: Model-diverse versus single-backbone performance: L3/L4 versus L1/L2 averages by task, workflow and agent count. L1/L2 average three single-backbone systems; L3/L4 use cyclic assignment from Qwen-2.5-7B, Llama-3.1-8B and Mistral-7B ($N = 2$: two backbones).

Our contributions are summarized as follows:

- We develop an evidence-coverage framework separating marginal coverage from joint non-coverage. Under explicit assumptions, we derive exact residual entropy and show how shared coverage propensities reduce recovery at matched marginal rates.
- We systematically compare four combinations of model and persona diversity across seven benchmarks, two workflows, and five agent counts. We characterize their aggregate gains, how persona gains depend on model composition, and task and workflow differences, including comparisons with the best single-backbone pool on ARC and Formal Logic.
- We characterize semantic output diversity using effective rank and correctness-conditioned trajectory profiles. We report its aggregate association with accuracy, differences between semantic-diversity and accuracy rankings, and sensitivity to task, workflow, and agent-count controls.

2 Problem Formulation

This section defines the MAS setting and distinguishes agent count from call count (Section 2.1). We then introduce an evidence-coverage model to analyze the value of overlapping and complementary observations (Section 2.2).

2.1 LLM-based Multi-Agent Systems

We begin by formally defining the class of systems we study.

Definition 2.1 (LLM-based MAS)**.** An *LLM-based multi-agent system* consists of N agents, each defined by a *configuration* specifying its backbone model, system prompt or persona, and tool access. Given an input $X \in \mathcal{X}$, the system makes a total of n calls through a workflow (e.g., voting or debate) and aggregates the outputs into a final answer estimating the ground-truth label $Y \in \mathcal{Y}$.

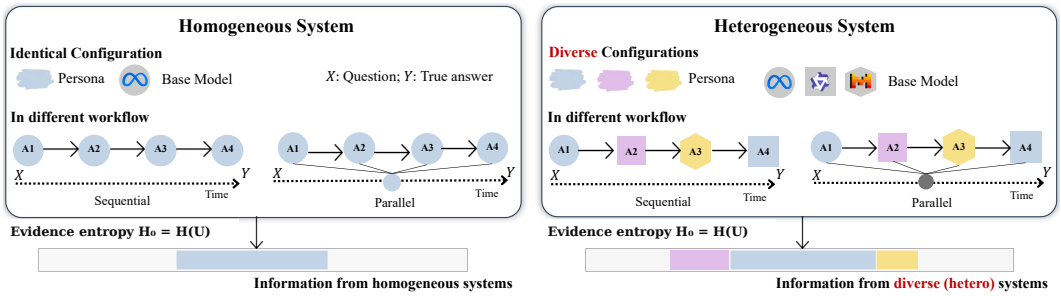


Figure 3: Illustration of the evidence-coverage model motivating our analysis. Repeated observations may overlap in the evidence they reveal, whereas complementary observations expand coverage. Model and persona diversity are design choices that can change this overlap; the illustration does not establish independence between heterogeneous agents.

Notation. We distinguish the *number of agents* N from the *number of calls* n: in single-round workflows such as voting, $n = N$, a workflow with R rounds and one call per agent per round uses $n = NR$, excluding any additional initialization or aggregation calls. This lets us analyze extracted information independently of which agent produces it.

Definition 2.2 (Agent configuration type)**.** A call's type $b(i) \in \mathcal{B}$ specifies its backbone, prompt and tool access. The allocation $m_b = |\{i : b(i) = b\}|$, with $\sum_b m_b = n$, records calls per type.

2.2 An Evidence-Coverage Model

For an input X, the MAS produces a transcript

$$Z_{1:n} := (Z_1, \ldots, Z_n), \tag{1}$$

and a final answer $\hat{Y} = g(X, Z_{1:n})$. We evaluate this answer against the benchmark label Y. If Y is determined by the complete input, standard Shannon entropy gives $H(Y \mid X) = 0$; computational difficulty does not change this identity. We therefore study complementary evidence through a separate observation model.

Assumption 2.3 (Evidence components and joint erasure observations)**.** Let $U = (U_1, \ldots, U_J)$ consist of independent discrete evidence components, with $h_j := H(U_j)$ and

$$0 < H_0 := \sum_{j=1}^{J} h_j < \infty.$$

For a specified collection of K observation channels, let $C_{jk} \in \{0, 1\}$ indicate whether channel k reveals component j. The entire coverage matrix C is independent of U. Each observation consists exactly of its coverage mask and the revealed component values:

$$W_k = \big((C_{jk}, V_{jk})\big)_{j=1}^{J}, \qquad V_{jk} = \begin{cases} U_j, & C_{jk} = 1, \\ \perp, & C_{jk} = 0, \end{cases} \tag{2}$$

where $\perp$ is a distinguished erasure symbol.

The model separates evidence components from the process that reveals them. Coverage events may be dependent across channels; their dependence determines how often additional observations revisit already covered components. We define residual uncertainty and recovered evidence as

$$R(K) := H(U \mid W_{1:K}), \qquad B(K) := I(U; W_{1:K}) = H_0 - R(K). \tag{3}$$

Scope. This reveal-or-erasure model isolates evidence coverage: revealed values are correct, while missed values are absent. It omits misleading evidence and aggregation errors. Its coverage variables are modeling quantities, and evidence recovery alone does not specify benchmark accuracy.

3 Coverage Dependence and the Role of Diversity

We use the coverage model to examine how overlap limits the value of additional observations. We first express evidence recovery in terms of joint non-coverage, then compare independent coverage with shared coverage propensities.

3.1 Information Limits and the Role of Diversity

In the observation model, additional channels contribute evidence when they reveal components missed by earlier observations. We characterize this contribution through the probability of jointly missing each component.

Let

$$q_j(K) := \Pr(C_{j1} = \cdots = C_{jK} = 0) \tag{4}$$

denote the probability that all K observations miss component j. At matched channel counts and marginal coverage rates, a collection is *more complementary for component j* if it has smaller $q_j(K)$. Componentwise smaller joint non-coverage gives greater total recovery.

Proposition 3.1 (Exact residual uncertainty)**.** *Under Assumption 2.3,*

$$R(K) = \sum_{j=1}^{J} h_j q_j(K), \qquad B(K) = \sum_{j=1}^{J} h_j[1 - q_j(K)]. \tag{5}$$

Consequently, $0 \le B(K) \le H_0$.

The proof is in Appendix C.1.1, which also gives the identity for dependent evidence components. Equation 5 shows that recovery improves by reducing joint non-coverage, with ceiling H_0.

3.2 Effective Channels: From Compute to Usable Evidence

We parameterize observations by their count K and marginal coverage probabilities. Proofs appear in Appendix C.1.

Definition 3.2 (Modeled channel count)**.** The count K specifies evidence observations $W_1, \ldots, W_K$. In a single-round abstraction, one observation per agent call gives $K = n = N$. In interactive Debate, changing N changes the transcript distribution, so the connection is conceptual rather than a fixed decomposition.

Definition 3.3 (Coverage rate)**.** For component j and channel k, define the marginal coverage probability

$$\alpha_{jk} := \Pr(C_{jk} = 1). \tag{6}$$

The uniform-rate model sets $\alpha_{jk} = \alpha \in (0, 1)$ for all j, k.

Assumption 3.4 (Independent coverage across channels)**.** For each fixed component j, the indicators $C_{j1}, \ldots, C_{jK}$ are mutually independent.

This assumption concerns coverage events in the observation model. Different backbone models or persona prompts do not, by themselves, imply that it holds for a real MAS.

Theorem 3.5 (Evidence recovery under independent coverage)**.** *Under Assumptions 2.3 and 3.4,*

$$R(K) = \sum_{j=1}^{J} h_j \prod_{k=1}^{K} (1 - \alpha_{jk}). \tag{7}$$

In the uniform-rate model,

$$\boxed{\begin{aligned} R(K) &= H_0(1-\alpha)^K, \\ B(K) &= H_0[1-(1-\alpha)^K]. \end{aligned}} \tag{8}$$

The result factors the joint miss probabilities in Proposition 3.1. For a one-observation-per-call abstraction ($K = n$), let type b have coverage rate $\alpha_{j,b}$. Under the same independence assumption, the allocation in Definition 2.2 gives

$$q_j(n) = \prod_{b:m_b>0} (1 - \alpha_{j,b})^{m_b}. \tag{9}$$

Thus composition changes recovery through coverage rates and, outside this independent case, their dependence.

3.3 Coverage Dependence and the Value of Diversity

To isolate the effect of dependence, we compare designs with the same marginal coverage rates. For each component j, a shared-propensity design draws $P_j \in [0,1]$ with $\mathbb{E}[P_j] = \alpha_j$. The vector P is independent of U. Conditional on P, coverage indicators are generated independently of U, and the m indicators for component j are independent Bernoulli(P_j) draws. An independent design instead uses m independent Bernoulli(α_j) draws for that component. Both designs use the observations in Assumption 2.3.

Proposition 3.6 (Effect of shared coverage propensity)**.** *For $m \geq 2$, the two designs satisfy*

$$\begin{aligned} R_{\text{shared}}(m) &= \sum_{j=1}^{J} h_j \mathbb{E}[(1-P_j)^m] \\ &\geq \sum_{j=1}^{J} h_j (1-\alpha_j)^m = R_{\text{independent}}(m). \end{aligned} \tag{10}$$

The inequality is strict if P_j is nonconstant for at least one component with $h_j > 0$.

The proof uses Jensen's inequality and is given in Appendix C.1.3. Shared variation in coverage propensity increases the chance that all observations miss the same component, even when individual observations have the same marginal coverage rate.

Implications for heterogeneous MAS. Proposition 3.6 compares shared-propensity dependence with independence at matched marginal rates. It shows that collective recovery can differ even when individual marginal coverage is unchanged. Model and persona diversity are practical design choices that may alter both individual evidence coverage and the overlap in what agents miss. This motivates examining how these choices affect performance as agent count grows; configuration differences alone do not establish independent coverage.

Diminishing marginal evidence recovery. Under uniform independent coverage, Theorem 3.5 gives

$$B(K+1) - B(K) = H_0 \alpha (1-\alpha)^K. \tag{11}$$

Under these assumptions, marginal evidence recovery decreases as coverage accumulates. Translating recovered evidence into task accuracy also requires an aggregation or interaction procedure that uses the evidence effectively. In particular, increasing the number of agents in Debate can change the existing responses, so different agent counts need not correspond to nested observations. The coverage result therefore does not assert monotone or concave accuracy scaling for every MAS.

3.4 A Semantic Diversity Diagnostic: $\widehat{K}$

We use entropy effective rank (Friedman and Dieng, 2023) to describe the geometry of agent outputs. It measures semantic variation in an embedding space; its association with accuracy is an empirical question, separate from the modeled channel count K.

For any nonzero row-embedding matrix $E \in \mathbb{R}^{n \times d}$, set

$$G = EE^\top, \qquad \bar{G} = G/\mathrm{Tr}(G). \tag{12}$$

If λ_j are the eigenvalues of $\bar{G}$, define

$$\widehat{K}(E) := 2^{-\sum_j \lambda_j \log_2 \lambda_j}, \qquad 0 \log_2 0 := 0. \tag{13}$$

For unit-norm output embeddings, G is a cosine Gram matrix. In the trajectory analysis, the rows of E are across-round average embeddings without further normalization (Section 4.3.2). We use $\widehat{K}$ for the resulting effective rank.

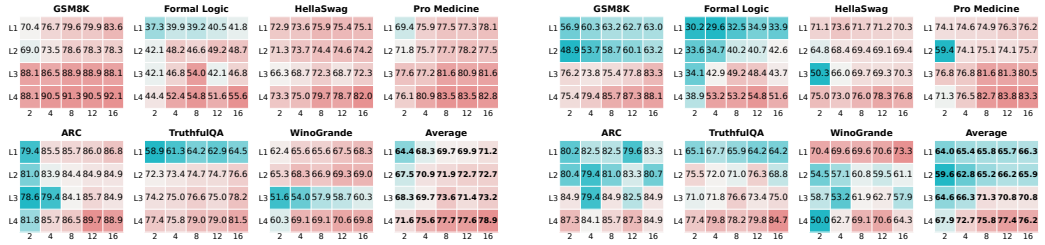
Interpretation. A concentrated spectrum has low effective rank; a balanced spectrum has higher rank:

$$1 \leq \widehat{K}(E) \leq \mathrm{rank}(E) \leq \min(n, d). \tag{14}$$

Rank one gives $\widehat{K}(E) = 1$; attaining n requires mutually orthogonal rows with equal nonzero norms. Proofs are in Appendix C.2.

Table 1: Effect of adding persona diversity to model-diverse systems. L3 uses model diversity, whereas L4 combines model and persona diversity. Comparisons use the same agent count within each workflow. Δ denotes L4−L3 in percentage points. The final rows are unweighted seven-task macro-averages. Single Agent is the three-backbone mean with a shared persona after self-refinement. Deltas and averages use unrounded entries.

Dataset	Single Agent	N	Vote			Debate			Dataset	Single Agent	N	Vote			Debate		
			L3	L4	Δ	L3	L4	Δ				L3	L4	Δ	L3	L4	Δ
GSM8K	50.8	2	88.1	88.1	0.0	76.2	75.4	-0.8	**ARC**	77.8	2	78.6	81.8	+3.2	84.9	87.3	+2.4
		4	86.5	90.5	+4.0	73.8	79.4	+5.6			4	79.4	85.7	+6.3	79.4	84.1	+4.8
		8	88.9	91.3	+2.4	75.4	85.7	+10.3			8	84.1	86.5	+2.4	84.9	85.7	+0.8
		12	88.9	90.5	+1.6	77.8	87.3	+9.5			12	85.7	89.7	+4.0	82.5	87.3	+4.8
		16	88.1	92.1	+4.0	83.3	88.1	+4.8			16	84.9	88.9	+4.0	84.9	84.9	0.0
Formal Logic	32.0	2	42.1	44.4	+2.4	34.1	38.9	+4.8	**Truthful QA**	71.8	2	74.2	77.4	+3.2	71.0	77.4	+6.5
		4	46.8	52.4	+5.6	42.9	53.2	+10.3			4	75.0	75.8	+0.8	71.8	79.8	+8.1
		8	54.0	54.8	+0.8	49.2	53.2	+4.0			8	76.6	79.0	+2.4	76.6	78.2	+1.6
		12	42.1	51.6	+9.5	48.4	54.8	+6.4			12	75.0	79.0	+4.0	73.4	79.8	+6.5
		16	46.8	55.6	+8.7	43.7	51.6	+7.9			16	78.2	81.5	+3.2	75.0	84.7	+9.7
HellaSwag	66.1	2	66.3	73.3	+7.0	50.3	75.0	+24.7	**Wino grande**	57.1	2	51.6	60.3	+8.7	58.7	50.0	-8.7
		4	68.7	75.0	+6.3	66.0	73.0	+7.0			4	54.0	69.1	+15.1	53.2	62.7	+9.5
		8	72.3	79.7	+7.3	69.7	76.0	+6.3			8	57.9	69.1	+11.1	61.9	69.1	+7.2
		12	68.7	78.7	+10.0	69.3	78.3	+9.0			12	58.7	70.6	+11.9	62.7	70.6	+7.9
		16	72.3	82.0	+9.7	70.3	76.8	+6.5			16	60.3	69.8	+9.5	57.9	64.3	+6.4
Pro Medicine	68.6	2	77.6	76.1	-1.5	76.8	71.3	-5.5	**Average**	60.6	2	68.3	71.6	**+3.3**	64.6	67.9	**+3.3**
		4	77.2	80.9	+3.7	76.8	76.5	-0.4			4	69.7	75.6	**+6.0**	66.3	72.7	**+6.4**
		8	81.6	83.5	+1.8	81.6	82.7	+1.1			8	73.6	77.7	**+4.0**	71.3	75.8	**+4.5**
		12	80.9	83.5	+2.6	81.3	83.8	+2.6			12	71.4	77.6	**+6.2**	70.8	77.4	**+6.7**
		16	81.6	82.8	+1.1	80.5	83.3	+2.8			16	73.2	78.9	**+5.7**	70.8	76.2	**+5.4**



(a) Vote (b) Debate

Figure 4: Diversity analysis of the **Vote** and **Debate** mechanisms across all datasets. Each dataset panel shows a 4×5 heatmap, where rows represent the four combinations of model and persona diversity (L1–L4) and columns denote the number of agents N. Colors indicate success rate values: cyan (lowest) to red (highest).

4 Experiments

Motivated by the distinction between additional observations and complementary evidence, we examine how agent count and model/persona composition affect MAS performance. We first study returns from homogeneous scaling, then compare the gains of model and persona diversity and their variation across tasks and workflows, and finally analyze how semantic output diversity relates to accuracy. These experiments evaluate practical configurations and observed outputs; they do not directly measure the coverage variables of the theoretical model. We compare four configurations: L1 uses a single backbone with a shared prompt; L2 adds persona diversity; L3 mixes backbones with the shared prompt; and L4 combines both diversity sources. L1/L2 results average three separately evaluated single-backbone systems, whereas L3/L4 mix models within each system. Vote aggregates separately generated answers; Debate allows agents to observe prior responses over multiple rounds. Within each workflow, comparisons fix N and the round schedule, matching agent and call counts. Implementation details appear in Section A.

4.1 Finding 1: Scaling Homogeneous MAS Exhibits Diminishing Returns

We first examine whether increasing the number of agents improves performance in homogeneous settings. Figure 1 shows homogeneous voting performance as a function of agent count across multiple tasks and base models. Many settings exhibit smaller gains at larger agent counts, while some

remain flat or deteriorate. Increasing N therefore does not reliably sustain accuracy improvements across the evaluated settings. The coverage model motivates examining whether additional agents revisit evidence already available to the system. Finding 3 examines output-level overlap and diversity.

4.2 Finding 2: Diverse Configurations Improve Scaling on Average

Figure 4 shows L4 leading the seven-task macro-average at all evaluated counts in both workflows. At $N = 16$, its gains over L1 are 7.78 percentage points for Vote and 9.93 for Debate, using unrounded averages.

Table 2: Smallest evaluated N reaching the L1 ($N = 16$) baseline using unrounded seven-task macro-averages. Peak: best evaluated accuracy (%); "–": baseline not reached.

Method	Config	Agents Reaching L1 ($N = 16$)	Accuracy at N	Peak Accuracy
Vote	L1	16	71.2	71.2
	L2	8	71.9	72.7
	L3	8	73.6	73.6
	L4	**2**	71.6	**78.9**
Debate	L1	16	66.3	66.3
	L2	–	–	66.2
	L3	8	71.3	71.3
	L4	**2**	67.9	**77.4**

Persona gains depend on composition. Across 35 task–count settings, persona gains occur in 20 single-backbone versus 33 mixed-model settings for Vote, and 14 versus 30 for Debate. For configuration accuracies A_ℓ, the descriptive contrast $I = (A_4 - A_3) - (A_2 - A_1)$ has seven-task means of 4.16 and 5.81 percentage points for Vote and Debate at $N = 16$ (Appendix D.6).

Task and workflow matter. On WinoGrande under Debate at $N = 16$, L4 trails L1 by 8.99 percentage points. On Formal Logic, persona-diverse MIX exceeds the best persona-diverse single-backbone pool at one of five counts for Vote and four for Debate. On ARC, it wins at one Vote count and none under Debate (Appendix D.5).

Table 2 additionally reports L4 with **2 agents** at 71.6% (Vote) and 67.9% (Debate), versus 71.2% and 66.3% for L1 with **16 agents**. These observed thresholds concern agent count; they establish neither statistical superiority nor equivalent token or latency savings.

4.3 Finding 3: Semantic Diversity Is Associated with Accuracy

We analyze recorded multi-round trajectories and the accuracy at the last recorded round. These examine how output diversity relates to performance across agent counts and configurations.

4.3.1 Output Overlap and Scaling Patterns

For each question, we concatenate each agent's responses across recorded rounds and embed this trajectory using NV-Embed-v2 (Lee et al., 2025). For $n = N \geq 2$ normalized trajectory embeddings, mean pairwise cosine similarity is

$$\bar{s} := \frac{2}{n(n-1)} \sum_{i<j} \langle \hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j \rangle. \qquad (15)$$

Higher $\bar{s}$ indicates greater average similarity in the chosen representation. It is a measure of semantic overlap, not a correlation coefficient between agent correctness indicators.

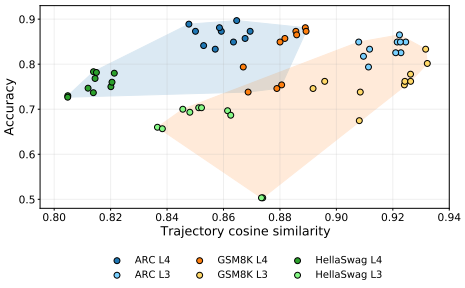


Figure 5: Illustrative L3/L4 results on ARC, GSM8K and HellaSwag (both trajectory types, five N). Shading encloses each configuration's points.

Figure 5 compares mean pairwise similarity $\bar{s}$ and terminal accuracy across persona settings. The plotted persona-diverse settings generally combine lower similarity with higher terminal accuracy. Figure 6 shows how similarity varies with agent count and configuration. Increasing agent count does not, by itself, increase expected mean pairwise similarity under independent sampling from a fixed distribution. We therefore interpret these trends within their generation and interaction settings. We next examine $\widehat{K}$, computed from across-round average embeddings.

Table 3: ARC trajectory statistics averaged over $N \in \{2, 4, 8, 12, 16\}$. Acc. uses the last recorded round; main Vote results use the initial round. ΔAcc is relative to L1 (pp), and c/w denote correct/incorrect final answers.

Config	Method	Acc.	ΔAcc	$\widehat{K}$	$\widehat{K}_c$	$\widehat{K}_w$	Method	Acc.	ΔAcc	$\widehat{K}$	$\widehat{K}_c$	$\widehat{K}_w$
L1	Debate	81.6	–	1.197	1.184	1.176	Vote	81.3	–	1.201	1.183	1.173
L2		81.0	-0.7	1.348	1.315	1.234		81.5	+0.2	1.349	1.318	1.222
L3		83.3	+1.7	1.246	1.220	1.160		83.8	+2.5	1.245	1.223	1.161
L4		**85.9**	**+4.2**	**1.517**	**1.472**	**1.288**		**87.5**	**+6.1**	**1.521**	**1.484**	**1.297**

4.3.2 Semantic Diversity and Task Accuracy

For each question, let $\hat{z}_{ir}$ be agent i's normalized response embedding at recorded round r, including initialization. We define

$$e_i = |\mathcal{T}_i|^{-1} \sum_{r \in \mathcal{T}_i} \hat{z}_{ir}, \qquad \widehat{K} = \widehat{K}([e_i]_i). \tag{16}$$

We retain averaged-vector norms, so within-agent cross-round agreement affects the Gram spectrum along with between-agent directions. We average per-question ranks; L1/L2 then average the three specified backbones.

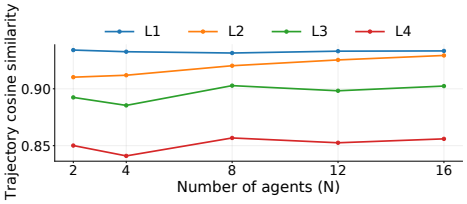


Figure 6: Mean trajectory similarity across available ARC, GSM8K and HellaSwag task–trajectory-type summaries, weighted equally. L1/L2 first average backbones. Lines connect means.

Configuration choices change semantic diversity. Table 3 reports the highest $\widehat{K}$ for L4 in both workflows on ARC. However, semantic diversity does not consistently rank configurations by accuracy. Under Debate, L2 has a higher effective rank than L3 but lower accuracy. Model diversity and persona diversity therefore affect output semantics differently.

Association with accuracy. Trajectory $\widehat{K}$ is positively associated with terminal accuracy across the evaluated settings. Under a linear baseline using agent count and configuration indicators, adding $\widehat{K}$ increases in-sample R^2 from 0.062 to 0.205 ($\Delta R^2 = 0.143$). On the same 220 observations, the increment falls to 0.0062 after adding task and workflow controls, and to 0.0014 when agent count is also treated categorically. The incremental fit therefore depends on the regression specification (Appendix D.3). Figure 7 describes the L4 trend by fitting ten means across the six tasks with available ranks; its R^2 evaluates these means.

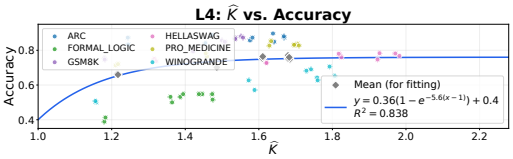


Figure 7: L4: 60 task-level points and 10 means, each averaging six tasks at fixed N and workflow. The exponential curve is fitted to these 10 means ($R^2 = 0.838$).

Correctness-conditioned diagnostic: $\widehat{K}_c$ and $\widehat{K}_w$. Among agents with extracted terminal answers, define $\mathcal{I}_c = \{i : \hat{y}_i = Y\}$ and $\mathcal{I}_w = \{i : \hat{y}_i \neq Y\}$ for gold answer Y. Apply the same operator to each nonzero submatrix E_a ($a \in \{c, w\}$):

$$\widehat{K}_a = \widehat{K}(E_a), \qquad 1 \leq \widehat{K}_a \leq \min(|\mathcal{I}_a|, d). \tag{17}$$

These statistics describe trajectories grouped by final-answer correctness; their ranges depend on subgroup sizes.

Correctness-conditioned response profiles. Figure 8 compares semantic diversity among trajectories with correct and incorrect final answers. These profiles complement overall effective rank by showing how diversity is distributed between the two groups.

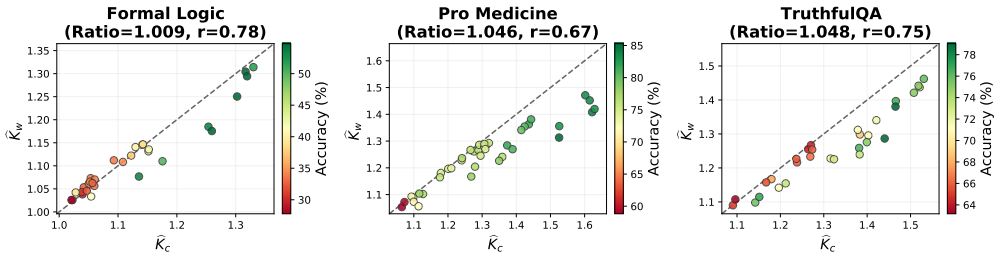


Figure 8: Correctness-conditioned trajectory profiles on three tasks, colored by terminal accuracy. Points below the diagonal satisfy $\widehat{K}_c > \widehat{K}_w$; the reported ratio is the mean of $\widehat{K}_c/\widehat{K}_w$ across settings; r is its Pearson correlation with terminal accuracy.

5 Related Work

LLM-based multi-agent systems. LLM-based multi-agent systems coordinate multiple LLM agents through sampling, communication, and aggregation mechanisms (Xi et al., 2023; Wang et al., 2024b; Guo et al., 2024). Representative designs include independent sampling and voting (Wang et al., 2023), decentralized debate and role-playing (Du et al., 2023; Khan et al., 2024; Li et al., 2024b; 2023), centralized orchestration such as AutoGen and MetaGPT (Wu et al., 2024; Hong et al., 2024), and more adaptive or self-improving coordination strategies (Dang et al., 2025; Zhao et al., 2025). Recent work also documents systematic failure modes in multi-agent systems (Cemri et al., 2025), highlighting that performance depends not only on agent count but also on how agents are configured and coordinated.

Multi-agent scaling and diversity. Studies of voting, debate, and broader agent architectures show that increasing the number of agents does not uniformly improve performance (Wang et al., 2024b; Chen et al., 2024; Qian et al., 2025; Choi et al., 2025; Kim et al., 2025). Heterogeneous model ensembles and diverse agent populations provide an alternative to homogeneous scaling (Wang et al., 2024a; Zhang et al., 2024). Related work suggests that preserving disagreement can be preferable to premature consensus (Wu and Ito, 2025), that the benefits of diversity depend on task complexity (Tang et al., 2025), and that persona-based diversification has important limitations (Samuel et al., 2024; Taillandier et al., 2025). Our empirical contribution is a joint analysis of model and persona composition across seven benchmarks, two workflows, and five agent counts. We examine how persona gains vary with model composition, scale, and workflow through task-level interaction contrasts and failure cases, and compare mixed-model configurations with the best single-backbone pool on ARC and Formal Logic.

Information and output diversity. Information-theoretic analyses have studied information gain and error accumulation in LLM reasoning (Ton et al., 2024; Gan et al., 2025), as well as dependence and collaborative intelligence in multi-agent systems (Riedl, 2025; Chang, 2025). Our evidence-coverage framework differs by isolating the role of joint non-coverage in evidence recovery under explicit assumptions; the experiments examine agent composition and output behavior rather than directly measuring latent evidence coverage. For output diagnostics, we apply the entropy effective-rank form used by the Vendi Score (Friedman and Dieng, 2023) to response embeddings and multi-round trajectories. This application relates semantic output diversity to observed accuracy while explicitly examining the sensitivity of its incremental association to task and workflow.

6 Conclusion

Agent count alone does not characterize the value of MAS scaling. Under explicit observation assumptions, our evidence-coverage framework shows how joint non-coverage limits recovery and why equal marginal coverage need not yield equal collective recovery. Across seven benchmarks and two workflows, combining model and persona diversity achieves the highest macro-average accuracy at every evaluated agent count, while task-level results and best-backbone comparisons show that gains depend on the setting. Persona gains also depend on model composition, highlighting the value of studying the two diversity sources jointly. Trajectory effective rank describes semantic output variation and its aggregate association with accuracy, with limited incremental fit after task, workflow, and agent-count controls. Together, the framework and experiments motivate considering the information supplied by additional agents and evaluating agent composition and workflow alongside scale.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Mert Cemri, Melissa Z Pan, Shuyi Yang, Lakshya A Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Why do multi-agent LLM systems fail? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Edward Y Chang. *Multi-LLM Agent Collaborative Intelligence: The Path to Artificial General Intelligence*. Edward Y. Chang, 2025.

Lingjiao Chen, Jared Quincy Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. Are more llm calls all you need? towards scaling laws of compound inference systems. *arXiv preprint arXiv:2403.02419*, 2024.

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? *arXiv preprint arXiv:2508.17536*, 2025.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Yufan Dang, Chen Qian, Xueheng Luo, Jingru Fan, Zihao Xie, Ruijie Shi, Weize Chen, Cheng Yang, Xiaoyin Che, Ye Tian, Xuantang Xiong, Lei Han, Zhiyuan Liu, and Maosong Sun. Multi-agent collaboration via evolving orchestration. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning, 2023. URL https://arxiv.org/abs/2210.02410.

Zeyu Gan, Yun Liao, and Yong Liu. Rethinking external slow-thinking: From snowball errors to probability of correct reasoning. *arXiv preprint arXiv:2501.15602*, 2025.

Aaron Grattafiori et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.

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

Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. Metagpt: Meta programming for a multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2024.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet, 2024. URL https://arxiv.org/abs/2310.01798.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive LLMs leads to more truthful answers. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24, 2024.

Yubin Kim, Ken Gu, Chanwoo Park, Chunjong Park, Samuel Schmidgall, A. Ali Heydari, Yao Yan, Zhihan Zhang, Yuchen Zhuang, Mark Malhotra, Paul Pu Liang, Hae Won Park, Yuzhe Yang, Xuhai Xu, Yilun Du, Shwetak Patel, Tim Althoff, Daniel McDuff, and Xin Liu. Towards a science of scaling agent systems. *arXiv preprint arXiv:2512.08296*, 2025.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models, 2025.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large language model society. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NeurIPS '23, 2023.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More agents is all you need. *arXiv preprint arXiv:2402.05120*, 2024a.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. Improving multi-agent debate with sparse communication topology. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7281–7294, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.427.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186. Association for Computational Linguistics, 2024.

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 *The Thirteenth International Conference on Learning Representations*, 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, 2025. URL `https://arxiv.org/abs/2412.15115`.

Christoph Riedl. Emergent coordination in multi-agent language models. *arXiv preprint arXiv:2510.05174*, 2025.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. 2019.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. Personagym: Evaluating persona agents and llms. *arXiv preprint arXiv:2407.18416*, 2024.

Patrick Taillandier, Jean Daniel Zucker, Arnaud Grignard, Benoit Gaudou, Nghi Quang Huynh, and Alexis Drogoul. Integrating llm in agent-based social simulation: Opportunities and challenges. *arXiv preprint arXiv:2507.19364*, 2025.

Bohan Tang, Huidong Liang, Keyue Jiang, and Xiaowen Dong. On the importance of task complexity in evaluating llm-based multi-agent systems. *arXiv preprint arXiv:2510.04311*, 2025.

Jean-Francois Ton, Muhammad Faaiz Taufiq, and Yang Liu. Understanding chain-of-thought in llms through information theory. *arXiv preprint arXiv:2411.11984*, 2024.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024a.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024b. doi: 10.1007/s11704-024-40231-1.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024.

Zengqing Wu and Takayuki Ito. The hidden strength of disagreement: Unraveling the consensus-diversity tradeoff in adaptive multi-agent systems. *arXiv preprint arXiv:2502.16565*, 2025.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.

Sizhe Yuen, Francisco Gomez Medina, Ting Su, Yali Du, and Adam J Sobey. Intrinsic memory agents: Heterogeneous multi-agent llm systems through structured contextual memory. *arXiv preprint arXiv:2508.08997*, 2025.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Kexun Zhang, Weiran Yao, Zuxin Liu, Yihao Feng, Zhiwei Liu, Rithesh Murthy, Tian Lan, Lei Li, Renze Lou, Jiacheng Xu, et al. Diversity empowers intelligence: Integrating expertise of software engineering agents. *arXiv preprint arXiv:2408.07060*, 2024.

Wanjia Zhao, Mert Yuksekgonul, Shirley Wu, and James Zou. Sirius: Self-improving multi-agent systems via bootstrapped reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

A EXPERIMENTAL SETUP

Tasks. We consider a diverse set of reasoning and knowledge benchmarks, including GSM8K (Cobbe et al., 2021), ARC (Clark et al., 2018), Formal Logic (Hendrycks et al., 2021), TruthfulQA (Lin et al., 2022), HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2019), and Pro Medicine (Hendrycks et al., 2021). These benchmarks cover arithmetic reasoning, formal deduction, commonsense reasoning, and domain-specific knowledge. For the main configuration comparison, GSM8K and ARC-Challenge use 126 test questions each; HellaSwag uses 300 validation questions; WinoGrande uses 126 validation questions from the XL subset. TruthfulQA uses the multiple-choice MC1 validation set: 126 questions are selected and 124 remain after input formatting. These subsets are selected by shuffling with `random_state=0` and taking the required prefix. Formal Logic and Pro Medicine use the full MMLU `formal_logic` and `professional_medicine` test subsets, containing 126 and 272 questions, respectively. Dataset selection and formatting are implemented in `src/data/` of the accompanying code.

Models. For the main open-weight experiments, agents are instantiated using three open-source LLMs: Qwen-2.5-7B (Qwen et al., 2025), Llama-3.1-8B (Grattafiori et al., 2024), and Mistral-7B (Jiang et al., 2023). In the *single-model* setting, all agents within a MAS share the same base model; in the *MIX* setting, agents within a single MAS can use different base models, enabling model-level heterogeneity. The Qwen and Llama checkpoints are `Qwen/Qwen2.5-7B-Instruct` and `meta-llama/Meta-Llama-3.1-8B-Instruct`, respectively. Task-specific persona prompts are defined in `src/model/model_utils.py`; their assignment and insertion into agent messages are implemented in `src/main.py`. Persona-conditioned settings use temperature 0. The Single Agent baseline uses a shared persona and terminal self-refinement accuracy averaged over the three backbones. The evaluated configurations vary in model assignment and persona prompts; Debate additionally allows agents' responses to depend on interaction history. For L3/L4, deterministic cyclic assignment to three model slots gives counts $(1, 1, 0)$, $(2, 1, 1)$, $(3, 3, 2)$, $(4, 4, 4)$ and $(6, 5, 5)$ at $N = 2, 4, 8, 12, 16$. Thus, only $N = 12$ is balanced and $N = 2$ uses two backbones. In Debate, agent identities and persona assignments remain fixed across all rounds for a given run.

MAS Workflows. We consider two representative collaboration mechanisms (Choi et al., 2025). In **Vote**, agents generate answers without observing one another's responses, followed by majority aggregation. In **Debate**, agents revise their initial answers over three interaction rounds before final aggregation. Ties between the most frequent extracted answers are broken uniformly at random. The main configuration comparison uses $N \in \{2, 4, 8, 12, 16\}$; Figure 1 is a separate extended sweep. Threshold comparisons and macro-averages use unrounded accuracies; displayed entries are rounded. Diversity configurations are compared at the same agent count and with the same round schedule within a workflow. These comparisons control agent count and the prescribed call schedule; they do not establish equal token usage, latency, or computational cost across models or workflows.

Diversity Configurations. We evaluate four configurations that combine the presence or absence of model diversity and persona diversity:

- **L1: No Diversity.** All agents share the same base model and the same shared system prompt. This serves as the homogeneous baseline. Results are averaged across the three backbone-specific settings.
- **L2: Persona Diversity Only.** All agents share the same base model, with personas assigned cyclically from the task-specific role set. Results are averaged across the three backbone-specific settings.
- **L3: Model Diversity Only.** Agents are drawn from different base models (Qwen, Llama, Mistral) but all use the same shared system prompt.
- **L4: Full Diversity.** Agents differ in both base model and persona prompt, combining model-level and prompt-level heterogeneity.

This design compares complete configurations at each task, workflow, and agent count. Model comparisons jointly vary backbone composition and member capability; persona comparisons jointly vary role content and between-agent prompt variation, without isolating variation from prompt quality. Mixed-model scaling follows the stated cyclic allocation, so increasing N changes both system size and backbone proportions.

B Design Guidelines for LLM-based MAS

Our theoretical and empirical analyses motivate the following guidelines for MAS design:

- **Evaluate diversity within the target setting.** Compare model and persona diversity at the same agent count within the intended task and workflow. Their effects differ across settings, so improvements from one diversity source should not be assumed to transfer to another.
- **Use output diversity alongside accuracy.** The label-free statistic $\widehat{K}$ describes semantic diversity and can complement accuracy-based evaluation. A larger value does not by itself establish more useful evidence. When labels are available, correctness-conditioned response profiles provide retrospective diagnostics of the outputs associated with successful aggregation.
- **Select agent count from observed marginal gains.** Compare improvements from additional agents with improvements from changing agent composition. The appropriate agent count depends on the task, workflow, configuration, and deployment cost; our results do not imply a universal saturation threshold.

C Proofs and Technical Details

We provide proofs for the evidence-coverage model defined in Assumption 2.3. Section C.1.1 derives the exact residual entropy, Section C.1.2 establishes recovery under independent coverage, and Section C.1.3 compares shared coverage propensities with independent coverage. Section C.2 establishes basic properties of the semantic effective-rank diagnostic $\widehat{K}$.

C.1 Proofs for the Evidence-Coverage Model

C.1.1 Exact Residual Entropy

Proof of Proposition 3.1. Let $C^{(K)} = (C_{jk})_{1\le j\le J,\, 1\le k\le K}$. For a mask realization c with positive probability, define

$$S(c) := \{j : \text{there exists } k \le K \text{ such that } c_{jk} = 1\}.$$

The observations $W_{1:K}$ contain exactly the coverage mask and the component values revealed by that mask. Conditional on $C^{(K)} = c$, observing $W_{1:K}$ is therefore equivalent to observing $U_{S(c)}$.

Since the entire coverage mask is independent of U,

$$H(U \mid C^{(K)} = c, U_{S(c)}) = H(U \mid U_{S(c)}).$$

Averaging this conditional entropy first yields, even for dependent evidence components,

$$R(K) = \mathbb{E}_{C^{(K)}}[H(U \mid U_{S(C^{(K)})})].$$

Here only independence of the mask from U is needed. Under Assumption 2.3, independence of the evidence components gives

$$H(U \mid U_{S(c)}) = \sum_{j \notin S(c)} h_j.$$

Averaging over mask realizations yields

$$\begin{aligned} R(K) &= \mathbb{E}\left[\sum_{j=1}^{J} h_j \mathbf{1}\{C_{j1} = \cdots = C_{jK} = 0\}\right] \\ &= \sum_{j=1}^{J} h_j q_j(K). \end{aligned}$$

Finally, $H(U) = H_0$, so

$$B(K) = H_0 - R(K) = \sum_{j=1}^{J} h_j[1 - q_j(K)].$$

Since $q_j(K) \in [0, 1]$, we obtain $0 \le B(K) \le H_0$. □

C.1.2 Recovery Under Independent Coverage

Proof of Theorem 3.5. For each component j, Assumption 3.4 gives

$$q_j(K) = \Pr(C_{j1} = \cdots = C_{jK} = 0) = \prod_{k=1}^{K}(1 - \alpha_{jk}).$$

Substitution into Proposition 3.1 yields

$$R(K) = \sum_{j=1}^{J} h_j \prod_{k=1}^{K}(1 - \alpha_{jk}).$$

When $\alpha_{jk} = \alpha$ for all j, k, this becomes

$$R(K) = H_0(1 - \alpha)^K, \qquad B(K) = H_0[1 - (1 - \alpha)^K].$$

□

Marginal recovery. For a nested sequence of observations in the uniform-rate independent-coverage model,

$$B(K + 1) - B(K) = H_0\alpha(1 - \alpha)^K.$$

For $\alpha \in (0, 1)$, this increment is positive and decreases with K. The result concerns recovered evidence in the specified observation model; it does not establish monotonicity or concavity of benchmark accuracy.

Exponential representation. Using the natural logarithm, define

$$\tau := -K \log(1 - \alpha).$$

Then

$$\frac{R(K)}{H_0} = e^{-\tau} \leq e^{-\alpha K},$$

because $-\log(1 - \alpha) \geq \alpha$. When α varies across designs, ordering αK need not order their exact residuals.

C.1.3 Shared Coverage Propensities

Proof of Proposition 3.6. In the shared-propensity design, the m coverage indicators for component j are independent Bernoulli(P_j) variables conditional on the propensity vector P. Therefore,

$$\Pr(C_{j1} = \cdots = C_{jm} = 0 \mid P) = (1 - P_j)^m.$$

Taking expectations and applying Proposition 3.1 gives

$$R_{\text{shared}}(m) = \sum_{j=1}^{J} h_j \mathbb{E}[(1 - P_j)^m].$$

For $m \geq 2$, the function $f(p) = (1 - p)^m$ is strictly convex on $[0, 1]$. Jensen's inequality implies

$$\mathbb{E}[(1 - P_j)^m] \geq (1 - \mathbb{E}[P_j])^m = (1 - \alpha_j)^m.$$

The independent design has the same marginal coverage rate α_j and joint non-coverage probability $(1 - \alpha_j)^m$. Consequently,

$$R_{\text{shared}}(m) \geq \sum_{j=1}^{J} h_j(1 - \alpha_j)^m = R_{\text{independent}}(m).$$

The inequality is strict if P_j is nonconstant for at least one component with $h_j > 0$. □

Interpretation. At matched marginal coverage rates, shared variation in coverage propensity increases joint non-coverage and reduces recovered evidence. This comparison isolates dependence within the stated model. Model or persona heterogeneity may change both marginal coverage and dependence, but configuration differences alone do not establish either coverage structure.

Asymptotic behavior. Since $(1-P_j)^m$ converges to $\mathbf{1}\{P_j = 0\}$ and is bounded by one, dominated convergence gives

$$\lim_{m\to\infty} R_{\text{shared}}(m) = \sum_{j=1}^{J} h_j \Pr(P_j = 0).$$

Thus, shared propensities can slow recovery without producing a positive asymptotic residual. A persistent residual requires positive probability of zero coverage propensity for a component with positive entropy.

C.2 Properties of Semantic Effective Rank

For a nonzero $E \in \mathbb{R}^{n\times d}$, let $G = EE^\top$, $\bar{G} = G/\mathrm{Tr}(G)$ and $r = \mathrm{rank}(G)$. The operator $\widehat{K}(E)$ in Eq. 13 applies to both normalized output embeddings and averaged trajectory embeddings.

Proposition C.1 (Spectral properties)**.** *1.* $1 \le \widehat{K}(E) \le r \le \min(n, d)$.

2. $\widehat{K}(E) = 1$ *exactly when* E *has rank one.*

3. $\widehat{K}(E) = n$ *exactly when* $G = sI_n$ *for some* $s > 0$*: the rows are mutually orthogonal with equal nonzero norms.*

4. $\widehat{K}(E)$ *is continuous on nonzero matrices and invariant under row permutations and multiplication by a nonzero scalar.*

Proof. The normalized eigenvalues form a probability vector with r positive entries, whose entropy lies in $[0, \log_2 r]$. Zero entropy is equivalent to rank one. Entropy $\log_2 n$ is attained exactly by n equal eigenvalues, giving $G = sI_n$. The Gram matrix and trace normalization are continuous for $E \neq 0$, as is entropy with $0 \log 0 = 0$. Row permutations conjugate G by a permutation matrix; scalar multiplication cancels in trace normalization. Both preserve its normalized spectrum. □

For unit-norm rows $s = 1$ in the third statement. For averaged trajectory rows the equal-norm condition is also necessary: orthogonal rows of different lengths need not give rank n in the entropy sense. Each nonzero correctness-conditioned submatrix with n_a rows likewise has $1 \le \widehat{K}(E_a) \le \min(n_a, d)$. Empty answer groups are excluded from their respective means. The numerical implementation returns 1 when the Gram trace is at most 10^{-12}; the stated spectral results concern nonzero matrices.

D Supplementary Experiments

D.1 Closed-Source Model Experiments

We extend the Formal Logic evaluation to two closed-source models, gpt-4.1-mini and gpt-5-mini. Table 4 reports shared-prompt (Base) and persona-diverse (Persona) configurations at $N = 2$ and $N = 16$, alongside the open-weight models. Each comparison fixes the backbone model; heterogeneity in this table concerns prompting rather than model mixing.

Key findings. The results extend the evidence for persona-diversity gains to the two evaluated closed-source models on Formal Logic.

- **Persona-diversity gains depend on the setting.** Both closed-source models have positive Δ_{Persona} under Vote and Debate. Across all five models, the average configuration gain is positive under both workflows.

Table 4: Formal Logic results for closed-source and open-weight models. Vote results use the initial round (Round 0), whereas Debate results use the final round after three interaction rounds (Round 3). Accuracy is reported in percent and differences in percentage points. Δ_{Persona} is the average Persona-minus-Base difference over $N \in \{2, 16\}$. Δ_N is the Persona accuracy at $N = 16$ minus that at $N = 2$. Differences are computed from the displayed accuracies.

Model	Method	Base		Persona		Δ_{Persona}	Δ_N
		$N = 2$	$N = 16$	$N = 2$	$N = 16$		
Closed-source							
gpt-4.1-mini	vote	55.56	62.70	61.90	69.05	+6.35	+7.15
	debate	46.03	39.68	50.79	42.86	+3.97	−7.93
gpt-5-mini	vote	1.59	5.56	41.27	47.62	**+40.87**	+6.35
	debate	0.79	6.35	34.92	54.76	**+41.27**	+19.84
Open-weight							
Qwen-2.5-7B	vote	39.68	42.86	46.83	51.59	+7.94	+4.76
	debate	30.95	35.71	25.40	42.06	+0.40	+16.66
Llama-3.1-8B	vote	37.30	47.62	44.44	57.14	+8.33	+12.70
	debate	24.60	26.19	35.71	39.68	+12.30	+3.97
Mistral-7B	vote	34.92	34.92	34.92	37.30	+1.19	+2.38
	debate	34.92	39.68	39.68	46.03	+5.56	+6.35

- **Configuration gains and scaling gains are distinct.** For gpt-4.1-mini, persona-diverse configurations outperform Base at both evaluated agent counts under both workflows. Their accuracy increases from 61.90% to 69.05% under Vote but decreases from 50.79% to 42.86% under Debate as N increases from 2 to 16. Thus, a configuration advantage at a fixed agent count does not guarantee a gain from increasing that count.
- **The magnitude of configuration effects varies.** Under the evaluated settings, gpt-5-mini has Base accuracies of 0.79–6.35% and Persona accuracies of 34.92–54.76%. Its average configuration gains are 40.87 percentage points for Vote and 41.27 for Debate. These results describe sensitivity to the evaluated configurations and do not establish a general relationship between intrinsic model capability and diversity gains.

D.2 Comparison Across Embedding Models

We compare semantic-diversity results from NV-Embed-v2 (4096 dimensions) and gte-Qwen2-1.5B-instruct (1536 dimensions) across the evaluated tasks, configurations, workflows and agent counts. Ranking the $(\text{workflow}, N)$ settings within each configuration–dataset group gives a mean Spearman correlation of $\rho = 0.91$; over 95% of groups have $\rho > 0.5$.

The reported accuracy associations are positive for both pipelines: $r = 0.40$ for NV-Embed-v2 and $r = 0.23$ for gte-Qwen2. This comparison describes agreement between the two embedding pipelines, including their preprocessing and aggregation, rather than isolating embedding-model choice alone.

The NV correlation uses 260 observations with available total effective rank. Table 6 uses 220 observations that also contain both correctness-conditioned ranks, giving $r = 0.383$. All regressions in Table 5 use that same complete-case set.

D.3 Is $\widehat{K}$ More Than a Proxy for Scale and Configuration?

Figure 7 fit. For L4, we average both axes over six tasks at each fixed $(N, \text{workflow})$, producing 10 mean points from 60 observations. We fit $y = 0.4 + c[1 - \exp(-a(x - 1))]$ by nonlinear least squares, with $0 \le a \le 20$ and $0 \le c \le 1$. Its $R^2 = 0.838$ is evaluated on these means.

Incremental regression. Since $\widehat{K}$ is computed from agent outputs whose diversity naturally varies with agent count N and configuration type, a key question is whether $\widehat{K}$ captures information *beyond* these design variables, or merely serves as a redundant proxy for them. We fit a baseline regression of terminal accuracy on agent count N and configuration labels, then evaluate changes in model fit after adding $\widehat{K}$ or correctness-conditioned statistics. All ΔR^2 values in Table 5 use the same baseline and the same 220 complete observations: 40 each for ARC, HellaSwag, Pro Medicine and WinoGrande, and 30 each for Formal Logic and TruthfulQA. GSM8K lacks the conditioned ranks used here. Each observation is a task–configuration–trajectory-type–N aggregate. The baseline has an intercept, linear N and configuration indicators; it does not control for task or trajectory type.

Table 5: **Incremental Regression Fit of Semantic Diversity Statistics.** All ΔR^2 values are relative to the baseline using N and configuration labels. Adding $\widehat{K}$ increases terminal-accuracy R^2 from 0.062 to 0.205. Correctness-conditioned statistics use answer labels and are included for retrospective analysis.

Model	R^2	**Adj.** R^2	ΔR^2	**AIC**
Baseline (N + Config)	0.062	0.045	–	1789.1
Baseline + $\widehat{K}$	0.205	0.187	+0.143	1754.7
Baseline + $\widehat{K}_c$	0.393	0.379	+0.331	1695.3
Baseline + $\widehat{K}_c + \widehat{K}_w$	0.399	0.382	+0.337	1695.2
Baseline + $\widehat{K}_c / \widehat{K}_w$	0.336	0.321	+0.274	1715.1

Table 5 shows that $\widehat{K}$ provides additional statistical information about accuracy beyond the fitted baseline. Adding $\widehat{K}$ increases R^2 from 0.062 to 0.205, an improvement of 0.143.

Using $\widehat{K}_c$ instead yields $R^2 = 0.393$. Adding $\widehat{K}_w$ to that model increases R^2 to 0.399, an additional improvement of 0.006. Within these regressions, the semantic diversity of responses with correct final answers has a stronger association with accuracy than total output diversity.

This analysis concerns model fit on the evaluated observations; it does not establish out-of-sample predictive performance or a causal effect of diversity. In particular, $\widehat{K}_c$ uses correctness labels and is bounded by the number of correct responses n_c. The reported regressions do not separate its association with accuracy from the contribution of n_c. Moreover, final-answer correctness does not certify the validity of every intermediate reasoning step.

Sensitivity to the regression specification. Controlling additionally for task and workflow reduces the incremental R^2 from adding $\widehat{K}$ to 0.0062 on the same 220 observations, and treating agent count categorically reduces it to 0.0014. Across all 260 observations with available total effective rank, the corresponding increments are 0.0022 and approximately zero. The incremental association therefore depends on the baseline specification and observation set.

D.4 Sanity Checks: Are $\widehat{K}$–Performance Relations Accidental?

We use permutation-based randomization tests to assess the observed correlations between trajectory diversity statistics and terminal accuracy. The observed correlations are compared with null distributions obtained from 1000 unrestricted permutations of each diversity statistic across the same 220 complete observations used in Table 5 (seed 42). For a two-sided test we use

$$p = \frac{1 + \#\{|r_{\text{perm}}| \geq |r_{\text{obs}}|\}}{1001}.$$

The z-score standardizes observed r by the permutation mean and standard deviation. Permutations are pooled across tasks and configurations.

As shown in Table 6, the observed correlations have permutation z-scores above 5 and reported p-values below 10^{-3}. They are therefore unusual under the implemented permutation null. Among the evaluated statistics, $\widehat{K}_c$ has the largest observed correlation. These results support an association with accuracy under the tested null, but do not isolate a causal mechanism or rule out confounding by dataset, workflow, or correct-response count.

Table 6: **Permutation Sanity Check (1000 shuffles).** Observed correlations between semantic diversity statistics and terminal accuracy are compared with their permutation null distributions. The test assesses association under the specified permutation scheme.

Metric	**Observed** r	z**-score**	p
$\widehat{K}$	0.383	5.79	<0.001
$\widehat{K}_c$	0.536	7.77	<0.001
$\widehat{K}_c/\widehat{K}_w$	0.507	7.28	<0.001

Table 7: Model Ablation on Formal Logic: Impact of Heterogeneity from $N = 2$ to $N = 16$. Default and Persona denote shared-prompt and persona-diverse settings, respectively. MIX combines backbone models in both settings. Δ_{P} denotes Persona minus Default in percentage points.

Base Model	**Agents** (N)	**Vote (Round 0)**			**Debate (Final)**		
		Default	**Persona**	Δ_P	**Default**	**Persona**	Δ_P
Qwen-2.5-7B	2	39.68	46.83	+7.15	30.95	25.40	-5.55
	4	42.06	53.97	+11.91	30.16	34.92	+4.76
	8	43.65	50.00	+6.35	27.78	35.71	+7.93
	12	43.65	54.76	+11.11	34.13	38.10	+3.97
	16	42.86	51.59	+8.73	35.71	42.06	+6.35
Llama-3.1-8B	2	37.30	44.44	+7.14	24.60	35.71	+11.11
	4	42.86	53.97	+11.11	23.02	24.60	+1.58
	8	41.27	52.38	+11.11	29.37	40.48	+11.11
	12	44.44	57.14	+12.70	31.75	39.68	+7.93
	16	47.62	57.14	+9.52	26.19	39.68	+13.49
Mistral-7B	2	34.92	34.92	0.00	34.92	39.68	+4.76
	4	34.92	36.51	+1.59	35.71	44.44	+8.73
	8	32.54	37.30	+4.76	40.48	44.44	+3.96
	12	33.33	35.71	+2.38	38.89	44.44	+5.55
	16	34.92	37.30	+2.38	39.68	46.03	+6.35
MIX	2	42.06	44.44	+2.38	34.13	38.89	+4.76
	4	46.83	52.38	+5.55	42.86	53.17	+10.31
	8	53.97	54.76	+0.79	49.21	53.17	+3.96
	12	42.06	51.59	+9.53	48.41	54.76	+6.35
	16	46.83	55.56	+8.73	43.65	51.59	+7.94

D.5 Model Comparisons on Formal Logic and ARC

The mean single-backbone baseline summarizes the three-model pool. The best comparator asks whether MIX exceeds its strongest member at the same task, workflow, and agent count. On ARC and Formal Logic, this comparator uses the maximum test accuracy rather than validation-based model selection.

Table 7 reports a comprehensive ablation study on the Formal Logic benchmark, varying base models, agent counts ($N = 2$–16), and interaction mechanisms. Persona-diverse configurations often improve accuracy at a fixed agent count, with effects that vary across backbones and workflows. For MIX, persona diversity improves accuracy at every evaluated agent count under both Vote and Debate. Single-backbone configurations show more variable effects, including negative differences in some settings.

Scaling trends also depend on configuration. For example, MIX with persona diversity reaches its highest reported accuracy under Vote at $N = 16$ (55.56%). Under Debate, it peaks at $N = 12$ (54.76%) and decreases to 51.59% at $N = 16$. The table therefore supports comparing agent composition and agent count jointly, rather than assuming that either change produces uniform gains.

Table 8: Formal Logic under Debate: MIX compared with the best persona-diverse single-backbone pool at each agent count. Accuracy is reported in percent and differences in percentage points.

Agents (N)	**Best Single (Persona)**	**MIX (Persona)**	$\Delta_{\text{MIX vs. Best}}$	**MIX (Default)**	Δ_{P}
2	39.68	38.89	-0.79	34.13	+4.76
4	44.44	53.17	**+8.73**	42.86	+10.31
8	44.44	53.17	**+8.73**	49.21	+3.96
12	44.44	54.76	**+10.32**	48.41	+6.35
16	46.03	51.59	**+5.56**	43.65	+7.94

Table 7 also permits a Vote comparison against the best persona-diverse single-backbone pool. MIX exceeds this comparator only at $N = 8$ (54.76% versus 52.38%); it is lower at $N = 2, 4, 12, 16$. At $N = 16$, for example, MIX achieves 55.56% versus 57.14% for Llama-3.1-8B.

Under Debate, Table 8 compares persona-diverse MIX with the best persona-diverse single-backbone pool at each evaluated agent count. MIX achieves higher accuracy at $N \in \{4, 8, 12, 16\}$, with the largest reported difference of 10.32 percentage points at $N = 12$. At $N = 2$, MIX is 0.79 percentage points below this comparator.

The contrast between one of five Vote settings and four of five Debate settings shows that the advantage over the best single-backbone pool depends on workflow in this case study. The comparison does not isolate the mechanism of improvement or distinguish evidence coverage from classical ensemble explanations.

Best-backbone comparison on ARC. Using the same three persona-diverse single-backbone systems, the L4-minus-best differences at $N = 2, 4, 8, 12, 16$ are $(-6.35, -0.80, -1.59, +0.79, 0.00)$ percentage points for Vote and $(0.00, -0.79, -0.80, 0.00, -1.59)$ for Debate. Vote therefore has one win, one tie, and three losses; Debate has two ties and three losses. The Vote win occurs at $N = 12$, with 89.68% for MIX versus 88.89% for the best single-backbone pool. These results show that a lead over the mean single-backbone baseline need not extend to the best member of that pool.

D.6 Interactions and Task-Level Variation

For each task, workflow and agent count, let A_ℓ be the accuracy of configuration Lℓ in percentage points. The changes in accuracy with persona prompts are $A_2 - A_1$ for single-backbone systems and $A_4 - A_3$ for mixed-model systems. We summarize their difference as

$$I = (A_4 - A_3) - (A_2 - A_1).$$

This is a descriptive interaction contrast between the evaluated configurations, including their backbone and prompt choices. All contrasts use unrounded accuracies and equal task weights.

Frequency and scale. Across the 35 task–count settings per workflow, positive/zero/negative persona gains occur in 20/0/15 single-backbone and 33/1/1 mixed-model settings for Vote; the corresponding Debate counts are 14/0/21 and 30/1/4. The contrast I is positive in 26/35 Vote settings and 30/35 Debate settings. Its seven-task mean is positive at all five evaluated counts: for Vote it is 0.13, 3.37, 1.81, 3.41 and 4.16 percentage points at $N = 2, 4, 8, 12, 16$; for Debate it is 7.70, 9.05, 5.04, 6.15 and 5.81. These frequencies summarize task–count settings, not independent experimental replicates.

Interaction and final accuracy can diverge. Table 9 shows task-dependent effects. TruthfulQA has negative mean I in both workflows, despite positive mean L4–L1 accuracy differences. Conversely, WinoGrande under Debate has positive I at every count while L4 has lower accuracy than L1 at every count. At $N = 16$, its single-backbone persona gain is -12.16 percentage points and its mixed-model gain is $+6.35$, giving $I = 18.51$, but L4 still trails L1 by 8.99 points. Positive I thus describes a difference in persona gains, not a guarantee that the combined configuration is best. Across all tasks and counts, L4–L1 differences are positive/zero/negative in 34/0/1 Vote settings and 28/0/7 Debate settings. These patterns motivate evaluating the complete configuration in the intended task and workflow.

Table 9: Task-level interaction and accuracy differences, averaged over $N \in \{2, 4, 8, 12, 16\}$. I compares persona gains across model compositions; n_+ counts scales with $I > 0$ out of five. Differences are in percentage points.

	Vote			Debate		
Task	Mean I	Mean $(A_4 - A_1)$	n_+	Mean I	Mean $(A_4 - A_1)$	n_+
GSM8K	+4.88	+12.44	5	+10.16	+21.96	5
ARC	+4.81	+1.85	5	+3.23	+4.23	5
Formal Logic	-1.80	+12.01	2	+0.53	+18.09	3
HellaSwag	+9.00	+3.15	5	+14.04	+4.25	5
Pro Medicine	+0.98	+5.72	4	+3.65	+4.29	5
TruthfulQA	-9.23	+16.18	0	-0.86	+14.57	2
WinoGrande	+9.38	+1.91	5	+16.51	-7.35	5

AI USE STATEMENT

We used generative AI tools to improve the clarity and fluency of the manuscript's language. The authors take full responsibility for the final content.