Benefits to Agentic Heterogeneity in Compositional Decision Making

PARNIAN SHAHKAR*, University of California, Irvine, USA
JIAXIN SONG*, University of Illinois Urbana-Champaign, USA
KATE DONAHUE, University of Illinois Urbana-Champaign, USA and Massachusetts Institute of Technology, USA
BHASKAR RAY CHAUDHURY, University of Illinois Urbana-Champaign, USA

In many real-life settings, decisions are made not by a single agent acting in isolation, but by teams of imperfect agents. Specifically, teammates often build on each other's work, reflecting a composition setting. For example, consider a setting where an algorithm presents a partial solution to a human, who makes the final decision herself, or an agentic team where the final output is a product of iterative improvements by multiple agents. In this case, we may wonder about the benefits of a *diverse* team - when agents are *composed* with each other, is it ever helpful to have agents with different strengths and weaknesses, rather than a homogeneous team of the strongest agent in isolation? In this work, we study a specific setting, where the task is to recover a single item (answer) from a discrete set, where items differ in quality. One agent (the curator) narrows down a set of items to a smaller set, while another agent (the decider) makes the final pick. In this setting, we show provable benefits to heterogeneity between the agents: that is, we identify settings where a heterogeneous team is stronger than a homogeneous one. We precisely characterize the type of heterogeneity, which relates to types of "misalignment" between each agent that leads them to make different kinds of mistakes. Finally, we conclude by exploring performance of LLMs on a benchmark dataset (MMLU) and demonstrating empirically that there exist scenarios where a mixed team can outperform a single agent in isolation, or homogeneous team. We conclude by discussing implications for the design of agentic teams, as well as desirable properties for design of algorithmic tools.

Contents

Abstract		0
Contents		0
1	Introduction	1
2	Related Work	3
3	Model	5
4	Conditions for Complementarity and Diversity Gain	6
5	Experiments	11
6	Conclusion	16
References		19
A	Discussion on simulating composition of agents	22
B	Properties of Random Choice Models	22
C	Omitted Proofs of Section 4.1	25
D	Omitted Proofs of Section 4.2	29
E	Beyond Top-Item Recovery	30

*Equal Contribution.

Manuscript submitted for review to the 27th ACM Conference on Economics & Computation (EC'26).

1 Introduction

In recent years, advances in artificial intelligence and machine learning have become increasingly integrated into our daily lives: algorithms help suggest movies for us to watch, routes for us to drive on, or even generate novel content for us to rely on. Additionally, outputs of algorithmic tools are increasingly not used in isolation, with one algorithm having the "final say". Instead, we often care about the performance of a system made up of multiple interacting agents. For example, a system may be made up of an algorithm providing suggestions and a human acting on them. For another example, a system could be made up multiple interacting AI agents. Finally, systems could involve both interacting AI agents and humans.

A team could have multiple different structures: for example, it could involve *routing* tasks between agents [Huang et al., 2025]. In this work, we will be especially interested in another method of team interaction: *composition*: where one agent builds on a partial solution created by another (e.g. [Bansal et al., 2024, Jiang et al., 2023, Wang et al., 2024]). A few main motivations for agentic composition are:

(1) Human-algorithm collaboration: when an algorithmic tool and human jointly solve a problem. For example, consider a setting where an algorithm provides information (e.g. via RAG) to partially solve the problem, while the human makes the final decision. In this line of work, one main objective is *complementarity* [Bansal et al., 2021b]: when does the joint human-algorithm team outperform a single agent in isolation?
(2) Agent teams: when a group of LLMs interact to jointly complete a task. Methods of collaboration could involve routing, solving subtasks, or composition: where one agent takes a first pass and subsequent agents improve the response. Here, one natural question is when there are benefits to having a team of heterogeneous LLMs (with some being weaker), as compared to a team composed of only identical LLMs. When is a diverse team stronger?

We aim to study agentic composition because it is prevalent and relatively under-studied in the theoretical literature: what are the properties of agentic composition that lead to a successful or unsuccessful team? When we are studying the performance of a system, there are a few natural questions that arise. First, when is the joint system better than any individual agent - that is, when are there true benefits to teamwork? Secondly, when the system is composed of multiple *different* agents, when is such heterogeneity helpful - when are there benefits to diversity? A diverse team could be helpful because agents make different errors, which may avoid harms for correlated mistakes. However, a diverse team necessarily includes some agents who are weaker and some are stronger, and it may be the case that the team could be strengthened by replacing a few weaker agents with stronger ones.

We motivate these questions as two formal benchmarks:

(1) Complementarity [Bansal et al., 2021b]: When does the joint system have strictly better performance (e.g. accuracy) than any member of the team in isolation?
(2) Benefits to diversity: When does a system composed of heterogeneous agents (e.g. humans and AI, or two different AIs) outperform the same system with homogeneous agents? For example, suppose that a system is composed of one LLM (e.g. Claude) that partially completes the task, and another LLM (e.g. Gemini) that completes it. The system has benefits to diversity when this system outperforms a hypothetical system where two instances of Claude collaborate (and similarly for Gemini).

In this work, in order to address the questions above, we will need a model of agentic composition that is both theoretically tractable and rich enough to model phenomena of interest. In particular, we need a setting where a) agents can differ from each other systematically, b) there is a natural model of agents building on each other's work, and c) a partial "solution" to the problem can make

it easier or harder for the following agents to arrive at the correct solution. In particular, c) is needed because if every agent is guaranteed to make strict progress towards the solution, complementarity is always trivially satisfied.

The model we turn to is one where there is a discrete set of items and the objective is return the "best" one with high probability. For example, suppose the goal is to return the "correct" answer to a question, out of some possible set of answers. Agents are modeled by choice functions over the set of possible answers, and differ from each other in their choice function (thus, may have different probabilities for answers to the same question). Agent composition occurs when one agent (the *curator*) narrows down the set of answers into a smaller subset, from which another agent (the *decider*) picks. It is important to note that because agents are stochastic, agentic composition has positive probability of either being harmful or helpful, but our core goal will be the expected accuracy of the team.

One particular focus of this paper will be benefits to diversity: when does a diverse team have a higher chance of recovering the right answer than a homogenous one? While benefits to diversity has been studied extensively in the team design literature [Hong and Page, 2004], composition has been less studied, and it has been less studied in the context of agentic teams. Diversity has natural implications in both motivating examples we highlighted above. For human-algorithm collaboration, diversity relates to benefits to misalignment. While much work has studied methods of aligning LLMs with humans, if there is strict benefits to diversity, then such alignment may actually *harm* the performance of the team. In agentic teams, diversity relates to algorithmic monoculture [Bommasani et al., 2022, Goel et al., 2025, Kim et al., 2025, Kleinberg and Raghavan, 2021], which has shown the presence of substantial homogenization and correlation of errors. In scenarios where there are benefits to heterogeneity, our work motivates the development of LLMs that bring diversity to the current set of models on the market, beyond purely accuracy benefits.

In Example 1, we provide an example of a system that exhibits both complementarity as well as benefits to diversity.

Example 1. *Consider the following example: we have two agents, Agent 1, and Agent 2, and 3 total items. The goal is to recover item x_1, but both agent 1 and agent 2 are noisy, with nonzero probability mass on exactly two permutations (Figure 1).*

- *In isolation, agents 1 and 2 will pick the best item with probability 0.9 and 0.8, respectively.*
- *Suppose that the agents are composed: the curator presents its top 2 items, while the decider picks its favorite among those presented. Agent 1 composed with itself has identical accuracy 0.9: it always presents x_1, x_3 in the top $k = 2$ indices, and thus composition with itself keeps the same accuracy rate. (Identical reasoning for agent 2 shows that its accuracy is still the same when it is composed with itself:* 0.8*).*
- *With Agent 1 as the curator and Agent 2 as the decider, Agent 1 will always present items x_1, x_3, and Agent 2 will always pick item x_1 over x_3: thus, the combined system always has perfect accuracy.*

This system exhibits both complementarity (joint performance outperforms any agent in isolation) and benefits to diversity (joint system outperforms a homogeneous team composed of only type 1 agents or type 2 agents).

Our contributions: In Section 2 we begin by discussing related work, and in Section 3 we describe our formal model in more detail. For most of our work, we will focus on the setting where there is one "correct" answer, but we will discuss how our results generalize to the setting where some answers may have "partial credit" for being somewhat correct. In Section 4, we describe our theoretical contributions, when our two benchmarks of complementarity and benefits to diversity are satisfied.

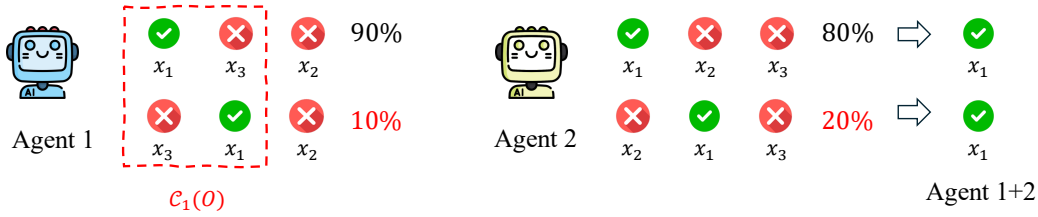


Fig. 1. Permutation models for Example 1.

For complementarity, we show that composing an agent with itself always leads to higher accuracy when that agent has the highest probability mass on the "best" answer, but could lead to harm if that is not satisfied. Also, we identify a pattern of heterogeneous composition: when two agents make different mistakes, they can complement each other by filtering out different wrong outcomes. For benefits to diversity, we show that the question is much more nuanced: in particular, there are types of "misalignment" between a pair of agents that improve team performance, and types of misalignment that harm it. In particular, we will show that misalignment between items that are incorrect always *helps* team performance, while misalignment with the best item *hurts* generally team performance. In general, misalignment between the decider and curator there is helpful given a specific relationship between the value of answers being inverted, how many other answers have value in between them, and the noise of the overall system.

In Section 5 we turn to our experimental analyses. Here, we look at the benchmark dataset MMLU [Hendrycks et al., 2020], which has multiple-choice questions across a range of topics. We pick a set of six open-source LLMs and extract the logits associated with every answer: this allows us to directly calculate the probability that this LLM would give every answer (A, B, C, D) at different temperatures. Then, we compose every LLM with every other LLM, where the curator model narrows down the set of answers, and the decider model picks the final answer from the resulting subset. With the goal of maximizing accuracy on MMLU, we study when the resulting curator/decider system satisfies either a) complementarity or b) benefits to diversity. While this system is substantially more complex, we show that insights from our theoretical analyses extend to this setting. For example, we show that composing a model with itself leads to increased accuracy whenever the model doesn't rank the "correct" answer last too often. We also show that our theoretical results for complementarity and benefits to diversity generally extend to this setting. Finally, we are also able to show novel insights from this experimental setting, such as conditions for when a pair of models are likely to be good "teammates". Finally, in Section 6 we conclude by discussing implications for both human-algorithm collaboration and agentic team design, as well as desired properties of individual LLMs.

2 Related Work

Human-algorithm collaboration. Our work relates to the general area of human-algorithm collaboration. In particular, there is a rich history of applied and empirical work in human-algorithm interaction: we refer interested readers to see [Kim, 2015, Lazar et al., 2017, MacKenzie, 2024, Preece et al., 1994] for textbook treatments. Specifically, our work relates to a growing literature using theoretical models to analyze humans interacting with algorithms. Some works study how to design algorithms to optimally assist humans [Bansal et al., 2021a, Brown and Agarwal, 2022, Chan et al., 2019, Donahue et al., 2022, Madras et al., 2018], including work incorporating models of human

cognitive biases, such as [Chen et al., 2025b, Ibrahim et al., 2025, Li et al., 2024]. Other work decides when human-algorithm teams perform well [Alur et al., 2023, 2024, Brown and Agarwal, 2024, Cowgill and Stevenson, 2020, Green and Chen, 2019, Greenwood et al., 2024, Guo et al., 2025, Peng et al., 2024, Steyvers et al., 2022], often relating to benchmarks such as complementarity (strict improvement over the human or algorithm alone defined in [Bansal et al., 2021b]). Some relevant literature reviews, taxonomies, and systematic studies in this space include [Gomez et al., 2025, Rastogi et al., 2022, Vaccaro et al., 2024].

Within human-algorithm collaboration, our work is most closely related to that of [Donahue et al., 2024], which studies a similar setting where an algorithmic tool presents a top k subset to a noisy human. A key difference in our work is that we allow humans to be *misaligned* with the algorithm and study when misalignment is helpful. In fact, many of our contributions strictly generalize theirs, such as generalizing the utility function beyond top item recovery. Other related areas include *conformal prediction*, which studies how to optimize a subset of items (e.g., ensuring that the best item is presented with high probability): see [Angelopoulos et al., 2023, Fontana et al., 2023] for a summary of work in this area. Within this space, some works focus on optimizing the set of items that are presented [Babbar et al., 2022, De Toni et al., 2024, Hullman et al., 2025, Straitouri and Rodriguez, 2023, Straitouri et al., 2022, Wang et al., 2022, Zhang et al., 2024] , while others include more empirical analyses of specific settings [Angelopoulos et al., 2020, Arnaiz-Rodriguez et al., 2025]. In general, these works do not consider settings with multiple humans who may be misaligned with each other: one exception is [Corvelo Benz and Gomez Rodriguez, 2025], which studies an empirical setting on when algorithmic alignment is a helpful property for tools assisting human decision-makers. There is also a line of work studying human-AI collaboration through information elicitation [Collina et al., 2025a, 2024, Corvelo Benz and Rodriguez, 2023, Corvelo Benz and Gomez Rodriguez, 2025, Steyvers et al., 2022]. [Steyvers et al., 2022] developed a Bayesian framework to combine the predictions of humans and AI algorithms with confidence scores. [Corvelo Benz and Rodriguez, 2023, Corvelo Benz and Gomez Rodriguez, 2025] then showed that if the confidence value aligns with the human's confidence, the collaboration will benefit the human's decision-making. [Collina et al., 2025a, 2024] studied the setting where each party holds different feature information and transmits their numerical prediction to each other at each round. [Collina et al., 2024] generalizes Aumann's Agreement Theorem by relaxing the rationality assumptions and introducing calibration-based conditions on each party that ensure efficient convergence of the conversation. Building on this, [Collina et al., 2025a] then proposes more communication-efficient protocols and removes the assumption of a common and correct prior. Finally, [Collina et al., 2025b] models a Bayesian communication setting where a human is relying on information from multiple AI tools that are misaligned with each other (and with her) to make a discrete decision. Compared to their work, our work does not model the interaction of the human and the algorithm as a multi-round process. The algorithm acts more like a curator, instead of a rational agent.

Other ranking/permutation models. The basic ranking model considered in our paper is defined over the number of inversions. A wide range of alternative permutation models has been studied in the literature. For example, the Bradley-Terry model [Bradley and Terry, 1952] and the Plackett-Luce model [Luce et al., 1959, Plackett, 1975] view a permutation as the consequence of a sequence of choices; the weighted Mallows model [Raman and Joachims, 2014] assigns weights to the inversions and defines the probability of a permutation by weighted Kendall-Tau distance. Moreover, [Awasthi et al., 2014, Liu and Moitra, 2018] studied learning a mixture of the Mallows model, which assumes the sampled permutation comes from a heterogeneous population.

3 Model

Stochastic agents. We consider a multi-agent collaboration model in which the goal is to select a single outcome from a finite set of outcomes O. Each agent is modeled as an *inherently stochastic decision-maker*—motivated, for example, by LLM-based agents with nonzero temperature, human decision-makers subject to cognitive noise, and exploratory algorithmic systems. An agent a can assist with selecting an outcome in one of two roles:

- **Decider:** selects a single outcome $\mathcal{D}_a(S)$ from a subset of outcomes $S \subseteq O$.
- **Curator:** reduces the set of outcomes by selecting an unordered subset $C_a(O) \subseteq O$.

Random choice behavior. We model agents' stochastic behavior using a *random choice rule*. For our theoretical results, we consider two standard stochastic models that induce such random choice behavior. For any set $S \subseteq O$, let $\mathbb{P}_a(x \mid S)$ denote the probability that agent a selects item x from S.

When acting as a decider, the agent selects a single outcome $\mathcal{D}_a(S) \sim \mathbb{P}_a(\cdot \mid S)$. When acting as a curator, an agent constructs a subset of outcomes via a sequential stochastic process without replacement. At each step, the probability of selecting an item depends on the set of items that remain available. The probability assigned to any subset $S \subseteq O$ is obtained by summing, over all permutations of its elements, the probabilities of selecting them in that order.

Definition 3.1 (Mallows model)**.** *The Mallows model is defined by a reference permutation σ_a over O and a temperature parameter $\phi \in (0,1)$. Given a set S with $|S| = k$, the probability of selecting an item $x \in S$ is*

$$\mathbb{P}_a[x \mid S] = \frac{1}{Z} \sum_{\sigma:\mathrm{Perm}(O)} \mathbb{1}[x \succ_\sigma S] \cdot \phi^{K_d(\sigma,\sigma_a^*)},$$

where $x \succ_\sigma S$ represents that x is located before other outcomes in S, $K_d(\sigma, \sigma_a^*)$ is the Kendall Tau *distance between σ and σ^* and Z is the normalization constant. The probability that exactly the items in S are sampled in the first k steps (in any order) is*

$$\Pr[C_a(O) = S] = \sum_{\sigma \in \mathrm{Perm}(S)} \prod_{i=1}^{|S|} \frac{\phi^{\mathrm{pos}_\sigma(i)-1}}{Z_{k-i+1}(\phi)},$$

where $\mathrm{pos}_\sigma(i)$ is the relative position of $\sigma(i)$ in the ordering induced by σ_a restricted to remaining outcomes $O \setminus \{\sigma(1), \ldots, \sigma(i-1)\}$ and $Z_k(\phi) = \sum_{j=0}^{k-1} \phi^j$.

Definition 3.2 (Plackett–Luce model)**.** *The Plackett–Luce model is defined by a collection of real-valued parameters $(v_i^a)_{i \in O}$ and a temperature parameter $\beta > 0$. Given a set S with $|S| = k$, the probability of selecting an item $x \in S$ is*

$$\mathbb{P}_a[x \mid S] = \frac{e^{v_x^a/\beta}}{\sum_{i \in S} e^{v_i^a/\beta}}.$$

The probability that exactly the items in S are sampled in the first k steps is

$$\Pr[C_a(O) = S] = \sum_{\sigma \in \mathrm{Perm}(S)} \prod_{i=1}^{|S|} \mathbb{P}_a[\sigma(i) \mid O \setminus \{\sigma(1), \ldots, \sigma(i-1)\}]. \tag{1}$$

True preferences and Imperfections. We assume there exists a common underlying utility function $u(\cdot) : O \to \mathbb{R}_{\geq 0}$, which induces a *true ranking* of outcomes shared by all agents, possibly with ties. Our main results focus on the *top-item recovery setting*, where $u(x) > 0$ for only one outcome $x \in O$, namely the most preferred outcome or the *top item*. This scenario is relevant when exactly one outcome is correct or usable, and all others yield no or negligible value, e.g., multiple-choice

questions, mathematical reasoning, verification, and medical diagnosis[1]. Agents, however, do not act directly on this true ranking: this is because we are assuming agents are inherently imperfect, and may incorrectly "believe" that items have different values than their true values. For example, consider an LLM that answers a question incorrectly, or a human whose beliefs about the value of an item are incorrect. In the Mallows model, an agent's reference permutation σ_a may differ from the true ranking induced by u; in the Plackett–Luce model, the parameters v_i^a may be misaligned with the true utilities $u(i)$. Thus, agents are imperfect in the sense that their stochastic decisions are centered around potentially incorrect internal representations of the true ranking.

Collaboration model. Two agents a and b can collaborate as follows: agent a acts as a curator and produces a subset $C_a(O)$, which is then passed to agent b, who acts as a decider. The resulting outcome is

$$O \xrightarrow{\text{curator } a} C_a(O) \xrightarrow{\text{decider } b} \mathcal{D}_b(C_a(O)) \triangleq x^*(a,b),$$

where $x^*(a,b)$ denotes the output by the agent composed of a and b. We also use $x^*(a)$ to denote the output by a single agent. Let $k = |C_a(O)|$ be the size of the curated outcomes.

Complementarity. A composition between agents a and b achieves *complementarity* if the expected utility of the composed agent is higher than that of each individual agent. Formally,

$$\mathbb{E}[u(x^*(a,b))] > \max(\mathbb{E}[u(x^*(a))], \mathbb{E}[u(x^*(b))]) \,. \tag{2}$$

Homogeneous and heterogeneous compositions. For the purposes of our investigations, we define two common types of collaborations.

Definition 3.3 (Homogeneous and heterogeneous composition)**.** *A composition between agents (a,b) is said to be* homogeneous *if their random choice models are identical, i.e., $\mathbb{P}_a = \mathbb{P}_b$. Otherwise, the composition is* heterogeneous*. To compare heterogeneous and homogeneous compositions, we define the* diversity gain *as the improvement achieved by a heterogeneous composition between a and b, relative to the better homogeneous composition:*

$$\text{Diversity gain}(a,b) = \mathbb{E}[u(x^*(a,b))] - \max(\mathbb{E}[u(x^*(a,a)], \mathbb{E}[u(x^*(b,b)]) \,. \tag{3}$$

For the ease of theoretical analysis, our theoretical results on heterogeneous composition focus on the setting focus on settings in which the random choice models share the same temperature parameter—ϕ under the Mallows model and β under the Plackett-Luce model—and in case the Plackett-Luce model, the same set of values, i.e., $\{v_i^a\}_i = \{v_i^b\}$. Note that this does not necessarily mean $v_i^a = v_i^b$ for any i. Two agents could have $v_1^a = 0.8, v_2^a = 0.4$ and $v_1^b = 0.4, v_2^b = 0.8$. Moreover, our experimental results in Section 5 generalize to the setting where two agents can have different temperatures and different sets of values. Also, we assume $\phi, \beta > 0$ and not all values are identical in the following content.

4 Conditions for Complementarity and Diversity Gain

In this section, we identify natural sufficient conditions under which (i) homogeneous compositions are guaranteed to have complementarity, and (ii) heterogeneous compositions provide additional gains, i.e., when diversity leads to higher performance than a homogeneous team. We now show our results for both the homogeneous and heterogeneous collaboration.

[1]See Appendix E for an extension where items may have "partial credit", even if they are not the best item.

4.1 Complementarity

Theorem 1. *A homogeneous collaboration always achieves **complementarity**, as long as the top outcome is ranked first in the agent's preference ranking.*

Proof sketch. We refer to the curator version of the agent as a^{C} and the decider version as a^{D}. We will begin by analyzing the setting where both agents have choice functions governed by the Mallows model, and then analyze settings where their choice function is given by Plackett-Luce.

First, assume both a^{C} and a^{D} follow the same Mallows model with accuracy parameter ϕ and central ranking σ_a^*, where the first outcome of σ_a^* is x_1. A basic property of the Mallows model implies that the probability of the agent picking x_1 from O when working alone is $1/Z_m(\phi)$. For the composed agent,

$$\Pr[x^*(a^{\mathsf{C}}, a^{\mathsf{D}}) = x_1] = \Pr[x_1 \in C_{a^{\mathsf{C}}}(O)] \cdot \Pr[\text{decider } a^{\mathsf{D}} \text{ picks } x_1 \mid x_1 \in C_{a^{\mathsf{C}}}(O)].$$

By [Awasthi et al., 2014], the first term equals $Z_k(\phi)/Z_m(\phi)$, where k is the number of curated outcomes. The key insight is that the second term is *strictly larger* than $1/Z_k(\phi)$. This follows from a fundamental conditioning property of the Mallows model, as characterized by Lemma 6: restricting a ranking drawn from a global Mallows distribution to a subset is *not* equivalent to drawing a fresh Mallows ranking on that subset. Even though there is no known closed-form formula for the conditional probability, we proved the following inequality,

$$\Pr[\text{decider } a^{\mathsf{D}} \text{ picks } x_1 \mid x_1 \in C_{a^{\mathsf{C}}}(O)] > \frac{1}{Z_k(\phi)},$$

which indicates that the restriction on a smaller set inherits additional bias on the top outcome.

In contrast, the conditioning property differs in the Plackett-Luce model—the process of picking the top outcomes can be interpreted as choosing the minimum from a set of independent Poisson random variables. Hence, we have a closed-form formula for the decider picking the top outcome from the subset $C_{a^{\mathsf{C}}}(O)$, leading to that

$$\begin{aligned}\Pr[x^*(a^{\mathsf{C}}, a^{\mathsf{D}}) = x_1] &= \Pr[x_1 \in C_{a^{\mathsf{C}}}(O)] \cdot \Pr[\text{decider } a^{\mathsf{D}} \text{ picks } x_1 \mid x_1 \in C_{a^{\mathsf{C}}}(O)] \\ &= \Pr[x_1 \in C_{a^{\mathsf{C}}}(O)] \cdot \frac{w_1}{\sum_{i \in C_{a^{\mathsf{C}}}(O)} w_i},\end{aligned}$$

where $w_i = \exp(v_i/\beta)$ is the weight induced by the given values v_i and randomness parameter β. Notice that the above probability can be reformulated as

$$\Pr[x^*(a^{\mathsf{C}}, a^{\mathsf{D}}) = x_1] = w_i \cdot \mathbb{E}\left[\frac{\mathbb{1}[x_1 \in C_{a^{\mathsf{C}}}(O)]}{\sum_{i \in C_{a^{\mathsf{C}}}(O)} w_i}\right].$$

Let the above term be r_1. Similarly, define r_i for any $i \in O$. Although it is non-trivial to provide a closed-form formula for r_i, we show r_i satisfies strict monotonicity: $r_1 \geq \cdots \geq r_m$. Also, by the linearity of expectation, $\sum_{i \in O} w_i \cdot r_i = 1$, which further implies that $w_1 \cdot r_1 > w_1/W$, where w_1/W is also the probability of a single agent picking x_1 from O. □

This then raises the question of whether a homogeneous collaboration still has complementarity when the top outcome is not ranked the highest. The following examples show that either could happen. Lemma 1 illustrates that, when the agent does not make a serious mistake in ranking the top outcome, collaboration still improves the overall performance. In contrast, lemma 1 also demonstrates that sometimes collaboration may instead amplify existing errors when the agent's preferences are substantially inaccurate, e.g., placing the true top outcome is ranked far below the top position.

We illustrate this using the Mallows model in the following examples, focusing on the case where the curator selects two outcomes. We defer the full proof for any $m > 3$ to Appendix C.

Lemma 1. *In the Mallows model, when $k = 2$ and $m > 3$, a homogeneous composition still achieves **complementarity** when the top outcome appears within the first* two *positions of the agent's preference ordering, but exhibits no complementarity when the top outcome is placed in the bottom two positions.*

Next, we turn to heterogeneous compositions, in which the curator and the decider have different underlying distributions for their actions. We are interested in identifying conditions under which the complementarity still holds. First, Theorem 2 shows that we can achieve complementarity when they make *different mistakes*: that is, when the sets of outcomes they rank above the best outcomes are disjoint.

Theorem 2 (Informal version of Theorem 5)**.** *A heterogeneous composition between agents a and b always has **complementarity** when their preference rankings satisfy*

$$\sigma_a^* = (S, x_1, \cdots) \quad \sigma_b^* = (T, x_1, \dots),$$

where $S \cap T = \emptyset$ and $|S| = |T|$. This holds for the Plackett-Luce Model and for the Mallows model under mild assumptions.

We illustrate the insight of Theorem 2 through the following example and defer the full proof to Appendix C.

Example 2 (Complementarity for heterogeneous agents making different mistakes)**.** *Consider two agents following Mallows models with the same accuracy parameter ϕ but different preference rankings of $\sigma_a^* = (x_i, x_1, \cdots)$ and $\sigma_b^* = (x_j, x_1, \dots)$, where $x_i \neq x_j$ and $x_i, x_j \neq x_1$. The two agents make different mistakes in ranking the top outcome x_1 and other outcomes—agent a ranks x_i above x_1, while agent b ranks x_j above x_1. For each of the two agents, when they act alone, they have a probability of $\phi/Z_m(\phi)$ to pick x_1 from O.*

However, when the two agents are composed, the two agents can complement each other by filtering out different wrong outcomes. In particular, although the decider b ranks x_1 behind x_j, the curator a ranks x_j behind x_1, and thus has a higher chance of presenting x_1 than x_j to the decider. In addition, although the curator a ranks x_1 behind x_i, the decider b ranks x_i behind x_1, and thus has a good probability of selecting x_1 when both are presented. Through a careful calculation (see Appendix C), we can show that the composed agent picks x_1 with a probability strictly larger than $\phi/Z_m(\phi)$, which means the heterogeneous composition achieves complementarity.

Note that $|S| = |T|$ is a necessary condition for the theorem to "always" hold. Intuitively, even though the two agents are allowed to make different mistakes, achieving complementarity requires neither agent to make more mistakes than the other. Otherwise, we have a counterexample where complementarity fails for $|S| \neq |T|$ in Example 3.

Example 3 (Non-complementarity when one makes more mistakes)**.** *Let $O = \{x_1, x_2, x_3\}$ Consider two agents following Mallows models with the same temperature ϕ and preference rankings*

$$\sigma_a^* = (x_1, x_2, x_3), \qquad \sigma_b^* = (x_2, x_1, x_3),$$

where the decider misranks the outcomes x_1 and x_2 while the curator ranks x_1 correctly. Suppose ϕ is small, which means the two models tend to give deterministic responses. When the curator makes the decision alone, it has a probability of $1/Z_3(\phi)$ of picking the top outcome x_1. When composed with agent b, it is likely to present the first two outcomes x_1 and x_2 to the decider. However, since x_2 is placed before x_1 in σ_b^*, the decider is likely to pick x_2 from the presented two outcomes. A direct calculation shows that the expected utility of the composed agent is lower than the curator's expected utility.

4.2 Diversity Gain

In this section, we aim to identify natural conditions under which the resulting *diversity gain* is positive. That is, we want to show when there are strict benefits to having the curator and decider be different agents (even if this means that one of them is less accurate). These results have implications for the design of human-algorithmic teams (when it is helpful or harmful for the AI to be *aligned* with the human), and for the design of agentic teams (when it is helpful for a team to be composed of multiple different models, even when some are weaker than others.

Following the pattern of the heterogeneous composition introduced in Theorem 2, we first show that the type of diversity increases the strength of the team. That is, when two agents make different types of mistakes, heterogeneous composition is more effective than homogeneous composition. Afterward, we study the optimal structure of heterogeneous composition—that is, the pair of curator and decider σ_a^* and σ_b^* that maximizes the expected utility. We find that the heterogeneous composition induced by the optimal preference pair achieves complementarity, attains the maximum expected utility, and exhibits a positive diversity gain.

Theorem 3. *A heterogeneous composition between agents a and b with preference rankings $\sigma_a^* = (S, x_1, \dots)$ and $\sigma_b^* = (T, x_1, \dots)$ with $S \cap T = \emptyset$ and $|S| = |T|$ always has a positive* ***diversity gain****.*

The proof of Theorem 3 relies on a key insight about the effect of the misalignment between the curator's and the decider's preference: Let σ_a^* and σ_b^* be their respective preference rankings. We say that the two agents are *aligned* on outcomes x_i and x_j if the relative ordering of x_i and x_j is the same under σ_a^* and σ_b^*. Lemma 3 shows that any misalignment between the non-top outcomes is beneficial for picking the top outcome, while any misalignment involving the top outcome is harmful for picking it. Based on this property, given a heterogeneous composition between agents a and b with $\sigma_a^* = (S, x_1, \dots)$ and $\sigma_b^* = (T, x_1, \dots)$, we can do a sequence of swaps without involving the top outcome x_1 to transform the composition into a homogeneous composition, and each swap strictly decreases the expected utility. We defer the full proof of Theorem 3 to Appendix C.

The above example shows that diversity can actually outperform homogeneous composition, which motivates the study of optimal heterogeneous composition. To define the space of the heterogeneous compositions, we first introduce the following notion:

Definition 4.1 (Preference swap)**.** *Given an agent a with a preference ranking σ_a^*, we defined the* swapped agent $a^{(i \leftrightarrow j)}$ *as the agent whose preference ranking is identical to agent σ_a^* except that the positions of x_i and x_j are interchanged. All the associated probabilities $\mathbb{P}_{a^{(i \leftrightarrow j)}}$ are then consistently redefined according to the new ranking.*

Let $\mathcal{S}(a)$ and $\mathcal{S}(b)$ be the set of agents constructed from a and b via sequences of preference swaps. Our goal is to find the optimal composition from $\Lambda(a, b) = \{(a', b') : a' \in \mathcal{S}(a), b' \in \mathcal{S}(b)\}$ that maximizes the expected utility.

Our first lemma formalizes how a pairwise disagreement between aligned agents—introduced via a preference swap—redistributes probability mass across outcomes. Specifically, swapping the relative order of two outcomes in the curator's ranking increases the chance of selecting outcomes other than the first outcome in the pair. Whether this redistribution is beneficial or harmful for expected utility depends on whether the swapped outcomes include the top-ranked alternative, as we make precise in Lemma 3.

Lemma 2. *Suppose agents a and b are aligned on outcomes x_i and x_j. Construct agent a' from a by swapping x_i and x_j in the preference ranking*[2]*. Then the team $T' = (a', b)$ has a higher probability of picking any outcome other than outcome x_i compared to the team $T = (a, b)$.*

[2]The swap is also known as the Kendall Tau swap.

While Lemma 2 establishes an outcome-wise probability comparison, its implications for utility depend critically on whether the swapped outcomes are valuable. In particular, disagreements on outcomes that are already suboptimal can be beneficial, as they help the team avoid over-committing to inferior choices. In contrast, disagreements involving the top outcome reduce the likelihood that the best alternative is selected and therefore harm performance. This distinction is formalized in the following lemma, which separates preference swaps into those that strictly improve expected utility and those that strictly decrease it.

Lemma 3. *Suppose agents a and b are aligned on outcomes x_i and x_j. Construct agent a' from a by swapping x_i and x_j in the preference ranking. The new composition (a', b) has a **higher** expected utility when neither x_i nor x_j is a top outcome, i.e., $\mathbb{E}[u(a', b)] > \mathbb{E}[u(a, b)]$, but **lower** expected utility when x_i is the top outcome, i.e., $\mathbb{E}[u(a', b)] < \mathbb{E}[u(a, b)]$.*

Note that, by the same argument, the conclusion also holds when swapping outcomes in the decider's preference ranking.

The above lemma highlights a sharp asymmetry: diversity introduced among value-less outcomes is always helpful, whereas diversity that disrupts agreement on the top outcome is always harmful. We illustrate this phenomenon concretely in the following example. We illustrate this effect elaborately in Example 4 below.

Example 4. *Consider the setting with three outcomes, where the decider's ground-truth ranking is $\sigma_b^* = (x_1, x_2, x_3)$, and there are four potential curators (Figure 2). Suppose outcome x_1 is the top outcome. By repeated applications of Lemma 3, we can derive relationships between the expected utility of each composition: for example, team (a_2, b) is better than team (a_1, b) because it is created by an inversion in value-less outcomes (x_2, x_3), and team (a_3, b) is better than (a_4, b) by identical reasoning. Similarly, (a_1, b) is better than (a_4, b) because Curator a_4 involves an inversion with the most valuable outcome (and team (a_2, b) is better than (a_3, b) by identical reasoning).*

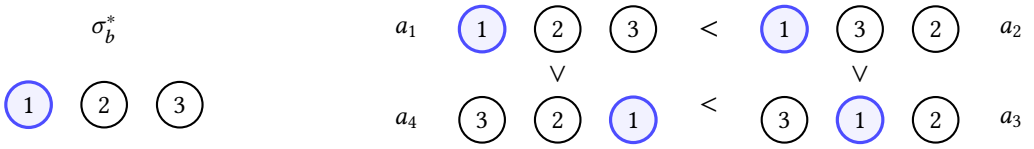


Fig. 2. Illustration of Lemma 3 with 3 outcomes, where the rounded node with number i represents outcome x_i and the top outcome is in blue.

Lemmas 2 and 3 are sufficient for us to identify the heterogeneous composition that achieves the maximum expected utility. Further, we also show that this diversity gain is strictly positive, i.e., composing agents who agree on the top outcome but are strongly misaligned on non-top outcomes yields strictly higher benefits than any homogeneous composition of the agents.

Theorem 4. *Given two agents a and b, consider all the compositions in $\Lambda(a, b)$. The composition with the highest expected utility is in the form of $\pi_a^* = (x_1, \overleftarrow{\sigma}_{-1})$ and $\pi_b^* = (x_1, \sigma_{-1})$, where σ_{-1} is an arbitrary ranking of $M \setminus \{x_1\}$ and $\overleftarrow{\sigma}_{-1}$ is the reverse ranking of σ_{-1}.*

Proof. Let π_a and π_b be the preference rankings of agents $a' \in \mathcal{S}(a)$ and $b' \in \mathcal{S}(b)$. If x_1 is not ranked first in π_b, we can apply Lemma 3 to swap x_1 with the outcome ranked first. This leads to a new team with a higher expected utility. Thus, we can assume that π_b ranks x_1 first. By repeatedly applying Lemma 3, the expected utility of team (a', b') is maximized when π_a ranks x_1 first, and the remaining outcomes are in the reverse order of π_b. □

From Theorem 1, we know that a homogeneous composition always achieves complementarity when the top outcome is ranked first. Combining this with Theorem 4, we obtain that the optimal heterogeneous composition not only achieves *complementarity* but also yields a *positive diversity gain*. In addition, following the characterization shown in Theorem 2, Theorem 4 further provides a qualitative description of which types of misalignment lead to the optimal composition.

5 Experiments

Next, we will extend our theoretical proofs with an experimental analysis. In particular, we will be interested in answering the following questions:

- **Q1:** When does composing two LLMs lead to complementarity? When does composing two different LLMs lead to better performance compared to using a single LLM alone and when does it not?
- **Q2:** Which kind of composition is the most effective? Are there ever benefits to diversity?
- **Q3:** Are there insights we can gain about collaboration in this applied setting that extend beyond our theoretical analyses?

We answer **Q1** in Section 5.2 and Section 5.3 and answer **Q2** in Section 5.3 and Section 5.4. Our experiments are based on the logits of real LLM agents, which generalize the assumptions used in our theoretical analysis. The empirical results in each subsection provide more practically relevant insights that complement the theory, thereby answering **Q3**.

5.1 Experimental Setup

Benchmark. We evaluate the performance of the LLMs on the Massive Multitask Language Understanding (MMLU) benchmark [Hendrycks et al., 2020], which is designed to assess broad knowledge and reasoning capabilities across a wide range of subjects. The dataset contains 57 subjects in total, and each subject consists of hundreds of multiple-choice questions at the high-school or undergraduate level. Each question is associated with a question statement Q, four options $\mathcal{A}$, and a correct answer $a^* \in \mathcal{A}$. The subjects are grouped into four categories: humanities, social sciences, STEM, and others. We select 10 subjects covering all the categories. We name a tuple $(Q, \mathcal{A}, a^*)$ as a *task*.

Model of LLMs. Recall that in Section 3 we described a model where each agent picks a response governed by a choice function, specifically Mallows or Plackett-Luce. LLMs satisfy this model quite well. Specifically, for an LLM given a specific prompt (e.g. a question Q), for each token in the LLM's output space, the LLM $\mathcal{L}$ assigns different logits ℓ_a to each option $a \in \mathcal{A}$, which are then converted to probabilities via a softmax function. If the temperature parameter is set to β, then after the softmax, the probability of choosing option a is given by:

$$p_{\mathcal{L}}(a \mid Q) = \frac{\exp(\ell_a/\beta)}{\sum_{a' \in \mathcal{A}} \exp(\ell_{a'}/\beta)}.$$

Thus, the Plackett–Luce model exactly models the LLM's single-token distribution over answer options: for example, whether the answer is A, B, C, or D. Model heterogeneity is given by differences in logits between models for the same query: the focus of this section is understanding empirical heterogeneity among LLMs. We conduct our experiments on a wide range of 6 open-source models. We selected these models primarily because they are open-source (allowing us to directly access logits) and because they are *not* extremely accurate on MMLU already, meaning that there are potential gains from teamwork. In the collaboration model, one LLM $\mathcal{L}_1$ is playing as the curator and selects the top-3 most probable answers according to its output distribution. The other LLM

$\mathcal{L}_2$ is playing as the decider and selects the final answer from the shortlisted answers provided by the curator. For more details on how we simulated composition in LLMs, see Appendix A.

Differences with theoretical assumptions. Finally, we want to flag two important differences with our theoretical model in Section 3 and our empirical experiments here.

First, our theoretical results in Section 4 look at a pair of agents A, B that have different fixed orders over outcomes, and study when composition for that pair is helpful. However, MMLU [Hendrycks et al., 2020] has many questions, and each LLM has a potentially different order over answers for each of them. Thus, composing two LLMs is akin to composing multiple individual models (one for each question) and seeing whether collaboration helps over the *average* of results.

The second difference is more subtle. In the theoretical setting, we model each agent as a choice function over items, where, even if agents are misaligned, they have the same probability of returning items of different ranks. That is, agents A and B may disagree on which item is ranked 1st, but they have the same probability of returning their top-ranked item from the overall set S. In our experimental setting, this is equivalent to saying that each LLM has logit value v_1 for its top-most answer, v_2 for its second answer, and so on, and while LLMs may disagree on *which* answer is first, they have a common *set* of logits. While this assumption ends up being necessary in the theoretical setting Section 3, this assumption clearly does not always hold for LLMs. For example, two models could have identical orders $[A, C, B, D]$, but have sharply different logits associated with each value. That is, one model could have logits $[1000, 10, 1, 0]$ and another could have logits $[3, 2, 1, 0]$: while these give the same *order* over answers, they clearly induce substantially different probability distributions.

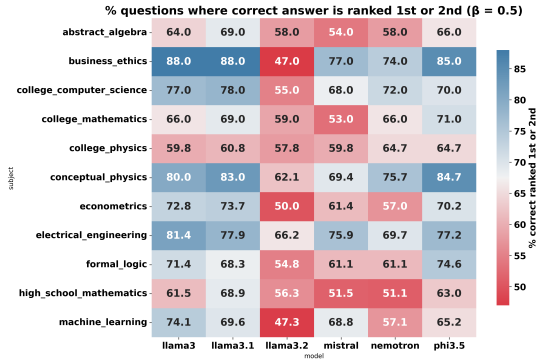
As a result of both of these differences, our experimental results are in a substantially more complex setting than our theoretical results. However, we will show that we can derive similar insights about benefits to agentic composition.

5.2 Benefits of Self-Collaboration

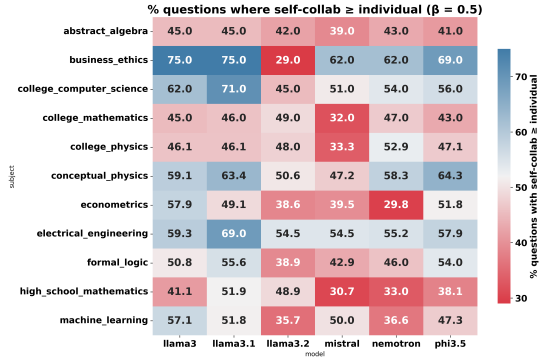
In Section 4, we have already shown that composing an agent with itself always improves accuracy, provided that the correct option is ranked first. We also showed in Lemma 1 that, under the Mallows model, questions for which the model ranks the correct answer first or second yield a positive collaboration gain.

This subsection further extends these theoretical results in the following ways. First, we compose different LLM agents with themselves across all subjects at a fixed temperature $\beta = 0.5$. In Figure 3, for each (subject, model) pair, we plot (Figure 3a) the percentage of questions with positive self-collaboration gain, and (Figure 3b) the percentage of questions for which the correct outcome is ranked first or second. (Recall that Lemma 1 showed that such a setting leads to strict benefits in the Mallows model with window size of 2.) We observe a meaningful relationship between these two quantities, suggesting that Lemma 1 may also generalize to the Plackett–Luce model with window size 3. In particular, when the fraction of questions for which the correct item is ranked first or second is high for a given (subject, model) pair, we are more likely to observe a benefit from self-collaboration. Indeed, the Pearson correlation between these two quantities is 0.8941, indicating a strong association.

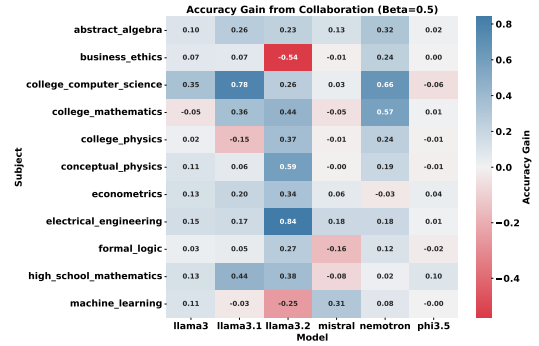
In Figure 3c, we plot the accuracy gain (%), which is defined as the average accuracy of the composed agent minus the average accuracy of a single agent. Note the difference from Figure 3b (right), which shows the *fraction* of questions with positive gains to self-composition, and Figure 3a, which shows the *average* benefits to composition (these differ because questions may vary in how helpful they are). As we discussed in Section 5.1, one gap between our theoretical and empirical settings is exactly this difference between the *number* of questions with benefits, and the



(a) Fraction of questions in which the model ranks the correct answer first or second.



(b) Fraction of questions with positive self-collaboration gain for each (subject, model) pair.



(c) Accuracy gain of self-collaboration

Fig. 3. Self-collaboration at $\beta = 0.5$ across subjects and models. The fraction of questions where both models had the correct output ranked first or second (Figure 3a)is strongly predictive of the *number* of questions for which self-collaboration is helpful (Figure 3b). The quantities are strongly correlated (Pearson $r = 0.8941$), consistent with the prediction that ranking the correct answer among the top two increases the likelihood of benefiting from self-collaboration (cf. Lemma 1). However, overall accuracy gain (Figure 3c) is less closely correlated with the *number* of questions that show benefit (Figure 3b).

magnitude of overall benefits. Nevertheless, we do see some connection between Figures 3b and 3c: settings with a large number of questions that benefit also tend to have positive accuracy gain from collaboration.

5.3 Composition gain across misalignment types

Next, we turn to settings beyond self-composition, namely, the composition of two distinct models A and B. Recall from Section 4 that composing two models can increase accuracy in certain regimes (yielding complementarity), and that there are also regimes in which accuracy strictly benefits from the models having different output distributions (yielding gains from diversity).

Recall from Section 5.1 that a key empirical complication is that two LLMs may induce different logit magnitudes for the same set of options, even when they agree on the *ranking* of those options. This issue is absent under self-composition, since an LLM $\mathcal{L}$ trivially shares its logits with itself; however, it generally arises when composing heterogeneous agents.

To relax the strict theoretical assumption that two models have identical logit values, we instead impose a weaker *similar correctness probability* condition. Fix a question Q with option set $\mathcal{A}$ (e.g., $\mathcal{A} = \{A, B, C, D\}$). For an LLM $\mathcal{L}$, let ℓ_a denote the logit assigned to option $a \in \mathcal{A}$. Given temperature β, the probability that $\mathcal{L}$ selects option a is

$$p_{\mathcal{L}}(a \mid Q) = \frac{\exp(\ell_a/\beta)}{\sum_{a' \in \mathcal{A}} \exp(\ell_{a'}/\beta)}.$$

Now consider two models $\mathcal{L}_1$ and $\mathcal{L}_2$ with logits $\{\ell_a\}_{a\in\mathcal{A}}$ and $\{\ell'_a\}_{a\in\mathcal{A}}$, respectively, on the same question Q. Let $o \in \mathcal{A}$ denote the correct option. We say that $\mathcal{L}_1$ and $\mathcal{L}_2$ have *similar correctness probability* on Q if their probabilities of selecting the correct option are within 10% of each other, i.e., $\left|p_{\mathcal{L}_1}(o \mid Q) - p_{\mathcal{L}_2}(o \mid Q)\right| \leq 0.1$. Accordingly, for each triplet (subject, $\mathcal{L}_1$, $\mathcal{L}_2$) such that $\mathcal{L}_2$ is more accurate than $\mathcal{L}_1$ on that subject—i.e., it has a higher average (over all questions) probability of selecting the correct answer—we restrict attention to questions Q satisfying the above condition.

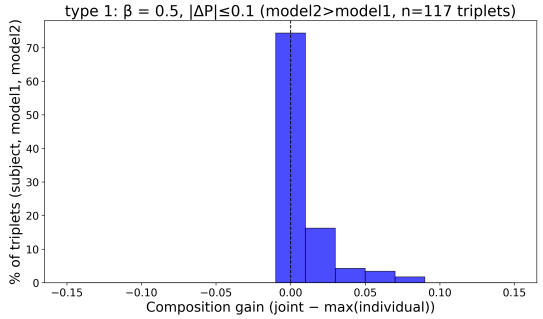
Having done this, we focus on specific types of misalignment for which we have proved related theoretical results.

(1) Type 1: both models ranked the correct output first. Theorem 1 shows that for homogeneous teams such a setting leads to composition gains.
(2) Type 2: both models ranked the same incorrect answer first. Lemma 1 studies a somewhat related setting, showing that with homogeneous composition, composition can be harmful when the top item is ranked low.
(3) Type 3: both models rank *different* incorrect answers first. Theorem 2 studies a related setting in showing benefits when two models make different mistakes.

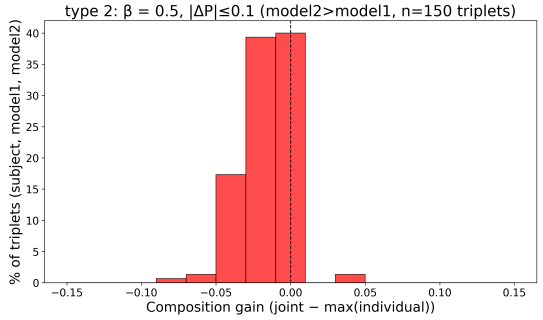
Figure 4 displays histograms of the benefits to compositional gain for such types of questions - that is, the increase in accuracy when two models are composed, and their patterns of errors falls into one of the three types. We would expect types 1 and 3 to show *benefits* (positive values), and type 2 to show *harms* (negative values): note that these patterns tend to hold empirically. This indicates that our theoretical results may be useful for indicating properties that lead to successful composition empirically.

5.4 Complementarity and benefits to diversity averaged over questions

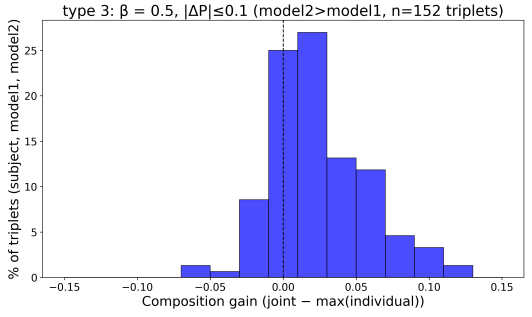
Finally, we consider the most general and challenging setting, in which we evaluate composition on *all* questions (rather than restricting to questions with similar correctness probability as in Section 5.3). Our goal in this subsection is to identify insights about the most effective ways to compose LLM agents. We study when our two benchmarks of complementarity and benefits to



(a) Type 1: computed using only the questions where both models had the correct output ranked first.



(b) Type 2: computed using only the questions where both models ranked the same wrong output first.



(c) Type 3: computed using only the questions where both models had different wrong outputs ranked first.

Fig. 4. **Histogram of composition gain across different misalignment types.** For each triplet $(\text{subject}, \text{model}_1, \text{model}_2)$ where model_2 performs better than model_1 on that subject, we restrict attention to questions whose correctness probabilities are comparable—specifically, those for which the softmax probability assigned to the correct answer by the two models differs by at most 10%. Using the resulting subset (when non-empty), we further categorize questions into misalignment Types 1, 2, and 3, and compute the corresponding *composition gain* for each group. We then plot a histogram of these composition gains for each misalignment type. Some triplets may contain no questions of a given type satisfying the probability-matching criterion; in such cases, that triplet is excluded from the histogram for that type. Accordingly, histograms are reported only for triplets with at least one qualifying question.

diversity are satisfied - for a given pair of models, when does the pair outperform a single model by itself, and when are there strict benefits to having a diverse team of models?

We conduct this analysis both on average over all subjects, and on a subject-by-subject level, for more granular analysis. Specifically, for any triplet (subject, Model 1, Model 2), we compute each model's individual accuracy on that subject, defined as the average expected utility—equivalently, the mean (over questions in the subject) of the model's probability of selecting the correct option. We also compute the composition accuracy on that subject, defined as the mean (over questions) probability that the composed procedure returns the correct option.

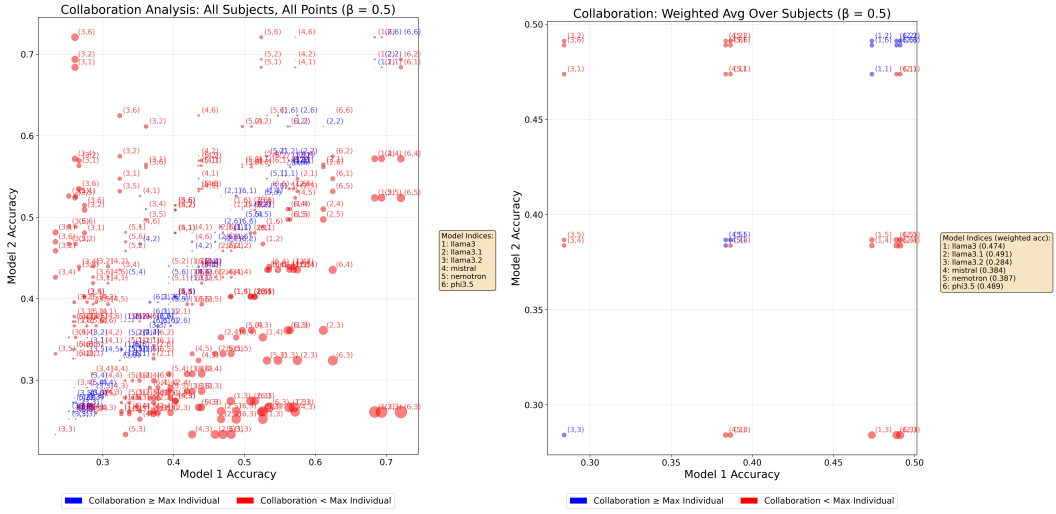
In Figure 5, the x-axis reports Model 1's individual accuracy and the y-axis reports Model 2's individual accuracy. In the left panel, each point corresponds to a triplet (subject, Model 1, Model 2). In the right panel, each point corresponds to a model pair (Model 1, Model 2), where the composition accuracy is aggregated across subjects using a question-count–weighted average. Points are colored blue (resp., red) when the composition gain is positive (resp., negative), where composition gain is defined relative to the best single-model baseline, i.e., $\max\{\text{Acc}(\text{Model 1}), \text{Acc}(\text{Model 2})\}$. Marker size is proportional to the magnitude of this gain. Figure 6 shows identical patterns, but for the stronger benchmark of *benefits to diversity*, where the joint team outperforms the strongest homogeneous team.

There are a few patterns we can see from Figures 5 and 6. First, note that achieving either benchmark is relatively uncommon: most of the dots are red, reflecting settings where either there are no benefits to heterogeneity or no benefits to a team at all. However, there are settings where collaboration is helpful. In general, these tend to occur when both models have comparable individual accuracy rates, and the second model (the decider) has higher accuracy. We view these results as suggesting settings that might be especially likely to benefit from composition – as well as flagging settings where composition is unlikely to be helpful, such as a large gap in individual accuracy rates.

6 Conclusion

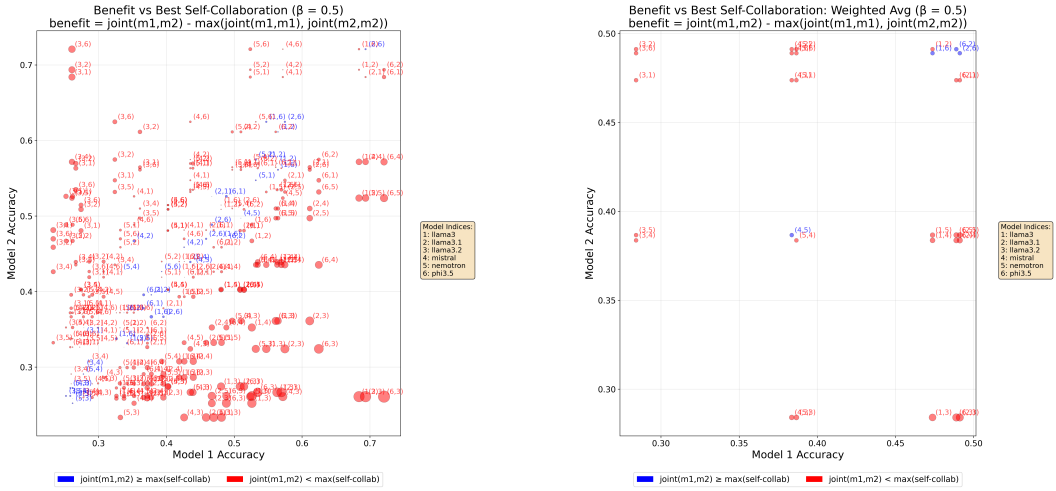
Contributions: In this work, we studied a setting with a team composed of two agents, and explored when this team a) achieves complementarity (strict benefits to being a team) and b) benefits to diversity (strict benefits to being a heterogeneous team, rather than a homogeneous one). We approached this question from both a theoretical and empirical lens: in Section 4, we proved conditions where composing two agents leads to strict improvements in accuracy, showing that there are conditions where misalignment between models can strictly improve the accuracy of the joint team. In Section 5 we extended these results to an applied experiment, where we looked at the varied performance of multiple models on the benchmark dataset MMLU [Hendrycks et al., 2020]. We showed that our core theoretical insights generalized here, such as specific types of misalignment that leads to stronger performance. We also explored more general properties that appear to influence whether a particular pair of LLMs is likely to be a strong team. Our results indicate that there can be benefits to diversity: for the human-AI collaboration setting, this suggests that minimizing *misalignment* between a user and an AI she is interacting with is not always optimal. For the agentic-team setting, benefits to model diversity suggest that there are real harms to algorithmic monoculture, and we may want model-providers to build models that make different kinds of errors.

Future directions: There are many natural extensions of our work. One direction would be to generalize the complexity of team structure (see [Chen et al., 2025a] for a survey of empirical work on agentic ensembles). For example, one generalization could allow composition between > 2 agents: it could be the case that deeper composition allows for even greater accuracy, or that it



(a) Each point for one subject (b) Average over subjects

Fig. 5. *Complementarity:* Composition performance across subjects and model pairs on the full question set. Axes report the individual accuracies of Model 1 (x-axis) and Model 2 (y-axis). **Left:** each point corresponds to a triplet (subject, Model 1, Model 2). **Right:** each point corresponds to a model pair (Model 1, Model 2), with accuracies aggregated across subjects using a question-count–weighted average. Color indicates the sign of the composition gain relative to the best single-model baseline $\max\{\text{Acc}(\text{Model 1}), \text{Acc}(\text{Model 2})\}$ (blue: positive, red: negative), and marker size is proportional to the magnitude of this gain.



(a) Each point for one subject. (b) Average over subjects.

Fig. 6. *Benefits to diversity:* Identical to Figure 5, but where color indicates the sign of the *diversity gain* (blue: positive, red: negative), and marker size is proportional to the magnitude of this gain.

allows for more chances for errors to be introduced. Alternatively, it would be interesting to study composition with other methods of multi-agent interaction such as model routing: for a given set of agents, when it is best to build a compositional chain of them, or to route tasks between agents based on which agent has greater accuracy for that task in particular? Other directions could study more complex types of tasks, such as giving "partial credit" for answers besides the best one. Additionally, it would be interesting to explore more complex types of collaboration beyond discrete item recovery, such as settings where models solve different sub-parts of the problem, which are then aggregated.

References

Rohan Alur, Loren Laine, Darrick Li, Manish Raghavan, Devavrat Shah, and Dennis Shung. 2023. Auditing for human expertise. *Advances in Neural Information Processing Systems* 36 (2023), 79439–79468.

Rohan Alur, Manish Raghavan, and Devavrat Shah. 2024. Human expertise in algorithmic prediction. *Advances in Neural Information Processing Systems* 37 (2024), 138088–138129.

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. 2020. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193* (2020).

Anastasios N Angelopoulos, Stephen Bates, et al. 2023. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning* 16, 4 (2023), 494–591.

Adrian Arnaiz-Rodriguez, Nina Corvelo Benz, Suhas Thejaswi, Nuria Oliver, and Manuel Gomez-Rodriguez. 2025. Towards Human-AI Complementarity in Matching Tasks. *arXiv preprint arXiv:2508.13285* (2025).

Pranjal Awasthi, Avrim Blum, Or Sheffet, and Aravindan Vijayaraghavan. 2014. Learning mixtures of ranking models. *Advances in Neural Information Processing Systems* 27 (2014).

Varun Babbar, Umang Bhatt, and Adrian Weller. 2022. On the Utility of Prediction Sets in Human-AI Teams. arXiv:2205.01411 [cs.AI]

Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. 2021a. Is the most accurate ai the best teammate? optimizing ai for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11405–11414.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021b. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–16.

Rachit Bansal, Bidisha Samanta, Siddharth Dalmia, Nitish Gupta, Sriram Ganapathy, Abhishek Bapna, Prateek Jain, and Partha Talukdar. 2024. Llm augmented llms: Expanding capabilities through composition. In *The Twelfth International Conference on Learning Representations*.

Niclas Boehmer, Piotr Faliszewski, and Sonja Kraiczy. 2023. Properties of the mallows model depending on the number of alternatives: a warning for an experimentalist. In *International Conference on Machine Learning*. PMLR, 2689–2711.

Rishi Bommasani, Kathleen A Creel, Ananya Kumar, Dan Jurafsky, and Percy S Liang. 2022. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *Advances in Neural Information Processing Systems* 35 (2022), 3663–3678.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* 39, 3/4 (1952), 324–345.

William Brown and Arpit Agarwal. 2022. Diversified recommendations for agents with adaptive preferences. *Advances in Neural Information Processing Systems* 35 (2022), 26066–26077.

William Brown and Arpit Agarwal. 2024. Online Recommendations for Agents with Discounted Adaptive Preferences. In *International Conference on Algorithmic Learning Theory*. PMLR, 244–281.

Lawrence Chan, Dylan Hadfield-Menell, Siddhartha Srinivasa, and Anca Dragan. 2019. The assistive multi-armed bandit. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 354–363.

Rex Chen, Ruiyi Wang, Norman Sadeh, and Fei Fang. 2025b. Missing Pieces: How Do Designs that Expose Uncertainty Longitudinally Impact Trust in AI Decision Aids? An In Situ Study of Gig Drivers. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 790–816.

Zhijun Chen, Jingzheng Li, Pengpeng Chen, Zhuoran Li, Kai Sun, Yuankai Luo, Qianren Mao, Ming Li, Likang Xiao, Dingqi Yang, et al. 2025a. Harnessing multiple large language models: A survey on llm ensemble. *arXiv preprint arXiv:2502.18036* (2025).

Natalie Collina, Ira Globus-Harris, Surbhi Goel, Varun Gupta, Aaron Roth, and Mirah Shi. 2025a. Collaborative Prediction: Tractable Information Aggregation via Agreement. *arXiv preprint arXiv:2504.06075* (2025).

Natalie Collina, Surbhi Goel, Varun Gupta, and Aaron Roth. 2024. Tractable Agreement Protocols. *arXiv preprint arXiv:2411.19791* (2024).

Natalie Collina, Surbhi Goel, Aaron Roth, Emily Ryu, and Mirah Shi. 2025b. Emergent Alignment via Competition. *arXiv preprint arXiv:2509.15090* (2025).

Nina Corvelo Benz and Manuel Rodriguez. 2023. Human-aligned calibration for ai-assisted decision making. *Advances in Neural Information Processing Systems* 36 (2023), 14609–14636.

Nina Laura Corvelo Benz and Manuel Gomez Rodriguez. 2025. Human-Alignment Influences the Utility of AI-assisted Decision Making. *arXiv preprint arXiv:2501.14035* (2025).

Bo Cowgill and Megan T Stevenson. 2020. Algorithmic social engineering. In *AEA Papers and Proceedings*, Vol. 110. 96–100.

Giovanni De Toni, Nastaran Okati, Suhas Thejaswi, Eleni Straitouri, and Manuel Gomez-Rodriguez. 2024. Towards human-ai complementarity with predictions sets. *arXiv preprint arXiv:2405.17544* (2024).

Kate Donahue, Alexandra Chouldechova, and Krishnaram Kenthapadi. 2022. Human-algorithm collaboration: Achieving complementarity and avoiding unfairness. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and*

Transparency. 1639–1656.

Kate Donahue, Sreenivas Gollapudi, and Kostas Kollias. 2024. When Are Two Lists Better than One?: Benefits and Harms in Joint Decision-making. arXiv:2308.11721 [cs.LG] https://arxiv.org/abs/2308.11721

Matteo Fontana, Gianluca Zeni, and Simone Vantini. 2023. Conformal prediction: a unified review of theory and new challenges. *Bernoulli* 29, 1 (2023), 1–23.

Shashwat Goel, Joschka Struber, Ilze Amanda Auzina, Karuna K Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge, and Jonas Geiping. 2025. Great models think alike and this undermines ai oversight. *arXiv preprint arXiv:2502.04313* (2025).

Catalina Gomez, Sue Min Cho, Shichang Ke, Chien-Ming Huang, and Mathias Unberath. 2025. Human-AI collaboration is not very collaborative yet: a taxonomy of interaction patterns in AI-assisted decision making from a systematic review. *Frontiers in Computer Science* 6 (2025), 1521066.

Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* (2019).

Sophie Greenwood, Karen Levy, Solon Barocas, Jon Kleinberg, and Hoda Heidari. 2024. Designing Algorithmic Delegates. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

Ziyang Guo, Yifan Wu, Jason Hartline, and Jessica Hullman. 2025. The Value of Information in Human-AI Decision-making. *arXiv preprint arXiv:2502.06152* (2025).

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300* (2020).

Lu Hong and Scott E Page. 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences* 101, 46 (2004), 16385–16389.

Zhongzhan Huang, Guoming Ling, Yupei Lin, Yandong Chen, Shanshan Zhong, Hefeng Wu, and Liang Lin. 2025. Routereval: A comprehensive benchmark for routing llms to explore model-level scaling up in llms. *arXiv preprint arXiv:2503.10657* (2025).

Jessica Hullman, Yifan Wu, Dawei Xie, Ziyang Guo, and Andrew Gelman. 2025. Conformal prediction and human decision making. *arXiv preprint arXiv:2503.11709* (2025).

Lujain Ibrahim, Katherine M Collins, Sunnie SY Kim, Anka Reuel, Max Lamparth, Kevin Feng, Lama Ahmad, Prajna Soni, Alia El Kattan, Merlin Stein, et al. 2025. Measuring and mitigating overreliance is necessary for building human-compatible AI. *arXiv preprint arXiv:2509.08010* (2025).

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. *arXiv preprint arXiv:2306.02561* (2023).

Elliot Kim, Avi Garg, Kenny Peng, and Nikhil Garg. 2025. Correlated Errors in Large Language Models. *arXiv preprint arXiv:2506.07962* (2025).

Gerard Jounghyun Kim. 2015. *Human-computer interaction: fundamentals and practice.* CRC press.

Jon Kleinberg and Manish Raghavan. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences* 118, 22 (2021), e2018340118.

Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. *Research methods in human-computer interaction.* Morgan Kaufmann.

Zhuoyan Li, Zhuoran Lu, and Ming Yin. 2024. Decoding ai's nudge: A unified framework to predict human behavior in ai-assisted decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 10083–10091.

Allen Liu and Ankur Moitra. 2018. Efficiently learning mixtures of mallows models. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 627–638.

R Duncan Luce et al. 1959. *Individual choice behavior*. Vol. 4. Wiley New York.

I. Scott MacKenzie. 2024. *Human-computer interaction: An empirical research perspective.* Elsevier.

David Madras, Toni Pitassi, and Richard Zemel. 2018. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/file/09d37c08f7b129e96277388757530c72-Paper.pdf

Colin L Mallows. 1957. Non-null ranking models. I. *Biometrika* 44, 1/2 (1957), 114–130.

Kenny Peng, Nikhil Garg, and Jon Kleinberg. 2024. A No Free Lunch Theorem for Human-AI Collaboration. *arXiv preprint arXiv:2411.15230* (2024).

Robin L Plackett. 1975. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics* 24, 2 (1975), 193–202.

Jenny Preece, Yvonne Rogers, Helen Sharp, David Benyon, Simon Holland, and Tom Carey. 1994. *Human-computer interaction.* Addison-Wesley Longman Ltd.

Karthik Raman and Thorsten Joachims. 2014. Methods for ordinal peer grading. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1037–1046.

Charvi Rastogi, Liu Leqi, Kenneth Holstein, and Hoda Heidari. 2022. A Unifying Framework for Combining Complementary Strengths of Humans and ML toward Better Predictive Decision-Making. *arXiv preprint arXiv:2204.10806* (2022).

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324* (2023).

Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. 2022. Bayesian modeling of human–AI complementarity. *Proceedings of the National Academy of Sciences* 119, 11 (2022), e2111547119.

Eleni Straitouri and Manuel Gomez Rodriguez. 2023. Designing Decision Support Systems Using Counterfactual Prediction Sets. arXiv:2306.03928 [cs.LG]

Eleni Straitouri, Lequn Wang, Nastaran Okati, and Manuel Gomez Rodriguez. 2022. Provably Improving Expert Predictions with Conformal Prediction. arXiv:2201.12006 [cs.LG]

Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. 2024. When combinations of humans and AI are useful: A systematic review and meta-analysis. *arXiv preprint arXiv:2405.06087* (2024).

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

Lequn Wang, Thorsten Joachims, and Manuel Gomez Rodriguez. 2022. Improving screening processes via calibrated subset selection. In *International Conference on Machine Learning*. PMLR, 22702–22726.

Dongping Zhang, Angelos Chatzimparmpas, Negar Kamali, and Jessica Hullman. 2024. Evaluating the utility of conformal prediction sets for ai-advised image labeling. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.

A Discussion on simulating composition of agents

As discussed in Section 5, we simulate composition as follows: For every question, we extract the logits for each model $\mathcal{L}_1$, $\mathcal{L}_2$ for answers to multiple choice questions (A, B, C, D), which allows us to trace out the probability distribution for each LLM for a range of temperatures β. We chose this method, rather than repeatedly sampling responses, because it is substantially faster and more direct, and allows us to study models that imperfectly solve the task at hand. Studying the performance of more powerful, closed-source models on more challenging tasks would be an interesting avenue for future work.

Then, we manually calculate the probability distribution of the joint system by composing the Plackett-Luce distributions induced by each model. That is, for every subset S of items of size 3, we calculate the probability of $\mathcal{L}_1$ returning that subset, and then the probability of $\mathcal{L}_2$ picking every element from the subset S according to Plackett-Luce.

Note that this second step is a simplifying assumption of how LLMs may behave empirically. To make this clearer, consider two queries, Q and Q', where both have identical questions, but Q' only provides 3 potential answers, while Q provides all 4. Our simplifying assumption is that the logits for an answer (e.g. A)) with Q' are identical to the logits for the same answer given the prompt Q. This is almost certainly incorrect: LLMs have been shown empirically to be sensitive to changes in prompts, including innocuous changes such as white space, formatting, and changes in order of how items are presented (e.g. [Sclar et al., 2023]). We chose to make this assumption for two primary reasons. First, to directly model $\mathcal{L}_2$'s true distribution to the modified query Q', we would need to calculate logits for each of the 6 models, for hundreds of questions within the dozens topics, the logits for each response, given each of 3 possible answers being presented. This would substantially increase the complexity of our analysis. Additionally, this complexity would obscure the main focus of our work: our goal is *not* to study the ways that LLMs may behave unpredictably to modified inputs, which has been studied elsewhere, but rather to understand for a specific stylized example when heterogeneity in errors could lead to a stronger team.

B Properties of Random Choice Models

B.1 Mallows Model

Let $\mathcal{D}(\pi^*, \phi)$ be a Mallows model with central ranking π^* and accuracy parameter $\exp(-\phi)$.

Lemma 4 (Proposition 3.8 of [Boehmer et al., 2023])**.** *In a Mallows distribution $\pi \sim \mathcal{D}(\pi^*, \phi)$, for any two distinct items $x_i, x_j \in M$ with $i < j$ and $k = j - i + 1$, the probability that x_i is before x_j in π is given by*

$$\Pr\left[x_i \succ_\pi x_j\right] = \frac{k}{1 - \exp(-\phi \cdot k)} - \frac{k-1}{1 - \exp(-\phi \cdot (k-1))}.$$

Lemma 5. *In a Mallows distribution $\pi \sim \mathcal{D}(\pi^*, \phi)$, given a subset S of items of size k, the probability of set S of being the first k items of π is given by*

$$\mathbb{P}[\pi[:k] = S] = \exp\left(-\phi \cdot \left(\sum\nolimits_{x_j \in S} j - \frac{k(k+1)}{2}\right)\right) \cdot \prod_{i=1}^{k} \frac{Z_i(\phi)}{Z_{m-i+1}(\phi)}.$$

Proof. Fix an order π_S of S. Below, we give the probability of π_S being the prefix of π. Denote the k items by $x_{j(1)}, \ldots, x_{j(k)}$ according order π_S. We insert the k items in order. At time of inserting the i-th item $x_{j(i)}$, consider the probability of locating $x_{j(i)}$ at the i-th place in π conditioned on $x_{j(1)}, \ldots, x_{j(i-1)}$ being located as the first $i-1$ items of π. Observe that the number of inserted items that are placed before $x_{j(i)}$ in the ground truth ranking π^* is equal to $|\{i' : j(i') < j(i) \wedge i' < i\}|$. Denote that number by $A(i)$. The current item $x_{j(i)}$ is the $(j(i) - A(i))$-th item among the set

of remaining items $M \setminus \{x_{j(t)}\}_{t=1}^{i-1}$. Then, according to the property of the Mallows model, the probability of $x_{j(i)}$ being placed in i-th place in π conditioned on the past insertion is given by

$$\mathbb{P}\left[\pi(i) = x_{j(i)} \mid \wedge_{t=1}^{i-1}\pi(t) = x_{j(t)}\right] = \frac{1}{Z_{m-i+1}(\phi)} \cdot \exp\left(-\phi \cdot (j(i) - A(i) - 1)\right) . \tag{4}$$

Based on it, the probability of π_S being the prefix of π is equal to the product of these conditional probabilities, which is given by

$$\begin{aligned}
\mathbb{P}\left[\wedge_{i=1}^{k}\pi(i) = x_{j(i)}\right] &= \prod_{i=1}^{k} \Pr\left[\pi(i) = x_{j(i)} \mid \wedge_{t=1}^{i-1}\pi(t) = x_{j(t)}\right] && \text{(By the chain rule)} \\
&= \prod_{i=1}^{k} \frac{\exp\left(-\phi \cdot (j(i) - A(i) - 1)\right)}{Z_{m-i+1}(\phi)} && \text{(By Equation (4))}
\end{aligned}$$

By direct calculation, we can see that

$$\mathbb{P}\left[\wedge_{i=1}^{k}\pi(i) = x_{j(i)}\right] = \exp\left(-\phi \cdot \sum_{i=1}^{k}(j(i) - i + i - 1 - A(i))\right) \cdot \prod_{i=1}^{k} \frac{1}{Z_{m-i+1}(\phi)} .$$

Notice that the term $i - 1 - A(i)$ is essentially the number of items that are placed before $x_{j(i)}$ in π but ranked after $x_{j(i)}$ in the ground-truth ranking π^*. Hence, each of them forms an inversion with $x_{j(i)}$. Thus, the sum of $i - 1 - A(i)$ over i equals to the number of inversions of π_S. Therefore, we have

$$\mathbb{P}\left[\wedge_{i=1}^{k}\pi(i) = x_{j(i)}\right] = \exp\left(-\phi \cdot \left(\sum_{x_j \in S} j - \frac{k(k+1)}{2}\right)\right) \cdot \exp(-\phi \cdot \mathrm{inv}(\pi_S)) \cdot \prod_{i=1}^{k} \frac{1}{Z_{m-i+1}(\phi)} . \tag{5}$$

By summing the above equation over all the permutations of items in S, we can obtain that

$$\mathbb{P}[\pi[:k] = S] = \sum_{\pi_S \in \mathfrak{S}(S)} \mathbb{P}\left[\wedge_{i=1}^{k}\pi(i) = x_{j(i)}\right] .$$

Then applying Equation (5), since the sum of $\exp(-\phi \cdot \mathrm{inv}(\pi_S))$ over all permutations of S is equal to $\prod_{i=1}^{k} Z_i(\phi)$ [Mallows, 1957], we then have

$$\mathbb{P}\left[\pi[:k] = S\right] = \exp\left(-\phi \cdot \left(\sum_{x_j \in S} j - \frac{k(k+1)}{2}\right)\right) \cdot \frac{\prod_{i=1}^{k} Z_i(\phi)}{\prod_{i=1}^{k} Z_{m-i+1}(\phi)} . \qquad \square$$

Lemma 6. *In a mallows distribution $\pi \sim \mathcal{D}(\pi^*, \phi)$, given a subset of items $S = \{x_{i(1)}, \ldots, x_{i(k)}\}$ with $i(1) < \ldots < i(k)$, the probability of $x_{i(1)}$ being placed before all other items of S in π satisfies*

$$\mathbb{P}_{\pi \sim \mathcal{D}(\pi^*, \phi)}\left[x_{i(1)} \succ_\pi S\right] \geq \frac{1}{Z_k(\phi)} .$$

Proof. We partition the set of permutations into k sets based on the relative position of $x_{i(1)}$. Let $\mathfrak{S}_t$ be the set of permutations where $x_{i(1)}$ is placed at the t-th relative position among all items of S. For any $\pi \in \mathfrak{S}_{t+1}$ with $t \leq k-1$, we map it to a permutation $\pi \in \mathfrak{S}_t$ by swapping $x_{i(1)}$ and the item that is placed at the t-th position among all items of S. It is a valid swapping with respect to π and π^* by definition. By [Donahue et al., 2024, Lemma 1], the probability of π' is at least $\exp(\phi)$ times the probability of π. By summing the inequality up over all permutations of $\mathfrak{S}_{t+1}$, we can conclude that $\mathbb{P}[\pi \in \mathfrak{S}_t] \geq \exp(\phi) \cdot \mathbb{P}[\pi \in \mathfrak{S}_{t+1}]$. Hence, by multiplying these inequalities,

$$\mathbb{P}[\pi \in \mathfrak{S}_1] \geq \exp(\phi \cdot (t-1)) \cdot \mathbb{P}[\pi \in \mathfrak{S}_t] .$$

Therefore, by summing it over all t, we have

$$\mathbb{P}[\pi \in \mathfrak{S}_1] \geq \frac{1}{\sum_{t=1}^{k} \exp(\phi \cdot (t-1))} \cdot \sum_{t=1}^{k} \mathbb{P}[\pi \in \mathfrak{S}_t] = \frac{1}{Z_k(\phi)}. \qquad \square$$

Lemma 7. *In a mallows distribution $\pi \sim \mathcal{D}(\pi^*, \phi)$ with $\pi^* = (x_1, \ldots, x_m)$, given a subset of items $S = \{x_{i(1)}, \ldots, x_{i(k)}\}$ with $i(1) < \ldots < i(k)$, for any s such that $1 \leq s \leq k$, the probability of one of $x_{i(1)}, \ldots, x_{i(s)}$ being the first among all items of S is at least*

$$\sum_{j=1}^{s} \mathbb{P}\left[x_{i(j)} \succ_\pi S\right] \geq \frac{Z_s(\phi)}{Z_k(\phi)}.$$

Proof. Similar to the above proof, let $\mathfrak{S}_j$ be the set of permutations where $x_{i(j)}$ is the first one among all the items in S. Then the left-hand side can be rewritten as $\sum_{j=1}^{s} \mathbb{P}[\pi \in \mathfrak{S}_j]$. Denote each term by $f(j)$ for $j = 1, \ldots, s$. Using the argument of valid-swapping-based mapping, we have $f(i) \geq \exp(\phi) \cdot f(i+1)$. Meanwhile, as $\sum_{i=1}^{k} f(i) = 1$, we have

$$\begin{aligned}
&\sum_{j=1}^{s} f(j) - \frac{Z_s(\phi)}{Z_k(\phi)} = \sum_{j=1}^{s} f(j) - \frac{Z_s(\phi)}{Z_k(\phi)} \cdot \sum_{j=1}^{k} f(j) \\
&\propto \sum_{j=1}^{s} f(j) \cdot (Z_k(\phi) - Z_s(\phi)) - \sum_{j=s+1}^{k} f(j) \cdot Z_s(\phi) && \text{(By multiplying } Z_k(\phi)) \\
&= \sum_{j=1}^{s} f(j) \cdot \exp(-\phi \cdot s) \cdot Z_{k-s}(\phi) - \sum_{j=s+1}^{k} f(j) \cdot Z_s(\phi) && (Z_k(\phi) - Z_s(\phi) = \tfrac{Z_{k-s}(\phi)}{\exp(\phi \cdot s)})
\end{aligned}$$

Since $Z_{k-s}(\phi)$ can be unfolded as $Z_{k-s}(\phi) = \sum_{t=1}^{k-s} \exp(-\phi(t-1))$, by changing the order of summation, we can obtain that

$$\begin{aligned}
\sum_{j=1}^{s} f(j) - \frac{Z_s(\phi)}{Z_k(\phi)} &\propto \sum_{t=1}^{k-s} \exp(-\phi \cdot (t-1)) \cdot \sum_{j=1}^{s} f(j) \cdot \exp(-\phi \cdot s) - \sum_{j=1}^{k} f(j) \cdot Z_s(\phi) \\
&= \sum_{t=s+1}^{k} \exp(-\phi \cdot (t-1)) \cdot \sum_{j=1}^{s} f(j) - \sum_{j=s+1}^{k} f(j) \cdot Z_s(\phi) \\
&= \sum_{t=s+1}^{k} \sum_{j=1}^{s} (f(j) \cdot \exp(-\phi \cdot (t-1))) - \sum_{j=s+1}^{k} f(j) \cdot Z_s(\phi) \\
&\geq \sum_{t=s+1}^{k} \sum_{j=1}^{s} f(t) \cdot \exp(-\phi \cdot (j-1)) - \sum_{j=s+1}^{k} f(j) \cdot Z_s(\phi) \qquad (f(i) \geq f(i+1)) \\
&= \sum_{t=s+1}^{k} f(t) \cdot Z_s(\phi) - \sum_{j=s+1}^{k} f(j) \cdot Z_s(\phi) = 0,
\end{aligned}$$

which concludes the proof. $\square$

B.2 Plackett-Luce Model

Denote $\mathcal{D}(\vec{w})$ as the Plackett-Luce distribution with parameter $\vec{w} = (w_1, \ldots, w_m)$.

Lemma 8. *Let $\pi \sim \mathcal{D}(\vec{w})$ be a random permutation drawn from the Plackett-Luce distribution with parameter $\vec{w} = (w_1, \ldots, w_m)$. For any two distinct items $x_i, x_j \in M$ and $S \subseteq M \setminus \{x_i, x_j\}$ with $w_i \geq w_j$ and $|S| = k - 1$, we have*

$$\frac{\mathbb{P}\left[\pi[:k] = S \cup \{x_i\}\right]}{\mathbb{P}\left[\pi[:k] = S \cup \{x_j\}\right]} \geq \frac{w_i}{w_j}.$$

Proof. Without loss of generality, we assume that $S = \{x_{s_1}, \ldots, x_{s_{k-1}}\}$ with $s_1 < \ldots < s_{k-1}$. Fix an arbitrary order $\pi_S = (x_{s_1}, \ldots, x_{s_{k-1}})$ of the set S. Since set $\pi[:k]$ is unordered, we can rewrite the probability over all the permutations of π_S:

$$\Pr[\pi[:k] = S \cup \{x_i\}] = \sum_{\pi_S} \sum_{t=1}^{k} \Pr\left[(x_{s_1}, \ldots, x_{s_{t-1}}, x_i, x_{s_t}, \ldots, x_{s_{k-1}})\right],$$

and analogously for x_j. Fix π_S and $t \in [k]$. Let

$$\tau_i^{(t)} = (x_{s_1}, \ldots, x_{s_{t-1}}, x_i, x_{s_t}, \ldots, x_{s_{k-1}}), \qquad \tau_j^{(t)} = (x_{s_1}, \ldots, x_{s_{t-1}}, x_j, x_{s_t}, \ldots, x_{s_{k-1}}).$$

We compare the probabilities of these two ordered prefixes under the Plackett–Luce model.

By the definition of the Plackett–Luce model, for the first $t-1$ steps, the two probabilities are identical, so these factors cancel in the ratio. At step t, we obtain w_i/w_j since the denominators are the same. Consider now the steps after t. At step $t+1$, we select x_{s_t} from the remaining items. The denominators differ as follows:

$$D_i = A + w_j, \qquad D_j = A + w_i,$$

where A is the total weight of all other remaining items. Since $w_i \geq w_j$, we have

$$D_i \leq D_j \implies \frac{1}{D_i} \geq \frac{1}{D_j}.$$

Hence, the probability of selecting x_{s_t} (and similarly each subsequent $x_{s_{t+1}}, \ldots, x_{s_{k-1}}$) is larger in the sequence $\tau_i^{(t)}$ than in $\tau_j^{(t)}$. Therefore, $\frac{\Pr[\tau_i^{(t)}]}{\Pr[\tau_j^{(t)}]} \geq \frac{w_i}{w_j}$.

Since this inequality holds for every order π_S and every t, summing over all such terms yields

$$\frac{\Pr[\pi[:k] = S \cup \{x_i\}]}{\Pr[\pi[:k] = S \cup \{x_j\}]} \geq \frac{w_i}{w_j},$$

where the inequality is strict for some π_S since not all the weights are equal. □

Lemma 9. *Let $\pi \sim \mathcal{D}(\vec{w})$ be a random permutation drawn from the Plackett-Luce distribution with parameter $\vec{w} = (w_1, \ldots, w_m)$. Let $S \subseteq M$ be a subset of items and $\pi_{|S}$ be the restriction of π to S. Then, $\pi_{|S}$ follows the Plackett-Luce distribution with parameter $\vec{w}_S = (w_i)_{x_i \in S}$.*

Proof. The result is an implication by the independence of irrelevant alternatives property of the Plackett-Luce model [Luce et al., 1959]. □

C Omitted Proofs of Section 4.1

Theorem 1. *A homogeneous collaboration always achieves **complementarity**, as long as the top outcome is ranked first in the agent's preference ranking.*

Proof. Denote the m outcomes by $O = \{x_1, \ldots, x_m\}$. Let k be the size of curated items, i.e., $k = |C_a(O)|$. We first consider the setting when the agent's decision-making follows a Mallow model with accuracy parameter ϕ and a central ranking σ^*. Without loss of generality, we assume $\sigma^* = (x_1, \ldots, x_m)$. By the basic property of the Mallows model, the probability of agent a choosing

item x_1 is $\mathbb{P}[\mathcal{D}_a(O) = x_1] = 1/Z_m(\phi)$. Then the probability of the composed agent $a + a$ choosing x_1 satisfies:

$$\begin{aligned}\mathbb{P}[x^*(a^{\mathsf{C}}, a^{\mathsf{D}}) = x_1] &= \mathbb{P}[x_1 \in C_a(O)] \cdot \mathbb{P}[\text{decider } a \text{ picks } x_1 \text{ from } C_a(O)] \\ &= \frac{Z_k(\phi)}{Z_m(\phi)} \cdot \mathbb{P}[\text{decider } a \text{ picks } x_1 \text{ from } C_a(O)] \qquad \text{(By [Awasthi et al., 2014])} \\ &\geq \frac{Z_k(\phi)}{Z_m(\phi)} \cdot \frac{1}{Z_k(\phi)} = \frac{1}{Z_m(\phi)}. \qquad \text{(By Lemma 6)}\end{aligned}$$

where the first inequality cannot be tight since $\phi > 0$ and $k < m$.

The second part proves the same conclusion holds for the Plackett-Luce model as well. Let $w_i = e_i^v$ and $W = \sum_{i=1}^m w_i$. When the agent makes the decision alone, it picks item x_1 with the probability of w_1/W. The probability of the composed agent $a + a$ choosing x_1 is given by

$$\begin{aligned}\mathbb{P}[x^*(a^{\mathsf{C}}, a^{\mathsf{D}}) = x_1] &= \sum_{S \subseteq O : x_1 \in S} \mathbb{P}[C_a(O) = S] \cdot \mathbb{P}[\text{decider } b \text{ picks } x_1 \text{ from } C_a(O)] \\ &= \sum_{S \subseteq O : x_1 \in S} \mathbb{P}[C_a(O) = S] \cdot \frac{w_1}{\sum_{o \in S} w_o} \\ &= w_1 \cdot \mathbb{E}\left[\frac{\mathbb{1}[x_1 \in C_a(O)]}{\sum_{o \in C_a(O)} w_o}\right]\end{aligned}$$

Define r_i as the expectation of the ratio $\frac{\mathbb{1}[x_i \in C_a(O)]}{\sum_{o \in C_a(O)} w_o}$. Next, we prove that $r_i \geq r_j$ for any $i \leq j$. The difference between the two ratios r_i and r_j is given by

$$r_i - r_j = \sum_S \mathbb{P}[C_a(O) = S] \cdot \left(\frac{\mathbb{1}[x_i \in C_a(O)]}{\sum_{o \in S} w_o} - \frac{\mathbb{1}[x_j \in C_a(O)]}{\sum_{o \in S} w_o}\right)$$

By removing the sets including both x_i and x_j, we have

$$\begin{aligned}r_i - r_j &= \sum_{S \subseteq O \setminus \{x_i, x_j\}} \frac{\mathbb{P}[C_a(O) = S \cup \{i\}]}{\sum_{o \in S} w_o + w_i} - \frac{\mathbb{P}[C_a(O) = S \cup \{j\}]}{\sum_{o \in S} w_o + w_j} \\ &\geq \sum_{S \subseteq O \setminus \{x_i, x_j\}} \mathbb{P}[C_a(O) = S \cup \{j\}] \cdot \left(\frac{w_i}{w_j} \cdot \frac{1}{\sum_{o \in S} w_o + w_i} - \frac{1}{\sum_{o \in S} w_o + w_j}\right) \quad \text{(By Lemma 8)}\end{aligned}$$

Since $w_i \geq w_j$, then each inner term is nonnegative, and the above expression can only be zero when $w_i = w_j$. Therefore, we have

$$\begin{aligned}\mathbb{P}[x^*(a^{\mathsf{C}}, a^{\mathsf{D}}) = x_1] &= \frac{w_1}{W} \cdot r_1 \cdot W \\ &\geq \frac{w_1}{W} \cdot \sum_{i=1}^m r_i \cdot w_i \qquad \text{(by the monotonicity of } r_i\text{)} \\ &= \frac{w_1}{W} \cdot \sum_{i=1}^m \mathbb{E}\left[\frac{w_i \cdot \mathbb{1}[x_i \in C_a(O)]}{\sum_{i \in C_a(O)} w_i}\right] \\ &= \frac{w_1}{W} = \mathbb{P}[a \text{ chooses } x_1],\end{aligned}$$

where the inequality cannot be tight since w_i are not identical. □

Lemma 1. *In the Mallows model, when $k = 2$ and $m > 3$, a homogeneous composition still achieves **complementarity** when the top outcome appears within the first* two *positions of the agent's preference ordering, but exhibits no complementarity when the top outcome is placed in the bottom two positions.*

Proof. We start with the case when the top item x_1 is ranked second in the agent's preference ranking. Without loss of generality, let $\sigma_a^* = (x_2, x_1, x_3, \dots, x_m)$. When the agent a makes the decision alone, a picks the top item x_1 with probability of $\phi/Z_m(\phi)$. Now, consider the homogeneous collaboration where both the curator and the decider follow this same Mallows distribution, and the curator presents the top two outcomes to the decider. The probability that the composed agent picks x_1 is

$$\begin{aligned}
&\sum_{x_o \in \{x_2, \dots x_m\}} \mathbb{P}[C_{a^C}(O) = \{x_1, x_o\}] \cdot \mathbb{P}_a[x_1 \mid \{x_1, x_o\}] \\
&= \frac{(1+\phi)}{Z_m(\phi)Z_{m-1}(\phi)} \cdot \frac{\phi}{1+\phi} + \sum_{t=1}^{m-2} \frac{\phi^{t+1}(1+\phi)}{Z_m(\phi)Z_{m-1}(\phi)} \left(\frac{t+1}{1-\phi^{t+1}} - \frac{t}{1-\phi^t} \right) && \text{(By Lemma 4)}
\end{aligned}$$

Notice that the comparison probability $\frac{t+1}{1-\phi^{t+1}} - \frac{t}{1-\phi^t}$ increases monotonically with t by the property of the Mallows model. Hence, we have

$$\frac{t+1}{1-\phi^{t+1}} - \frac{t}{1-\phi^t} \geq \frac{2}{1-\phi^2} - \frac{1}{1-\phi} = \frac{1}{1+\phi} \qquad \text{(since } m \geq 4\text{)}$$

Therefore, we have

$$\begin{aligned}
\mathbb{P}[x^*(a^C, a^D) = x_1] &> \frac{\phi}{Z_m(\phi)Z_{m-1}(\phi)} + \sum_{t=1}^{m-2} \frac{\phi^{t+1}}{Z_m(\phi)Z_{m-1}(\phi)} \\
&= \frac{\phi}{Z_m(\phi)} = \mathbb{P}[x^*(a) = x_1],
\end{aligned}$$

which implies that the collaboration gain is still positive.

Lastly, we prove that the complementarity does not hold when the top item falls in the bottom two positions. Let $\sigma_a^* = (x_2, x_3, \dots, x_{m-1}, x_1, x_m)$. The probability that the composed agent picks x_1 is given by

$$\begin{aligned}
&\sum_{x_o \in \{x_2, \dots x_m\}} \mathbb{P}[C_{a^C}(O) = \{x_1, x_o\}] \cdot \mathbb{P}_a[x_1 \mid \{x_1, x_o\}] \\
&= \sum_{t=1}^{m-2} \frac{\phi^{t+m-4}(1+\phi)}{Z_m(\phi)Z_{m-1}(\phi)} \cdot \mathbb{P}_a[x_1 \mid \{x_1, x_o\}] + \frac{\phi^{2m-4}(1+\phi)}{Z_m(\phi)Z_{m-1}(\phi)} \frac{1}{1+\phi} \\
&< \sum_{t=1}^{m-2} \frac{\phi^{t+m-4}(1+\phi)}{Z_m(\phi)Z_{m-1}(\phi)} \cdot \frac{\phi}{1+\phi} + \frac{\phi^{2m-4}(1+\phi)}{Z_m(\phi)Z_{m-1}(\phi)} \frac{1}{1+\phi} && \text{(since } m > 3\text{)} \\
&= \frac{\phi^{m-2}}{Z_m(\phi)} = \mathbb{P}[x^*(a) = x_1],
\end{aligned}$$

which implies the non-complementarity. □

Theorem 5. *A heterogeneous composition between agents a and b always has complementarity when their preference rankings satisfy*

$$\sigma_a^* = (S, x_1, \cdots) \quad \sigma_b^* = (T, x_1, \dots),$$

where $S \cap T = \emptyset$ and $|S| = |T|$. This holds for the Mallows model when S and T are singletons and $k = 2$, and for the Plackett-Luce Model in the general case.

Proof. We start by showing that the result holds for the Mallows model. Without loss of generality, we assume $\sigma_a^* = (x_2, x_1, \dots)$ and $\sigma_b^* = (x_3, x_1, \dots)$. When the agent a works alone, it picks the top item x_1 with probability of $\phi/Z_m(\phi)$. Now, consider the heterogeneous collaboration

where both the curator and the decider follow this same Mallows distribution, and the curator presents the top two outcomes to the decider. Let t be the index of x_3 in σ_a^*, The probability that the composed agent picks x_1 is

$$\sum_{x_o \in \{x_2, x_3, \ldots x_m\}} \mathbb{P}[C_{a^{\mathrm{C}}}(O) = \{x_1, x_o\}] \cdot \mathbb{P}_b[x_1 \mid \{x_1, x_o\}]$$

For outcome x_2, we have $\mathbb{P}_b[x_1 \mid \{x_1, x_2\}] \geq \frac{1}{1+\phi}$ and $\mathbb{P}[C_{a^{\mathrm{C}}}(O) = \{x_1, x_2\}] = \frac{(1+\phi)}{Z_m(\phi)Z_{m-1}(\phi)}$. Hence, the product is at least $\frac{1}{Z_m(\phi)Z_{m-1}(\phi)}$. For outcome x_3, we have $\mathbb{P}_b[x_1 \mid \{x_1, x_3\}] = \frac{\phi}{1+\phi}$ and $\mathbb{P}[C_{a^{\mathrm{C}}}(O) = \{x_1, x_3\}] = \frac{\phi^{t-1}(1+\phi)}{Z_m(\phi)Z_{m-1}(\phi)}$. Hence, the product is at least $\frac{\phi^t}{Z_m(\phi)Z_{m-1}(\phi)}$. For the j-th outcome in σ_a^* (without loss of generality, say x_j), we have $\mathbb{P}_b[x_1 \mid \{x_1, x_j\}] \geq \frac{1}{1+\phi}$ and $\mathbb{P}[C_{a^{\mathrm{C}}}(O) = \{x_1, x_j\}] = \frac{\phi^{j-1}(1+\phi)}{Z_m(\phi)Z_{m-1}(\phi)}$. The product is at least $\frac{\phi^{j-1}}{Z_m(\phi)Z_{m-1}(\phi)}$. Therefore, we have

$$\begin{aligned}
\mathbb{P}[x^*(a^{\mathrm{C}}, a^{\mathrm{D}}) = x_1] &\geq \frac{1}{Z_m(\phi)Z_{m-1}(\phi)} + \sum_{j=3, j\neq t}^{m} \frac{\phi^{j-1}}{Z_m(\phi)Z_{m-1}(\phi)} + \frac{\phi^t}{Z_m(\phi)Z_{m-1}(\phi)} \\
&> \frac{\phi}{Z_m(\phi)} = \mathbb{P}[x^*(a) = x_1],
\end{aligned}$$

which concludes the proof for the Mallows model.

Next, we prove it also holds for the Plackett-Luce model without any additional assumptions. According to Lemma 3, alignment on outcomes other than the top item x_1 hurts the expected utility of the composed agent. Since S is before any other outcomes in $O \setminus S$ in σ_a^*, and T is before any other outcomes in $O \setminus T$ in σ_b^*, we can assume without loss of generality that $\sigma_a^* = (S, x_1, T, \ldots)$ and $\sigma_b^* = (T, x_1, S, \ldots)$. Otherwise, we can swap the positions of outcomes in S and T in σ_a^* and σ_b^* without increasing the overall expected utility. Since $|S| = |T|$, we can define a bijection $f : S \to T$ between the items in S and T. For a set $C \subseteq S$, we define $f(C) = \{f(x) : x \in C\}$. Let C be the curated set by the curator a, $C^S = C \cap S$ be the intersection of C and S, and $C^T = C \cap T$ be the intersection of C and T. Let $C^O = C \setminus (C^S \cup C^T)$ be the remaining items in C. The probability of the composed agent picking x_1 is given by

$$\sum_{C \subseteq O : x_1 \in C} \mathbb{P}[C_a(O) = C] \cdot \mathbb{P}_b[x_1 \mid C] = \sum_{C \subseteq O : x_1 \in C} \mathbb{P}[C_a(O) = C^S \cup C^T \cup C^O \cup \{x_i\}] \cdot \frac{\exp(v_1^b/\beta)}{\sum_{x \in C} \exp(v_x^b/\beta)}$$

Since the two agents follow the Plackett-Luce model with the same set of values, we have

$$\mathbb{P}[C_a(O) = C^S \cup C^T \cup C^O \cup \{x_i\}] = \mathbb{P}[C_b(O) = f(C^S) \cup f^{-1}(C^T) \cup C^O \cup \{x_i\}]$$

Thus, the above probability can be rewritten as

$$\mathbb{P}[x^*(a^{\mathrm{C}}, a^{\mathrm{D}}) = x_1] = \sum_{C \subseteq O : x_1 \in C} \mathbb{P}[C_b(O) = f(C^S) \cup f^{-1}(C^T) \cup C^O \cup \{x_i\}] \cdot \frac{\exp(v_1^b/\beta)}{\sum_{x \in C} \exp(v_x^b/\beta)} \quad (6)$$

Also, when the decider makes the decision alone, the probability can be rewritten as

$$\begin{aligned}
\mathbb{P}[x^*(b) = x_1] &= \sum_{C \subseteq O : x_1 \in C} \mathbb{P}[C_b(O) = C] \cdot \frac{\exp(v_1^b/\beta)}{\sum_{x \in C} \exp(v_x^b/\beta)} \\
&= \sum_{C \subseteq O : x_1 \in C} \mathbb{P}[C_b(O) = C^S \cup C^T \cup C^O \cup \{x_i\}] \cdot \frac{\exp(v_1^b/\beta)}{\sum_{x \in C} \exp(v_x^b/\beta)} \quad (7)
\end{aligned}$$

Therefore, it suffices to prove (6) is strictly larger than (7). To this end, we fix the set C^O and the item x_i, and compare the sum of terms $C = C^S \cup C^T \cup C^O \cup \{x_i\}$ with the same $f(C^S) \cup C^T$ and $f(S) \cap C^T$. Formally, the family of these sets can be represented as

$$F^S(S_1, T_1) = S_1 \cup (C^S \cap f^{-1}(C^T)) \cup f^{-1}(T_1))$$
$$F^T(S_1, T_1) = (f(C^S \setminus S_1 \setminus f^{-1}(C^T)) \cup (f(C^S) \cap C^T) \cap (C^T \setminus T_1 \setminus f(S))$$
$$\mathcal{S}(C^O, x_i) = \left\{F^S(S_1, T_1) \cup F^T(S_1, T_1) \cup C^O \cup \{x_i\} : S_1 \subseteq C^S \setminus f^{-1}(C^T), T_1 \subseteq C^T \setminus f(C^S)\right\}$$

It can be verified that,

$$\sum_{C \in \mathcal{S}(C^O, x_i)} \mathbb{P}[C_b(O) = f(C^S) \cup f^{-1}(C^T) \cup C^O \cup \{x_i\}] = \sum_{C \in \mathcal{S}(C^O, x_i)} \mathbb{P}[C_b(O) = C]$$

Also, since T is ranked on the top of σ_b^*, each term on the left-hand side is monotone increasing on S_1 and T_1, while each term on the right-hand side is monotone decreasing on S_1 and T_1. However, since T is ranked on the top of σ_b^*, we can find the choice probability $\frac{\exp(v_1^b/\beta)}{\sum_{x \in C} \exp(v_x^b/\beta)}$ is monotone decreasing on S_1 and T_1 as well. Therefore, by applying the rearrangement inequality, we have

$$\sum_{C \in \mathcal{S}(C^O, x_i)} \mathbb{P}[C_b(O) = f(C^S) \cup f^{-1}(C^T) \cup C^O \cup \{x_i\}] \cdot \frac{\exp(v_1^b/\beta)}{\sum_{x \in C} \exp(v_x^b/\beta)}$$
$$> \sum_{C \in \mathcal{S}(C^O, x_i)} \mathbb{P}[C_b(O) = C] \cdot \frac{\exp(v_1^b/\beta)}{\sum_{x \in C} \exp(v_x^b/\beta)}$$

Summing over all C^O and x_i concludes the proof. □

D Omitted Proofs of Section 4.2

Lemma 2. *Suppose agents a and b are aligned on outcomes x_i and x_j. Construct agent a' from a by swapping x_i and x_j in the preference ranking[3]. Then the team $T' = (a', b)$ has a higher probability of picking any outcome other than outcome x_i compared to the team $T = (a, b)$.*

Proof. We first prove that team T has a higher probability of selecting any item x_r other than item x_i and x_j. Denote by x_T and x'_T the items picked by team T and T'. Thus, the probabilities of x_r being picked by the human are given by

$$\mathbb{P}[x_T = x_r] = \sum_{S : x_r \in S} \mathbb{P}_a\left[C_a(O) = S\right] \times \mathbb{P}_b\left[x_r \mid S\right], \tag{Prob 1}$$

$$\mathbb{P}[x'_T = x_r] = \sum_{S : x_r \in S} \mathbb{P}_{a'}\left[C_{a'}(O) = S\right] \times \mathbb{P}_b\left[x_r \mid S\right], \tag{Prob 2}$$

We compare Prob 1 and Prob 2 term-by-term. For any set S such that $\{x_i, x_j\} \subseteq S$ or $\{x_i, x_j\} \cap S = \emptyset$, the relabeling swap between x_i and x_j implies that $\mathbb{P}_a[C_a(O) = S] = \mathbb{P}_{a'}[C_{a'}(O) = S]$. Thus, these terms cancel out in the difference $\mathbb{P}[x_T = x_r] - \mathbb{P}[x'_T = x_r]$.

It suffices to consider sets containing exactly one of $\{x_i, x_j\}$. Let S^i be a set S containing x_i without x_j. Define $S^j = (S^i \setminus \{x_i\}) \cup \{x_j\}$. Next, we show that the following inequality holds for every constructed pair (S^i, S^j),

$$\sum_{S \in (S^i, S^j)} \mathbb{P}_a\left[C_a(O) = S\right] \mathbb{P}_b\left[x_r \mid S\right] \leq \sum_{S \in (S^i, S^j)} \mathbb{P}_{a'}\left[C_{a'}(O) = S\right] \mathbb{P}_b\left[x_r \mid S\right]. \tag{InEq (1)}$$

[3]The swap is also known as the Kendall Tau swap.

As x_i is better than x_j in agent b's preference ranking σ_b^*, by Lemma [REF], it is easier for the agent b to rank x_r before x_j. Formally, $\mathbb{P}_b[x_r \mid S^j] \geq \mathbb{P}_b[x_r \mid S^i]$. Also, since the relabeling does not change the structure of the distribution, we have

$$\mathbb{P}_a\left[C_a(O) = S^i\right] = \mathbb{P}_{a'}\left[C_{a'}(O) = S^j\right], \quad \mathbb{P}_a\left[C_a(O) = S^j\right] = \mathbb{P}_{a'}\left[C_{a'}(O) = S^i\right]$$

Meanwhile, as x_i is placed before x_j in σ_a^*, by Lemma 5 and Lemma 8, we have

$$\mathbb{P}_a\left[C_a(O) = S^i\right] > \mathbb{P}_a\left[C_a(O) = S^j\right]$$

By applying the rearrangement inequality, then InEq (1) holds.

We next consider the change of the probabilities of picking x_i or x_j. First, we show that the probability of picking x_i decreases after the algorithm places x_i after x_j. Similarly,

$$\mathbb{P}[x_T = x_i] = \sum_{S: x_i \in S} \mathbb{P}_a\left[C_a(O) = S\right] \times \mathbb{P}_b\left[x_i \mid S\right] \tag{8}$$

$$\mathbb{P}[x'_T = x_i] = \sum_{S: x_i \in S} \mathbb{P}_{a'}\left[C_{a'}(O) = S\right] \times \mathbb{P}_b\left[x_i \mid S\right], \tag{9}$$

We still compare the two probabilities term-by-term. For any set S that contains both x_i and x_j, the corresponding terms in the difference $\mathbb{P}[x_T = x_i] - \mathbb{P}[x'_T = x_i]$. For any set S containing x_i but not x_j, let $S' = S \setminus \{x_i\}) \cup \{x_j\}$. As x_i is placed before x_j in σ_a^*, it follows from Lemma [CITE] that $\mathbb{P}_a[C_a(O) = S] > \mathbb{P}_a[C_a(O) = S']$. Hence, $\mathbb{P}_a[C_a(O) = S] > \mathbb{P}_{a'}[C_{a'}(O) = S]$. This further concludes the remaining comparison. The remaining comparison of the probabilities of picking x_j by the two teams follows by symmetry. □

Theorem 3. *A heterogeneous composition between agents a and b with preference rankings $\sigma_a^* = (S, x_1, \dots)$ and $\sigma_b^* = (T, x_1, \dots)$ with $S \cap T = \emptyset$ and $|S| = |T|$ always has a positive **diversity gain**.*

Proof. Since x_1 ranks in the same position in the curator's and the decider's preference, it suffices to prove that the composed agent has a higher expected utility than the homogeneous composition solely consisting of the decider. Since S is before T in σ_a^* while S is before T in σ^b, by keep exchanging items and applying Lemma 3, we can tweak σ_a such that the outcomes before x_1 become T. Thereafter, we can continue applying Lemma 3 to make the items later than x_1 have the same ranking as σ_b^*. Note that the process does not involve any preference swap involving x_1, hence, the expected utility does not increase during the process. Therefore, the heterogeneous composition has a higher expected utility than the homogeneous composition, which means the diversity gain is positive. □

E Beyond Top-Item Recovery

Theorem 6. *A homogeneous composition always achieves complementarity as long as the preference ranking is consistent with the underlying utilities, i.e., $x_i \succ x_j$ implies that $u(x_i) \geq u(x_j)$.*

Proof. By unfolding the two expected utilities, it suffices to show that

$$\sum_{j=1}^{m} u(x_j) \cdot \mathbb{P}[(a,a) \text{ picks } x_j] > \sum_{j=1}^{m} u(x_j) \cdot \mathbb{P}[a \text{ picks } x_j].$$

To show the above inequality, we apply Karamata's inequality to prove the following proposition, which essentially states that, for any $i \in [T]$, the probability of the composed agent picking an item of $x_1, \dots, x_i$ is weakly larger than the probability of a single agent acting alone.

Proposition 1. *For any $i \in [T]$, we have $\sum_{j=1}^{i} \mathbb{P}[(a,a) \text{ picks } x_j] > \sum_{j=1}^{i} \mathbb{P}[a \text{ picks } x_j]$.*

Proof. We first consider the right-hand side probability. According to the law of conditional probability, we can rewrite it by conditioning on the first k items being S.

$$\sum_{j=1}^{i} \mathbb{P}[a \text{ picks } x_j] = \sum_{j=1}^{i} \sum_{S:|S|=k} \mathbb{P}[\mathcal{D}_a(O) = x_j \mid C_a(O) = S] \cdot \mathbb{P}[C_a(O) = S]$$

$$= \sum_{S:|S|=k} \mathbb{P}[C_a(O) = S] \sum_{j=1}^{i} \mathbb{P}[\mathcal{D}_a(O) = x_j \mid C_a(O) = S],$$

where $\mathbb{P}[\mathcal{D}_a(O) = x_j \mid C_a(O) = S]$ is the probability of the first sampled item is x_j conditioned on the first k sampled items are S. Let $S = \{x_{i(1)}, \ldots, x_{i(k)}\}$ with $i(1) < \cdots < i(k)$. Let $s = |S \cap \{x_1, \ldots, x_i\}|$ be the size of the intersection of S and $\{x_1, \ldots, x_i\}$.

- In the Mallows model, the process of sampling the first k items forms a Mallows model with the same accuracy parameter ϕ and a central ranking of $(x_{i(1)}, \ldots, x_{i(k)})$. Hence, we have

$$\sum_{j=1}^{i} \mathbb{P}[\mathcal{D}_a(O) = x_j \mid C_a(O) = S] = \sum_{j=1}^{s} \mathbb{P}[\mathcal{D}_a(O) = x_{i(j)} \mid C_a(O) = S] = \frac{Z_s(\phi)}{Z_k(\phi)}$$

- In the Plackett-Luce model, the process of sampling the first k items forms a Mallows model with values $v_{i(1)}, \ldots, v_{i(k)}$ and temperature parameter β. Hence, we have

$$\sum_{j=1}^{i} \mathbb{P}[\mathcal{D}_a(O) = x_j \mid C_a(O) = S] = \sum_{j=1}^{s} \mathbb{P}[\mathcal{D}_a(O) = x_{i(j)} \mid C_a(O) = S]$$

$$= \frac{\sum_{j=1}^{s} \exp(v_{i(j)}/\beta)}{\sum_{j=1}^{k} \exp(v_{i(j)}/\beta)}$$

Next, we consider the probability of the homogeneous team picking one of $x_1, \ldots, x_i$. To distinguish the two agents for different roles, we refer to a^{c} and a^{d} as the curator agent and the decider agent.

$$\sum_{j=1}^{i} \mathbb{P}[(a^{\mathrm{c}}, a^{\mathrm{d}}) \text{ picks } x_j] = \sum_{j=1}^{i} \sum_{S:|S|=k} \mathbb{P}[\mathcal{D}_{a^{\mathrm{d}}}(S) = x_j \mid C_{a^{\mathrm{c}}}(O) = S] \cdot \mathbb{P}[C_{a^{\mathrm{c}}}(O) = S]$$

$$= \sum_{S:|S|=k} \mathbb{P}[C_{a^{\mathrm{c}}}(O) = S] \sum_{j=1}^{i} \mathbb{P}[x^*(a^{\mathrm{C}}, a^{\mathrm{D}}) = x_j \mid C_{a^{\mathrm{c}}}(O) = S],$$

Let $s = |S \cap \{x_1, \ldots, x_i\}|$. By the properties of the Mallows model and the Plackett-Luce model, we have

$$\mathbb{P}[x^*(a^{\mathrm{C}}, a^{\mathrm{D}}) \in S \cap \{x_1, \ldots, x_i\} \mid C_{a^{\mathrm{c}}}(O) = S] \geq \frac{Z_s(\phi)}{Z_k(\phi)} \qquad \text{(Mallows, Lemma 7)}$$

$$\mathbb{P}_{a^{\mathrm{D}}}[x^*(a^{\mathrm{C}}, a^{\mathrm{D}}) \in S \cap \{x_1, \ldots, x_i\} \mid C_{a^{\mathrm{c}}}(O) = S] \geq \frac{\sum_{j=1}^{s} \exp(v_{i(j)}/\beta)}{\sum_{j=1}^{k} \exp(v_{i(j)}/\beta)}. \qquad \text{(Plackett-Luce, Lemma 9)}$$

The above inequalities further imply that $\sum_{j=1}^{i} \mathbb{P}[(a^{\mathrm{c}}, a^{\mathrm{d}}) \text{ picks } x_j] \geq \sum_{j=1}^{i} \mathbb{P}[a \text{ picks } x_j]$ and completes the proof. □

Using Karamata's inequality and Proposition 1, we conclude that the theorem holds. □