On the Importance of Task Complexity in Evaluating LLM-Based Multi-Agent Systems

Bohan Tang[1]*, Huidong Liang[1]*, Keyue Jiang[2]*, Xiaowen Dong[1]
[1]University of Oxford, [2]University College London
bohan.tang@eng.ox.ac.uk

Abstract

Large language model multi-agent systems (LLM-MAS) offer a promising paradigm for harnessing collective intelligence to achieve more advanced forms of AI behaviour. While recent studies suggest that LLM-MAS can outperform LLM single-agent systems (LLM-SAS) on certain tasks, the lack of systematic experimental designs limits the strength and generality of these conclusions. We argue that a principled understanding of task complexity, such as the degree of sequential reasoning required and the breadth of capabilities involved, is essential for assessing the effectiveness of LLM-MAS in task solving. To this end, we propose a theoretical framework characterising tasks along two dimensions: depth, representing reasoning length, and width, representing capability diversity. We theoretically examine a representative class of LLM-MAS, namely the multi-agent debate system, and empirically evaluate its performance in both discriminative and generative tasks with varying depth and width. Theoretical and empirical results show that the benefit of LLM-MAS over LLM-SAS increases with both task depth and width, and the effect is more pronounced with respect to depth. This clarifies when LLM-MAS are beneficial and provides a principled foundation for designing future LLM-MAS methods and benchmarks.

1 Introduction

Recent advances in AI technologies, exemplified by autonomous agents powered by large language models (LLMs), offer an exciting opportunity to explore the notion of a future society where machines assist and interact with humans in a wide range of tasks [1, 2, 3, 4, 5]. Rather than operating in isolation, multiple specialised agents can be deployed together, communicating and collaborating to address complex challenges. Consequently, LLM multi-agent systems (LLM-MAS) represent a promising paradigm for leveraging collective intelligence to achieve advanced forms of AI behaviours [6, 7]. Understanding the effectiveness of such systems in task-solving is key to their deployment.

Current effort for understanding LLM-MAS in task solving predominantly focuses on empirically assessing the capability of the system in solving a wide range of tasks, including both discriminative (e.g., problem solving) and generative tasks (e.g., creative writing) [8, 9, 10]. Given a particular task, there is an increasingly large literature that investigates how better task-solving performance can be achieved via the optimisation of prompt engineering, as well as communication networks among AI agents [11, 12, 13, 14, 15]. While recent empirical results have shown that LLM-MAS can outperform LLM single-agent systems (LLM-SAS) on certain tasks, these findings are based solely on downstream performance.Therefore, we still lack a principled understanding of when and why an LLM-MAS system can be effective, i.e., it can outperform an LLM-SAS in task-solving.

*Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Scaling Environments for Agents (SEA).

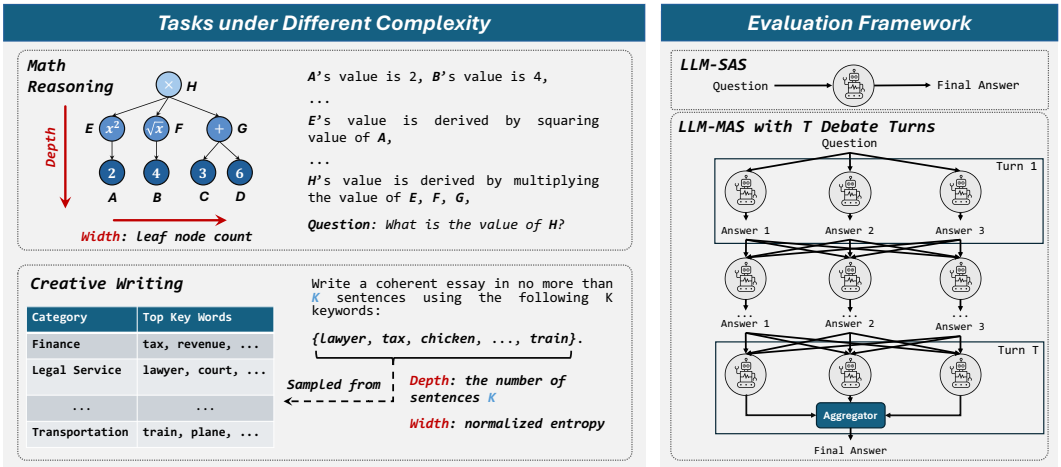


Figure 1: **Left**: math reasoning and creative writing tasks with controllable complexity in terms of width and depth. **Right**: Exemplar LLM-SAS and LLM-MAS framework. For simplicity, the input question to each agent is omitted from the presentation starting with the second debate turn.

In this work, we argue that a deeper understanding of the task complexity is the key to assessing the effectiveness of LLM-MAS in task-solving. Specifically, we focus on a representative class of LLM-MAS, namely a multi-agent debate system, in which multiple LLM agents collaboratively address a single problem, exchanging both their reasoning pathways and solution proposals throughout the process [8, 12, 13, 14]. Drawn from the literature on complex systems and collective intelligence [16, 17], we introduce a two-dimensional measure of task complexity: **depth** and **width**. The depth of a task corresponds to the length of the reasoning chain an agent must follow, namely, the number of sequential inference or problem-solving steps required. The width of a task corresponds to the breadth of knowledge or skills required at each step, i.e., the range of capabilities that the agent needs to draw on per step. Figure 1 illustrates how LLM-SAS and LLM-MAS address tasks with varying levels of width and depth. Based on this two-dimensional measure, we propose a novel *theoretical framework* to analyse how both task width and depth influence the performance of an LLM-MAS. We further demonstrate that, relative to the increase in task width, an increase in task depth tends to yield greater performance gains for LLM-MAS over LLM-SAS.

We empirically validate our theoretical findings in two representative tasks: math reasoning (a discriminative task) and creative writing (a generative task). First, in the math reasoning task, we employ the framework provided by the DyVal benchmark [18] to dynamically adjust the task complexity, where the depth and width are naturally defined as those of the tree-based directed acyclic graph (DAG) used to generate the problem. Second, in the creative writing task, we propose a novel benchmark, named Depth-Width Writing (DW^2) benchmark, where the depth and width are the number of coherent sentences that are required to be written and the Shannon entropy of the domains of keywords to be used, respectively. Across both tasks, we observe two consistent patterns: 1) the benefit of LLM-MAS over LLM-SAS increases with task complexity; and 2) gains in task depth appear more significant compared to those in task width. These results affirm our theoretical findings and our main argument: a deeper understanding of the task complexity is the key to assessing the effectiveness of LLM-MAS in task-solving. Based on this insight, we discuss the open challenges and opportunities for designing future LLM-MAS. Our contributions are summarised as follows:

- We propose a novel theoretical framework to analyse the effectiveness of LLM-MAS. The framework is grounded in a two-dimensional measure of task complexity based on **depth** and **width**, which capture the degree of sequential reasoning and breadth of capabilities required for a task.

- We propose a simple framework to theoretically show that the performance gain of an LLM-MAS over an LLM-SAS grows with task complexity. Crucially, our analysis reveals that this advantage is more sensitive to increases in task depth than to that in task width.

- We provide extensive empirical validation for our theoretical findings in two representative tasks: math reasoning (a discriminative task) and creative writing (a generative task). For the latter, we propose a novel Depth-Width Writing (DW^2) benchmark, where depth is the number of coherent

sentences required and width is the Shannon entropy of keyword domains. The empirical results affirm our theoretical findings and offer insights for designing more efficient LLM-MAS.

2 Analysis Framework

In this section, we develop a theoretical framework for analysing the effectiveness of LLM-MAS. We first set up our problem formulation in Subsection 2.1. Subsequently, in Subsection 2.2, we show how both task width and task depth affect the performance gain of LLM-MAS relative to LLM-SAS. Our analysis reveals that although increasing either dimension improves such performance gain of LLM-MAS, task depth plays a more decisive role than task width.

2.1 Problem Formulation

LLM multi-agent systems. We define LLM multi-agent systems (LLM-MAS) as computational systems comprised of multiple autonomous agents, each empowered by LLMs, that perceive, reason, and collaborate to accomplish complex tasks at scale [19, 6, 7, 20]. Unlike the LLM single-agent systems (LLM-SAS), where only one LLM operates in isolation, LLM-MAS utilises natural language as a communication medium, enabling agents to coordinate, share context, and jointly refine their outputs toward a common goal. In this work, we focus specifically on the multi-agent debate system, a particularly prominent instantiation of LLM-MAS. In this framework, multiple LLM agents engage in proposing, critiquing, and refining arguments [8, 9, 10].

Problem statement. While LLM-MAS offer the potential to improve problem-solving through collaboration, they also introduce additional computational overhead due to multi-agent coordination and repeated communication rounds [21]. Theron, we formalise effectiveness as the relative performance gain of LLM-MAS over its LLM-SAS counterparts on a given task. Our research problem can be formulated as investigating how intrinsic task properties influence the comparative effectiveness of LLM-MAS, to establish a principled understanding of when and why LLM-MAS shows advantages.

2.2 Task Complexity for LLM-MAS Effectiveness

In this subsection, we first introduce a two-dimensional measure for characterising task complexity in terms of **depth** and **width**. We then formalise the success rates of both LLM-SAS and LLM-MAS. Finally, we provide a theoretical analysis showing that: 1) increasing either task width or depth enhances the effectiveness of LLM-MAS relative to their single-agent counterparts; and 2) compared with width, increases in task depth generally yield greater performance gains for LLM-MAS.

Width and depth. Building on prior works in complex systems and collective intelligence [16, 17], we formalise task complexity using a two-dimensional measure defined by **depth** and **width**. The *depth* of a task characterises the length of the reasoning chain an agent must traverse, that is, the number of sequential inference or problem-solving steps required. In contrast, the *width* of a task captures the breadth of knowledge, skills, or alternatives that must be considered at each step, i.e., the range of capabilities an agent must draw upon. Figure 2 provides a visualisation for task depth and width under LLM-SAS and LLM-MAS, which we will introduce as follows.

Definition 2.1 (The task defined by depth and width)**.** Consider a task requiring a reasoning chain of depth $d \in \mathbb{N}$ (i.e., d steps). At each step, the agent must complete $w \in \mathbb{N}$ capability-specific micro-operations, which capture the width of the task[2]. Let the set of involved capabilities at a step be indexed by $j = 1, \ldots, w$, and denote by $q_j \in (0, 1)$ the success rate associated with capability j. Since tasks relying on the same capability have highly correlated outcomes, we treat the success of each capability as a single event, independent across different capabilities. The per-step success rate is thus $s(w) \triangleq \prod_{j=1}^{w} q_j$. Theron, the task can be formalized as the sequence $\mathcal{T} = \{s(w)\}_{t=1}^{d}$.

To illustrate the intuition behind Definition 2.1, consider the two tasks in Figure 1. In the math reasoning task, depth is the number of steps along the reasoning chain, and width corresponds to applying different operators (e.g., addition, multiplication) within a step. Success on addition is independent of success on multiplication, so the per-step probability is the product of their accuracies. In the creative writing task, depth is the number of coherent sentences to be written, while width

[2]While the width can vary across reasoning steps, we assume the width to be fixed across steps for simplicity.

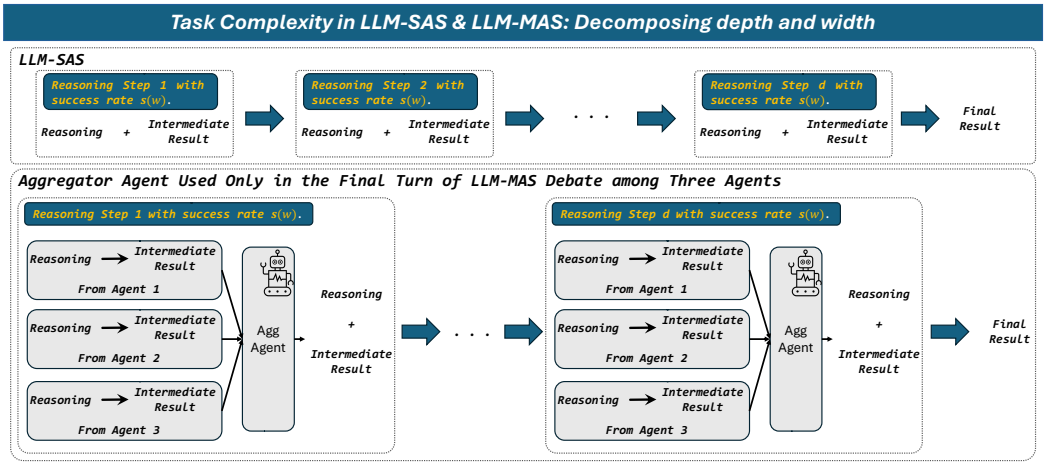


Figure 2: Visualization of task complexity defined by depth and width. The pipeline represents one round of multi-agent debate. "Agg" stands for aggregator.

corresponds to correctly integrating each required keywords. Keywords from different domains are treated as independent, since correctly using an astronomy keyword does not make it easier or harder to integrate a culinary keyword. Based on this definition, we formalise the success rates of LLM-SAS and LLM-MAS. Our formulation focuses on the final round of discussion among the LLM agents. In this round, the answers produced by individual agents are summarised by one aggregator agent. Each agent conditions its response on both its own previous output and the prior responses of others, but the subsequent text generation is executed independently by the underlying LLM. Accordingly, we assume that the final-round outputs of different agents are independent. An illustration of such independent generation is provided in turn T of LLM-MAS in Figure 1. Moreover, for simplicity in subsequent derivations, we assume $q_j = q$ for all j, so that $s(w) = q^w$.

Definition 2.2 (The success rate of LLM-SAS)**.** For a given task defined as $\mathcal{T} = \{s(w)\}_{t=1}^{d}$. A single agent must correctly execute every step. Therefore, the success rate of an LLM-SAS across depth d with width w can be defined as $S_{\text{single}}(d, w) \triangleq \left[s(w)\right]^d$.

Definition 2.3 (The success rate of LLM-MAS)**.** For a given task defined as $\mathcal{T} = \{s(w)\}_{t=1}^{d}$. Suppose we deploy $N \geq 2$ agents and an aggregator agent that, upon observing the outputs, selects the correct answer with probability $r \in (0, 1]$ whenever at least one agent produced it. As a result, the success rate of an LLM-MAS across depth d with width w can be defined as $S_{\text{multi}}(d, w, N, r) \triangleq r\left[1 - \left(1 - s(w)\right)^N\right]^d$.

Definition 2.4 (LLM-MAS Performance Gain over LLM-SAS)**.** For the scenarios with the LLM-MAS outperforming the LLM-SAS, we define the performance gain of the LLM-MAS over the LLM-SAS as the following relative-improvement function: $\Delta(d, w, N, r) \triangleq \frac{S_{\text{multi}} - S_{\text{single}}}{S_{\text{single}}}$.

Theoretical analysis. With Definition 2.4, we establish two key propositions to capture the respective roles of task width and depth in shaping the performance gain of LLM-MAS over LLM-SAS. The proofs for these propositions can be found in the Appendix.

Proposition 2.1 (Increase of LLM-MAS Performance Gain with Depth d and Width w)**.** *Let $\mathcal{T} = \{s(w)\}_{t=1}^{d}$ denote a given task. According to Definition 2.4, we have $\Delta(d, w, N, r) \triangleq \frac{S_{\text{multi}} - S_{\text{single}}}{S_{\text{single}}}$. Then, we have $\frac{\partial \Delta}{\partial d} > 0$ and $\frac{\partial \Delta}{\partial w} > 0$.*

Proposition 2.1 demonstrates that increasing either dimension of task complexity, depth or width, amplifies the relative benefit of LLM-MAS over LLM-SAS. We attribute this to two complementary mechanisms: 1) increasing width lowers the per-step accuracy of any individual agent, thereby heightening the value of redundancy and error mitigation through multiple cooperative agents; and 2) increasing depth compounds this effect across a longer chain of reasoning steps, making single-agent error accumulation more detrimental, while LLM-MAS benefit from collaborative coverage to counteract this compounding failure risk. Consequently, both dimensions independently enhance the performance gain of LLM-MAS over LLM-SAS.

Proposition 2.2 (Unbounded Growth in Depth vs. Finite Saturation in Width)**.** *Let $\mathcal{T} = \{s(w)\}_{t=1}^{d}$ denote a given task. According to Definition 2.4, we have $\Delta(d, w, N, r) \triangleq \frac{S_{\text{multi}} - S_{\text{single}}}{S_{\text{single}}}$. Then, we have $\lim_{w \to +\infty} \Delta(d, w, N, r) = (rN)^d - 1$ and $\lim_{d \to +\infty} \Delta(d, w, N, r) = +\infty$.*

Proposition 2.2 highlights an asymmetry between the proposed two dimensions of task complexity. As the width grows, the benefit of having multiple agents eventually saturates: LLM-MAS cannot gain more than the N-fold redundancy adjusted by the aggregator's reliability r. By contrast, as depth grows, the advantage of LLM-MAS scales without bound. Intuitively, this arises because LLM-SAS suffer from compounding errors that degrade performance exponentially across long reasoning chains, whereas LLM-MAS can mitigate this compounding effect by distributing intermediate reasoning across agents and cross-validating partial results.

Together, Propositions 2.1 and 2.2 establish our central claim: both width and depth improve the relative performance of LLM-MAS, but depth exerts a greater influence.

3 Related Work

Analysis of LLM-MAS effectiveness. The effectiveness of LLM-MAS has been empirically investigated in many recent work. [22] conceptualises LLM-MAS as a form of test-time computational scaling and finds that, while LLM-MAS yields limited benefits over strong LLM-SAS baselines in mathematical reasoning, its benefit grows when tasks become more difficult, model capabilities are limited, and agent diversity contributes minimally. [21] demonstrates that as foundation models improve in reasoning and tool use, the performance gains from LLM-MAS diminish, motivating a hybrid agentic paradigm that selectively uses LLM-MAS and LLM-SAS for different user requests. In contrast, [23] provides a systematic diagnosis of LLM-MAS performance limitations, introducing the Multi-Agent System Failure Taxonomy (MAST), identifying fourteen distinct failure modes. We complement and extend these empirical and diagnostic studies by introducing a theoretical framework linking LLM-MAS effectiveness directly to task complexity, characterised in two dimensions: depth (reasoning chain length) and width (the breadth of capabilities required per step). We further prove that performance gains from increasing width saturate, whereas those from increasing depth can grow unbounded, offering principled insight into when and why LLM-MAS outperform LLM-SAS.

4 Experiments

Following the theoretical findings introduced in Section 2, we empirically evaluate the behaviour of LLM-MAS and LLM-SAS under different task complexities.

4.1 Experiment Setup

We conduct extensive experiments across two complementary tasks: discriminative math reasoning and generative creative writing, as detailed below.

Math reasoning dataset. We consider the DyVal benchmark [18], a synthetic math reasoning dataset in which problems are defined via a tree-based directed acyclic graph (DAG). As illustrated in Figure 1, the root nodes denote the values of variables, where the other nodes represent mathematical operations $\{+, -, \times, \cdot^2, \sqrt{\cdot}\}$. The DAGs are later described in natural language with different task options, and we adopt the `linear equation` problems in our experiment. The task complexity naturally resides in the DAG structure, where ***depth*** (i.e., number of reasoning steps) governs the sequential breakdown of sub-problems, and ***width*** (i.e., number of children per node) controls the variety of operations required at each sub-problem. Therefore, we design our dataset to cover different depth and width levels from 2 to 4, where each depth-width pair contains 100 questions, leading to a total of 900 questions. We select the depth and width values to ensure that questions fit within the Qwen-2.5-32B-Instruct maximum token length and can be fully processed.

Creative writing dataset. Based on the common creative writing tasks [24], we further propose a novel Depth-Width Writing (DW^2) benchmark, in which the task is to write a coherent essay using K sentences with K keywords drawn from across 23 major occupation groups [25] in the United States. The ***depth*** of the task is then captured by the number of sentences K, which dominates the length

of the reasoning path. For ***width***, we use the *normalized Shannon entropy* $\bar{H}(\mathcal{S})$ of the occupation categories in each keyword set $\mathcal{S} = \{c_1, c_2, ..., c_K\}$ with the following expression:

$$\bar{H}(\mathcal{S}) \;=\; \frac{H(\mathcal{S})}{H_{\max}} \;=\; -\frac{1}{H_{\max}} \sum_{i=1}^{K} p(c_i)\, \log_2 p(c_i),$$

where $H_{\max} = \log_2(K)$ is the maximum entropy when all keywords are from different categories, and $p(c_i) = \frac{1}{K}\sum_{j=1}^{K} \mathbb{1}_{c_i = c_j}$ is the probability of category c_i present in the set. Intuitively, a higher $\bar{H}(\mathcal{S})$ suggests a wider range of domain knowledge across occupation groups, hence making the completion of the task requiring broader knowledge. We then randomly sample 500 K-keyword sets for each $K \in \{4, 8, 12, 16, 20\}$, and then split them into 5 quintiles by their $\bar{H}$ scores, i.e., 100 questions per depth-width pair, which yields a total of $2,500$ questions. We select the depth and width values to ensure that any question fits within the Qwen-2.5-32B-Instruct maximum token length.

Implementation. For LLM-SAS, we implement the method based on the chain-of-thought (CoT) technique [26], where a single LLM is prompted to decompose each problem into a sequence of clear, step-by-step reasoning steps before reaching the final answer. For LLM-MAS, we implement the multi-agent debate framework [8], which involves multiple LLM agents independently attempting a task and exchanging their reasoning along with their answers. After the discussion, a dedicated LLM agent will synthesize the information into a final response via summary prompting. For consistency, we use Qwen-2.5-32B-Instruct as the base model across all experiments.

Evaluation. Since our goal is to compare the behaviours of LLM-SAS and LLM-MAS under different task complexities, we will focus on the performance gain (Definition 2.4) at each depth-width pair and monitor how it changes along the depth and width dimensions. Specifically, we use the standard accuracy within each depth-width pair as the metric for the math reasoning task. For creative writing, we define a composite writing score, which is a product of a standard score ranging from 0 to 1 (e.g., completion within the target sentence count and inclusion of required keywords) and an LLM-generated quality score ranging from 0 to 10 (e.g., fluency, coherence, creativity). To quantify which dimension exerts greater influence on performance gain, we adopt a Shapley-R^2 decomposition [27]. Concretely, let $R^2(S)$ denote the coefficient of determination obtained from an ordinary least squares regression where the set of predictors is S. In our setting, the predictor set is $\{\text{depth}, \text{width}\}$. We therefore consider four regression fits: (i) the empty model $R^2(\emptyset)$, (ii) the depth-only model $R^2(\{\text{depth}\})$, (iii) the width-only model $R^2(\{\text{width}\})$, and (iv) the full model $R^2(\{\text{depth}, \text{width}\})$. The **Shapley score (S-Score)** for a predictor $x \in \{\text{depth}, \text{width}\}$ is defined as

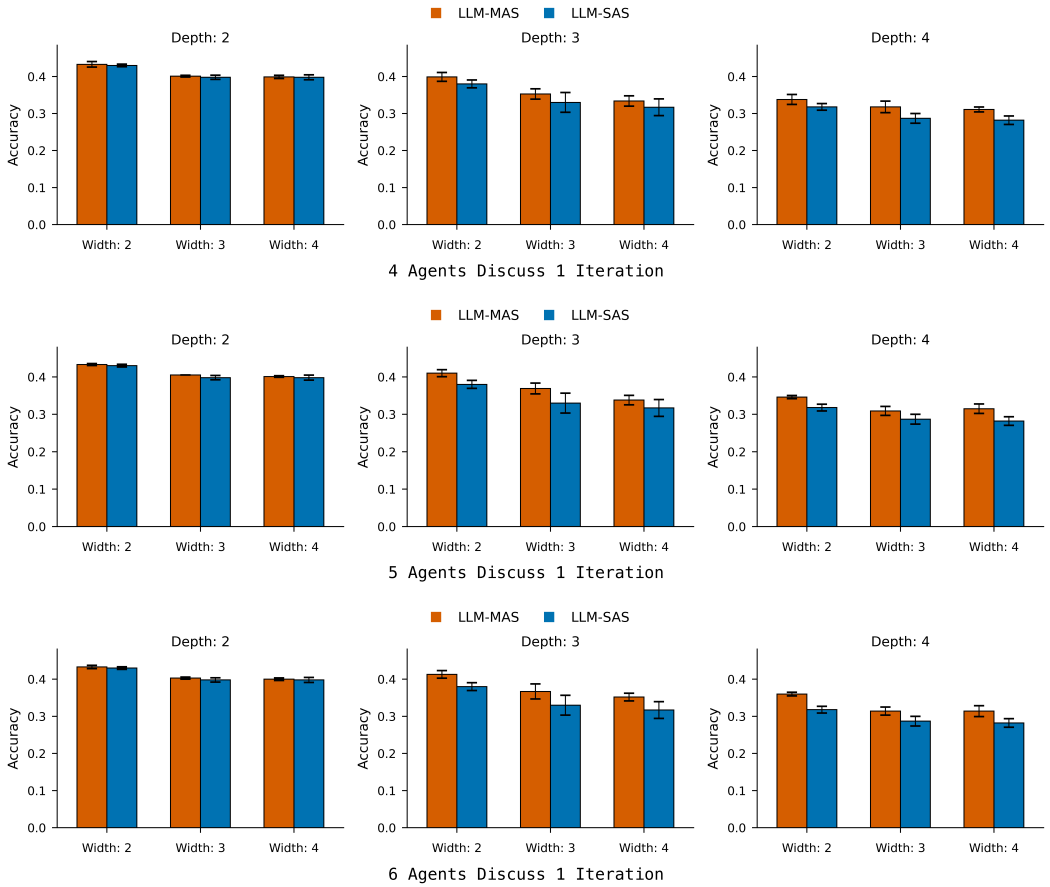
$$S(x) \;=\; \tfrac{1}{2}\Big(R^2(\{x\}) - R^2(\emptyset)\Big) \;+\; \tfrac{1}{2}\Big(R^2(\{\text{depth}, \text{width}\}) - R^2(\{\bar{x}\})\Big),$$

where $\bar{x}$ is the complementary predictor, e.g., if $x = \text{depth}$, then $\bar{x} = \text{width}$. The larger the S-Score for a dimension, the more dominant that dimension is considered in driving the performance gain.

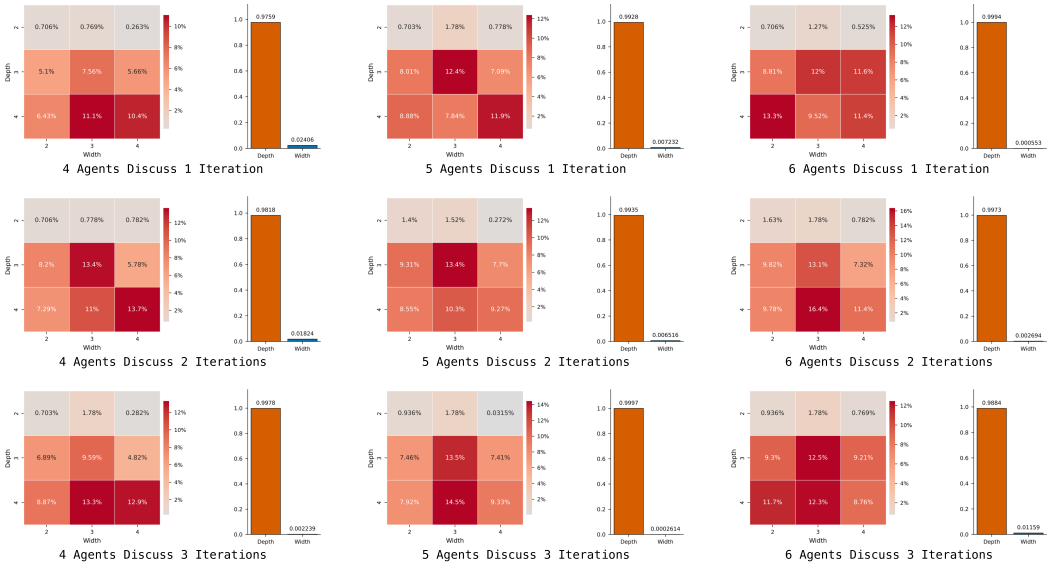
4.2 Results and Analysis

Math reasoning. The results are summarised in Figure 3. More results about the writing score are in the Appendix. We conduct experiments on an LLM-MAS with a varying number of agents from 4 to 6 (including a summarising agent). Our findings reveal three key patterns that align with our theoretical analysis. First, as illustrated in Figure 3a, we observe that increasing task depth and width heightens task complexity, leading to a decrease in absolute performance for both single-agent and multi-agent systems. This validates our complexity measures as effective indicators of problem difficulty. Second, the heatmaps in Figure 3b show that the performance advantage of an LLM-MAS over a single-agent baseline (LLM-SAS) grows with task complexity; the collaborative benefit becomes more significant as problems become deeper and wider. Third, and most crucially, the bar charts within Figure 3b provide quantitative evidence that this advantage is more sensitive to increases in task depth than to comparable increases in task width. This result empirically confirms our theoretical analysis, highlighting that the depth of the task is a dominant factor in the efficacy of LLM-MAS.

Creative writing. The results are summarised in Figure 4. More results about the writing score are in the Appendix. We conduct experiments on the DW^2 benchmark, employing the same LLM-MAS setup with 4 to 6 agents. The outcomes are highly consistent with those from the math reasoning experiments: the benefit of LLM-MAS over LLM-SAS grows with task complexity, and this gain is more pronounced with increases in depth than width. Notably, the magnitude of the performance



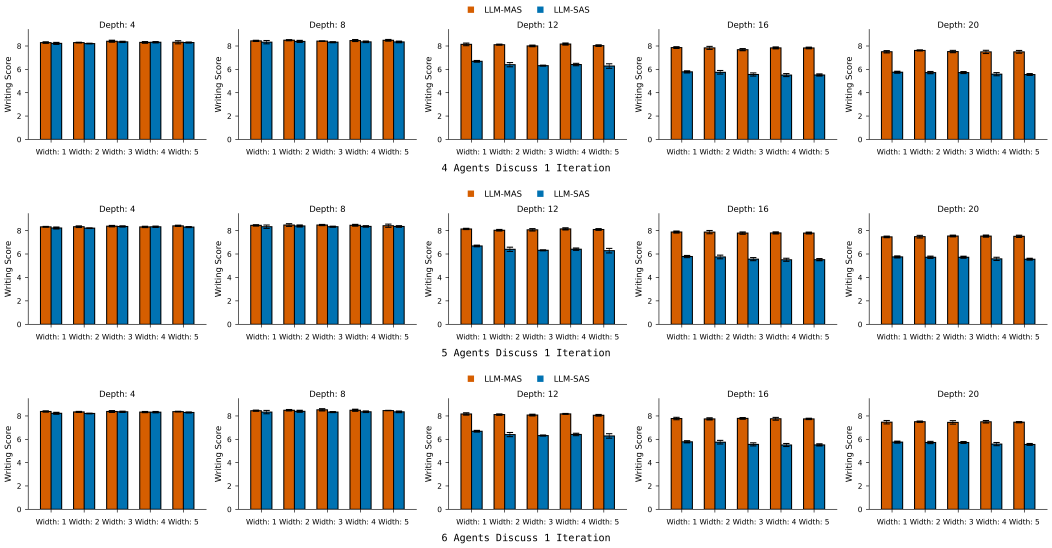
(a) The accuracy of LLM-SAS and LLM-MAS.



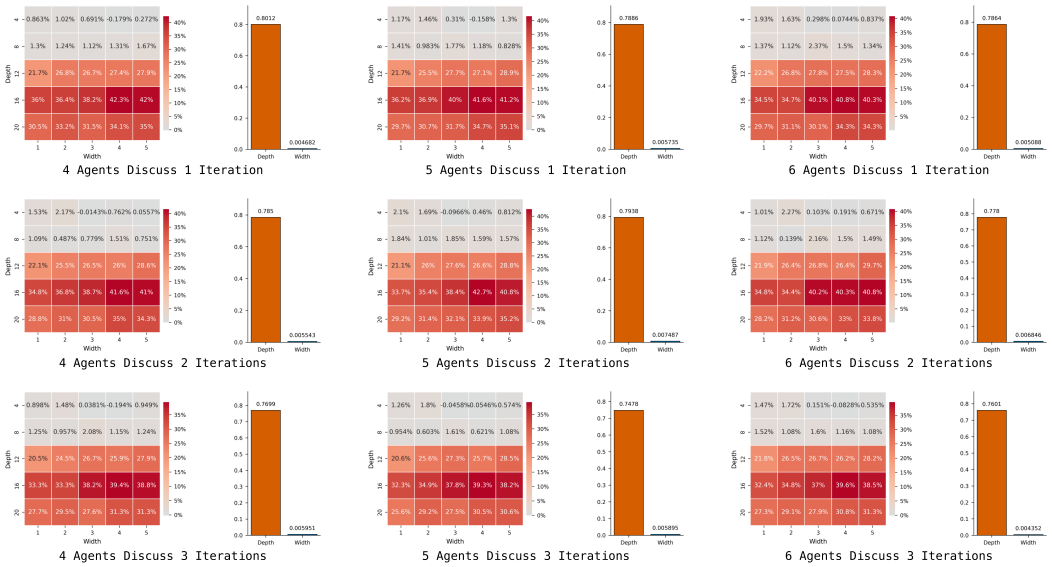
(b) The performance gain and the S-Score.

Figure 3: Results on the math reasoning benchmark.

improvement is substantially larger in this domain compared to the math reasoning task. We hypothesise that this is because creative writing is a generative task with a vast, open-ended solution



(a) The writing score of LLM-SAS and LLM-MAS.



(b) The performance gain and the S-Score.

Figure 4: Results on the creative writing benchmark.

space while satisfying a complex web of interacting constraints. As task complexity increases, we observe that the quality scores of LLM-MAS and LLM-SAS remain closely aligned, typically within a margin of about 5%, but LLM-SAS more frequently fail due to difficulties in meeting constraints, such as ensuring complete keyword coverage. LLM-MAS, however, excels by distributing this cognitive load across multiple agents and the debate process between them. Individual agents can propose solutions that champion different constraints, and the collaborative process allows for a more effective synthesis, leading to a final output that is both creative and holistically compliant.

5 Challenges and Opportunities

Evaluation of LLM-MAS necessity. Current evaluation of LLM-MAS is dominated by empirical performance in downstream tasks, such as answer accuracy in mathematical reasoning, success rates in safety-critical reasoning [22], or heuristic-based assessments in generative modeling [9]. While

our work attempts to characterise task complexity as a complementary lens, this formulation, like pure performance metrics, remains agnostic to the intrinsic properties of MAS themselves, including model scale, agent capacity, and coordination overhead. Since LLM-MAS incur higher computational costs than LLM-SAS, a key challenge is to develop evaluation frameworks that jointly consider task complexity and system complexity, thereby providing a more principled measure of when the deployment of LLM-MAS is truly warranted.

Dynamic task-oriented agent systems. The configuration of MAS, including communication protocols, agent profiles, and agent capacities, is typically predetermined by system designers. This static and manually specified design often leads to limited adaptability, suboptimal performance, and significant inefficiencies in inference costs [14]. Our analysis of task complexity highlights why this is problematic: the effectiveness of MAS depends critically on the alignment between system configuration and the intrinsic demands of the task. A better understanding of how task complexity shapes the utility of higher-order interactions suggests that MAS should not be fixed, but rather adaptive. Building on this insight, we advocate for dynamic optimisation frameworks that allocate communication structures and agent capacities in response to the depth and width of the task at hand [28]. By tailoring MAS configurations to task complexity, such systems can improve both efficiency and task effectiveness, offering a principled pathway towards more adaptive and scalable LLM-based multi-agent architectures.

Datasets and benchmarks. Analogous to human collaboration, MAS collaboration is most beneficial when tasks are inherently difficult, when individual expertise is insufficient, or when diverse perspectives are essential [22]. However, most existing MAS benchmarks are adapted from single-agent settings and therefore fail to adequately capture these conditions. As a result, they often do not necessitate or reward genuine multi-agent collaboration, with only limited exceptions [28] that are explicitly designed for MAS. Looking forward, we argue that future benchmarks should be tailored to debate-style interactions, explicitly encoding the need for reasoning chains that benefit from critique, counter-arguments, and consensus-building. Such datasets would better reflect the unique challenges of LLM debate frameworks and enable systematic evaluation of when and how collaboration among agents translates into tangible performance gains.

Beyond depth and width. Finally, while our analysis formalises task complexity in terms of depth (sequential reasoning steps) and width (capabilities required within a step), these are not the only possible dimensions. Alternative characterisations may consider, for example, the interaction complexity, which means the extent to which sub-tasks interdepend or conflict. Exploring these alternatives remains an open direction that could yield a richer taxonomy of complexity, bridging theoretical formulations with practical challenges faced by LLM-MAS.

6 Conclusion

In this work, we present a theoretical and empirical investigation into the effectiveness of LLM multi-agent systems (LLM-MAS) over single-agent systems (LLM-SAS). To formulate our analysis, we introduce a novel measure of task complexity, defined in terms of reasoning sequentiality (depth) and knowledge diversity (width). We begin with a theoretical analysis, showing that while both dimensions yield a monotonically increasing performance gap between LLM-MAS and LLM-SAS, the gains from width are fundamentally bounded, whereas depth can provide unbounded improvements. We then validate these insights through empirical studies, demonstrating that: (1) both depth and width consistently enhance the relative performance of LLM-MAS over LLM-SAS, and (2) the benefits of increasing depth scale substantially, while the advantages from width quickly plateau once a finite bound is reached. Together, these results establish a principled foundation for the design of more efficient, effective, and theoretically grounded LLM-MAS. Our analysis focused on debate-based frameworks, which underpin many existing LLM-MAS. For instance, sub-task decomposition methods can be viewed as extensions of debate, tailored for project-level tasks where a large problem is divided into smaller questions, each solved by single agents or debate modules before being integrated into a coherent solution. Exploring such variants across different scenarios is non-trivial, as it requires not only generalising our theoretical results, but also designing new benchmarks with controllable depth and width for their specific settings. Similarly, extending our study to base models larger than Qwen-2.5-32B-Instruct would also be valuable but demands substantial computational resources. We therefore leave these as future work.

References

[1] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *NeurIPS*, 36:51991–52008, 2023.

[2] Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, et al. Androidworld: A dynamic benchmarking environment for autonomous agents. *arXiv preprint arXiv:2405.14573*, 2024.

[3] Shuo Tang, Xianghe Pang, Zexi Liu, Bohan Tang, Rui Ye, Tian Jin, Xiaowen Dong, Yanfeng Wang, and Siheng Chen. Synthesizing post-training data for llms through multi-agent simulation. *The 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*, 2025.

[4] Dezhao Luo, Bohan Tang, Kang Li, Georgios Papoudakis, Jifei Song, Shaogang Gong, Jianye Hao, Jun Wang, and Kun Shao. Vimo: A generative visual gui world model for app agents. *arXiv preprint arXiv:2504.13936*, 2025.

[5] Bohan Tang, Dezhao Luo, Jingxuan Chen, Shaogang Gong, Jianye Hao, Jun Wang, and Kun Shao. Beyond syntax: Action semantics learning for app agents. *arXiv preprint arXiv:2506.17697*, 2025.

[6] Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*, 2025.

[7] Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, 1(1):9, 2024.

[8] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2023.

[9] Tian Yu, Ken Shi, Zixin Zhao, and Gerald Penn. Multi-agent based character simulation for story writing. In *Proceedings of the Fourth Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2025)*, pages 87–108, 2025.

[10] Fantine Huot, Reinald Kim Amplayo, Jennimaria Palomaki, Alice Shoshana Jakobovits, Elizabeth Clark, and Mirella Lapata. Agents' room: Narrative generation through multi-step collaboration. In *The Thirteenth International Conference on Learning Representations*, 2025.

[11] Han Zhou, Xingchen Wan, Ruoxi Sun, Hamid Palangi, Shariq Iqbal, Ivan Vulić, Anna Korhonen, and Sercan Ö Arık. Multi-agent design: Optimizing agents with better prompts and topologies. *arXiv preprint arXiv:2502.02533*, 2025.

[12] Guibin Zhang, Yanwei Yue, Xiangguo Sun, Guancheng Wan, Miao Yu, Junfeng Fang, Kun Wang, Tianlong Chen, and Dawei Cheng. G-designer: Architecting multi-agent communication topologies via graph neural networks. In *Forty-second International Conference on Machine Learning*, 2025.

[13] Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. Gptswarm: Language agents as optimizable graphs. In *Forty-first International Conference on Machine Learning*, 2024.

[14] Rui Ye, Shuo Tang, Rui Ge, Yaxin Du, Zhenfei Yin, Siheng Chen, and Jing Shao. Mas-gpt: Training llms to build llm-based multi-agent systems. In *Forty-second International Conference on Machine Learning*, 2025.

[15] Yuanshuo Zhang, Yuchen Hou, Bohan Tang, Shuo Chen, Muhan Zhang, Xiaowen Dong, and Siheng Chen. Gnns as predictors of agentic workflow performances. *arXiv preprint arXiv:2503.11301*, 2025.

[16] Soumyakant Padhee and Babak Heydari. The echo chambers of complexity: how task complexity influences team groupthink and individual exploration. *Available at SSRN 4892421*, 2024.

[17] Vincent J Straub, Milena Tsvetkova, and Taha Yasseri. The cost of coordination can exceed the benefit of collaboration in performing complex tasks. *Collective Intelligence*, 2(2):26339137231156912, 2023.

[18] Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. Dyval: Dynamic evaluation of large language models for reasoning tasks. In *The Twelfth International Conference on Learning Representations*, 2024.

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

[20] Rui Ye, Keduan Huang, Qimin Wu, Yuzhu Cai, Tian Jin, Xianghe Pang, Xiangrui Liu, Jiaqi Su, Chen Qian, Bohan Tang, et al. Maslab: A unified and comprehensive codebase for llm-based multi-agent systems. *arXiv preprint arXiv:2505.16988*, 2025.

[21] Mingyan Gao, Yanzi Li, Banruo Liu, Yifan Yu, Phillip Wang, Ching-Yu Lin, and Fan Lai. Single-agent or multi-agent systems? why not both? *arXiv preprint arXiv:2505.18286*, 2025.

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

[23] Mert Cemri, Melissa Z Pan, Shuyi Yang, Lakshya A Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, et al. Why do multi-agent llm systems fail? *arXiv preprint arXiv:2503.13657*, 2025.

[24] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

[25] National Center for O*NET Development. O*net-soc 2019 taxonomy. `https://www.onetcenter.org/taxonomy.html#listings`. Accessed: 2025-08-21.

[26] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[27] Stan Lipovetsky and Michael Conklin. Analysis of regression in game theory approach. *Applied stochastic models in business and industry*, 17(4):319–330, 2001.

[28] Rui Ye, Xiangrui Liu, Qimin Wu, Xianghe Pang, Zhenfei Yin, Lei Bai, and Siheng Chen. X-mas: Towards building multi-agent systems with heterogeneous llms, 2025.

[29] Bertrand Russell. *Principles of mathematics*. Routledge, 2020.

A Proof for Proposition 2.1

Proof. By definition, we have:

$$\Delta(d, w, N, r) \triangleq \frac{S_{\text{multi}} - S_{\text{single}}}{S_{\text{single}}} = \left(\frac{r\,[1 - (1 - s(w))^N]}{s(w)} \right)^d - 1.$$

Let $s = s(w) = q^w \in (0, 1)$ and define:

$$A(s) \triangleq r\big[1 - (1 - s)^N\big], \quad f(s) \triangleq \frac{A(s)}{s}.$$

Hence $\Delta(d, w, N, r) = f(s)^d - 1$. Moreover, by assumption, we have $f(s) = \frac{A(s)}{s} > 1$. Therefore, for the partial derivative in d, we have

$$\frac{\partial \Delta}{\partial d} = \ln(f(s))\, f(s)^d > 0.$$

For the partial derivative in w, we have:

$$\frac{\partial \Delta}{\partial w} = d\, f(s)^{d-1}\, f'(s)\, s'(w).$$

Let $t = 1 - s \in (0, 1)$. Then, we have:

$$A'(s)s - A(s) = r\left[Ns(1 - s)^{N-1} - (1 - (1 - s)^N)\right] = -r\left(1 - t^N - N(1 - t)t^{N-1}\right).$$

By the Mean Value Theorem [29] applied to $\varphi(x) = x^N$ on $[t, 1]$, there exists $\xi \in (t, 1)$ such that

$$1 - t^N = N\xi^{N-1}(1 - t) \geq Nt^{N-1}(1 - t).$$

Thus the bracketed expression is nonnegative, making $A'(s)s - A(s) \leq 0$, so $f'(s) = \frac{A'(s)s - A(s)}{s^2} \leq 0$. Since $s'(w) = q^w \ln q < 0$, it follows that

$$\frac{\partial \Delta}{\partial w} = d\, f(s)^{d-1}\, f'(s)\, s'(w) > 0.$$

□

B Proof for Proposition 2.2

Proof. By definition, we have:

$$\Delta(d, w, N, r) \triangleq \frac{S_{\text{multi}} - S_{\text{single}}}{S_{\text{single}}} = \left(\frac{r\,[1 - (1 - s(w))^N]}{s(w)} \right)^d - 1.$$

Let $s = s(w) = q^w \in (0, 1)$ and define:

$$A(s) \triangleq r\big[1 - (1 - s)^N\big], \quad f(s) \triangleq \frac{A(s)}{s}.$$

As $w \to \infty$, $s(w) = q^w \to 0$. By the Binomial Theorem [29], for integer N, we have $(1 - s)^N = 1 - Ns + \binom{N}{2}s^2 - \cdots + (-s)^N$. Therefore, we can have the following representation: $1 - (1 - s)^N = Ns + O(s^2)$. Therefore, we can have:

$$\lim_{w \to \infty} \Delta(d, w, N, r) = (rN)^d - 1.$$

Since $f(s) > 1$, $f(s)^d \to +\infty$ as $d \to +\infty$, therefore we can have:

$$\lim_{d \to \infty} \Delta(d, w, N, r) = +\infty.$$

□

C Additional Experimental Results

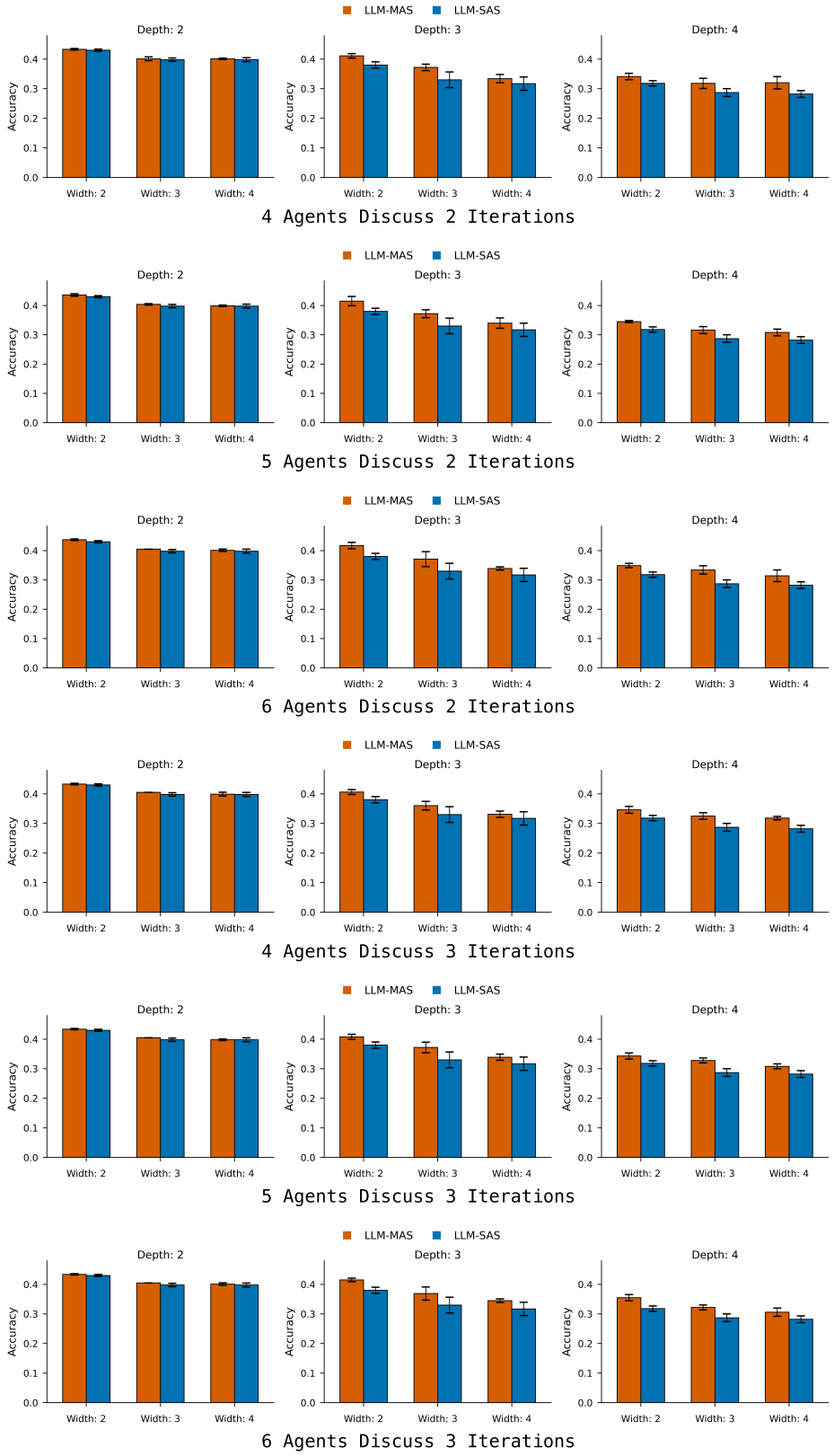


Figure 5: The accuracy of LLM-SAS and LLM-MAS on math reasoning.

Figure 6: The writing score of LLM-SAS and LLM-MAS on creative writing.