UNDERSTANDING AGENT SCALING IN LLM-BASED MULTI-AGENT SYSTEMS VIA DIVERSITY

Yingxuan Yang[1], **Chengrui Qu**[2], **Muning Wen**[1], **Laixi Shi**[3]
Ying Wen[1], **Weinan Zhang**[1], **Adam Wierman**[2], **Shangding Gu**[4*]
[1]Shanghai Jiao Tong University [2]California Institute of Technology
[3]Johns Hopkins University [4]UC Berkeley

ABSTRACT

LLM-based multi-agent systems (MAS) have emerged as a promising approach to tackle complex tasks that are challenging for a single LLM. A natural way to improve performance is to increase the number of agents. However, prior empirical findings suggest that in homogeneous settings, such scaling yields diminishing returns, whereas introducing heterogeneity (e.g., different models, prompts, or tools) can lead to more substantial gains. This motivates a fundamental question: *what limits multi-agent scaling, and why does diversity help?* In this paper, we develop an information-theoretic framework suggesting that MAS performance is fundamentally constrained by the intrinsic task uncertainty, rather than increasing monotonically with the number of agents. We derive architecture-agnostic bounds showing that performance improvements depend on the number of *effective channels* through which the system acquires non-redundant information. Under this view, homogeneous agents saturate early because their outputs are strongly correlated, while heterogeneous agents can provide more complementary evidence. To make this perspective operational, we provide a label-free proxy $\widehat{K}$ that estimates the number of *effective channels* from semantic similarity patterns in agent outputs. Empirically, we find that heterogeneous configurations consistently outperform homogeneous scaling, and in our experiments, 2 diverse agents can match or exceed the performance of 16 homogeneous agents. Overall, our results offer a principled perspective on the limits of multi-agent scaling and suggest diversity-aware design as a promising direction for building more efficient MAS. Code and dataset are available at the link: `https://github.com/SafeRL-Lab/Agent-Scaling`.

1 INTRODUCTION

Large language models (LLMs) have achieved remarkable performance across diverse tasks, including reasoning, coding, and open-domain question answering (Wei et al., 2022; Achiam et al., 2023). However, individual LLMs still struggle with complex problems that require multi-step reasoning, diverse perspectives, or complementary expertise (Huang et al., 2023). To address these limitations, LLM-based multi-agent systems (MAS) have emerged as a promising paradigm. By orchestrating multiple LLM agents through communication, coordination, or aggregation mechanisms, MAS can tackle challenges that are difficult for single models (Wu et al., 2024; Hong et al., 2024; Du et al., 2023). Recent studies have demonstrated that multi-agent collaboration can yield substantial improvements 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 multi-agent systems, a natural question arises: *can we improve MAS performance simply by scaling the number of agents?* Intuitively, one might expect ensemble-style gains from aggregating more agent outputs (Li et al., 2024a; Wang et al., 2023). However, recent work (Kim et al., 2025) and our experiments reveal a more nuanced picture. As shown in Figure 2, scaling homogeneous agents (identical models, prompts, and configurations) exhibits strong diminishing returns: accuracy improves at small agent counts, but the marginal gain per additional agent

[*] Corresponding author: `shangding.gu@berkeley.edu`

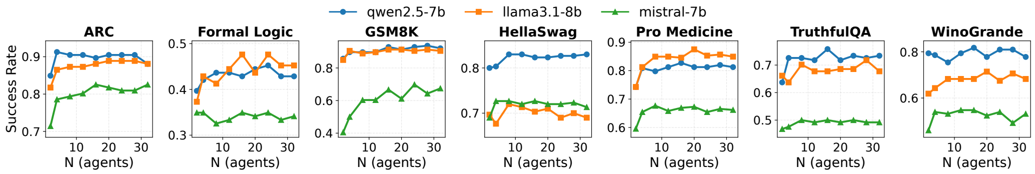


Figure 2: Scaling behavior of homogeneous multi-agent voting. Success rate versus agent count N on seven tasks for three base models. Performance improves with N but saturates, indicating clear diminishing marginal gains at larger agent counts.

rapidly collapses toward zero. This suggests that simply adding more homogeneous agents (or allocating more test-time compute) does not reliably introduce new *usable evidence* into the system, but may instead produce increasingly redundant trajectories.

In contrast, our experiments (Figure 1) show that introducing *diversity* yields sustained performance improvements. Here, diversity broadly refers to heterogeneity in agent configurations, such as backbone models, prompts or personas, and tool access, which empirically leads to more complementary, rather than redundant, information being introduced into the system (Wang et al., 2024a; Zhang et al., 2024; Qian et al., 2025). We illustrate a comprehensive comparison between homogeneous and heterogeneous systems in Figure 3. Motivated by these observations, we ask: **what fundamentally limits MAS scaling, and why does diversity help?**

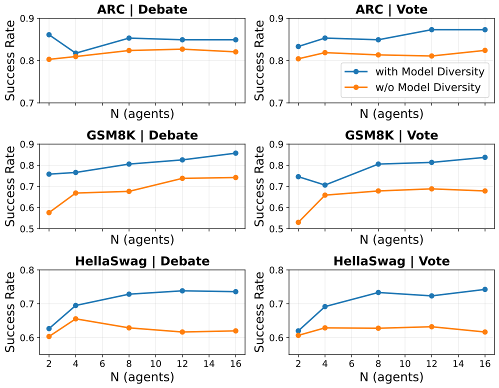


Figure 1: Effect of model diversity. We compare a mixture of three LLMs (Qwen-2.5-7B, Llama-3.1-8B, Mistral-7B) with the average of independent single-LLM runs.

Prior work has documented diminishing returns from scaling homogeneous agents (Wang et al., 2024b; Kim et al., 2025; Chen et al., 2024; Choi et al., 2025) and shown that diversity can mitigate this effect in specific settings (Wang et al., 2024a; Zhang et al., 2024; Wu & Ito, 2025; Yuen et al., 2025; Tang et al., 2025). These results are largely empirical and setting-specific: they demonstrate *that* diversity helps, but do not characterize *why* it helps or *when* its benefits plateau. In this paper, we develop an information-theoretic framework that shifts the analysis from the raw agent count n to the *effective channel count* K, a quantity capturing the number of independently informative reasoning paths present in agent outputs.

Our contributions are summarized as follows:

- We introduce an information-theoretic framework that quantifies the effectiveness of a MAS by asking: how much task-relevant information can its agent outputs collectively reveal? We show that MAS information gain is upper-bounded by the intrinsic task uncertainty $H(Y \mid X)$ and governed by the product αK of the complementarity rate α and the effective channel count K, both constrained by agent configuration types rather than raw agent count. The resulting lower bound on information gain exhibits the $1-(1-\alpha)^K$ shape, providing a formal account consistent with the observed fast-then-slow scaling pattern and the advantage of heterogeneous designs. On the practical side, we introduce $\widehat{K}$, a label-free proxy for the effective channel count, and show that gains in $\widehat{K}$ correlate positively with task performance, providing empirical support that diversity drives improvement.
- We conduct experiments across seven benchmarks and two representative MAS workflows (vote, debate) to study the effectiveness of diversity in MAS performance. Homogeneous MAS with an increasing number of agents exhibit strong diminishing returns, while heterogeneous configurations consistently outperform homogeneous ones under identical compute cost. With only **2 diverse agents**, the heterogeneous configuration matches or exceeds the performance of **16 homogeneous agents**.

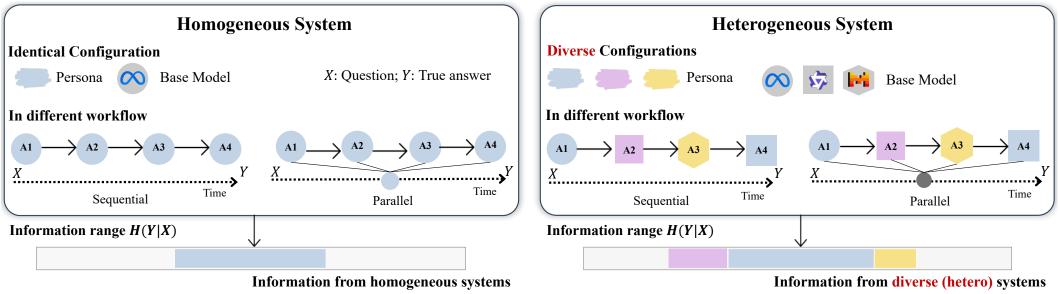


Figure 3: A comparison between homogeneous and heterogeneous systems. In a homogeneous system, agents with identical configurations result in redundant behavior and limited information coverage. In contrast, heterogeneous agents, through diverse configurations (e.g., varying models or personas), provide complementary coverage and better diversity in the information processed, allowing for more effective problem-solving across different workflows.

2 Problem Formulation

This section formalizes the setup of LLM-based multi-agent systems and introduces the key information-theoretic quantities used throughout the paper. We first define the system and notation, then introduce the central quantity that governs MAS performance: *usable evidence*.

2.1 LLM-based Multi-Agent Systems

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

Definition 2.1 (LLM-based Multi-Agent System)**.** We consider an *LLM-based multi-agent system* that consists of N agents, each characterized by a *configuration* that specifies its backbone model, system prompt or persona, and tool access. Given a task input $X \in \mathcal{X}$, the system executes a total of n calls (i.e., queries to LLM agents) through a specified workflow (e.g., parallel voting, sequential debate) and aggregates the outputs to produce a final answer as an estimate of the ground-truth answer $Y \in \mathcal{Y}$.

Notation. We distinguish between the *number of agents* N and the *number of agent calls* n. In single-round workflows such as majority voting, $n = N$. In multi-round workflows such as debate with R rounds, $n = N \times R$. This distinction is important because we focus on how much information is extracted, regardless of which agent produces it.

Agent Configuration Types Agent diversity is operationalized through variations in backbone model, system prompt/persona, and tool access. We index these choices by *configuration types*.

Definition 2.2 (Agent Configuration Type)**.** Consider a finite set of types $\mathcal{B}$. Each call $i \in \{1, \dots, n\}$ is associated with a type $b(i) \in \mathcal{B}$. For each $b \in \mathcal{B}$, define the number of calls

$$m_b := \big|\{i \in \{1, \dots, n\} : b(i) = b\}\big|, \qquad \sum_{b \in \mathcal{B}} m_b = n. \tag{1}$$

The multiset $\{(b, m_b)\}_{b \in \mathcal{B}}$ summarizes the agent configuration of the system.

2.2 Quantifying Information Gain

To begin, we introduce the task's *intrinsic uncertainty*, i.e., the remaining uncertainty about the ground-truth answer Y after observing the input X alone, defined as the conditional entropy:

$$H(Y \mid X). \tag{2}$$

During inference, the MAS executes n agent calls and produces a dialogue transcript:

$$Z_{1:n} := (Z_1, \dots, Z_n), \tag{3}$$

where each output Z_i may depend on the input X and all preceding outputs $Z_{<i} := (Z_1, \dots, Z_{i-1})$.

The central question to answer is: *how much information about the answer Y can the MAS include from its agent calls?* We quantify this through the conditional mutual information:

$$I_{\mathrm{MAS}}(n) \;:=\; I(Z_{1:n}; Y \mid X) \;=\; H(Y \mid X) - H(Y \mid X, Z_{1:n}). \tag{4}$$

We refer to this quantity as *usable evidence*. It measures the reduction in uncertainty about Y achieved by observing the transcript beyond what is already contained in X.

3 Why Diversity Matters

We now present our theoretical results explaining the role of diversity in MAS scaling. We first derive a fundamental limit on information gain and show that it depends on agent configuration types rather than raw agent count, then introduce the *effective channel product* αK as the key quantity governing MAS performance.

3.1 Information Limits and the Role of Diversity

By the chain rule for mutual information, $I_{\mathrm{MAS}}(n)$ decomposes into incremental contributions $\Delta_i := I(Z_i; Y \mid X, Z_{<i})$, where each Δ_i measures the new information provided by the i-th call given all previous outputs (see Appendix C.2 for the formal statement).

This means that MAS performance depends not merely on the total number of calls n, but critically on how much *new* evidence each call contributes. When agents produce highly correlated outputs, the incremental contributions Δ_i diminish rapidly, causing the total gain to saturate well below the information budget. As illustrated in Figure 3, heterogeneous agents provide complementary coverage and better diversity in information processing compared to homogeneous configurations.

We begin with a basic but important observation: the total information any MAS can extract about Y is bounded by the intrinsic task uncertainty. Because conditional entropy is nonnegative, for any transcript $Z_{1:n}$,

$$I_{\mathrm{MAS}}(n) \;=\; H(Y \mid X) - H(Y \mid X, Z_{1:n}) \;\leq\; H(Y \mid X). \tag{5}$$

While elementary, this ceiling has a concrete design implication: scaling benefits must eventually plateau once this budget is approached. The key question is *how quickly* different MAS designs approach this limit. As show in Section 3.2, homogeneous systems reach saturation much earlier than heterogeneous ones because their redundant outputs fail to increase the effective evidence coverage.

Beyond this universal ceiling, workflow-specific bounds further clarify the role of agent configuration types. For both parallel voting and sequential debate, the achievable information gain is upper-bounded by quantities that depend on the multiset of configuration types $\{(b, m_b)\}_{b\in\mathcal{B}}$, rather than the raw call count n (see Appendix C.3–C.4 for formal statements and proofs).

Since these upper bounds depend on the composition of configuration types rather than the raw call count n alone, the raw call count is not the right quantity for characterizing MAS performance limits. This motivates us to identify a new quantity, the *effective channel count*, that more directly governs how much usable evidence a MAS can extract.

3.2 Effective Channels: From Compute to Usable Evidence

To mathematically quantify the usable evidence in MAS, we introduce two modeling quantities: the *effective channel count* K and the *complementarity rate* α. Both are defined within an idealized *evidence-coverage model* that we now summarize; full details are deferred to Appendix C.5.

Evidence-coverage model. We model the MAS transcript $Z_{1:n}$ as providing K *effective channels* $\tilde{Z}^{(1)}, \ldots, \tilde{Z}^{(K)}$, each independently resolving part of the uncertainty about the answer Y given X. The formal construction and the assumptions under which the geometric bound is derived are deferred to Appendix C.5; here we state the two key quantities at the entropy level.

Definition 3.1 (Effective Channel Count)**.** An *effective channel decomposition* of the transcript $Z_{1:n}$ of size K is a collection of channels $\tilde{Z}^{(1)}, \ldots, \tilde{Z}^{(K)}$, each derived from $(X, Z_{1:n})$, such that the channels provide *independently informative* evidence about Y given X (formalized via cross-channel conditional independence in Assumption C.11).

Definition 3.2 (Complementarity Rate)**.** Given a valid K-channel decomposition $\tilde{Z}^{(1)}, \dots, \tilde{Z}^{(K)}$ (Definition 3.1), the *complementarity rate* $\alpha \in (0, 1)$ is the constant such that each successive channel reduces the residual uncertainty by the factor $(1 - \alpha)$:

$$\frac{H(Y \mid X,\, \tilde{Z}_{1:k})}{H(Y \mid X,\, \tilde{Z}_{1:k-1})} \;=\; 1 - \alpha, \qquad k = 1, \dots, K. \tag{6}$$

Remark 3.3 (Joint Modeling Parameters (K, α)). For a fixed system, K and α are *jointly determined* parameters of the evidence-coverage representation, not independently defined quantities. A coarser decomposition (smaller K) concentrates more evidence per channel (larger α), while a finer decomposition (larger K) distributes evidence more thinly (smaller α). The product αK captures the total effective coverage and is the fundamental composite quantity governing the information-recovery bound (Theorem 3.4); the bound is tightest when optimized over all valid decompositions.

Structurally, $K \leq n$ and is informally constrained by the number of distinct configuration types $|\mathcal{B}|$: within each type, additional same-type calls yield diminishing marginal coverage due to output correlation (Lemma C.17), so heterogeneous systems generally admit decompositions with a larger K (or higher α, or both) than homogeneous ones under the same call budget.

The product αK governs the rate of information recovery: since $(1 - \alpha)^K \leq e^{-\alpha K}$, larger αK yields tighter geometric contraction of the residual uncertainty:

Theorem 3.4 (Geometric Contraction with Effective Channels)**.** *Under Assumptions C.10 and C.11, the residual uncertainty after observing K effective channels satisfies*

$$\boxed{\begin{aligned} &H(Y \mid X) - I(\tilde{Z}_{1:K}; Y \mid X) \\ &\quad \leq\; (1-\alpha)^K\, H(Y \mid X) \;\leq\; e^{-\alpha K}\, H(Y \mid X). \end{aligned}} \tag{7}$$

Equivalently, the normalized residual *satisfies* $H(Y \mid X, \tilde{Z}_{1:K})/H(Y \mid X) \leq (1 - \alpha)^K \leq e^{-\alpha K}$.

Remark 3.5 (Uniformity over Inputs). Because the coverage model (Assumption C.11) posits a uniform per-channel coverage rate for every input instance, the geometric contraction holds for $H(Y \mid X, \tilde{Z}_{1:K})$ directly, without an outer expectation. The proof (Appendix C.5.3) decomposes the residual uncertainty instance by instance, bounds each term using the per-instance non-coverage rate $(1 - \alpha)^K$, and re-averages. If the coverage rate were allowed to vary with input difficulty, the bound would require an additional expectation (see Appendix C.5.2 for discussion).

3.3 K AS THE STATE VARIABLE OF MAS SCALING

The central question in MAS scaling is not whether n increases, but whether additional agents contribute independently informative channels, i.e., whether the system admits a valid evidence-coverage representation with a larger effective channel product αK.

This follows from Section 3.1: ceilings are fixed by intrinsic uncertainty $H(Y \mid X)$ and the multiset of configuration types (Appendix C.3–C.4), while achievability improves with the product αK (Section 3.2).

Why heterogeneous channels outperform homogeneous ones. Consider m outputs from agents sharing the same configuration type b. These outputs are conditionally independent given both the input and the answer (X, Y) (Assumption C.4), but when Y is unobserved they become *positively correlated*: if the answer makes a piece of evidence easy to find, all same-type agents are likely to find it; if the answer makes it hard, all are likely to miss it. By the law of total covariance, this answer-induced correlation inflates the probability that all agents simultaneously fail to resolve the same component, yielding a strictly weaker residual than the independent effective channels $\tilde{Z}^{(1)}, \dots, \tilde{Z}^{(K)}$ of Definition 3.1 (see Lemma C.17 in Appendix C.5.5).

Proposition 3.6 (Complementarity Advantage)**.** *Let $Z^{(b)}_{1:m} := (Z^{(b)}_1, \dots, Z^{(b)}_m)$ denote m outputs from agents of the same configuration type b. Under Assumptions C.10, C.4, and C.11, the normalized residual uncertainty satisfies*

$$\frac{H(Y \mid X,\, Z^{(b)}_{1:m})}{H(Y \mid X)} \;>\; (1-\alpha)^m \tag{8}$$

for any non-trivial task (i.e., one where the channel's resolution probability depends on the ground-truth answer Y for at least one evidence component). In contrast, if m channels from m distinct configuration types form a valid effective channel decomposition of size $K = m$ (Definition 3.1), Theorem 3.4 yields the tight residual $(1-\alpha)^m$. Thus, at the same marginal coverage rate α, same-type calls are provably less efficient than cross-type effective channels.

The proof is given in Appendix C.5.5. This gap translates directly into a full-system comparison:

A direct heterogeneous–homogeneous advantage bound. Consider two designs under the same compute budget n, with effective channel parameters $(K_{\text{heterog}}, \alpha_{\text{heterog}})$ and $(K_{\text{homog}}, \alpha_{\text{homog}})$ in the evidence-coverage model.

Corollary 3.7 (Heterogeneity Advantage)**.** *Under Assumptions C.10, C.4, and C.11, Theorem 3.4 provides each design with a guaranteed information recovery level:*

$$I_{\text{heterog}} \;\geq\; H(Y \mid X)\big(1 - e^{-\alpha_{\text{heterog}} K_{\text{heterog}}}\big), \tag{9}$$

$$I_{\text{homog}} \;\geq\; H(Y \mid X)\big(1 - e^{-\alpha_{\text{homog}} K_{\text{homog}}}\big). \tag{10}$$

Since $1 - e^{-t}$ is strictly increasing, $\alpha_{\text{heterog}} K_{\text{heterog}} > \alpha_{\text{homog}} K_{\text{homog}}$ implies that the heterogeneous guarantee strictly exceeds the homogeneous one.

Justification of the premise. The condition $\alpha_{\text{heterog}} K_{\text{heterog}} > \alpha_{\text{homog}} K_{\text{homog}}$ is not merely assumed but follows from Proposition 3.6 and Lemma C.17: same-type channels are positively correlated through the unobserved answer Y, so m same-type calls cannot form an effective decomposition of size m with the same per-channel rate α; in the coverage model, this forces the homogeneous system into a representation with a smaller αK product (fewer independently informative channels, or lower per-channel coverage, or both). Our empirical findings are directionally consistent: as shown in Figure 1 and Table 1, heterogeneous configurations consistently outperform homogeneous ones under matched compute.

Fast-then-slow scaling: the $1 - e^{-\alpha K}$ shape. Rearranging Theorem 3.4 (see Corollary C.14 in Appendix C.5), the recoverable information grows at least as

$$I(\tilde{Z}_{1:K}; Y \mid X) \;\geq\; H(Y \mid X)\big(1 - e^{-\alpha K}\big). \tag{11}$$

The shape of equation 11 provides a formal account of diminishing returns: the marginal improvement of the lower bound from one additional effective channel satisfies

$$\big(1 - e^{-\alpha(K+1)}\big) - \big(1 - e^{-\alpha K}\big) \;=\; (1 - e^{-\alpha})\, e^{-\alpha K}, \tag{12}$$

which is largest at small K and decays exponentially thereafter. This yields a clean explanation for the empirically observed *fast-then-slow* improvement pattern as the number of agents n increases: early gains occur when the achievable αK is still growing with n, while later gains diminish once the system's effective coverage saturates.

3.4 Measuring Effective Channels Without Labels: $\widehat{K}$

The effective channel count K is a parameter of the evidence-coverage model and depends on the unknown ground-truth Y; it therefore cannot be computed directly at inference time. We introduce $\widehat{K}$, a *label-free proxy* that measures semantic diversity in agent outputs via embedding space: $\widehat{K}$ is large when outputs are diverse and approaches 1 when outputs are similar. We emphasize that $\widehat{K}$ is a *proxy* for, rather than a direct estimate of, the theoretical K: it captures semantic diversity in embedding space, which is a necessary but not sufficient condition for task-relevant information diversity. The empirical correlation between $\widehat{K}$ and accuracy (Section 4.4.2) provides indirect validation that semantic diversity tracks effective channels in practice.

Definition. Let $\mathrm{Emb}(\cdot)$ be an embedding model. Given outputs $\{Z_i\}_{i=1}^n$, define normalized embeddings

$$\hat{\mathbf{z}}_i \;:=\; \frac{\mathrm{Emb}(Z_i)}{\|\mathrm{Emb}(Z_i)\|_2} \in \mathbb{R}^d, \tag{13}$$

Table 1: Effect of persona diversity. Δ denotes improvement from heterogeneity. All agents share the same base model pool (Qwen-2.5-7B, Llama-3.1-8B, and Mistral-7B); only persona assignments differ between Homog and Heterog.

Dataset	**Single Agent**	N	**Vote**			**Debate**			**Dataset**	**Single Agent**	N	**Vote**			**Debate**		
			Homog	Heterog	Δ	Homog	Heterog	Δ				Homog	Heterog	Δ	Homog	Heterog	Δ
GSM8K	50.8	2	86.5	87.3	+0.8	76.2	75.4	-0.8	**ARC**	77.8	2	78.6	81.8	+3.2	84.9	87.3	+2.4
		4	84.9	88.1	+3.2	73.8	79.4	+5.6			4	79.4	85.7	+6.3	79.4	84.1	+4.8
		8	90.5	93.7	+3.2	75.4	85.7	+10.3			8	84.1	86.5	+2.4	84.9	85.7	+0.8
		12	86.5	90.5	+4.0	77.8	87.3	+9.5			12	85.7	89.7	+4.0	82.5	87.3	+4.8
		16	89.7	92.1	+2.4	83.3	88.1	+4.8			16	84.9	88.9	+4.0	84.9	84.9	0.0
Formal Logic	32.0	2	45.2	48.4	+3.2	34.1	38.9	+4.8	**Truthful QA**	71.8	2	74.2	77.4	+3.2	71.0	77.4	+6.4
		4	47.6	52.4	+4.8	42.9	53.2	+10.3			4	75.0	75.8	+0.8	71.8	79.8	+8.0
		8	47.6	55.6	+7.9	49.2	53.2	+4.0			8	76.6	79.0	+2.4	76.6	78.2	+1.6
		12	48.4	57.9	+9.5	48.4	54.8	+6.4			12	75.0	79.0	+4.0	73.4	79.8	+6.4
		16	50.0	54.0	+4.0	43.6	51.6	+8.0			16	78.2	81.5	+3.3	75.0	84.7	+9.7
HellaSwag	66.1	2	62.3	73.7	+11.4	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	75.3	+6.6	66.0	73.0	+7.0			4	54.0	69.1	+15.1	53.2	62.7	+9.5
		8	70.0	79.0	+9.0	69.7	76.0	+6.3			8	57.9	69.1	+11.2	61.9	69.1	+7.2
		12	72.3	79.0	+6.7	69.3	78.3	+9.0			12	58.7	70.6	+11.9	62.7	70.6	+7.9
		16	72.0	79.9	+7.9	70.3	76.4	+6.1			16	60.3	69.8	+9.5	57.9	64.3	+6.4
Pro Medicine	68.6	2	78.3	78.7	+0.4	76.8	71.3	-5.5	**Average**	60.6	2	68.1	72.5	**+4.4**	64.6	67.9	**+3.3**
		4	80.5	81.6	+1.1	76.8	76.5	-0.3			4	69.9	76.1	**+6.2**	66.3	72.7	**+6.4**
		8	81.3	83.5	+2.2	81.6	82.7	+1.1			8	72.6	79.6	**+7.0**	71.3	75.8	**+4.5**
		12	80.2	82.7	+2.5	81.3	83.8	+2.5			12	72.4	81.0	**+8.6**	70.8	77.4	**+6.6**
		16	80.5	81.8	+1.3	80.5	83.3	+2.8			16	73.6	81.1	**+7.5**	70.8	76.2	**+5.4**

and the cosine-similarity Gram matrix $G \in \mathbb{R}^{n \times n}$:

$$G_{ij} \;:=\; \langle \hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j \rangle. \tag{14}$$

Trace-normalize to obtain $\bar{G} := G/\mathrm{Tr}(G)$ with $\mathrm{Tr}(\bar{G}) = 1$, and let $\{\lambda_j\}_{j=1}^{n}$ be eigenvalues of $\bar{G}$.

We define the entropy effective rank

$$\widehat{K} := 2^{H(\bar{G})}, \quad \text{where} \quad H(\bar{G}) = -\sum_{j=1}^{n} \lambda_j \log_2 \lambda_j. \tag{15}$$

Interpretation. $\widehat{K}$ counts how many "independent directions" the agent outputs span in embedding space. When all agents produce nearly identical outputs (e.g., paraphrases of the same reasoning), their embeddings are collinear and $\widehat{K} \approx 1$: the system effectively has a single channel. When agents produce genuinely different outputs whose embeddings point in different directions with roughly equal magnitude, $\widehat{K}$ grows toward n: each agent contributes a distinct channel.

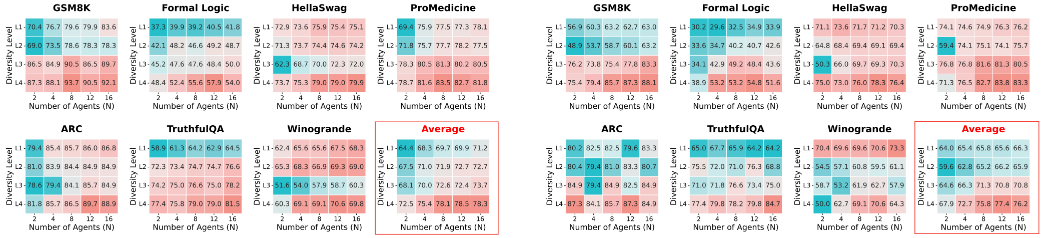
Remark 3.8 (Relationship between K, α, and $\widehat{K}$). The theoretical effective channel count K is defined through task-relevant evidence coverage (Definition 3.1), whereas $\widehat{K}$ measures semantic diversity in embedding space. These two quantities are related but not identical: high $\widehat{K}$ is a *necessary* condition for large K (semantically identical outputs cannot cover distinct evidence), but not a *sufficient* one (outputs may be semantically diverse yet irrelevant to the task). Our experiments (Section 4.4.2) provide indirect validation that $\widehat{K}$ tracks K in practice, via the strong positive correlation between $\widehat{K}$ and accuracy.

4 Experiments

This section validates three core claims: (i) scaling *homogeneous* MAS exhibits diminishing returns, (ii) *heterogeneity* consistently outperforms pure scaling under matched compute, and (iii) performance gains track the effective channel proxy $\widehat{K}$ (which reflects the theoretical product αK) rather than the raw agent count.

4.1 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 (ai2, 2019), and Pro Medicine (Hendrycks et al., 2021). These tasks span arithmetic reasoning, formal deduction, commonsense reasoning, and domain knowledge, covering both deterministic and ambiguous settings.



(a) Vote (b) Debate

Figure 4: Diversity analysis of the **Vote** and **Debate** mechanisms across all datasets. Each subfigure corresponds to one dataset and visualizes absolute performance as a 4×5 heatmap, where rows represent progressively enriched diversity layers (L1–L4) and columns denote the number of agents N. Colors indicate success rate values: cyan (lowest) to red (highest).

Models. Agents are instantiated using three open-source LLMs: Qwen-2.5-7B (Qwen Team, 2024), 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.

MAS Workflows. We consider two representative collaboration mechanisms (Choi et al., 2025): **Vote**, where agents independently generate answers and a majority decision is taken after 1 round, and **Debate**, where agents interact sequentially for 4 rounds before producing a final answer. For each mechanism, we vary the number of agents $N \in \{2, 4, 8, 12, 16\}$. Compute budgets are matched by fixing the total number of agent calls.

Diversity Configurations. We organize agent heterogeneity into four progressively enriched layers to isolate the contribution of each diversity source:

- **L1: No Diversity.** All agents share the same base model and the same default system prompt (no persona). This serves as the homogeneous baseline. Results are averaged over the three single-model runs.
- **L2: Persona Diversity Only.** All agents share the same base model, but each agent receives a distinct persona prompt (e.g., "You are an expert mathematician" vs. "You are a careful logician"). Results are averaged over the three single-model runs.
- **L3: Model Diversity Only.** Agents are drawn from different base models (Qwen, Llama, Mistral) but all use the same default system prompt.
- **L4: Full Diversity.** Agents differ in both base model and persona prompt, combining model-level and prompt-level heterogeneity.

This controlled design allows us to isolate and compare the contributions of model diversity and persona diversity.

4.2 Finding 1: Scaling Homogeneous MAS Exhibits Diminishing Returns

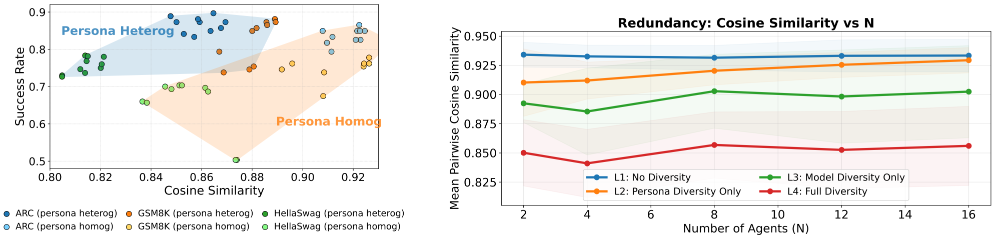
We first examine whether increasing the number of agents improves performance in homogeneous settings. Figure 2 shows success rates and marginal gains for both voting- and debate-based MAS across multiple tasks and base models.

Across all settings, we observe a consistent pattern: accuracy improves only at small counts, after which the marginal gain per additional agent rapidly collapses toward zero. In several cases, performance even degrades as N increases.

Table 2: Efficiency gains from diversity. Number of agents needed to match L1 (N=16). Higher diversity achieves equivalent performance with fewer agents.

Method	Config	Agents to Match L1 (N=16)	Accuracy at that N	Peak Accuracy (any N)
Vote	L1	16 (baseline)	65.34	65.49
	L2	8	65.44	66.01
	L3	4	67.29	71.54
	L4	**2**	67.71	**76.86**
Debate	L1	16 (baseline)	65.48	65.48
	L2	12	66.08	66.08
	L3	4	66.26	71.33
	L4	**2**	67.90	**77.43**

This saturation is consistent with our theoretical framework (Theorem 3.4): homogeneous agents produce highly correlated outputs, so additional calls are unlikely to increase the achievable product αK in the evidence-coverage model. In other words, allocating more test-time computation via homogeneous scaling does not reliably inject new usable evidence into the system.



(a) Correlation between cosine similarity and success rate.

(b) Mean pairwise cosine similarity vs. count. Higher diversity (L1→L4) consistently reduces redundancy.

Figure 5: Output redundancy analysis. (a) Higher cosine similarity correlates with lower success rate. (b) Redundancy increases with agent count N, but decreases from L1 to L4.

4.3 Finding 2: Diversity Consistently Beats Scale

We compare homogeneous scaling with heterogeneous designs under matched compute in Table 1, which reports the performance of Vote and Debate mechanisms across all tasks and agent counts. In most conducted cases, heterogeneous configurations significantly outperform homogeneous ones, with gains increasing as N grows. Figure 4 provides a detailed view of this effect. Enriching diversity from L1 to L4 yields consistent performance improvements for both Vote and Debate. Notably, model diversity (L3) and persona diversity (L2) each deliver non-trivial gains, while their combination (L4) consistently performs best.

Table 2 shows the minimum number of heterogeneous agents required to outperform homogeneous configurations. For both Vote and Debate, L4 (full diversity) with just **2 agents** surpasses the performance of L1 (no diversity) with **16 agents**. This represents an 8× reduction in agent count for equivalent or better accuracy. This observation is consistent with Corollary 3.7: a heterogeneous design can admit a higher αK product through more independently informative channels, so fewer agents may suffice to reach the same information-recovery level.

We also compare heterogeneous model mixtures against independent single-model runs. Figure 1 shows that a mixture of three LLMs outperforms the average performance of the individual models, suggesting that the improvements arise from complementary evidence rather than simple averaging.

4.4 Finding 3: Performance Gains Track the Effective Channel Proxy $\widehat{K}$

Our framework suggests that homogeneous agents produce highly correlated outputs, contributing few effective channels and leading to saturation. We now examine this empirically, proceeding from a simple redundancy proxy (pairwise cosine similarity) to the effective channel measure ($\widehat{K}$).

4.4.1 High output similarity hinders performance

A key reason homogeneous scaling saturates is that additional agent calls increasingly produce *correlated* outputs, yielding limited *new* evidence. To quantify this redundancy, we embed each agent output (the full reasoning trace) using NV-Embed-v2 (Lee et al., 2025) and compute the *mean pairwise cosine similarity*: for n agent outputs with normalized embeddings $\hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_n$, this is $\bar{\rho} = \frac{2}{n(n-1)} \sum_{i<j} \langle \hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j \rangle$. While $\bar{\rho}$ is not an information-theoretic quantity, it provides a consistent proxy for output overlap: higher $\bar{\rho}$ indicates that agents explore fewer independent directions, which constrains the growth of effective channels.

Figure 5a shows that homogeneous persona settings produce higher similarity yet do not translate this additional compute into higher success rates, whereas heterogeneous personas maintain lower similarity and achieve stronger performance. Moreover, Figure 5b reveals a systematic scaling trend: for every diversity layer, redundancy increases with agent count N, implying that larger homogeneous ensembles mainly amplify existing trajectories rather than introducing qualitatively new evidence. Crucially, redundancy decreases monotonically from L1 to L4, consistent with our hypothesis that heterogeneity mitigates output correlation and thus enlarges the number of effective channels. While these results confirm a qualitative relationship between output diversity and performance, pairwise cosine similarity is a coarse measure. To obtain a more precise and theoretically grounded characterization, we next turn to the effective channel count $\widehat{K}$ introduced in Section 3.4.

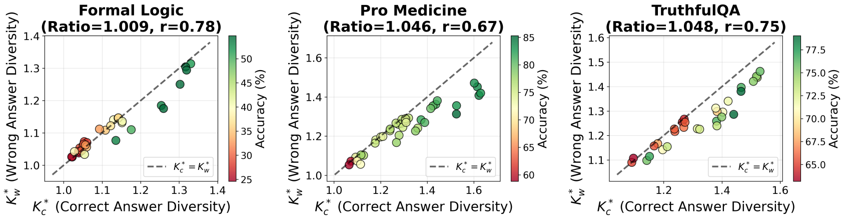


Figure 6: Decomposition of $\widehat{K}$ on three tasks. Points below the diagonal correspond to configurations where correct answer diversity dominates.

4.4.2 Diverse channels improve performance

We compute $\widehat{K}$ by embedding each agent output with NV-Embed-v2 (Lee et al., 2025), forming the cosine-similarity matrix G, trace-normalizing it to $\bar{G}$ with $\mathrm{Tr}(\bar{G}) = 1$, and defining $\widehat{K}$ as the entropy effective rank of $\bar{G}$ (Eq. 15).

Diversity increases $\widehat{K}$. In Table 3, $\widehat{K}$ consistently increases with diversity level from L1 to L4, validating $\widehat{K}$ as a robust indicator of system diversity without ground-truth labels.

Table 3: Relation between $\widehat{K}$ and Accuracy on ARC.

Method	Config	**Performance**		**Channels**		**Answer-Cond.**	
		Acc.	ΔAcc	$\widehat{K}$	$\Delta\widehat{K}$	$\widehat{K}_c$	$\widehat{K}_w$
Debate	L1	81.6%	–	1.197	–	1.184	1.177
	L2	81.0%	-0.7	1.348	+0.152	1.315	1.234
	L3	83.3%	+1.7	1.246	+0.049	1.220	1.160
	L4	**85.9%**	+4.2	**1.517**	**+0.320**	**1.472**	1.288
Vote	L1	81.3%	–	1.201	–	1.183	1.173
	L2	81.5%	+0.2	1.349	+0.149	1.318	1.222
	L3	83.8%	+2.5	1.245	+0.044	1.223	1.161
	L4	**87.5%**	**+6.1**	**1.521**	**+0.321**	**1.484**	1.297

Higher $\widehat{K}$ leads to better performance. The increase in $\widehat{K}$ is accompanied by higher accuracy in most cases (Table 3). Figure 7 further confirms this positive correlation, depicting a strong linear relationship between $\widehat{K}$ and task accuracy across configurations. Moreover, the marginal improvement in accuracy diminishes as $\widehat{K}$ grows, a pattern consistent with the geometric decay $(1-\alpha)^K$ in Theorem 3.4. We observe a minor anomaly in L2 under Debate, where $\widehat{K}$ increases but accuracy slightly decreases; we investigate this through the decomposition of $\widehat{K}$ below.

Mechanistic Decomposition: $\widehat{K}_c$ vs. $\widehat{K}_w$. To determine if the growth in $\widehat{K}$ represents useful evidence or merely increased noise, we decompose it into $\widehat{K}_c$ (correct reasoning diversity) and $\widehat{K}_w$ (incorrect reasoning diversity). Let $\hat{y}_i$ represent the final answer of agent i, and Y be the ground-truth label. We define:

$$\mathcal{I}_c = \{i : \hat{y}_i = Y\}, \qquad \mathcal{I}_w = \{i : \hat{y}_i \neq Y\}$$

Here, $\mathcal{I}_c$ is the set of correct agents, and $\mathcal{I}_w$ is the set of incorrect agents. We then compute the effective number of channels for each set:

$$\widehat{K}_c = \widehat{K}(\mathbf{E}_c), \qquad \widehat{K}_w = \widehat{K}(\mathbf{E}_w)$$

where $\mathbf{E}_c$ and $\mathbf{E}_w$ are the sub-matrices of the embedding matrix $\mathbf{E}$ (Section 3.4) corresponding to correct and incorrect agents, respectively.

The Empirical Boundary. Figure 6 suggests an empirical boundary in the $(\widehat{K}_c, \widehat{K}_w)$ plane: high-accuracy configurations concentrate in the region where $\widehat{K}_c > \widehat{K}_w$ (below the diagonal line). The intuition is as follows: when multiple agents arrive at the correct answer through genuinely *different* reasoning paths ($\widehat{K}_c$ is high), the correct answer receives support from independent evidence sources, making it more robust under aggregation. Conversely, when incorrect answers are also diverse ($\widehat{K}_w$ is high), the error "votes" are spread across many competing alternatives, which can dilute the correct signal. Additional design guidelines are provided in Appendix A.

5 Conclusion

This paper shows that simply increasing agent count in multi-agent systems results in diminishing returns, both for homogeneous and heterogeneous configurations. However, heterogeneity improves performance by introducing more diverse, non-redundant information, delaying saturation. We introduce $\widehat{K}$, a label-free measure of effective channels, which reveals that performance gains are driven by the balance between correct-path diversity and redundancy. These results suggest that the challenge in multi-agent scaling lies in the effective allocation of diverse information channels rather than just raw computational power.

REFERENCES

Winogrande: An adversarial winograd schema challenge at scale. 2019.

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.

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 Gu, Le Chen, Jiawei Han, and Tim Kraska. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*, 2023.

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 (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 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)*, pp. 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 Team. Qwen2.5: A party of foundation models. *arXiv preprint arXiv:2412.15115*, 2024.

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

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 Maeda, Ed Chi, Sharan Xia, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

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.

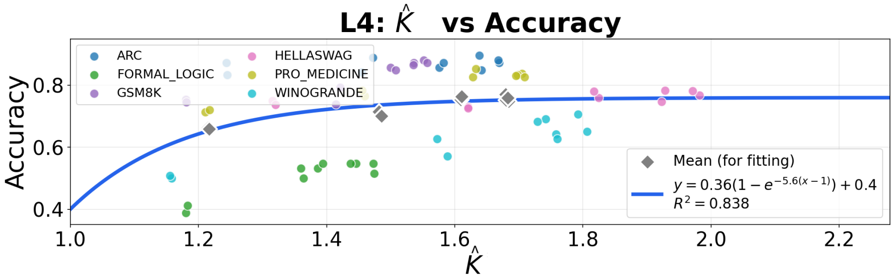


Figure 7: Correlation between $\widehat{K}$ and accuracy.

A Other Findings and Experiments

A.1 Design Guidelines for LLM-based MAS

Our analysis of effective channels yields several data-driven design guidelines for MAS development:

- **Match diversity to task type.** $\widehat{K}$ predicts accuracy strongly on reasoning tasks but weakly on knowledge-heavy tasks. For tasks requiring complex multi-step reasoning (e.g., *GSM8K*, *ARC*), investing in diversity yields significant performance gains. In contrast, for tasks dominated by factual retrieval (e.g., *Winogrande*), the diversity investment should be more conservative.
- **Ensure correct-path dominance.** Systems with high $\widehat{K}_c/\widehat{K}_w$ achieve substantially higher accuracy. In practice, this means that when introducing diversity, one should focus on increasing the diversity of *correct* reasoning paths, for example by using personas that encourage different valid problem-solving strategies (e.g., algebraic vs. geometric approaches in math tasks), rather than indiscriminately adding diversity that may also amplify incorrect reasoning (e.g., random temperature increases that introduce more errors).
- **Right-size agent count.** Homogeneous systems plateau at $N \approx 4$, while heterogeneous systems continue to benefit from scaling up to $N \approx 8$. Beyond this point, adding more agents results in diminishing returns and wasted compute resources. Thus, it is important to find a balance in agent count to avoid inefficiency.

B Related Works

Information-Theoretic Analysis of LLMs. Recent work has applied information theory to understand the behavior of LLMs. Ton et al. (2024) quantify information gain at each chain-of-thought step, showing that effective reasoning requires each step to contribute new information. Gan et al. (2025) analyze cascading failures through information loss accumulation, showing that conditional entropy can increase rather than decrease when per-step information loss grows super-linearly. In multi-agent settings, Riedl (2025) use Time-Delayed Mutual Information to distinguish genuine coordination from mere information sharing, and Chang (2025) track whether agent dialogues converge to consensus or maintain distributed information. These works provide valuable tools for analyzing information dynamics in LLM systems. Building on these foundations, our work focuses specifically on formalizing how agent diversity governs the diminishing returns observed in MAS scaling, and derives performance bounds that connect the effective channel count to information recovery.

LLM-based Multi-Agent Systems. LLM–based MAS instantiate multiple interacting LLM agents to perform compound inference through communication, coordination, or aggregation mechanisms (Xi et al., 2023; Wang et al., 2024b; Guo et al., 2024). Existing designs span independent sampling and voting schemes related to self-consistency (Wang et al., 2023), decentralized debate and role-playing frameworks (Du et al., 2023; Khan et al., 2024; Li et al., 2024b;a; 2023), centralized orchestration frameworks such as AutoGen (Wu et al., 2024) and MetaGPT (Hong et al., 2024), as well as hybrid, evolving, or self-improving coordination strategies (Dang et al., 2025; Zhao et al., 2025). Cemri et al. (2025) identify systematic failure modes across multi-agent systems. Taken together, these findings suggest that MAS performance is influenced by multiple design factors, among which we specifically focus on the role of agent diversity.

Empirical Studies of MAS Scaling and Diversity. It has been shown that naively scaling the number of agents yields limited benefits when agent behaviors are homogeneous (Wang et al., 2024b; Chen et al., 2024), across majority voting (Qian et al., 2025), debate (Choi et al., 2025), and more general coordination mechanisms (Kim et al., 2025). In contrast, a growing body of empirical evidence highlights the central role of diversity in MAS. Zhang et al. (2024) demonstrates that diversity leads to higher success rates in software engineering agents, while Wang et al. (2024a) finds that heterogeneous ensembles outperform homogeneous ones. Wu & Ito (2025) argue that preserving disagreement is preferable to enforcing early consensus. Related work shows that diversity benefits depend on task complexity (Tang et al., 2025) and that persona-based diversification has limitations (Samuel et al., 2024; Taillandier et al., 2025). These empirical findings collectively indicate the importance of diversity in MAS design. Building on these observations, our work provides an information-theoretic perspective that connects different forms of diversity to a common quantity, the effective channel count, offering a unified lens to understand when and why diversity improves performance across different settings.

C Proofs and Technical Details

This appendix provides full proofs and technical details. Section C.1 reviews standard information-theoretic identities. Section C.2 restates and proves the finite information budget equation 5. Sections C.3–C.4 derive upper bounds for common MAS workflows. Section C.5 formalizes the coverage model and proves Theorem 3.4 and Corollary C.14; Section C.5.5 proves Lemma C.17 and Proposition 3.6, which formally establish why same-type channels yield a strictly weaker coverage guarantee than independent cross-type channels, together with a generalization to non-uniform coverage. Section C.6 proves basic properties of $\widehat{K}$.

C.1 Information-Theoretic Preliminaries

We recall standard definitions and lemmas from information theory.

Definition C.1 (Conditional Mutual Information)**.** For random variables A, B, C,

$$I(A; B \mid C) = H(A \mid C) - H(A \mid B, C) = H(B \mid C) - H(B \mid A, C). \tag{16}$$

Lemma C.2 (Chain Rule for Mutual Information)**.** *For random variables $W_1, \ldots, W_n, A, B$,*

$$I(W_1, \ldots, W_n;\, A \mid B) = \sum_{i=1}^{n} I(W_i;\, A \mid B, W_{<i}). \tag{17}$$

Lemma C.3 (Incremental Information as Entropy Difference)**.** *Let $\Delta_i := I(Z_i; Y \mid X, Z_{<i})$. Then*

$$\Delta_i = H(Y \mid X, Z_{<i}) - H(Y \mid X, Z_{\le i}), \tag{18}$$

where $Z_{\le i} = (Z_1, \ldots, Z_i)$.

Proof. By the definition of conditional mutual information,

$$\begin{aligned} \Delta_i &= I(Z_i; Y \mid X, Z_{<i}) && (19)\\ &= H(Y \mid X, Z_{<i}) - H(Y \mid X, Z_{<i}, Z_i) && (20)\\ &= H(Y \mid X, Z_{<i}) - H(Y \mid X, Z_{\le i}). && \square \end{aligned}$$

C.2 Finite Information Budget (Upper Bound)

For completeness, the total information an MAS can extract is always upper-bounded by the intrinsic task uncertainty. As noted in Eq. equation 5, for any transcript $Z_{1:n}$,

$$I(Z_{1:n}; Y \mid X) \;=\; H(Y \mid X) - H(Y \mid X, Z_{1:n}) \;\le\; H(Y \mid X), \tag{21}$$

since conditional entropy is nonnegative. Moreover, writing $I(Z_{1:n}; Y \mid X) = \sum_{i=1}^{n} \Delta_i$ with $\Delta_i := I(Z_i; Y \mid X, Z_{<i})$ (Lemma C.2), the partial sums are uniformly bounded by $H(Y \mid X)$ and $\Delta_i \ge 0$, so $\Delta_i \to 0$ as $i \to \infty$.

C.3 Parallel Voting: Assumptions and Upper Bounds

This section derives the parallel-voting upper bounds (Section 3.1). The key message is that repeated sampling from the same configuration produces redundant evidence.

C.3.1 Conditional Independence for Parallel Sampling

Assumption C.4 (Conditional Independence for Parallel Sampling (All Types))**.** Consider a parallel MAS with agent configuration types $b(i) \in \mathcal{B}$. There exist channel kernels $\{\kappa_b(\cdot \mid x, y)\}_{b \in \mathcal{B}}$ such that, for every i,

$$P(Z_i = z \mid X = x, Y = y, Z_{<i}) \;=\; \kappa_{b(i)}(z \mid x, y), \tag{22}$$

and the outputs are mutually independent conditioned on (X, Y):

$$P(Z_{1:n} \mid X, Y) \;=\; \prod_{i=1}^{n} \kappa_{b(i)}(Z_i \mid X, Y). \tag{23}$$

Define the single-call information for type b:

$$I_b := I(Z^{(b)}; Y \mid X), \tag{24}$$

where $Z^{(b)}$ denotes one output from type b in isolation.

C.3.2 A Redundancy Identity

Lemma C.5 (Three-Way Mutual Information Decomposition)**.** *For any random variables A, B, C, D,*

$$I(A; B \mid C, D) = I(A; B \mid C) + I(A; D \mid B, C) - I(A; D \mid C). \tag{25}$$

Proof. Apply chain rule in two ways:

$$I(A; B, D \mid C) = I(A; B \mid C) + I(A; D \mid B, C), \tag{26}$$

$$I(A; B, D \mid C) = I(A; D \mid C) + I(A; B \mid C, D). \tag{27}$$

Equating and rearranging yields the claim. □

Corollary C.6 (Incremental Gain under Parallel Sampling)**.** *With $A = Z_i$, $B = Y$, $C = X$, $D = Z_{<i}$,*

$$I(Z_i; Y \mid X, Z_{<i}) = I(Z_i; Y \mid X) + I(Z_i; Z_{<i} \mid X, Y) - I(Z_i; Z_{<i} \mid X). \tag{28}$$

Under Assumption C.4, $I(Z_i; Z_{<i} \mid X, Y) = 0$, hence

$$I(Z_i; Y \mid X, Z_{<i}) = I(Z_i; Y \mid X) - I(Z_i; Z_{<i} \mid X) \leq I(Z_i; Y \mid X). \tag{29}$$

This formalizes redundancy: previous outputs can only reduce the new information.

Implication: redundancy controls early saturation. Upper bounds identify *what* limits the total information gain. To explain *when* saturation occurs, consider the incremental contribution $\Delta_i := I(Z_i; Y \mid X, Z_{<i})$. Eq. equation 29 provides an explicit decomposition:

$$\Delta_i = I(Z_i; Y \mid X) - I(Z_i; Z_{<i} \mid X), \tag{30}$$

where the *redundancy term* $I(Z_i; Z_{<i} \mid X)$ quantifies how much the i-th output overlaps with previous outputs. Thus, early saturation arises when repeated calls increase $I(Z_i; Z_{<i} \mid X)$, leaving little additional evidence to accumulate. Homogeneous agents typically induce large redundancy due to similar reasoning trajectories, while heterogeneity mitigates overlap and sustains Δ_i.

Since $I(Z_i; Y \mid X, Z_{<i}) \geq 0$, the identity also implies $I(Z_i; Z_{<i} \mid X) \leq I(Z_i; Y \mid X)$ under Assumption C.4.

C.3.3 Homogeneous Parallel Bound

Proposition C.7 (Homogeneous Parallel Upper Bound)**.** *Assume m parallel samples from a single type b under Assumption C.4. Then*

$$I(Z_{1:m}; Y \mid X) \leq mI_b. \tag{31}$$

Proof. By chain rule,

$$I(Z_{1:m}; Y \mid X) = \sum_{i=1}^{m} I(Z_i; Y \mid X, Z_{<i}). \tag{32}$$

Using Eq. equation 29 and $I(Z_i; Y \mid X) = I_b$ for all i,

$$I(Z_{1:m}; Y \mid X) \leq \sum_{i=1}^{m} I_b = mI_b. \tag{33}$$

□

C.3.4 Heterogeneous Parallel Bound

Theorem C.8 (Heterogeneous Parallel Upper Bound)**.** *Consider parallel voting with configuration types $\mathcal{B}$. Let type b be sampled m_b times, with total $n = \sum_{b \in \mathcal{B}} m_b$. Then*

$$I(Z_{1:n}; Y \mid X) \;\leq\; \sum_{b \in \mathcal{B}} m_b \, I_b. \tag{34}$$

Proof. Apply the chain rule:

$$I(Z_{1:n}; Y \mid X) = \sum_{i=1}^{n} I(Z_i; Y \mid X, Z_{<i}). \tag{35}$$

By Eq. equation 29, each term is bounded by $I(Z_i; Y \mid X) = I_{b(i)}$. Summing over steps grouped by type gives $\sum_{b \in \mathcal{B}} m_b \, I_b$. □

C.4 Sequential Pipelines and Debate: Upper Bounds

In sequential settings, each output conditions on the interaction history. This invalidates conditional independence, but the chain rule remains valid.

C.4.1 Maximal Per-Step Contribution

Define the maximal incremental contribution for agent configuration type b:

$$I_b^{\max} := \sup_{z_{<i}} I(Z_i; Y \mid X, Z_{<i} = z_{<i}, b(i) = b). \tag{36}$$

Proposition C.9 (Sequential Pipeline Upper Bound)**.** *For any sequential MAS with n steps and configuration types $\mathcal{B}$, where type b is invoked m_b times,*

$$I(Z_{1:n}; Y \mid X) \;\leq\; \sum_{b \in \mathcal{B}} m_b \, I_b^{\max}. \tag{37}$$

Proof. By chain rule,

$$I(Z_{1:n}; Y \mid X) = \sum_{i=1}^{n} I(Z_i; Y \mid X, Z_{<i}). \tag{38}$$

For each i, by definition of $I_{b(i)}^{\max}$ we have $I(Z_i; Y \mid X, Z_{<i}) \leq I_{b(i)}^{\max}$. Grouping the n terms by configuration type yields $\sum_{b \in \mathcal{B}} m_b \, I_b^{\max}$. □

Debate. Two-agent debate is a special case of sequential interaction and inherits the same bound. This formalizes why debate cannot systematically improve over voting if agents remain redundant.

Unified form. Both the parallel bound (Theorem C.8) and the sequential bound (Proposition C.9) share the same structure. Let $I_b^\star$ denote the per-call information ceiling for type b: $I_b^\star = I_b$ under parallel voting, and $I_b^\star = I_b^{\max}$ under sequential pipelines. Then

$$I(Z_{1:n}; Y \mid X) \;\le\; \sum_{b\in\mathcal{B}} m_b\, I_b^\star. \tag{39}$$

Combined with the finite information budget equation 5, the effective upper bound is $\min\bigl(H(Y \mid X),\; \sum_b m_b\, I_b^\star\bigr)$. Crucially, this bound depends on the multiset of configuration types $\{(b, m_b)\}_{b\in\mathcal{B}}$ rather than the raw call count n, which motivates the effective channel count introduced in Section 3.2.

C.5 Lower Bound via Independent Evidence-Bits Coverage

This section formalizes the "effective channels" view of Section 3.2. It proves Theorem 3.4 (geometric contraction of the residual uncertainty) and Corollary C.14 (the saturated lower bound), which together imply the characteristic $1 - e^{-\alpha K}$ improvement curve of Eq. equation 11.

C.5.1 Latent Component Model

Assumption C.10 (Independent Latent Components)**.** There exist latent random variables $U = (U_1, \dots, U_M)$ such that:

1. (**Sufficiency**) $H(Y \mid X, U) = 0$.
2. (**Conditional independence**) $U_1, \dots, U_M$ are independent conditioned on X.
3. (**Matching uncertainty scale**) $H(U \mid X) = H(Y \mid X)$.

Condition (iii) calibrates the latent components to exactly match the intrinsic task uncertainty: $H(U \mid X) = H(Y \mid X)$, so recovering all components eliminates residual uncertainty about Y.

Discussion of strength. Together, conditions (i) and (iii) imply that U and Y are in bijection given X: U determines Y (sufficiency) and Y determines U (since $H(U \mid X) = H(Y \mid X)$ combined with sufficiency forces $H(U \mid X, Y) = 0$). This means the latent decomposition contains no internal redundancy; each component U_j carries unique information. Condition (ii) further assumes that the components are conditionally independent given X, which is natural when they represent logically separable pieces of evidence (e.g., independent sub-steps of a multi-step reasoning problem), but may be violated for tasks with strong inter-step dependencies. These assumptions are idealizations that make the geometric contraction analysis tractable; extensions to correlated components (e.g., via Markov-chain or block-dependent models) are a natural direction for future work.

C.5.2 Fractional Coverage by Effective Channels

We say that channel k *resolves* component U_j if $H(U_j \mid X, \tilde{Z}^{(k)}) = 0$, i.e., the channel fully determines U_j given X.

Assumption C.11 (Fractional Coverage)**.** Let $\tilde{Z}_{1:K}$ denote K effective channels extracted from an MAS transcript.

1. (**Uniform resolution rate**) For each component U_j and channel k, the probability that channel k resolves U_j equals α, uniformly over all $j \in \{1, \dots, M\}$, $k \in \{1, \dots, K\}$, and all inputs x:
$$\Pr\bigl(H(U_j \mid X, \tilde{Z}^{(k)}) = 0 \bigm| X{=}x\bigr) \;=\; \alpha \quad \text{for some fixed } \alpha \in (0,1).$$
2. (**Cross-channel independence**) For each fixed j, the resolution events $\{H(U_j \mid X, \tilde{Z}^{(k)}) = 0\}_{k=1}^{K}$ are conditionally independent given X.

Consequence. If at least one channel resolves U_j, then by monotonicity of conditional entropy, $H(U_j \mid X, \tilde{Z}_{1:K}) = 0$: conditioning on additional channels cannot increase uncertainty. Thus, any component resolved by at least one channel has zero residual uncertainty given the full set of channels.

Connection to Definition 3.2. In Definition 3.2, the complementarity rate α is stated at the entropy level: $H(Y \mid X, \tilde{Z}_{1:k})/H(Y \mid X, \tilde{Z}_{1:k-1}) = 1 - \alpha$. This is a *consequence* of the per-component coverage model above. Indeed, as shown in the proof of Lemma C.12, when each channel independently resolves each component with probability α, the residual entropy after k channels satisfies $H(Y \mid X, \tilde{Z}_{1:k}) \le (1-\alpha)^k \, H(Y \mid X)$. Definition 3.2 captures this geometric contraction at the entropy level, abstracting away the per-component model.

Discussion of the uniformity assumption. Condition (i) assumes a single resolution rate α for all components j, all channels k, and all inputs x. In practice, certain components may be inherently easier to discover (higher α_j), different channels may have heterogeneous strengths (varying α_k), and harder inputs may have lower resolution rates (varying $\alpha(x)$). Under a non-uniform model where the resolution probability is $\alpha_{j,k}(x)$, the geometric bound generalizes to input-dependent products, which can be tighter or looser than the uniform-α result depending on the variance structure. The uniform model adopted here is an idealization that yields clean closed-form bounds while capturing the essential mechanism (geometric contraction with channel count). When the uniformity fails, the bound remains valid with the conservative choice $\alpha = \min_{x,j,k} \alpha_{j,k}(x)$.

C.5.3 Residual Contraction and Saturated Lower Bound

Lemma C.12 (Geometric Decay of Residual Uncertainty)**.** *Under Assumptions C.10 and C.11,*

$$H(Y \mid X, \tilde{Z}_{1:K}) \;\le\; (1-\alpha)^K \, H(Y \mid X). \tag{40}$$

Proof. By Assumption C.10(i), Y is a function of (X, U), hence

$$H(Y \mid X, \tilde{Z}_{1:K}) \le H(U \mid X, \tilde{Z}_{1:K}). \tag{41}$$

Subadditivity of conditional entropy yields

$$H(U \mid X, \tilde{Z}_{1:K}) \le \sum_{j=1}^{M} H(U_j \mid X, \tilde{Z}_{1:K}). \tag{42}$$

Fix j. Expanding the conditional entropy as an average over instances:

$$H(U_j \mid X, \tilde{Z}_{1:K}) = \sum_x p(x) \sum_{\tilde{z}} p(\tilde{z} \mid x) \, H(U_j \mid X{=}x, \tilde{Z}_{1:K}{=}\tilde{z}). \tag{43}$$

For each instance $(x, \tilde{z})$: if at least one channel resolves U_j (i.e., $H(U_j \mid X{=}x, \tilde{Z}^{(k)}{=}\tilde{z}^{(k)}) = 0$ for some k), then by monotonicity of conditional entropy $H(U_j \mid X{=}x, \tilde{Z}_{1:K}{=}\tilde{z}) = 0$; otherwise, $H(U_j \mid X{=}x, \tilde{Z}_{1:K}{=}\tilde{z}) \le H(U_j \mid X{=}x)$. Hence

$$H(U_j \mid X, \tilde{Z}_{1:K}) \le \sum_x p(x) \, H(U_j \mid X{=}x) \; \Pr\!\big(\text{no channel resolves } U_j \;\big|\; X{=}x\big). \tag{44}$$

By Assumption C.11(i)–(ii), the resolution events are independent with probability α each, so $\Pr(\text{no channel resolves } U_j \mid X{=}x) = (1-\alpha)^K$ for every x. Therefore,

$$H(U_j \mid X, \tilde{Z}_{1:K}) \le (1-\alpha)^K \sum_x p(x) \, H(U_j \mid X{=}x) = (1-\alpha)^K \, H(U_j \mid X). \tag{45}$$

Summing over j gives

$$H(U \mid X, \tilde{Z}_{1:K}) \le (1-\alpha)^K \sum_{j=1}^{M} H(U_j \mid X). \tag{46}$$

Finally, by Assumption C.10(ii), $H(U \mid X) = \sum_{j=1}^{M} H(U_j \mid X)$, and by (iii) $H(U \mid X) = H(Y \mid X)$. Combining completes the proof. □

Theorem C.13 (Geometric Contraction with Effective Channels). *Under Assumptions C.10 and C.11,*

$$H(Y \mid X) - I(\tilde{Z}_{1:K}; Y \mid X) \;=\; H(Y \mid X, \tilde{Z}_{1:K}) \;\leq\; (1-\alpha)^K \, H(Y \mid X). \tag{47}$$

Consequently,

$$\boxed{\begin{aligned} &H(Y \mid X) - I(\tilde{Z}_{1:K}; Y \mid X) \\ &\;\leq\; (1-\alpha)^K \, H(Y \mid X) \;\leq\; e^{-\alpha K} \, H(Y \mid X). \end{aligned}} \tag{48}$$

Equivalently, the normalized residual *satisfies* $H(Y \mid X, \tilde{Z}_{1:K})/H(Y \mid X) \leq (1-\alpha)^K \leq e^{-\alpha K}$.

Proof. By definition, $I(\tilde{Z}_{1:K}; Y \mid X) = H(Y \mid X) - H(Y \mid X, \tilde{Z}_{1:K})$, so $H(Y \mid X) - I(\tilde{Z}_{1:K}; Y \mid X) = H(Y \mid X, \tilde{Z}_{1:K})$. Apply Lemma C.12 to obtain equation 47. The exponential form follows from $(1-\alpha)^K \leq e^{-\alpha K}$. □

Corollary C.14 (Saturated Lower Bound). *Under the same assumptions,*

$$\boxed{I(\tilde{Z}_{1:K}; Y \mid X) \;\geq\; H(Y \mid X)\Big(1 - (1-\alpha)^K\Big) \;\geq\; H(Y \mid X)\Big(1 - e^{-\alpha K}\Big).} \tag{49}$$

Proof. Rearrange the identity $I(\tilde{Z}_{1:K}; Y \mid X) = H(Y \mid X) - H(Y \mid X, \tilde{Z}_{1:K})$ and apply Lemma C.12. The exponential form follows from $(1-\alpha)^K \leq e^{-\alpha K}$. □

C.5.4 Heterogeneity Advantage as an αK Comparison

This subsection provides a formal underpinning for Corollary 3.7: heterogeneity improves recoverable information whenever it increases the effective evidence term αK.

Lemma C.15 (Monotonicity in αK). *Define* $f(t) := 1 - e^{-t}$ *for* $t \geq 0$*. Then for any* $t_1, t_2 \geq 0$*,* $t_2 > t_1$ *implies* $f(t_2) > f(t_1)$*.*

Proof. $f'(t) = e^{-t} > 0$ for all $t \geq 0$, hence f is strictly increasing. □

Corollary C.16 (Heterogeneity Advantage from Corollary C.14). *Consider two designs summarized by* $(K_{\text{homog}}, \alpha_{\text{homog}})$ *and* $(K_{\text{heterog}}, \alpha_{\text{heterog}})$ *under Assumptions C.10–C.11. By Corollary C.14, the lower bounds on recoverable information for the two designs are:*

$$I_{\text{heterog}} \geq H(Y \mid X)\big(1 - e^{-\alpha_{\text{heterog}} K_{\text{heterog}}}\big), \tag{50}$$

$$I_{\text{homog}} \geq H(Y \mid X)\big(1 - e^{-\alpha_{\text{homog}} K_{\text{homog}}}\big). \tag{51}$$

When $\alpha_{\text{heterog}} K_{\text{heterog}} > \alpha_{\text{homog}} K_{\text{homog}}$*, the heterogeneous design admits a strictly higher* lower bound *on recoverable information, since by Lemma C.15 the function* $1 - e^{-t}$ *is strictly increasing in* t*.*

Proof. Apply Corollary C.14 to each design to obtain equation 50 and equation 51. Since $\alpha_{\text{heterog}} K_{\text{heterog}} > \alpha_{\text{homog}} K_{\text{homog}}$ and $f(t) = 1 - e^{-t}$ is strictly increasing (Lemma C.15), the lower bound for the heterogeneous design is strictly larger than that for the homogeneous design. □

C.5.5 Formal Basis for Heterogeneity Advantage

This subsection states and proves the Within-Type Saturation lemma and then proves Proposition 3.6 (Complementarity Advantage), which together establish why heterogeneous systems achieve a superior coverage guarantee. We also discuss the generalization to non-uniform coverage rates.

Lemma C.17 (Within-Type Saturation). *Under Assumptions C.10 and C.4, consider* m *i.i.d. outputs from a single configuration type* b*. For each component* U_j*, define the answer-conditional resolution rate* $p_b(j, y) := \Pr\big(H(U_j \mid X, \tilde{Z}^{(k)}) = 0 \;\big|\; X, Y{=}y\big)$*. Then:*

1. ***Positive correlation.*** *For same-type channels $k \neq k'$, the resolution events for U_j are positively correlated:*
$$\mathrm{Cov}\big(\mathbf{1}\{U_j \textit{ resolved by } k\},\ \mathbf{1}\{U_j \textit{ resolved by } k'\} \bigm| X\big) = \mathrm{Var}_{Y|X}[p_b(j,Y)] \geq 0.$$

2. ***Degraded non-resolution.*** *The probability that U_j remains unresolved after all m same-type channels satisfies*
$$\Pr\big(\textit{no channel resolves } U_j \bigm| X\big) \ \geq\ (1-\alpha)^m, \tag{52}$$
with equality iff $p_b(j,Y) = \alpha$ a.s., i.e., resolution is independent of the answer.

Proof of Lemma C.17. For each component U_j and channel k, the answer-conditional resolution rate is $p_b(j,y) := \Pr(H(U_j \mid X, \tilde{Z}^{(k)}) = 0 \mid X, Y{=}y)$, with marginal $\mathbb{E}_{Y|X}[p_b(j,Y)] = \alpha$.

Write $\mathbf{1}_{j,k} := \mathbf{1}\{H(U_j \mid X, \tilde{Z}^{(k)}) = 0\}$ for brevity (the indicator that channel k resolves U_j).

Part (i): Positive within-type correlation. Under Assumption C.4, outputs $Z_k^{(b)}$ and $Z_{k'}^{(b)}$ are conditionally independent given (X,Y). Since $\mathbf{1}_{j,k}$ is a deterministic function of $(X, Z_k^{(b)})$, the indicators $\mathbf{1}_{j,k}$ and $\mathbf{1}_{j,k'}$ are also conditionally independent given (X,Y). By the law of total covariance,
$$\begin{aligned}\mathrm{Cov}(\mathbf{1}_{j,k},\ \mathbf{1}_{j,k'} \mid X) &= \underbrace{\mathbb{E}_{Y|X}\big[\mathrm{Cov}(\mathbf{1}_{j,k}, \mathbf{1}_{j,k'} \mid X, Y)\big]}_{=0 \text{ (cond. indep. given } (X,Y))} + \mathrm{Var}_{Y|X}\big[\underbrace{\mathbb{E}[\mathbf{1}_{j,k} \mid X, Y]}_{=p_b(j,Y)}\big] \\ &= \mathrm{Var}_{Y|X}\big[p_b(j,Y)\big] \ \geq\ 0.\end{aligned} \tag{53}$$

Part (ii): Degraded non-resolution. Conditional on (X,Y), the m resolution indicators $\{\mathbf{1}_{j,k}\}_{k=1}^m$ are i.i.d. $\mathrm{Bernoulli}(p_b(j,Y))$:
$$\Pr\big(\text{no channel resolves } U_j \bigm| X,\, Y\big) \ =\ \big(1 - p_b(j,Y)\big)^m.$$
Marginalizing over Y:
$$\Pr\big(\text{no channel resolves } U_j \bigm| X\big) \ =\ \mathbb{E}_{Y|X}\big[(1 - p_b(j,Y))^m\big].$$
Since $t \mapsto (1-t)^m$ is strictly convex for $m \geq 2$ and convex for $m = 1$, and $\mathbb{E}_{Y|X}[p_b(j,Y)] = \alpha$, Jensen's inequality gives
$$\mathbb{E}_{Y|X}\big[(1-p_b(j,Y))^m\big] \ \geq\ \big(1 - \mathbb{E}_{Y|X}[p_b(j,Y)]\big)^m = (1-\alpha)^m.$$
Equality holds if and only if $p_b(j,Y)$ is a.s. constant given X, i.e., $p_b(j,Y) = \alpha$ a.s., meaning resolution of U_j does not depend on the ground-truth answer. □

Proof of Proposition 3.6. Following the proof of Lemma C.12: by sufficiency of U (Assumption C.10(i)) and subadditivity of conditional entropy,
$$H(Y \mid X, Z_{1:m}^{(b)}) \ \leq\ \sum_{j=1}^{M} H(U_j \mid X, Z_{1:m}^{(b)}).$$

For each j, exactly as in Lemma C.12:
$$H(U_j \mid X, Z_{1:m}^{(b)}) \ \leq\ H(U_j \mid X) \cdot \Pr\big(\text{no channel resolves } U_j \bigm| X\big).$$
Substituting the result of Lemma C.17(ii) and summing over j:
$$H(Y \mid X, Z_{1:m}^{(b)}) \ \leq\ \sum_{j=1}^{M} H(U_j \mid X)\, \mathbb{E}_{Y|X}\big[(1-p_b(j,Y))^m\big] \ =:\ R_{\text{same}}(m) \cdot H(Y \mid X).$$

The bound $R_{\text{same}}(m) \geq (1-\alpha)^m$ follows from applying Lemma C.17(ii) to each summand and using $\sum_j H(U_j \mid X) = H(U \mid X) = H(Y \mid X)$ (Assumption C.10(ii)–(iii)). The strict inequality $R_{\text{same}}(m) > (1-\alpha)^m$ holds whenever $p_b(j,Y)$ varies with Y for at least one j—which is the case for any non-trivial inference task where the channel's ability to resolve a component depends on the ground-truth answer. □

Remark C.18 (Generalization to Non-Uniform Coverage Rates). The uniform-α assumption (Assumption C.11(i)) is an idealization. In practice, different components may have heterogeneous resolution rates α_j, different channels may have varying strengths α_k, and harder inputs may have lower rates. Under a non-uniform model where the resolution probability is $\alpha_{j,k}(x)$, the within-type saturation mechanism of Lemma C.17 still applies:

$$\Pr\bigl(\text{no channel resolves } U_j \bigm| X{=}x\bigr) \;=\; \mathbb{E}_{Y|X=x}\bigl[(1-p_b(j,Y))^m\bigr] \;\geq\; (1-\alpha_j(x))^m,$$

where $\alpha_j(x) := \mathbb{E}_{Y|X=x}[p_b(j,Y)]$. The key insight—that same-type channels are positively correlated through Y—does not depend on the uniformity of α; it follows purely from the law of total covariance. Thus, the qualitative conclusion (homogeneous systems yield strictly weaker coverage guarantees than heterogeneous ones) is robust to the specific distributional assumptions. The uniform model (Assumption C.11) provides clean closed-form bounds while preserving this essential mechanism.

Interpretation. Lemma C.17 and Proposition 3.6 together formalize the key mechanism behind the heterogeneity advantage. The positive correlation (Lemma C.17(i)) arises because same-type channels share the same "view" of the task: when the answer Y makes component U_j easy to resolve (high $p_b(j,Y)$), *all* same-type channels are likely to resolve it; when Y makes it hard (low $p_b(j,Y)$), *all* channels are likely to miss it. This correlation inflates the probability of simultaneous non-resolution (Lemma C.17(ii)), yielding a strictly weaker residual bound (Proposition 3.6).

In contrast, channels from different configuration types may have *different* answer-conditional profiles: one type may succeed on instances where another fails. This complementarity is precisely what the cross-channel independence condition (Assumption C.11(ii)) captures, and it is what enables the tight $(1-\alpha)^K$ geometric contraction of Theorem 3.4. Thus, Lemma C.17 and Proposition 3.6 provide the formal basis connecting heterogeneity to a superior αK product: heterogeneous systems satisfy the independence condition that homogeneous systems provably violate.

C.6 Properties of the Effective Channel Count $\widehat{K}$

This section proves basic properties of the label-free proxy $\widehat{K}$ (Section 3.4). We restate the definition for completeness.

Setup. Given n outputs, let $\hat{\mathbf{z}}_i := \mathrm{Emb}(Z_i)/\|\mathrm{Emb}(Z_i)\|_2 \in \mathbb{R}^d$ be the normalized embeddings, and let $\mathbf{E} \in \mathbb{R}^{n\times d}$ be the embedding matrix whose i-th row is $\hat{\mathbf{z}}_i^\top$. Define the cosine-similarity Gram matrix $G_{ij} := \langle \hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j \rangle$ (equivalently $G = \mathbf{E}\mathbf{E}^\top$) and its trace-normalization

$$\bar{G} := \frac{G}{\mathrm{Tr}(G)}. \tag{54}$$

Let $\{\lambda_j\}_{j=1}^n$ be the eigenvalues of $\bar{G}$. The von Neumann entropy is

$$H(\bar{G}) := -\sum_{j=1}^{n} \lambda_j \log_2 \lambda_j, \tag{55}$$

and the effective channel count is $\widehat{K} := 2^{H(\bar{G})}$.

Proposition C.19 (Basic Properties of $\widehat{K}$)**.** *For any nonzero embedding matrix* $\mathbf{E}$*,*

1. $1 \leq \widehat{K} \leq n$.
2. $\widehat{K} = 1$ *iff* $\bar{G}$ *is rank-*1 *(all embeddings are collinear up to scaling).*
3. $\widehat{K} = n$ *iff* $\bar{G} = \frac{1}{n}\mathbf{I}_n$ *(embeddings are orthogonal with equal norm).*
4. $\widehat{K}$ *is continuous in* $\mathbf{E}$ *(when* $\mathrm{Tr}(G) > 0$*) and invariant to permutation of outputs.*

Proof. **(i) Bounds.** Entropy satisfies $0 \leq H(\bar{G}) \leq \log_2 n$, hence $1 = 2^0 \leq 2^{H(\bar{G})} \leq 2^{\log_2 n} = n$.

Table 4: Closed-source vs. Open-source on Formal Logic (%). Δ_{Het}: average Heterog gain over Base. Δ_N: accuracy change from $N=2$ to $N=16$ under the Heterog configuration.

Model	**Method**	**Base**		**Heterog**		Δ_{Het}	Δ_N
		N=2	N=16	N=2	N=16		
Closed-source							
gpt-4.1-mini	vote	46.83	48.41	55.56	52.38	+6.35	−3.18
	debate	46.03	39.68	50.79	42.86	+3.97	−7.93
gpt-5-mini	vote	0.00	6.35	35.71	49.21	**+39.29**	+13.50
	debate	0.79	6.35	34.92	54.76	**+41.27**	+19.84
Open-source							
Qwen-2.5-7B	vote	38.10	44.44	45.24	50.00	+6.35	+4.76
	debate	30.95	34.13	25.40	38.10	−0.79	+12.70
Llama-3.1-8B	vote	45.24	42.86	44.44	54.76	+5.55	+10.32
	debate	24.60	31.75	35.71	39.68	+9.52	+3.97
Mistral-7B	vote	35.71	42.06	34.92	41.27	−0.79	+6.35
	debate	34.92	44.44	39.68	46.83	+3.58	+7.15

(ii) $\widehat{K} = 1$**.** $H(\bar{G}) = 0$ iff the spectrum is $(1, 0, \ldots, 0)$, which holds iff $\bar{G}$ is rank-1. This corresponds to all rows of $\mathbf{E}$ being collinear, i.e., all embeddings are identical up to scaling.

(iii) $\widehat{K} = n$**.** $H(\bar{G}) = \log_2 n$ iff $\lambda_j = 1/n$ for all j, which occurs when $\bar{G} = \frac{1}{n}\mathbf{I}_n$. This corresponds to G being proportional to the identity, i.e., embeddings are orthogonal with equal norm.

(iv) Continuity and permutation invariance. The map $\mathbf{E} \mapsto G = \mathbf{E}\mathbf{E}^\top$ is continuous. Normalization by $\mathrm{Tr}(G)$ is continuous when $\mathrm{Tr}(G) > 0$. Eigenvalues of a symmetric matrix vary continuously with the entries, and entropy is continuous on the simplex. Permutation of outputs corresponds to $G \mapsto PGP^\top$ for a permutation matrix P, which preserves eigenvalues. □

D Supplementary Experiments

D.1 Closed-Source Model Experiments

We extend our analysis to closed-source models (gpt-4.1-mini, gpt-5-mini) on the Formal Logic benchmark to test whether the heterogeneity advantage generalizes across model families. Table 4 compares closed- and open-source models under homogeneous (Base) and heterogeneous (Heterog) configurations at $N=2$ and $N=16$.

Key findings. The results confirm that the heterogeneity advantage generalizes to closed-source models, while revealing that its magnitude and scaling behavior vary across model families.

- **Heterogeneity consistently improves over homogeneous baselines.** All five models exhibit positive Δ_{Het} in at least one interaction mechanism, confirming that the advantage is not specific to open-source settings.
- **Models with weaker homogeneous baselines benefit more from heterogeneity.** gpt-5-mini achieves near-zero accuracy under homogeneous settings (0–6%) but reaches 35–55% with heterogeneous prompting (Δ_{Het} of +39–41%). In contrast, gpt-4.1-mini and the open-source models, which already achieve 25–46% under homogeneous settings, show more modest gains (+3–10%).
- **Scaling trends diverge across model families.** Open-source models exhibit positive scaling ($\Delta_N > 0$) under both configurations. gpt-4.1-mini, however, shows *negative* scaling in debate: accuracy drops from 50.79% to 42.86% ($\Delta_N = -7.93$) even under heterogeneous settings, indicating that adding more agents can hurt when the base model is already strong. gpt-5-mini shows the opposite pattern: under heterogeneous settings it benefits substantially from more agents ($\Delta_N = +19.84$ for Debate), whereas its homogeneous scaling remains near-flat.

D.2 Robustness to Embedding Model Choice

A potential concern is whether our effective channel metric $\widehat{K}$ depends critically on the choice of embedding model. To address this, we recompute $\widehat{K}$ using a different embedding model, gte-Qwen2-1.5B-instruct (1536 dimensions), and compare the results against our primary model NV-Embed-v2 (4096 dimensions). We conduct this comparison across seven datasets (ARC, Formal Logic, GSM8K, HellaSwag, Pro Medicine, TruthfulQA, WinoGrande), varying agent counts $N \in \{2, 4, 8, 12, 16\}$ and interaction mechanisms (Vote and Debate).

Since embedding dimensionality affects absolute $\widehat{K}$ values, direct comparison of raw values across models is not meaningful. Instead, we assess robustness by measuring whether the two embeddings agree on *relative ordering*: within each (configuration type, dataset) pair, do both embeddings rank different (method, N) combinations consistently? Across all matched pairs, we observe an average Spearman correlation of $\rho = 0.91$, with over 95% of pairs showing $\rho > 0.5$. This indicates that both embeddings consistently identify which experimental settings produce more diverse outputs, even though their absolute scales differ.

Furthermore, both embeddings yield $\widehat{K}$ metrics that positively correlate with task accuracy (NV-Embed-v2: $r = 0.40$; gte-Qwen2: $r = 0.23$), confirming that our core finding, diversity predicts performance, is not an artifact of a particular embedding choice. We use NV-Embed-v2 in the main experiments as it achieves stronger predictive power.

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

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. To disentangle this, we fit a baseline regression that predicts task accuracy from N and configuration labels alone, then measure the incremental variance explained (ΔR^2) when $\widehat{K}$ or its components are added.

Table 5: **Incremental Explanatory Power of Effective Channels.** The baseline model using only agent count (N) and configuration labels explains little variance ($R^2 = 0.062$). Adding $\widehat{K}$ substantially improves fit ($\Delta R^2 = +0.147$), and conditioning on answer correctness ($\widehat{K}_c$) yields the largest gain ($\Delta R^2 = +0.331$), while $\widehat{K}_w$ contributes negligibly.

Model	R^2	**Adj.** R^2	ΔR^2	**AIC**
Baseline (N + Config)	0.062	0.044	–	1806.6
Baseline + $\widehat{K}$	0.209	0.190	+0.147	1771.1
Baseline + $\widehat{K}_c$	0.393	0.378	+0.331	1713.0
Baseline + $\widehat{K}_c + \widehat{K}_w$	0.396	0.379	+0.334	1713.8
Baseline + $\widehat{K}_c/\widehat{K}_w$	0.325	0.309	+0.263	1736.4

Table 5 reveals three findings. First, the baseline model with only N and configuration labels achieves $R^2 = 0.062$, confirming that scale and configuration alone are poor predictors of MAS performance. Second, adding $\widehat{K}$ raises ΔR^2 by $+0.147$, demonstrating that it captures structural information about output diversity that is not reducible to agent count or configuration choice. Third, and most importantly, replacing $\widehat{K}$ with its correctness-conditioned component $\widehat{K}_c$ more than doubles the incremental gain ($\Delta R^2 = +0.331$), while further adding $\widehat{K}_w$ yields negligible improvement (ΔR^2: $+0.331 \to +0.334$). This asymmetry directly supports our central thesis: what drives MAS performance is not output diversity in general, but specifically the diversity of *correct* reasoning paths. Increasing the number of distinct ways agents arrive at the right answer is far more predictive than total channel count or the diversity of incorrect responses.

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

We further test whether the observed relationship between effective channels and performance could arise by chance. To this end, we conduct permutation-based randomization tests that preserve the marginal distribution of accuracy while destroying any structural association with $\widehat{K}$.

Table 6: **Permutation Sanity Check (1000 shuffles).** Observed correlations between effective-channel metrics and accuracy lie far outside the null distribution, confirming that the relationship is not due to chance.

Metric	**Observed** r	z**-score**	p
$\widehat{K}$	0.388	5.87	<0.001
$\widehat{K}_c$	0.535	7.75	<0.001
$\widehat{K}_c/\widehat{K}_w$	0.503	7.23	<0.001

As shown in Table 6, all effective-channel metrics exhibit z-scores well above 5 under permutation testing, with $p < 10^{-3}$. This rules out the possibility that the observed correlations arise from random alignment or dataset-specific artifacts. Notably, $\widehat{K}_c$ again yields the strongest signal, reinforcing the interpretation that correct-path diversity is the dominant driver of multi-agent performance.

D.5 Case Study: Heterogeneity Effects Across Models and Workflows

Table 7 reports a comprehensive ablation study on the Formal Logic benchmark, varying base models, agent counts ($N = 2$–16), and interaction mechanisms. Across nearly all settings, heterogeneous configurations outperform homogeneous ones, often by substantial margins. Importantly, these gains do not arise from scaling alone. For example, in both Vote and Debate, increasing N beyond moderate values frequently yields diminishing or unstable returns in homogeneous settings, while heterogeneous systems maintain consistent improvements. This pattern holds across all three base models and their mixture, indicating that the benefit of heterogeneity is robust to model choice and interaction protocol.

Table 8 isolates the effect of model mixing by comparing a heterogeneous mixture (MIX) against the best-performing single model under the same agent count. At $N \geq 4$, MIX consistently outperforms the strongest individual model by large margins, reaching up to +14.28% absolute accuracy at $N = 8$.

Crucially, these gains cannot be explained by model selection alone. Even when the best single model is used with heterogeneous prompting, the MIX configuration achieves higher performance, demonstrating genuine synergy across models rather than simple averaging or dominance effects.

Table 7: Model Ablation on Formal Logic: Impact of Heterogeneity from $N = 2$ to $N = 16$

Base Model	**Agents (N)**	**Vote (Round 0)**			**Debate (Final)**		
		Homog	**Heterog**	Δ_{H-M}	**Homog**	**Heterog**	Δ_{H-M}
Qwen-2.5-7B	2	38.10%	45.24%	+7.14%	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%	28.57%	38.10%	+9.53%
	12	44.44%	52.38%	+7.94%	31.75%	35.71%	+3.96%
	16	44.44%	50.00%	+5.56%	34.13%	38.10%	+3.97%
Llama-3.1-8B	2	45.24%	44.44%	-0.80%	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%	27.78%	35.71%	+7.93%
	12	43.65%	53.97%	+10.32%	30.95%	38.89%	+7.94%
	16	42.86%	54.76%	+11.90%	31.75%	39.68%	+7.93%
Mistral-7B	2	35.71%	34.92%	-0.79%	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%	38.89%	-1.59%
	12	38.89%	38.10%	-0.79%	42.06%	42.86%	+0.80%
	16	42.06%	41.27%	-0.79%	44.44%	46.83%	+2.39%
MIX	2	45.24%	48.41%	+3.17%	34.13%	38.89%	+4.76%
	4	47.62%	52.38%	+4.76%	42.86%	53.17%	+10.31%
	8	47.62%	55.56%	+7.94%	49.21%	53.17%	+3.96%
	12	48.41%	57.94%	+9.53%	48.41%	54.76%	+6.35%
	16	50.00%	53.97%	+3.97%	43.65%	51.59%	+7.94%

Table 8: Formal Logic: Synergy of Model Mixing (MIX vs. Best Single Model)

Agents (N)	**Best Single (Heterog)**	**MIX (Heterog)**	$\Delta_{\text{MIX vs. Best}}$	**MIX (Homog)**	$\Delta_{\text{H-M (MIX)}}$
2	39.68%	38.89%	-0.79%	34.13%	+4.76%
4	44.44%	53.17%	**+8.73%**	42.86%	+10.31%
8	38.89%	53.17%	**+14.28%**	49.21%	+3.96%
12	42.86%	54.76%	**+11.90%**	48.41%	+6.35%
16	46.83%	51.59%	**+4.76%**	43.65%	+7.94%