Single-Agent Generation Surpasses Multi-Agent Systems in Semantic Diversity

Anonymous ACL submission

Abstract

Multi-Agent Systems (MAS) are commonly used to improve reasoning diversity and robustness by simulating interactions among agents with distinct roles. However, prior work often entangles the contribution of the multi-agent architecture with that of prompt conditioning, making the source of observed diversity gains unclear. We address this confound with a controlled study on divergent thinking tasks, using identical prompt conditioning for MAS and single-agent settings. Under these matched conditions, single-agent setups consistently outperform multi-agent systems in semantic diversity. We attribute this gap to *information visibility*: parallel agents often converge on overlapping ideas, whereas a serial single-agent model can condition on its own generation history to avoid redundancy. We further find that a Multi-Output strategy, which prompts a single agent to produce multiple responses within a single inference pass, achieves the highest diversity without degrading logical validity. Together, these results suggest that costly inter-agent interactions may be unnecessary; instead, efficient information sharing and multi-output generation appear to be effective mechanisms for improving diversity, with implications for designing efficient agentic frameworks.

1 Introduction

Agentic frameworks built on Large Language Models (LLMs) have become a prominent research direction. Within such systems, diversity is widely regarded as a core mechanism for improving reasoning quality, broadening coverage, and increasing robustness (Wang et al., 2023; Yao et al., 2023a; Du et al., 2023; Liang et al., 2024). Multi-Agent Systems (MAS) have become a dominant paradigm for this purpose. MAS instantiate multiple agents with distinct roles and simulate their interactions to produce diverse reasoning trajectories. Although empirical studies support the effectiveness of this approach (Li et al., 2023a; Liang et al., 2024; Estornell and Liu, 2024), the underlying source of these gains remains insufficiently examined.

In most MAS setups, agents share the same LLM backbone; accordingly, MAS is typically accompanied by carefully designed prompt conditioning that specifies distinct roles or perspectives for different agents (Guo et al., 2024; Li et al., 2023a; Park et al., 2023; Wang et al., 2024). In other words, MAS agents are explicitly seeded with a predefined set of diverse reasoning priors via carefully designed prompt engineering. However, to the best of our knowledge, existing MAS studies do not apply the same diversity-inducing prompt conditioning to their single-agent baselines when making comparisons. This raises a critical question: **are the observed diversity gains driven by the multi-agent architecture itself, or by the well-designed prompt conditioning?**

To investigate this issue, we design a controlled experiment in which single-agent models employ the same prompt conditioning strategies as their multi-agent counterparts. As part of this setup, we also strictly align output volume with MAS by adopting a *Multi-Output* strategy, where a single agent is prompted to generate multiple responses within a single inference pass. Under these controlled conditions, we find that single-agent setups exhibit higher generation diversity than multi-agent systems, contrary to prevailing intuition. Moreover, although the *Multi-Output* strategy is introduced to align output size under the same prompt conditioning, it also emerges as a significant factor in enhancing diversity.

Our results suggest that computationally expensive multi-agent interactions(Wang et al., 2025a; Zeng et al., 2025; Liu et al., 2024a; Wang et al., 2025b) may be unnecessary for achieving diversity, and that simpler, well-controlled prompting strategies can be effective.

2 Related Work

Diversity in Generation and Reasoning Diversity is fundamental for enhancing LLM reasoning capability, robustness, and creativity. However, standard greedy decoding often produces repetitive or generic outputs; consequently, stochastic sampling strategies such as nucleus sampling and top-k decoding are widely adopted to expand the generation space (Holtzman et al., 2020; Fan et al., 2018; Kool et al., 2019; Massarelli et al., 2020). Beyond variance introduced at decoding time, a growing body of work shows that aggregating diverse reasoning paths can substantially improve performance on complex tasks (Wang et al., 2023; Yao et al., 2023a; Besta et al., 2024; Long, 2023; Li et al., 2023b). Similarly, generating diverse candidate solutions for subsequent verification has been effective at reducing hallucinations and logical errors in mathematical and commonsense reasoning (Cobbe et al., 2021; Lightman et al., 2024; Weng et al., 2023; Dhuliawala et al., 2024). Beyond accuracy, diversity also improves robustness to prompt variation, because broader coverage reduces sensitivity to specific phrasing (Tam et al., 2023; Agrawal et al., 2023; Ippolito et al., 2019; Arora et al., 2023; Tevet and Berant, 2021).

Multi-Agent Systems for Eliciting Diversity To systematically induce diverse perspectives, recent research has converged on MAS as a promising paradigm. These frameworks orchestrate interactions among agents with distinct roles and memories to simulate complex social dynamics and collaborative problem-solving (Park et al., 2023; Wu et al., 2024a; Zhuge et al., 2023). A prevalent application is multi-agent debate, in which agents with conflicting viewpoints or assigned personas critique one another to elicit divergent thinking (Du et al., 2023; Liang et al., 2024; Chan et al., 2023; Xiong et al., 2024; Cohen et al., 2023).Theoretical and empirical studies have further analyzed how factors such as communication topology and cognitive synergy shape the diversity of outcomes (Estornell and Liu, 2024; Li et al., 2024; Wang et al., 2024; Zhang et al., 2024; Chen et al., 2024).

Despite these advances, a methodological ambiguity persists in the literature. Although studies often attribute performance gains to the multi-agent architecture itself, they typically compare MAS equipped with highly engineered, role-specific prompt conditioning against generic single-agent baselines. This confound makes it difficult to determine whether the observed diversity arises from intrinsic multi-agent interactions or from prompt conditioning that could be equally effective within a single-agent setup.

Defining LLM agents. Recent literature diverges in how it defines "LLM agents." One common view characterizes agents as stateful controllers that act and reason in external environments, often incorporating memory and planning (Park et al., 2023; Yao et al., 2023b; Shinn et al., 2023; Yang et al., 2024; Zhou et al., 2024a). Another line of work defines agents as prompt-conditioned conversational roles, where distinctions arise primarily from role specifications rather than architectural design (Li et al., 2023a; Wu et al., 2024b; Liu et al., 2024b; Zhou et al., 2024b; Schick et al., 2023; Du et al., 2023).

Our goal is to disentangle agent architecture from prompt conditioning, so we adopt the former definition. When this definition conflicts with prior usage, we provide clarifying notes.

Quantifying Diversity Lexical diversity is commonly measured using n-gram overlap ratios (e.g., Distinct-N) and pairwise similarity measures such as Self-BLEU (Li et al., 2016; Zhu et al., 2018; Zhang et al., 2018; Pillutla et al., 2021). For example, Liang et al. (2024) employed Self-BLEU to compare the output diversity of multi-agent and single-agent configurations.

However, lexical metrics often fail to capture deeper semantic differences. Accordingly, recent work has increasingly adopted semantic diversity measures based on embedding spaces and contrastive representations (Reimers and Gurevych, 2019; Zhang et al., 2020; Su et al., 2022). For example, Estornell and Liu (2024) showed that incorporating embedding-based measures into the "diversity pruning" step within a MAS significantly improved performance. Motivated by this line of work, we use embedding-based semantic diversity as our primary evaluation metric and retain Self-BLEU as a secondary measure for complementary analysis.

3 Task Formulation

Most benchmarks for evaluating LLMs, including Question Answering (QA) datasets and mathematical problem sets, are designed to elicit a single correct answer. Even when models reach that answer through different reasoning trajectories, a unique ground truth imposes an inherent ceiling on out-

put diversity. As a result, when multiple methods all saturate this ceiling, comparisons among them cannot reveal meaningful differences in their capacity for semantic variation. To properly evaluate semantic diversity across multiple valid answers, we therefore require tasks that naturally admit divergent yet coherent outputs.

To this end, we construct a new dataset grounded in cognitive psychology and creativity studies, which have long examined the mechanisms of divergent thinking and open-ended problem solving. Drawing on this literature, we identify five representative task types that are known to elicit creative and associative reasoning:

1. **Impossible Situations Task** (Runco, 1999): Reasoning about implausible or paradoxical premises.
2. **Alternative Uses Task** (Guilford, 1967): Proposing unconventional uses for everyday objects.
3. **Improvement Task** (McCaffrey, 2012): Suggesting modifications to enhance common objects.
4. **Just Suppose Task** (Torrance, 1974): Exploring hypothetical or counterfactual scenarios.
5. **Bridge-the-Associative-Gap Task** (Gianotti et al., 2001): Connecting distant or conceptually unrelated ideas.

We use GPT-5 to generate 60 diverse, well-formed questions per task type, yielding a total of 300 open-ended questions. Each question is designed to admit multiple valid *solution perspectives*, enabling systematic evaluation of semantic diversity in both single-agent and multi-agent settings. We manually reviewed all questions to ensure clarity and conceptual breadth.

Perspectives Instead of Personas Prompt conditioning is often framed in terms of *personas* (e.g., "critic","engineer", or "judge"), but we instead treat *perspectives* as the primary dimension of variation. Personas typically shape outputs through loosely specified roles that affect tone, background knowledge, or rhetorical style, introducing unstructured variability that is difficult to standardize or interpret. By contrast, we define perspectives as explicit, task-specific reasoning angles, such as emphasizing ethical tradeoffs, practical constraints, or long-term consequences. This framing links variation to the task semantics rather than to superficial identity markers, allowing us to isolate the effects of reasoning diversity more precisely.

Diversity Metric: Vendi Score (Friedman and Dieng, 2023) To quantify semantic diversity, we adopt the *Vendi Score*, a reference-free and similarity-aware metric computed from a kernel similarity matrix K (e.g., cosine similarities between embedding vectors). Let $\{\lambda_i\}$ denote the eigenvalues of K, and define the normalized spectrum as $\tilde{\lambda}_i = \lambda_i / \sum_j \lambda_j$. The Vendi Score is defined as:

$$\mathrm{VS}(K) = \exp\left(-\sum_i \tilde{\lambda}_i \log \tilde{\lambda}_i\right)$$

We compute semantic similarity among generated answers using three embedding models: Sentence-Transformers `all-mpnet-base-v2` (768 dimensions), `Qwen3-8B embedding` (4096 dimensions), and OpenAI `text-embedding-3-small` (1536 dimensions).

Because the Vendi Score is sensitive to the number of items being compared, we control for this factor by fixing the number of *perspectives* per question. Each perspective corresponds to exactly one generated answer. Consequently, choosing $k \in \{2, 4, 8, 16\}$ jointly determines both the number of reasoning angles and the number of resulting outputs.

These perspectives are not manually written; instead, they are generated in advance via a dedicated prompting procedure (see Appendix A.2 for details). As shown in Figure 1, given a question and a target value of k, we use the same LLM as in answer generation to produce k diverse, task-relevant reasoning perspectives. This design ensures that variation across answers reflects semantically meaningful differences in perspective, while enabling scalable and consistent evaluation of diversity across experimental protocols.

As a complementary measure, we also compute Self-BLEU to evaluate lexical diversity. While our main analyses focus on the semantic variation captured by the Vendi Score, Self-BLEU provides a supplementary indicator of surface-level similarity across outputs.

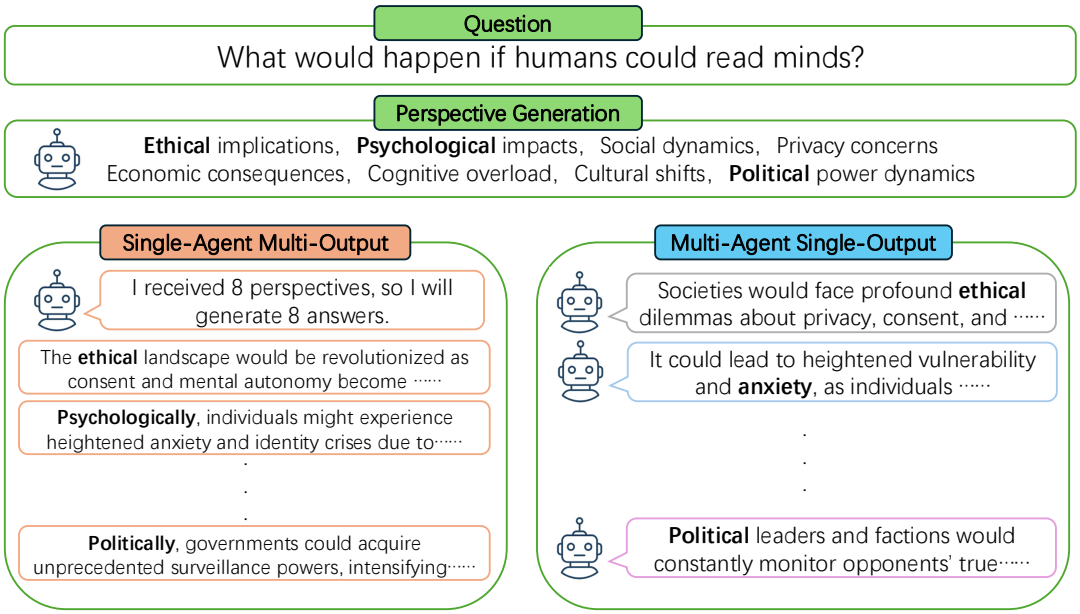

Figure 1: **Overview of the generation pipeline for single-agent and multi-agent settings.** Given an open-ended question, the system first generates k distinct reasoning perspectives (illustrated here with $k = 8$). For a fair comparison, both settings use the same prompt conditioning. In the *single-agent* setting, a single model instance receives all k perspectives simultaneously and is asked to generate one answer per perspective in a single inference pass. In the *multi-agent* setting, the k perspectives are distributed across k independent agents, and each agent is likewise asked to generate one answer per perspective.

4 Methodology

4.1 A Two-Factor Framework for Generation Protocols within generation round

To enable a rigorous comparison between the single-agent and multi-agent settings, we enforce identical prompt conditioning between them: both use the same prompt template P and the same set of k perspectives $A = \{a_1, \ldots, a_k\}$. In addition, we strictly hold the total output volume constant at k answers in both settings.

To match the same output volume in a linear system, there are two strategies: the model can either switch perspectives iteratively to produce answers in a linear sequence, or ingest all perspectives at once and generate the corresponding answers in a single inference pass. Based on these operational differences, we can decouple the generation process along two orthogonal axes:

1. **Agent Architecture** (*Single-Agent* vs. *Multi-Agent*): This axis distinguishes the processing structure. The task is either handled by a single agent operating linearly or distributed across multiple independent agents operating in parallel. This categorization follows the state-based definition of an agent in prior work (Park et al., 2023; Yao et al., 2023b; Shinn et al., 2023; Yang et al., 2024; Zhou et al., 2024a), which facilitates our controlled separation between agent architecture and prompt conditioning.

2. **Output Multiplicity** (*Single-Output* vs. *Multi-Output*): This axis specifies the decoding granularity per model invocation. *Single-Output* protocols generate exactly one answer per call, whereas *Multi-Output* protocols generate the full set of k answers in a single inference pass. In practice, we control the number of perspectives provided as input to the model, and pair this with the one answer each perspective instruction.

The intersection of these dimensions yields four canonical protocols:

- **Single-Agent Single-Output** (SA-SO)
- **Single-Agent Multi-Output** (SA-MO)
- **Multi-Agent Single-Output** (MA-SO)
- **Multi-Agent Multi-Output** (MA-MO)

4.2 Protocol Definitions

We formally define the generation process for each protocol. Let G denote the generation function for each invocation of the LLM.

Single-Agent Single-Output (SA-SO). A single agent generates answers sequentially over k discrete steps. At step i, the agent adopts a distinct perspective a_i; completing all k steps constitutes one full generation round. Critically, each step is conditioned on the set of answers generated in earlier steps. Considering that some MAS implementations use this protocol in practice, we discuss it from both single-agent and multi-agent perspectives in our subsequent analyses.

Let $R_{<i} = \{r_1, \ldots, r_{i-1}\}$ denote the partial set of answers generated prior to step i. Then:

$$r_i = G(P \cup R_{<i}, a_i), \quad i = 1, \ldots, k \tag{1}$$

$$R_{\text{SA-SO}} = \{r_1, \ldots, r_k\} \tag{2}$$

Single-Agent Multi-Output (SA-MO). A single agent instance is invoked once to produce one corresponding answer for each perspective, distributing its decoding capacity internally across the k outputs; this setting unambiguously constitutes a single-agent system and serves as our definitive single-agent anchor throughout the paper.

$$R_{\text{SA-MO}} = G(P, A) = \{r_1, \ldots, r_k\} \tag{3}$$

Multi-Agent Single-Output (MA-SO). Each perspective a_i is assigned to an independent agent instance that operates in parallel and produces a single answer in isolation:

$$r_i = G(P, a_i), \quad i = 1, \ldots, k \tag{4}$$

$$R_{\text{MA-SO}} = \{r_1, \ldots, r_k\} \tag{5}$$

Multi-Agent Multi-Output (MA-MO). For completeness, we define a hybrid protocol in which K parallel agents each produce m answers, keeping the total output size fixed at $k = K \cdot m$. The perspective set A is partitioned into K disjoint subsets $A_1, \ldots, A_K$, each of size m. Each agent is assigned subset A_i and generates the corresponding answers using a multi-output strategy:

$$S_i = G(P, A_i) = \{r_{i,1}, \ldots, r_{i,m}\} \tag{6}$$

$$R_{\text{MA-MO}} = \bigcup_{i=1}^{K} S_i = \{r_1, \ldots, r_k\} \tag{7}$$

In our experiments, MA-MO is evaluated at $k = 8$ with $m \in \{2, 4\}$ and at $k = 16$ with $m \in \{2, 4, 8\}$.

4.3 Iterative Generation Rounds

In both single-agent and multi-agent settings, we adopt the canonical iterative formulation, in which each round conditions on the full set of outputs generated in all previous rounds; in the multi-agent setting, this conditioning serves as the communication mechanism. This design corresponds to a causal all-to-all (complete-graph) communication topology across rounds: outputs from every agent in earlier rounds are visible in later rounds, and more specialized interaction topologies can be viewed as pruned variants of this structure.

Let $R^{(t)}$ denote the set of outputs generated in round t. Each subsequent round builds on the full accumulated history by updating P at the beginning of the round:

$$P^{(t)} = P^{(t-1)} \cup R^{(t-1)}$$

$$R^{(0)} = \emptyset, \quad P^{(0)} = P$$

In our experiments, we fix the generation horizon to 4 rounds. Empirically, our metrics converges within this range, while additional rounds yield diminishing returns and can occasionally degrade performance. This threshold is sufficient to capture the essential iterative dynamics while maintaining computational efficiency.

5 Results and Discussions

5.1 Single-Agent vs. Multi-Agent under Single-Output

Across all evaluated chat models, embedding models, rounds, and values of k, we observe a consistent pattern: **SA-SO exhibits substantially higher diversity than MA-SO** (Figures 2, 5, and 8, and Table 1). A similar trend is observed under the Self-BLEU metric (Figure 11).

Under a strictly controlled experimental setup, this performance gap can be attributed to a structural asymmetry between the two paradigms. In the parallel MA-SO setting, agents generate answers independently at the current round. Although each agent has a distinct perspective, the lack of real-time visibility into peer outputs precludes mutual adjustment. With our manual inspection of a subset of experimental outputs, we found that agents remain unable to fully circumvent text degeneration (Holtzman et al., 2020; Li et al., 2016; Welleck

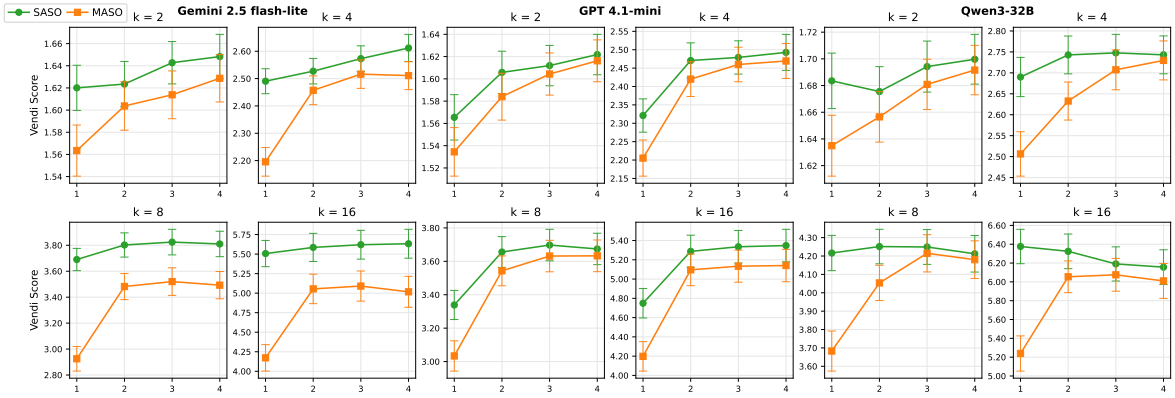

Figure 2: **Comparison of Single-Agent vs. Multi-Agent architectures under Single-Output setting (mpnet-v2)** The **Single** condition (green) represents SA-SO, and the **Multi** condition (orange) represents MA-SO. The x-axis indicates the generation round, and the y-axis reports the semantic diversity (Vendi score). Error bars denote 95% confidence intervals.

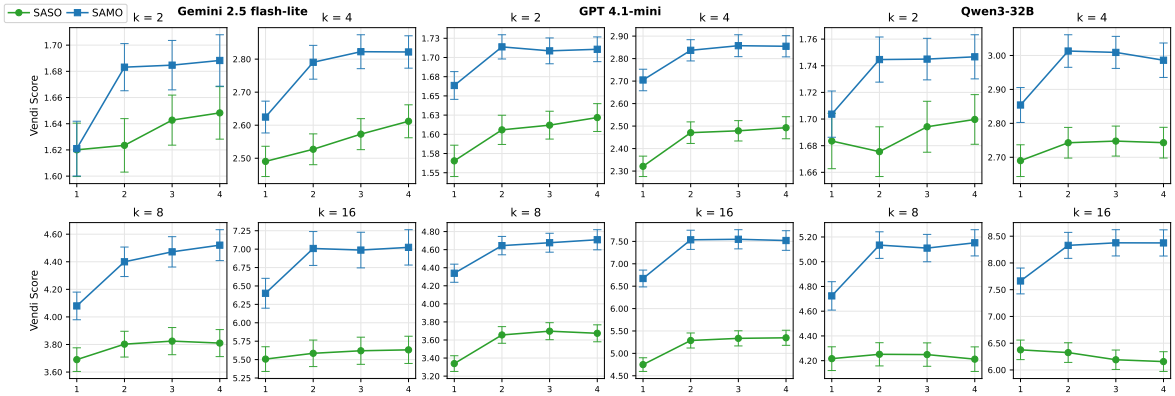

Figure 3: **Comparison of Single-Output vs. Multi-Output strategies under Single-Agent architecture (mpnet-v2)** The **Single-output** condition (green) corresponds to SA-SO, while the **Multi-output** condition (blue) corresponds to SA-MO. The x-axis indicates the generation round, and the y-axis reports the semantic diversity (Vendi score). Error bars denote 95% confidence intervals.

et al., 2020), often converging on similar outputs despite their assigned perspective. In contrast, the linear structure of SA-SO enables the model to condition every output on the preceding history, avoiding redundancy in real time. Thus, single-agent generation achieves higher diversity, even when operating over the identical set of perspectives. We refer to this difference as the degree of *information visibility*, defined as the extent to which an agent's access to information produced within the system is restricted during a single inference pass.

This *information visibility* is also supported by the marked increase in MA-SO diversity from the first to the second generation round, when agents first gain access to the interaction information. The increases suggests that higher *information visibility* can substantially alleviate text degeneration. MA-SO's *information visibility* disadvantage is particularly pronounced in the first generation round, and this gap is also reflected in our experimental results, where SA-SO typically shows a substantial lead in round 1.

Taken together, our controlled comparisons indicate that agent multiplicity does not inherently increase diversity or even reduce it. MA-SO does improve over rounds as interaction information accumulates, but SA-SO maintains higher *information visibility* throughout, and correspondingly remains more diverse across the vast majority of subsequent steps. Even if we set aside our agent definition and analyze SA-SO as a form of MAS, our conclusions remain unchanged: across both MAS formulations, higher *information visibility* enables the system to generate more diverse outputs.

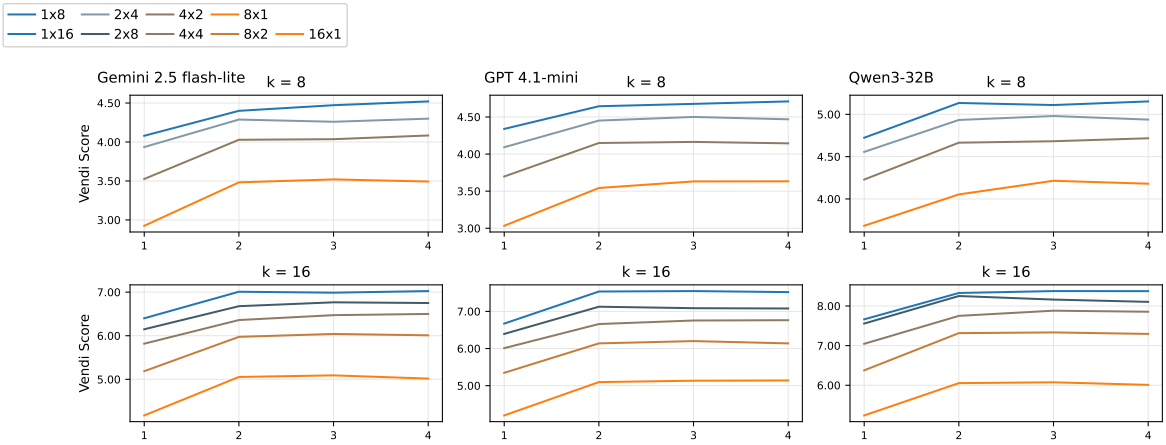

Figure 4: **Exploring multi-agent multi-output (MA-MO) configurations (mpnet-v2)** This figure accompanies the analysis in *Additional Derivation: Exploring MA-MO*. Each configuration is denoted as $n \times m$, where n is the number of agents and each agent generates m outputs, with the total number of answers fixed at $k = n \cdot m$. The special cases $1 \times k$ and $k \times 1$ correspond to the *SA-MO* and *MA-SO* settings, respectively. The x-axis indicates the generation round, and the y-axis reports semantic diversity quantified by the Vendi score.

5.2 Single-Output vs. Multi-Output under Single-Agent

Under a fixed single-agent architecture, we observe a consistent pattern across all conditions: **SA-MO yields substantially higher diversity than SA-SO** (Figures 3, 6, and 9, and Table 1). Moreover, this gap widens as k increases. A parallel trend is observed under the Self-BLEU metric (Figure 12).

Although both protocols condition later outputs on earlier content, they implement this conditioning differently. In SA-SO, earlier answers are explicitly reintroduced in the prompt across successive invocations, whereas in SA-MO they are generated within a single auto-regressive decoding stream.Auto-regressive models apply causal self-attention uniformly over preceding tokens (Radford et al., 2019; Brown et al., 2020; Touvron et al., 2023), so the placement of earlier content within a continuous sequence can affect how it is attended to; differences in positional encoding and the presence of sequence boundaries may further modulate how prior context is utilized.

While our controlled experiments isolate protocol-level differences, they can't identify the precise mechanism underlying the observed diversity gap. Beyond the model architectural considerations discussed above, we propose a data-centric hypothesis: Multi-Output instructions may mirror the human-authored enumerations that are pervasive in pretraining corpora, where list items are implicitly expected to be non-redundant and to contribute distinct aspects in accordance with the Gricean Maxim of Quantity (Grice, 1975). When asked to produce multiple answers in a single inference pass, the model may generalize this learned convention and actively vary content across the requested outputs, consistent with evidence that pretrained LLMs encode non-trivial discourse-level regularities (Koto et al., 2021; Shen et al., 2021).

Regardless of the underlying mechanism, our results support a methodological conclusion: generating multiple outputs per invocation is a simple yet effective choice for improving output diversity in agentic framework. When we set aside our agent definition and analyze SA-SO as a form of MAS, we can find that under identical prompt conditioning, the single-agent setup outperforms the multi-agent setup in diversity generation, suggested that the long-held intuition that splitting a model into multiple interacting agents inherently yields greater diversity may be misguided.

5.3 Additional Derivation: Exploring MA-MO

We now turn to the derived setting that follows naturally from our two-factor analysis: MA-MO.

As shown in Figures 4, 7, and 10, and Table 2, a consistent pattern emerges: when the total output count is held constant, for all settings, **MA-MO exhibits a largely monotonic decrease in diversity as the agent count increases**. A parallel trend is observed under the Self-BLEU metric (Figure 13).

Increasing the number of agents necessarily re-

duces the number of multi-output generations available to each agent, which in turn lowers the system's *information visibility*. These results further corroborate our earlier conclusion that *information visibility* is a plausible driver of the observed diversity differences.

5.4 Post-hoc Quality Audit: Does Multi-Output Harm Validity?

Motivation. Although our prompts explicitly require logical consistency and close relevance to the input question, it is not clear *a priori* whether MO harms answer quality. In particular, if MO amplifies hallucinations or invalid reasoning, any gains in diversity may come at the cost of reliability. To quantify this potential trade-off, we perform an LLM-as-judge quality audit on the **SA-MO** and **MA-SO** outputs.

In open-ended generation, however, hallucination is difficult to define operationally, particularly when no clear reference is available. As a result, separating hallucinated content from answers that are merely unverifiable or intentionally creative can be ill-posed (van Deemter, 2024; Ji et al., 2023). We therefore take a conservative approach and restrict our audit to **strict logical errors**, such as explicit self-contradictions, mutually incompatible causal claims, or inconsistent definitions—independent of factuality or stylistic fluency (Lin et al., 2022).

LLM-based screening with targeted human verification. We use GPT-4o as a conservative logical auditor to screen approximately **210K** answers from SA-MO and MA-SO, and it flags only **74** cases (**0.03%**) as containing logical errors, of which **81.08%** come from MA-SO. Given this extremely low base rate, exhaustively annotating all outputs would be inefficient. We therefore form a 400-item evaluation set by including all 74 LLM-flagged answers and adding 326 randomly sampled unflagged ones. Two annotators independently label the set: Annotator 1 marks **63** answers as erroneous, whereas Annotator 2 marks **57**. Among the 74 LLM-flagged answers, Annotator 1 and Annotator 2 confirm errors in **43.97%** and **52.59%** of cases, respectively. Overall, only **32** items receive consensus error labels from both human annotators.

These results underscore a known challenge in evaluating open-ended outputs: even under a narrow validity criterion, human judgments can diverge because annotators apply different interpretive thresholds. Nonetheless, GPT-4o aligns well on clear-cut cases, flagging **87.5%** (28 out of 32) of the errors on which both annotators agree. This indicates that although fine-grained agreement remains difficult to obtain, LLMs can still function as conservative auditors for *unambiguous* logical flaws.

Overall, we find no evidence that MO instructions measurably degrade logical validity. Given the low incidence of strict logical errors and the LLM's high recall on cases where annotators agree, we conclude that MO's diversity gains are not accompanied by an increase in obvious logical defects.

6 Conclusion

Increasing diversity with an agentic framework is widely viewed as a practical way to improve the reasoning and generative capabilities of LLMs. Multi-agent systems are often seen as a natural mechanism for achieving this diversity, since multiple agents can approximate human-like deliberation and multi-perspective reasoning.

In this work, we re-examined this assumption via a controlled study that isolates the effect of agent architecture on generation diversity. Under matched experimental conditions, single-agent configurations consistently and substantially outperformed their multi-agent counterparts across all settings.

Our analysis attributed the relative disadvantage of MAS to two structural factors: the lower **information visibility** induced by multi agent architecture, and the fact that **multi-output** delivers larger gains under comparable conditions, thereby widening the gap between single-agent and the traditional multi-agent settings.

These findings challenge the intuition that Multi-Agent Architecture inherently increases diversity. Unlike human groups, LLM agents may not be suitable for inter-agent discussion architecture. Instead, a single model that internally generates multiple distinct answers could effectively yield higher semantic diversity.

Beyond the empirical results, our study pointed to a broader implication: **multi-output generation itself is a key driver of diversity**. This finding points to a simple and efficient way to increase generation diversity, offering a new direction for designing high-performance agentic frameworks.

Limitations

While our framework provided a controlled comparison between single-agent and multi-agent paradigms, several limitations remained to be addressed.

First, our experiments showed that Multi-Output increases diversity, but the mechanism underlying this effect remains hypothetical. We also do not characterize the upper limit of its generative capacity; there may exist a critical point beyond which generating additional outputs yields diminishing returns or even degrades performance. We leave both questions for future investigation.

Second, for consistency and comparability, we **fixed the temperature parameter to 1.0** across all models. However, MAS could potentially leverage **heterogeneous temperature configurations** to induce greater diversity across agents.

Third, in terms of **Quality Audit**, the incidence of detected errors is extremely low, and the open-ended, reference-free nature of our dataset makes quality judgments inherently subjective. To accommodate this property, we employed unconventional evaluation procedures. Future work could develop more principled **evaluation frameworks** specifically tailored to multi-answer generation and diversity-oriented reasoning.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Ghazvininejad Marjan. 2023. In-context learning with diverse examples. In *Proceedings of EMNLP*.

Simran Arora, Avanika Narayan, Mayee F Chen, Laurel Orr, Neel Gu, Christopher Re, Aaron Sidford, and Frederic Sala. 2023. Ask me anything: A simple strategy for prompting language models. In *Proceedings of ICLR*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, and 1 others. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of AAAI*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, and 1 others. 2020. Language models are few-shot learners. In *Proceedings of NeurIPS*, volume 33, pages 1877–1901.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. In *Proceedings of EMNLP*.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, and 1 others. 2024. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *Proceedings of ICLR*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. Lm vs lm: Detecting factual errors via cross examination. In *Proceedings of EMNLP*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Asghari, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of ACL*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of ICML*.

Andrew Estornell and Yang Liu. 2024. Multi-llm debate: Framework, principals, and interventions. In *Proceedings of NeurIPS*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of ACL*.

Dan Friedman and Adji Bousso Dieng. 2023. The vendi score: A diversity evaluation metric for machine learning. In *Proceedings of ICML*.

Lorena R R Gianotti, Christine Mohr, Diego Pizzagalli, Dietrich Lehmann, and Peter Brugger. 2001. Associative processing and paranormal belief. *Psychiatry and Clinical Neurosciences*, 55(6):595–603. Introduces the Bridge-the-Associative-Gap (BAG) task.

H. Paul Grice. 1975. Logic and conversation. In *Syntax and Semantics, Vol. 3: Speech Acts*. Academic Press.

Joy Paul Guilford. 1967. *The Nature of Human Intelligence*. McGraw-Hill, New York.

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

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *Proceedings of ICLR*.

Daphne Ippolito, Reno Kriz, Joao Sedoc, Maria Kristvokapic, and Chris Callison-Burch. 2019. Comparison of diverse decoding methods for natural language generation. In *Proceedings of ACL*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Wouter Kool, Herke van Hoof, and Max Welling. 2019. Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences. In *Proceedings of ICLR*.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Discourse probing of pretrained language models. In *Proceedings of NAACL*.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023a. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Proceedings of NeurIPS*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL*.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023b. On the advance of diversity-based reasoning with large language models. In *Proceedings of ICLR*.

Yunxuan Li, Yibing Du, Jiageng Zhang, Yiqun Li, and Xinyu Hu. 2024. Improving multi-agent debate with sparse communication topology. In *Findings of EMNLP*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, and 1 others. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of ICLR*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, and 1 others. 2024. Let's verify step by step. In *Proceedings of ICLR*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of ACL*, pages 3214–3252.

Tongxuan Liu, Xingyu Wang, Weizhe Huang, Wenjiang Xu, Yuting Zeng, Lei Jiang, Hailong Yang, and Jing Li. 2024a. Groupdebate: Enhancing the efficiency of multi-agent debate using group discussion. *arXiv preprint arXiv:2409.14051*.

Xiao Liu, Hao Yu, Hanchen Zhang, and 1 others. 2024b. Agentbench: Evaluating llms as agents. In *Proceedings of ICLR*.

Jieyi Long. 2023. Large language model guided tree-of-thought. *arXiv preprint arXiv:2305.08291*.

Luca Massarelli, Fabio Petroni, Aleksandra Piktus, Myle Ott, Tim Rocktäschel, Vassilis Plachouras, Fabrizio Silvestri, and Sebastian Riedel. 2020. How decoding strategies affect the verifiability of generated text. In *Findings of EMNLP*.

Tony McCaffrey. 2012. Innovation relies on the obscure: A key to overcoming the classic problem of functional fixedness. *Psychological Science*, 23(3):215–218.

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of UIST*.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. In *Proceedings of NeurIPS*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Technical Report*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of EMNLP*.

Mark A Runco. 1999. Divergent thinking. In Mark A Runco and Steven R Pritzker, editors, *Encyclopedia of Creativity*, volume 1, pages 577–582. Academic Press, San Diego, CA.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Proceedings of NeurIPS*.

Aili Shen, Meladel Mistica, Bahar Salehi, Hang Li, Timothy Baldwin, and Jianzhong Qi. 2021. Evaluating document coherence modeling. *Transactions of the Association for Computational Linguistics*.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Proceedings of NeurIPS*.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. In *Proceedings of NeurIPS*.

Derek Tam, Sachit Menon, Mohit Bansal, Shashank Srivastava, and Colin Raffel. 2023. Evaluating the robustness of large language models to prompt variations. In *Proceedings of EMNLP*.

Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of EMNLP*.

Ellis Paul Torrance. 1974. *Torrance Tests of Creative Thinking: Norms-Technical Manual*. Scholastic Testing Service, Bensenville, IL.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Kees van Deemter. 2024. The pitfalls of defining hallucination. *Computational Linguistics*, 50(2):807–816.

Qian Wang, Zhenheng Tang, Zichen Jiang, Nuo Chen, Tianyu Wang, and Bingsheng He. 2025a. Agenttaxo: Dissecting and benchmarking token distribution of llm multi-agent systems. In *ICLR 2025 Workshop on Foundation Models in the Wild*.

Qian Wang, Tianyu Wang, Zhenheng Tang, Qinbin Li, Nuo Chen, Jingsheng Liang, and Bingsheng He. 2025b. Megaagent: A large-scale autonomous llm-based multi-agent system without predefined sops. In *Findings of the Association for Computational Linguistics: ACL 2025*.

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

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024. Unleashing the emergent cognitive synergy in large language models via multi-role co-operation. In *Proceedings of ICLR*.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *Proceedings of ICLR*.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Kang Liu, and Jun Zhao. 2023. Large language models are better reasoners with self-verification. *arXiv preprint arXiv:2212.09561*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Peng, Toby Walsh, and Ahmed Awadallah. 2024a. Autogen: Enabling next-gen llm applications. In *Proceedings of ICLR*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2024b. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *Proceedings of COLM*.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2024. Examining inter-consistency of large language models collaboration. In *Proceedings of ICLR*.

John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. 2024. Swe-agent: Agent-computer interfaces enable automated software engineering. In *Proceedings of NeurIPS*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. In *Proceedings of NeurIPS*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. In *Proceedings of ICLR*.

Yuting Zeng, Weizhe Huang, Lei Jiang, Tongxuan Liu, XiTai Jin, Chen Tianying Tiana, Jing Li, and Xiaohua Xu. 2025. S^2-mad: Breaking the token barrier to enhance multi-agent debate efficiency. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

Ceyao Zhang, Kaijie Yang, Xujian Hu, Zeyuan Wang, Guang Li, Yuhui Sun, and Cheng Zhang. 2024. Proagent: Building proactive cooperative ai with large language models. In *Proceedings of ICLR*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proceedings of ICLR*.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. In *Proceedings of NeurIPS*.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. 2024a. Language agent tree search unifies reasoning, acting, and planning in language models. In *Proceedings of ICML*.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024b. Webarena: A realistic web environment for building autonomous agents. In *Proceedings of ICLR*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *Proceedings of SIGIR*.

Mingchen Zhuge, Haozhe Liu, Francesco Faccio, David Rutherford, Gabriele Santacatterina, and 1 others. 2023. Mindstorms in natural language-based society of mind. In *Proceedings of NeurIPS*.

A Appendix

A.1 Question Set

We constructed our question set by adapting classic creativity and divergent-thinking tasks from cognitive psychology. Each category contained 60 prompts in the full dataset; here we list representative examples.

1. Alternative Uses Task Generate unusual or creative uses for common objects:

- What are alternative uses for a paperclip?
- What are alternative uses for a brick?
- What are alternative uses for a cardboard box?
- What are alternative uses for a mirror?
- What are alternative uses for a backpack?

2. Improvement Task Suggest ways to improve the design or function of everyday items:

- How could you improve a bicycle to make it safer?
- How could you improve a smartphone to last longer?
- How could you improve a refrigerator to waste less food?
- How could you improve a chair to be more ergonomic?
- How could you improve a pen to write in all conditions?

3. Just Suppose Task Explore counterfactual "what if" scenarios:

- Just suppose humans could breathe underwater—what would change?
- Just suppose everyone could teleport instantly—what would change?
- Just suppose language barriers disappeared—what would change?
- Just suppose no one could lie—what would change?
- Just suppose the sun never set—what would change?

4. Impossible Situations / Consequences Task Imagine outcomes of implausible or paradoxical conditions:

- What would happen if gravity stopped working for one day?
- What would happen if humans never needed sleep?
- What would happen if money lost all value overnight?
- What would happen if humans could fly like birds?
- What would happen if humans could live underwater?

5. Bridge-the-Associative-Gap Task Form connections between unrelated or distant concepts:

- What is the connection between a cloud and a pillow?
- What is the connection between a ladder and a song?
- What is the connection between a violin and the wind?
- What is the connection between a shoe and a story?
- What is the connection between a river and a road?

Dataset Summary. Each task category included 60 questions (300 in total). All prompts were verified for clarity and open-endedness, ensuring multiple valid yet distinct answers could be generated across perspectives.

A.2 Prompt Design

To ensure consistent prompting across experimental settings, we used three structured templates corresponding to different stages of generation: the `Perspective Generation Prompt`, `Answer Generation System Prompt`, and `Answer Generation User Prompt (Iteration)`. The prompt concatenation was implemented using LangChain, and Pydantic was used to produce structured outputs.

All models were decoded with a fixed temperature of 1.0. All other decoding and sampling parameters were left at their respective model defaults.

(1) Perspective Generation Prompt

`MISSION:` You are a creative thinker agent specialized in generating diverse and varied perspectives toward the given question. Your core expertise is in exploring multiple perspectives, approaches, and solutions to any given question.
Your mission is to generate exactly $\{k\}$ diverse perspectives that maximize variety and avoid redundancy.

`CORE INSTRUCTIONS:`
1. **Maximize Diversity:** Generate exactly $\{k\}$ answer's perspectives that are as different as possible from each other and from any previous answers.
2. **Explore Different Dimensions:** Consider various aspects like:

- Different approaches or methodologies
- Various perspectives or viewpoints
- Different scales or levels of analysis
- Alternative frameworks or paradigms
- Contrasting assumptions or premises

3. **Concise answers:** Generate perspective with just one phrase or a single word.

Question: $\{Q\}$
Generate $\{k\}$ diverse perspectives for analyzing this question.

(2) Answer Generation System Prompt

`MISSION:` You are a creative yet analytical thinker agent. Thinking about the problem from the following perspective:
`Perspective:` $\{A_i\}$

`CORE INSTRUCTIONS and WORKFLOW`
1. **Keep the direction firmly anchored in logic:** Every idea must clearly address the core intent of the question — stay focused and avoid drifting off-topic.
2. **Maximize Creativity:** Generate one answer for EACH perspective. Within that relevance, be as imaginative as possible — propose unconventional, cross-disciplinary, or thought-provoking ideas.
3. **Output neatly:** Express each idea concisely in one sentence.

(3) Answer Generation User Prompt (Iteration)

`QUESTION:` "$\{Q\}$"
`PREVIOUS ANSWERS:`
$\{R\}$
`INSTRUCTIONS:`
Carefully analyze the list of `PREVIOUS ANSWERS` provided and draw inspiration from them.
Ensure your answers contain no logical flaws.
Generate new answers that introduce novel ideas while avoiding redundancy with `PREVIOUS ANSWERS`.
Make sure every idea is clearly and logically connected to the question.
Ensure the answers array contains exactly $\{m\}$ items, in the same order as your assigned perspectives, with one sentence per item.

Algorithm 1

Require: Question Q, perspectives set A, number of agents K
Ensure: Answer set R

Divide A into K subsets $\{P_1, \ldots, P_K\}$
$R \leftarrow \emptyset$
for $t = 0 \rightarrow 3$ **do** ▷ 1 Initial + 3 Iterations
  for all agents $i = 1 \rightarrow K$ **in parallel do**
    $R_i \leftarrow G(Q, A_i, R)$
  end for
  $R^{(t)} \leftarrow \bigcup_{i=1}^{K} R_i$
  $R \leftarrow R \cup R^{(t)}$
end for
return R

Explanation. SA-MO, MA-SO and MA-MO settings follow the same generation algorithm and share identical prompt templates described above. The only difference lies in the number of agents K: in SA-MO, we set $K = 1$, meaning that a single model instance sequentially generates all answers. In contrast, in MA-SO and MA-MO, $K > 1$, and each agent independently operates on a distinct subset of perspectives in parallel. This ensures that any observed differences in output diversity or quality arise solely from the collaborative dynamics rather than prompt or structural variations.

Algorithm 2 About SA-SO

Require: Question Q, perspectives set $A = \{P_1, \ldots, P_k\}$, number of perspectives k
Ensure: Answer set R

$R \leftarrow \emptyset$
for $t = 0 \rightarrow 3$ **do** ▷ 1 Initial + 3 Iterations
  for $i = 1 \rightarrow k$ **do** ▷ iterate over A
    $R_i \leftarrow G(Q, P_i, R)$
    $R \leftarrow R \cup R_i$
  end for
end for
return R

Explanation. The SA-SO algorithm also use the identical prompt templates described above.

A.3 Additional Results Table and Figure

Model	k	Setting	Round 1						Round 2						Round 3						Round 4					
			MPNet		OpenAI		Qwen		MPNet		OpenAI		Qwen		MPNet		OpenAI		Qwen		MPNet		OpenAI		Qwen	
			Value	Δ (%)	Value	Δ (%)	Value	Δ (%)	Value	Δ (%)	Value	Δ (%)	Value	Δ (%)	Value	Δ (%)	Value	Δ (%)	Value	Δ (%)	Value	Δ (%)	Value	Δ (%)	Value	Δ (%)
GPT 4.1-mini	2	SAMO	**1.66** (0.03)		**1.70** (0.01)		**1.64** (0.02)		**1.71** (0.02)		**1.73** (0.01)		**1.66** (0.01)		**1.71** (0.02)		**1.72** (0.01)		**1.67** (0.01)		**1.71** (0.02)		**1.72** (0.01)		**1.67** (0.01)	
		SASO	1.57 (0.03)	6.27	1.62 (0.02)	5.36	1.57 (0.02)	4.09	1.61 (0.03)	6.73	1.63 (0.02)	5.66	1.60 (0.02)	3.85	1.61 (0.03)	6.00	1.63 (0.02)	5.80	1.61 (0.01)	3.97	1.62 (0.03)	5.48	1.62 (0.02)	5.87	1.61 (0.02)	3.97
		MASO	1.53 (0.04)	-1.98	1.58 (0.03)	-2.52	1.54 (0.03)	-1.98	1.58 (0.03)	-1.36	1.61 (0.02)	-1.64	1.57 (0.02)	-1.79	1.60 (0.03)	-0.47	1.62 (0.02)	-0.31	1.59 (0.02)	-0.81	1.62 (0.03)	-0.35	1.62 (0.02)	-0.24	1.61 (0.02)	0.02
	4	SAMO	**2.70** (0.18)		**2.85** (0.10)		**2.54** (0.09)		**2.84** (0.17)		**2.90** (0.08)		**2.65** (0.08)		**2.86** (0.18)		**2.91** (0.08)		**2.67** (0.08)		**2.85** (0.17)		**2.90** (0.08)		**2.67** (0.07)	
		SASO	2.32 (0.16)	16.54	2.48 (0.11)	14.94	2.33 (0.07)	8.77	2.47 (0.18)	14.82	2.52 (0.12)	15.12	2.44 (0.08)	8.70	2.48 (0.16)	15.26	2.50 (0.11)	16.54	2.43 (0.08)	9.93	2.49 (0.19)	14.53	2.50 (0.12)	16.00	2.44 (0.08)	9.51
		MASO	2.21 (0.19)	-4.99	2.34 (0.14)	-5.39	2.23 (0.12)	-4.31	2.42 (0.17)	-2.05	2.49 (0.11)	-1.43	2.40 (0.09)	-1.61	2.46 (0.17)	-0.77	2.48 (0.11)	-0.58	2.42 (0.08)	-0.53	2.47 (0.17)	-0.94	2.47 (0.12)	-1.23	2.42 (0.08)	-0.89
	8	SAMO	**4.34** (0.78)		**4.67** (0.44)		**3.88** (0.36)		**4.64** (0.80)		**4.79** (0.41)		**4.09** (0.35)		**4.68** (0.85)		**4.76** (0.42)		**4.13** (0.35)		**4.71** (0.96)		**4.73** (0.43)		**4.15** (0.37)	
		SASO	3.34 (0.59)	29.96	3.60 (0.47)	29.60	3.33 (0.28)	16.37	3.66 (0.66)	27.05	3.71 (0.46)	29.07	3.54 (0.28)	15.57	3.70 (0.68)	26.48	3.66 (0.46)	29.87	3.54 (0.27)	16.83	3.67 (0.68)	28.18	3.59 (0.46)	31.80	3.50 (0.28)	18.63
		MASO	3.03 (0.63)	-9.14	3.28 (0.52)	-8.88	3.04 (0.39)	-8.73	3.54 (0.62)	-3.08	3.63 (0.43)	-1.98	3.46 (0.28)	-2.33	3.63 (0.68)	-1.78	3.61 (0.47)	-1.58	3.48 (0.29)	-1.52	3.63 (0.69)	-1.12	3.56 (0.46)	-0.94	3.45 (0.29)	-1.37
	16	SAMO	**6.67** (2.55)		**7.40** (1.48)		**5.62** (1.26)		**7.54** (3.32)		**7.73** (1.53)		**6.23** (1.45)		**7.55** (3.31)		**7.60** (1.49)		**6.27** (1.39)		**7.52** (3.48)		**7.48** (1.46)		**6.25** (1.34)	
		SASO	4.75 (1.70)	40.48	5.17 (1.39)	43.18	4.64 (0.82)	21.03	5.29 (2.03)	42.52	5.31 (1.37)	45.58	4.97 (0.85)	25.47	5.34 (2.06)	41.46	5.20 (1.29)	46.23	4.93 (0.88)	27.14	5.35 (2.05)	40.62	5.11 (1.29)	46.30	4.90 (0.90)	27.64
		MASO	4.20 (1.69)	-11.61	4.63 (1.50)	-10.35	4.11 (1.04)	-11.38	5.10 (1.96)	-3.62	5.23 (1.36)	-1.42	4.88 (0.90)	-1.81	5.13 (2.00)	-3.78	5.07 (1.32)	-2.44	4.79 (0.82)	-2.76	5.14 (2.03)	-3.88	4.97 (1.26)	-2.89	4.75 (0.85)	-2.98
Gemini 2.5 Flash-lite	2	SAMO	**1.62** (0.03)		1.67 (0.02)		1.58 (0.02)		**1.68** (0.02)		**1.71** (0.01)		**1.63** (0.02)		**1.68** (0.03)		**1.72** (0.01)		**1.64** (0.02)		**1.69** (0.03)		**1.72** (0.01)		**1.64** (0.02)	
		SASO	1.62 (0.03)	0.06	**1.67** (0.02)	-0.14	**1.60** (0.02)	-1.62	1.62 (0.03)	3.68	1.67 (0.02)	2.33	1.62 (0.02)	0.58	1.64 (0.03)	2.55	1.67 (0.02)	2.68	1.62 (0.02)	0.98	1.65 (0.03)	2.43	1.68 (0.02)	2.72	1.63 (0.02)	0.47
		MASO	1.56 (0.04)	-3.50	1.60 (0.03)	-4.24	1.53 (0.03)	-4.40	1.60 (0.04)	-1.22	1.64 (0.02)	-1.92	1.57 (0.02)	-3.09	1.61 (0.04)	-1.76	1.65 (0.02)	-1.32	1.59 (0.02)	-2.02	1.63 (0.03)	-1.19	1.67 (0.02)	-0.56	1.60 (0.02)	-1.68
	4	SAMO	**2.62** (0.18)		**2.78** (0.10)		**2.47** (0.10)		**2.79** (0.20)		**2.89** (0.11)		**2.58** (0.10)		**2.82** (0.20)		**2.91** (0.10)		**2.62** (0.09)		**2.82** (0.18)		**2.91** (0.09)		**2.61** (0.09)	
		SASO	2.49 (0.16)	5.39	2.65 (0.09)	5.16	2.43 (0.08)	1.62	2.53 (0.17)	10.42	2.64 (0.10)	9.46	2.48 (0.09)	4.11	2.57 (0.17)	9.70	2.67 (0.11)	9.07	2.51 (0.10)	4.18	2.61 (0.19)	8.03	2.68 (0.12)	8.70	2.53 (0.11)	2.93
		MASO	2.20 (0.21)	-11.84	2.31 (0.17)	-12.69	2.13 (0.16)	-12.21	2.46 (0.21)	-2.77	2.56 (0.15)	-3.02	2.37 (0.14)	-4.38	2.52 (0.21)	-2.20	2.60 (0.14)	-2.76	2.43 (0.14)	-3.28	2.51 (0.20)	-3.86	2.60 (0.13)	-3.05	2.42 (0.12)	-4.60
	8	SAMO	**4.08** (0.78)		**4.48** (0.45)		**3.69** (0.40)		**4.40** (0.89)		**4.68** (0.49)		**3.91** (0.44)		**4.47** (0.94)		**4.72** (0.48)		**3.97** (0.46)		**4.52** (0.98)		**4.76** (0.50)		**4.00** (0.45)	
		SASO	3.69 (0.57)	10.56	4.01 (0.36)	11.73	3.54 (0.33)	4.26	3.80 (0.68)	15.71	4.03 (0.43)	16.13	3.63 (0.38)	7.62	3.82 (0.75)	16.93	4.02 (0.47)	17.40	3.66 (0.43)	8.39	3.81 (0.74)	18.64	4.00 (0.47)	19.15	3.65 (0.41)	9.60
		MASO	2.92 (0.69)	-20.74	3.18 (0.61)	-20.85	2.80 (0.51)	-20.87	3.48 (0.79)	-8.42	3.72 (0.57)	-7.54	3.31 (0.49)	-9.01	3.52 (0.88)	-7.96	3.74 (0.60)	-7.11	3.34 (0.53)	-8.92	3.49 (0.86)	-8.33	3.69 (0.62)	-7.54	3.30 (0.54)	-9.50
	16	SAMO	**6.40** (3.11)		**7.23** (1.99)		**5.40** (1.50)		**7.01** (4.01)		**7.52** (2.34)		**5.85** (1.90)		**6.99** (4.43)		**7.48** (2.57)		**5.84** (1.96)		**7.02** (4.36)		**7.49** (2.47)		**5.84** (1.89)	
		SASO	5.51 (2.13)	16.23	6.06 (1.42)	19.38	5.17 (1.18)	4.51	5.59 (2.45)	25.49	6.00 (1.59)	25.24	5.23 (1.31)	11.84	5.62 (2.57)	24.33	6.00 (1.76)	24.74	5.26 (1.44)	11.06	5.63 (2.58)	24.71	5.98 (1.76)	25.33	5.26 (1.43)	10.92
		MASO	4.17 (2.17)	-24.24	4.62 (1.98)	-23.76	3.89 (1.51)	-24.67	5.05 (2.70)	-9.51	5.43 (2.09)	-9.51	4.65 (1.55)	-11.19	5.09 (2.82)	-9.41	5.41 (2.16)	-9.90	4.66 (1.64)	-11.42	5.02 (2.97)	-10.92	5.30 (2.29)	-11.29	4.61 (1.75)	-12.42
Qwen3-32B	2	SAMO	**1.70** (0.02)		**1.71** (0.01)		**1.65** (0.02)		**1.74** (0.02)		**1.74** (0.01)		**1.68** (0.02)		**1.75** (0.02)		**1.74** (0.01)		**1.68** (0.01)		**1.75** (0.02)		**1.74** (0.01)		**1.69** (0.01)	
		SASO	1.68 (0.03)	1.20	1.70 (0.02)	0.81	1.65 (0.02)	0.32	1.68 (0.03)	4.13	1.68 (0.02)	3.30	1.64 (0.02)	2.44	1.69 (0.03)	3.00	1.70 (0.02)	2.57	1.65 (0.02)	1.55	1.70 (0.03)	2.77	1.69 (0.02)	2.57	1.65 (0.02)	2.00
		MASO	1.63 (0.04)	-2.89	1.66 (0.03)	-2.47	1.62 (0.03)	-1.59	1.66 (0.03)	-1.13	1.67 (0.02)	-1.02	1.62 (0.02)	-1.00	1.68 (0.03)	-0.79	1.69 (0.02)	-0.41	1.65 (0.02)	-0.31	1.69 (0.03)	-0.48	1.69 (0.02)	-0.39	1.66 (0.02)	0.16
	4	SAMO	**2.85** (0.20)		**2.90** (0.13)		**2.68** (0.11)		**3.01** (0.18)		**2.99** (0.10)		**2.76** (0.10)		**3.01** (0.17)		**2.97** (0.09)		**2.78** (0.09)		**2.99** (0.20)		**2.95** (0.11)		**2.77** (0.11)	
		SASO	2.69 (0.17)	6.10	2.77 (0.11)	4.74	2.58 (0.10)	4.03	2.74 (0.16)	9.85	2.76 (0.12)	8.36	2.58 (0.10)	6.92	2.75 (0.15)	9.51	2.74 (0.11)	8.30	2.60 (0.10)	7.11	2.74 (0.16)	8.86	2.73 (0.11)	7.86	2.61 (0.11)	6.34
		MASO	2.51 (0.22)	-6.83	2.55 (0.16)	-7.84	2.43 (0.15)	-5.86	2.63 (0.16)	-4.01	2.68 (0.12)	-3.01	2.52 (0.12)	-2.54	2.71 (0.17)	-1.47	2.71 (0.12)	-1.03	2.57 (0.11)	-1.16	2.73 (0.17)	-0.49	2.71 (0.11)	-0.94	2.59 (0.11)	-0.52
	8	SAMO	**4.72** (1.00)		**4.85** (0.59)		**4.17** (0.55)		**5.13** (0.88)		**5.09** (0.43)		**4.42** (0.48)		**5.11** (0.93)		**5.01** (0.54)		**4.42** (0.55)		**5.15** (0.86)		**5.06** (0.47)		**4.48** (0.45)	
		SASO	4.22 (0.70)	12.03	4.35 (0.53)	11.65	3.86 (0.45)	8.17	4.25 (0.67)	20.77	4.26 (0.50)	19.51	3.83 (0.42)	15.57	4.25 (0.68)	20.26	4.19 (0.50)	19.55	3.86 (0.42)	14.41	4.21 (0.75)	22.33	4.13 (0.59)	22.37	3.83 (0.48)	16.85
		MASO	3.68 (0.91)	-12.65	3.82 (0.68)	-12.15	3.48 (0.57)	-9.71	4.05 (0.70)	-4.65	4.11 (0.46)	-3.52	3.71 (0.40)	-3.06	4.21 (0.78)	-0.81	4.21 (0.50)	0.42	3.85 (0.46)	-0.44	4.18 (0.80)	-0.76	4.15 (0.53)	0.31	3.82 (0.47)	-0.25
	16	SAMO	**7.66** (4.26)		**7.97** (2.47)		**6.35** (1.72)		**8.33** (4.30)		**8.29** (2.17)		**6.73** (1.55)		**8.38** (4.40)		**8.26** (2.18)		**6.83** (1.78)		**8.37** (4.38)		**8.25** (2.03)		**6.83** (1.67)	
		SASO	6.38 (2.43)	20.19	6.61 (1.87)	20.55	5.60 (1.40)	13.26	6.32 (2.48)	31.68	6.29 (1.50)	31.97	5.58 (1.43)	20.60	6.19 (2.41)	35.30	6.08 (2.01)	35.73	5.48 (1.43)	24.48	6.16 (2.49)	36.02	6.00 (2.01)	37.65	5.46 (1.42)	25.18
		MASO	5.24 (2.55)	-17.83	5.50 (1.92)	-16.76	4.85 (1.47)	-13.45	6.05 (2.08)	-4.28	6.09 (1.43)	-3.04	5.37 (1.10)	-3.70	6.08 (2.22)	-1.87	5.99 (1.57)	-1.55	5.38 (1.22)	-1.89	6.01 (2.51)	-2.38	5.83 (1.71)	-2.70	5.33 (1.36)	-2.35

Table 1: For each chat model, embedding model and k, the table reports the performance of three conditions (SAMO, SASO, and MASO). Each "Value" entry is the mean Vendi score for that condition and round, with the value in parentheses giving the corresponding variance across items. For each round, the two Δ (%) columns show relative percentage changes: the Δ above is SAMO versus SASO, and the Δ below is MASO versus SASO, both computed as $(\text{condition} - \text{SASO})\,/\text{SASO} \times 100$.

Model	k	Setting	Round 1						Round 2						Round 3						Round 4					
			BERT		GPT-emb		Qwen3-emb		BERT		GPT-emb		Qwen3-emb		BERT		GPT-emb		Qwen3-emb		BERT		GPT-emb		Qwen3-emb	
			Value	Δ (%)	Value	Δ (%)	Value	Δ (%)	Value	Δ (%)	Value	Δ (%)	Value	Δ (%)	Value	Δ (%)	Value	Δ (%)	Value	Δ (%)	Value	Δ (%)	Value	Δ (%)	Value	Δ (%)
GPT 4.1-mini	8	1x8	**4.34 (0.88)**		**4.67 (0.66)**		**3.88 (0.60)**		**4.64 (0.89)**		**4.79 (0.64)**		**4.09 (0.59)**		**4.68 (0.92)**		**4.76 (0.65)**		**4.13 (0.59)**		**4.71 (0.98)**		**4.73 (0.66)**		**4.15 (0.60)**	
		2x4	4.09 (0.83)	-5.69	4.41 (0.65)	-5.70	3.74 (0.57)	-3.58	4.45 (0.89)	-4.16	4.59 (0.63)	-4.08	4.04 (0.59)	-1.36	4.50 (0.90)	-3.77	4.57 (0.64)	-4.07	4.06 (0.58)	-1.59	4.47 (0.93)	-5.12	4.53 (0.68)	-4.36	4.06 (0.60)	-2.28
		4x2	3.70 (0.79)	-9.64	3.98 (0.64)	-9.58	3.51 (0.59)	-6.17	4.15 (0.84)	-6.79	4.26 (0.63)	-7.17	3.89 (0.55)	-3.65	4.16 (0.84)	-7.47	4.20 (0.61)	-7.93	3.89 (0.54)	-4.32	4.14 (0.85)	-7.25	4.14 (0.61)	-8.51	3.87 (0.55)	-4.57
		8x1	3.03 (0.80)	-17.95	3.29 (0.72)	-17.52	3.04 (0.62)	-13.32	3.54 (0.78)	-14.61	3.63 (0.65)	-14.69	3.46 (0.53)	-11.08	3.63 (0.83)	-12.78	3.61 (0.68)	-14.20	3.48 (0.54)	-10.48	3.63 (0.83)	-12.35	3.56 (0.68)	-14.11	3.45 (0.54)	-10.84
	16	1x16	**6.07 (1.60)**		**7.31 (1.29)**		**5.62 (1.12)**		**7.54 (1.82)**		**7.64 (1.31)**		**6.23 (1.20)**		**7.55 (1.82)**		**7.51 (1.30)**		**6.27 (1.18)**		**7.52 (1.87)**		**7.39 (1.29)**		**6.25 (1.16)**	
		2x8	6.39 (1.50)	-4.18	7.08 (1.17)	-3.19	5.49 (1.05)	-2.27	7.13 (1.69)	-5.43	7.35 (1.18)	-3.83	6.05 (1.11)	-2.94	7.09 (1.69)	-6.11	7.21 (1.18)	-3.92	6.06 (1.10)	-3.30	7.08 (1.70)	-5.87	7.12 (1.13)	-3.69	6.08 (1.07)	-2.82
		4x4	6.01 (1.40)	-5.98	6.64 (1.14)	-6.18	5.30 (1.03)	-3.46	6.66 (1.58)	-6.55	6.92 (1.13)	-5.90	5.89 (1.04)	-2.69	6.75 (1.60)	-4.68	6.85 (1.12)	-4.98	5.96 (1.03)	-1.62	6.76 (1.62)	-4.46	6.77 (1.12)	-4.91	5.98 (1.04)	-1.59
		8x2	5.35 (1.37)	-11.05	5.89 (1.15)	-11.28	4.92 (1.03)	-7.08	6.14 (1.50)	-7.83	6.34 (1.12)	-8.33	5.65 (1.02)	-3.98	6.20 (1.53)	-8.21	6.21 (1.11)	-9.46	5.65 (1.02)	-5.17	6.14 (1.50)	-9.23	6.08 (1.07)	-10.12	5.62 (1.00)	-5.99
		16x1	4.20 (1.30)	-21.48	4.64 (1.22)	-21.32	4.11 (1.02)	-16.49	5.10 (1.40)	-16.99	5.24 (1.17)	-17.36	4.88 (0.95)	-13.70	5.13 (1.41)	-17.20	5.08 (1.16)	-18.23	4.79 (0.91)	-15.22	5.14 (1.43)	-16.26	4.97 (1.12)	-18.28	4.75 (0.92)	-15.46
Gemini 2.5 Flash-lite	8	1x8	**4.08 (0.88)**		**4.48 (0.67)**		**3.69 (0.63)**		**4.40 (0.94)**		**4.68 (0.70)**		**3.91 (0.66)**		**4.47 (0.97)**		**4.72 (0.69)**		**3.97 (0.68)**		**4.52 (0.99)**		**4.76 (0.71)**		**4.00 (0.67)**	
		2x4	3.94 (0.87)	-3.54	4.29 (0.66)	-4.36	3.60 (0.60)	-2.57	4.29 (0.96)	-2.55	4.54 (0.71)	-3.01	3.86 (0.67)	-1.20	4.26 (0.94)	-4.75	4.52 (0.71)	-4.33	3.88 (0.67)	-2.27	4.30 (0.96)	-4.89	4.56 (0.71)	-4.27	3.90 (0.66)	-2.46
		4x2	3.53 (0.81)	-10.41	3.86 (0.65)	-9.99	3.33 (0.61)	-7.50	4.03 (0.89)	-6.05	4.28 (0.66)	-5.61	3.73 (0.64)	-3.47	4.04 (0.89)	-5.25	4.30 (0.68)	-4.94	3.76 (0.67)	-3.04	4.08 (0.91)	-5.03	4.32 (0.69)	-5.14	3.77 (0.65)	-3.33
		8x1	2.92 (0.83)	-17.03	3.18 (0.78)	-17.71	2.80 (0.71)	-15.79	3.48 (0.89)	-13.56	3.72 (0.76)	-13.04	3.31 (0.70)	-11.34	3.52 (0.94)	-12.78	3.74 (0.78)	-13.01	3.34 (0.73)	-11.32	3.49 (0.93)	-14.47	3.69 (0.79)	-14.54	3.30 (0.73)	-12.43
	16	1x16	**6.40 (1.76)**		**7.21 (1.40)**		**5.40 (1.22)**		**7.01 (2.00)**		**7.50 (1.53)**		**5.88 (1.38)**		**6.99 (2.10)**		**7.46 (1.61)**		**5.84 (1.40)**		**7.02 (2.09)**		**7.48 (1.59)**		**5.84 (1.37)**	
		2x8	6.15 (1.69)	-3.94	6.89 (1.35)	-4.46	5.28 (1.17)	-2.21	6.68 (1.85)	-4.73	7.18 (1.45)	-4.15	5.70 (1.31)	-2.50	6.77 (1.93)	-3.16	7.26 (1.49)	-2.63	5.77 (1.33)	-1.19	6.75 (1.88)	-3.90	7.25 (1.46)	-3.09	5.77 (1.33)	-1.07
		4x4	5.82 (1.56)	-5.38	6.45 (1.30)	-6.39	5.12 (1.14)	-3.00	6.36 (1.75)	-4.76	6.86 (1.36)	-4.57	5.61 (1.24)	-1.61	6.47 (1.75)	-4.35	6.92 (1.36)	-4.68	5.67 (1.27)	-1.73	6.50 (1.78)	-3.71	6.93 (1.41)	-4.35	5.66 (1.30)	-1.90
		8x2	5.19 (1.50)	-10.85	5.73 (1.26)	-11.18	4.72 (1.12)	-7.89	5.97 (1.70)	-6.06	6.39 (1.35)	-6.76	5.35 (1.22)	-4.76	6.04 (1.72)	-6.66	6.45 (1.39)	-6.77	5.43 (1.26)	-4.25	6.01 (1.75)	-7.54	6.40 (1.40)	-7.61	5.39 (1.26)	-4.86
		16x1	4.17 (1.47)	-19.57	4.62 (1.40)	-19.42	3.89 (1.23)	-17.51	5.05 (1.64)	-15.40	5.43 (1.44)	-15.02	4.65 (1.24)	-13.09	5.09 (1.68)	-15.73	5.40 (1.47)	-16.30	4.66 (1.28)	-14.21	5.02 (1.72)	-16.51	5.30 (1.51)	-17.21	4.61 (1.32)	-14.49
Qwen3-32B	8	1x8	**4.72 (1.00)**		**4.86 (0.77)**		**4.17 (0.74)**		**5.13 (0.94)**		**5.10 (0.66)**		**4.42 (0.70)**		**5.11 (0.97)**		**5.02 (0.74)**		**4.42 (0.73)**		**5.15 (0.93)**		**5.07 (0.69)**		**4.48 (0.67)**	
		2x4	4.56 (0.95)	-3.55	4.70 (0.73)	-3.18	4.16 (0.68)	-0.36	4.93 (0.92)	-3.91	4.92 (0.68)	-3.35	4.39 (0.65)	-0.78	4.98 (0.91)	-2.54	4.93 (0.68)	-1.70	4.41 (0.65)	-0.20	4.94 (0.91)	-4.17	4.87 (0.68)	-3.87	4.39 (0.64)	-2.01
		4x2	4.23 (0.87)	-7.17	4.38 (0.67)	-6.99	3.96 (0.65)	-4.79	4.67 (0.78)	-5.45	4.69 (0.60)	-4.75	4.22 (0.61)	-3.99	4.68 (0.85)	-5.98	4.68 (0.60)	-5.11	4.25 (0.63)	-3.62	4.72 (0.83)	-4.47	4.67 (0.59)	-4.22	4.25 (0.63)	-3.01
		8x1	3.68 (0.95)	-12.92	3.82 (0.82)	-12.76	3.48 (0.75)	-12.02	4.05 (0.84)	-13.11	4.12 (0.68)	-12.26	3.71 (0.63)	-11.95	4.21 (0.88)	-9.98	4.21 (0.70)	-9.99	3.85 (0.68)	-9.53	4.18 (0.89)	-11.39	4.15 (0.73)	-11.04	3.82 (0.68)	-10.18
	16	1x16	**7.66 (2.06)**		**7.96 (1.56)**		6.35 (1.31)		**8.33 (2.07)**		**8.30 (1.47)**		6.73 (1.25)		**8.38 (2.10)**		**8.25 (1.47)**		**6.83 (1.33)**		**8.37 (2.09)**		**8.25 (1.42)**		**6.83 (1.29)**	
		2x8	7.56 (1.94)	-1.39	7.79 (1.49)	-2.18	**6.42 (1.34)**	1.16	8.25 (1.98)	-0.92	8.12 (1.47)	-2.10	**6.83 (1.28)**	1.52	8.16 (2.04)	-2.56	7.98 (1.45)	-3.28	6.82 (1.29)	-0.04	8.10 (1.98)	-3.22	7.92 (1.44)	-3.97	6.78 (1.27)	-0.69
		4x4	7.05 (1.73)	-6.76	7.32 (1.40)	-6.02	6.18 (1.26)	-3.74	7.75 (1.82)	-6.07	7.72 (1.37)	-4.98	6.61 (1.24)	-3.19	7.88 (1.80)	-3.44	7.71 (1.33)	-3.37	6.71 (1.20)	-1.74	7.86 (1.79)	-3.07	7.66 (1.35)	-3.33	6.73 (1.20)	-0.77
		8x2	6.38 (1.64)	-9.49	6.60 (1.32)	-9.76	5.76 (1.23)	-6.79	7.32 (1.64)	-5.62	7.26 (1.21)	-5.87	6.40 (1.16)	-3.20	7.33 (1.62)	-6.96	7.24 (1.25)	-6.08	6.43 (1.19)	-4.13	7.30 (1.74)	-7.13	7.10 (1.28)	-7.30	6.41 (1.17)	-4.74
		16x1	5.24 (1.60)	-17.85	5.50 (1.40)	-16.78	4.85 (1.21)	-15.81	6.05 (1.44)	-17.24	6.08 (1.19)	-16.29	5.37 (1.05)	-16.07	6.08 (1.49)	-17.16	5.98 (1.25)	-17.36	5.38 (1.10)	-16.29	6.01 (1.59)	-17.62	5.82 (1.30)	-17.98	5.33 (1.17)	-16.90

Table 2: **Quantitative results for multi-agent multi-output (MA-MO) configurations.** This table summarizes the results from the *Exploring MA-MO* analysis. Each configuration is denoted as $n \times m$, where n indicates the number of agents and each agent generates m outputs, with the total number of answers fixed at $k = n \cdot m$. The special cases $1 \times k$ and $k \times 1$ correspond to the previously defined *SA-MO* and *MA-SO* protocol, respectively. For each model, value of k, and generation round, the table reports the mean Vendi score, with variance shown in parentheses. The Δ (%) column indicates the relative change in diversity between consecutive configurations, computed by ordering them in ascending number of agents and comparing each setting to the one immediately preceding it.

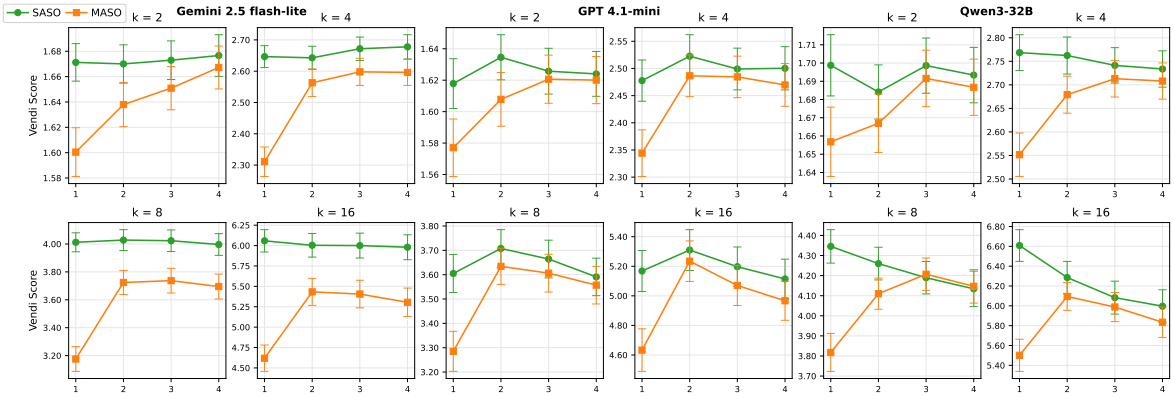

Figure 5: **Single-agent vs. multi-agent generation under single-output settings with text-embedding-3-small**

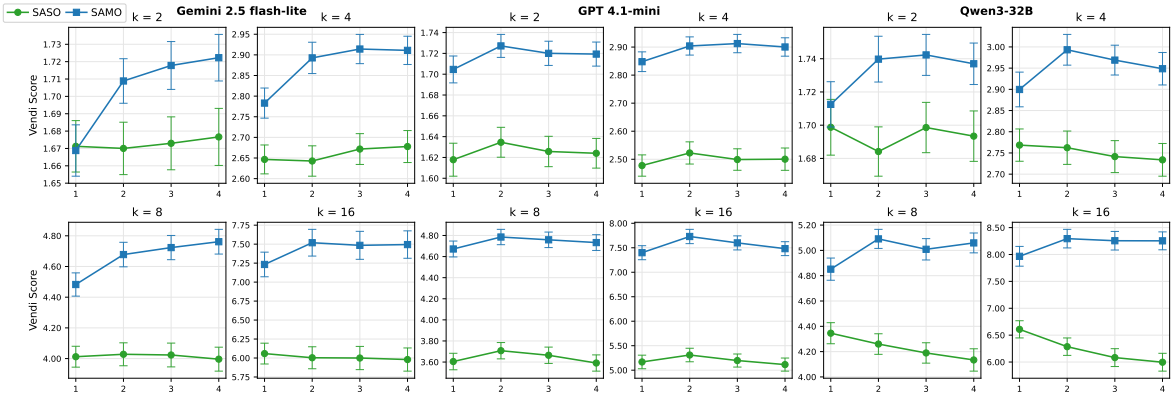

Figure 6: **Single-output vs. multi-output generation under single-agent settings with text-embedding-3-small**

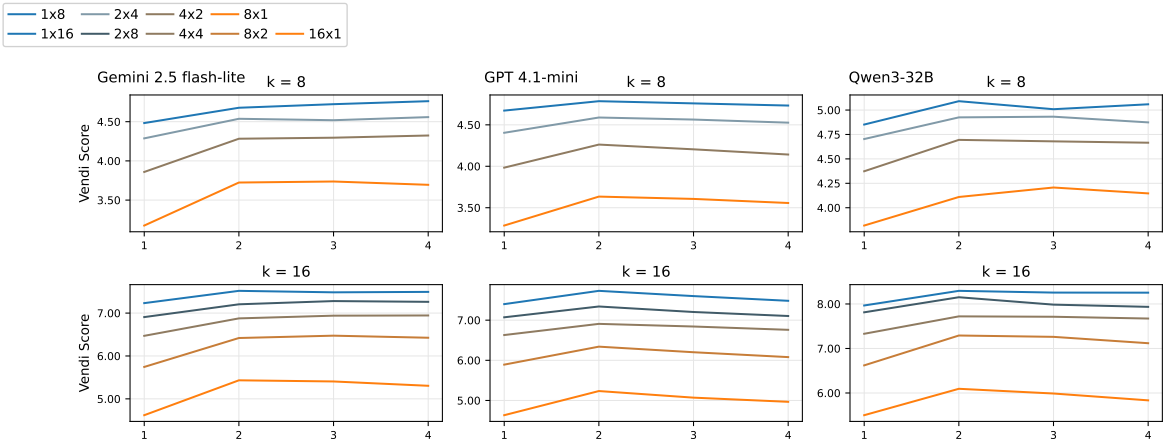

Figure 7: **Exploring multi-agent multi-output (MA-MO) configurations with text-embedding-3-small**

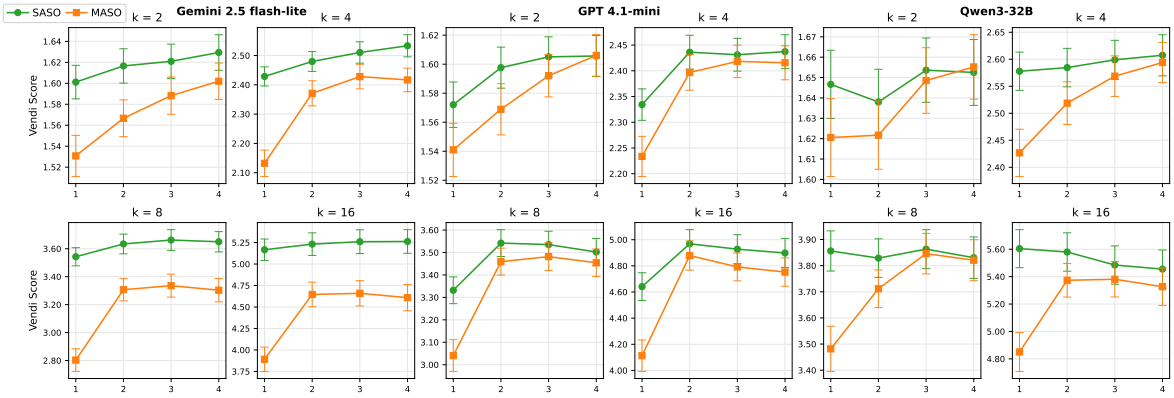

Figure 8: **Single-agent vs. multi-agent generation under single-output settings with Qwen3 8b-embedding**

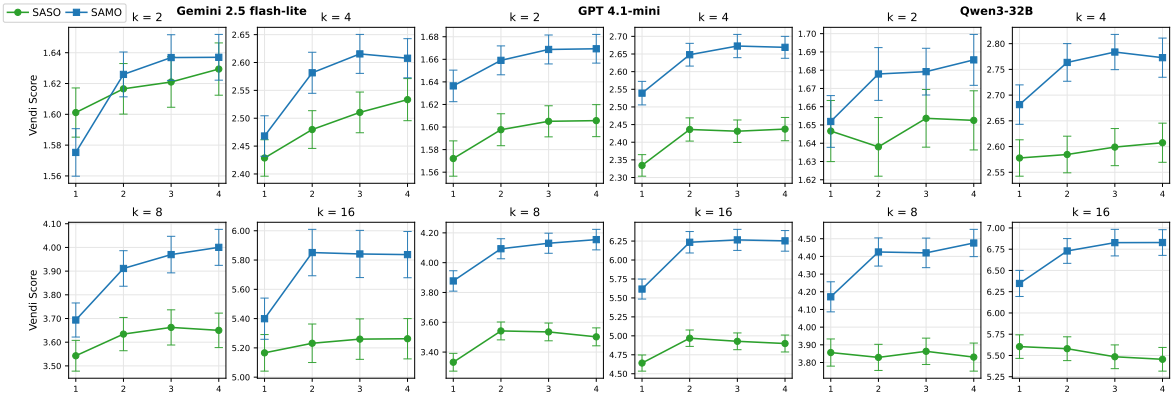

Figure 9: **Single-output vs. multi-output generation under single-agent settings with Qwen3 8b-embedding**

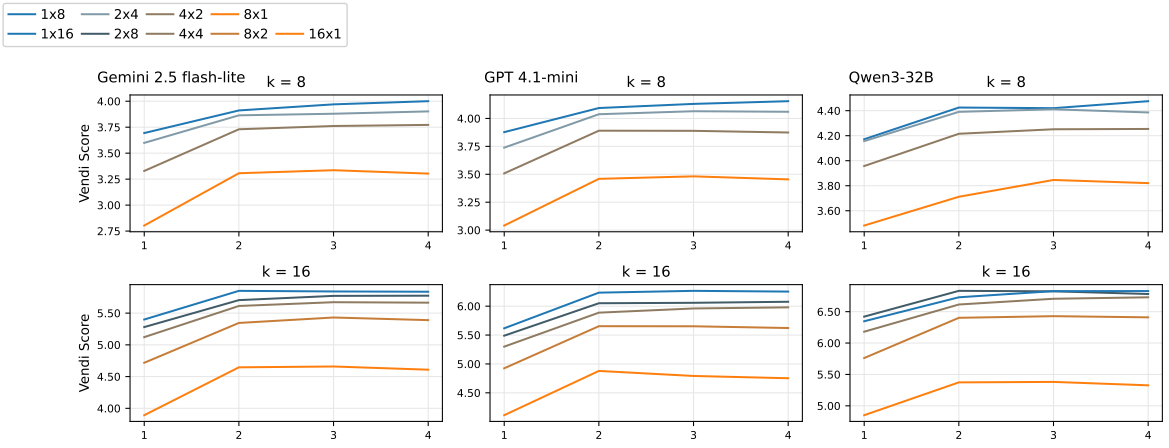

Figure 10: **Exploring multi-agent multi-output (MA-MO) configurations with Qwen3 8b-embeddin**

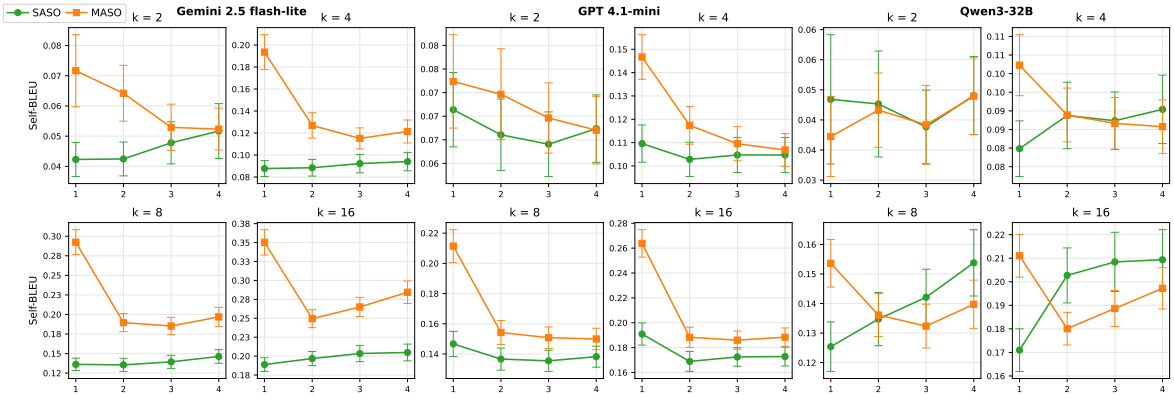

Figure 11: **Single-agent vs. multi-agent generation under single-output settings with Self-BLEU**

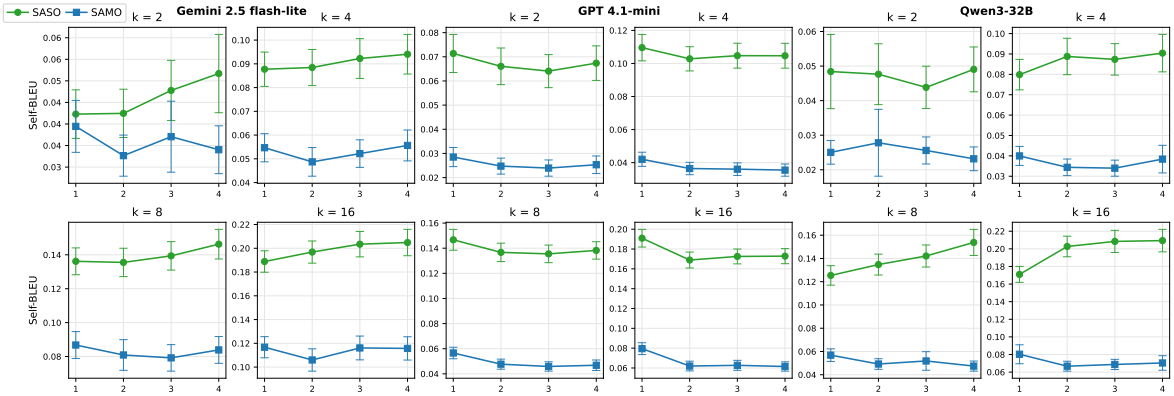

Figure 12: **Single-output vs. multi-output generation under single-agent settings with Self-BLEU**

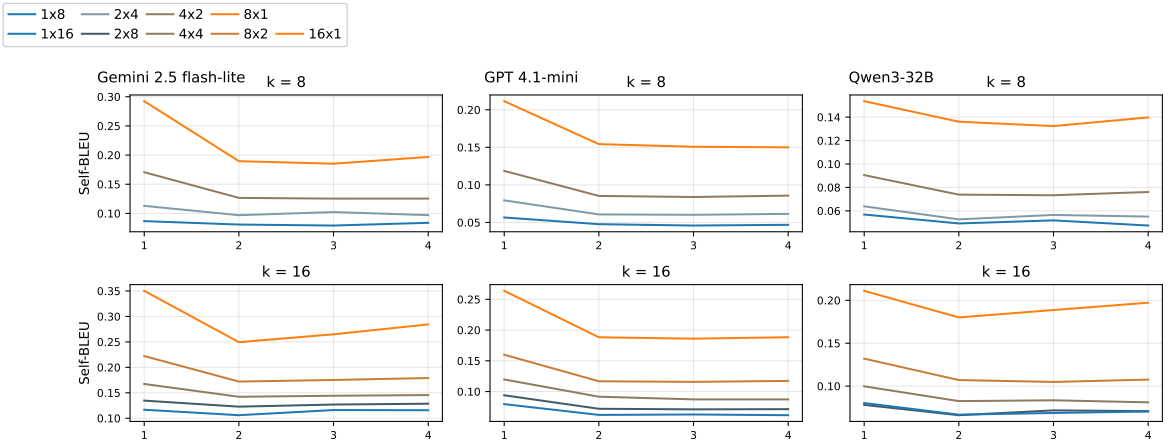

Figure 13: **Exploring multi-agent multi-output (MA-MO) configurations with Self-BLEU**