

Exploring Collaboration Mechanisms for LLM Agents: A Social Psychology View

Anonymous ACL submission

Abstract

As Natural Language Processing (NLP) systems are increasingly employed in intricate social environments, a pressing query emerges: *Can these NLP systems mirror human-esque collaborative intelligence, in a multi-agent society consisting of multiple large language models (LLMs)?* This paper probes the collaboration mechanisms among contemporary NLP systems by melding practical experiments with theoretical insights. We fabricate four unique ‘societies’ comprised of LLM agents, where each agent is characterized by a specific ‘trait’ (easy-going or overconfident) and engages in collaboration with a distinct ‘thinking pattern’ (debate or reflection). Through evaluating these multi-agent societies on three benchmark datasets, we discern that certain collaborative strategies not only outshine previous top-tier approaches, but also optimize efficiency (using fewer API tokens). Moreover, our results further illustrate that LLM agents manifest human-like social behaviors, such as conformity and consensus reaching, mirroring foundational social psychology theories. In conclusion, we integrate insights from social psychology to contextualize the collaboration of LLM agents, inspiring further investigations into the collaboration mechanism for LLMs. We commit to sharing our code and datasets¹, hoping to catalyze further research in this promising avenue.

1 Introduction

With the prevalence of LLMs (Zhao et al., 2023; Yin et al., 2023; Zhu et al., 2023) integral to daily social collaboration, there is a growing imperative to cultivate AI systems embodied with social intelligence. This also resonates with the Society of Mind (SoM) concept (Li et al., 2023a; Zhuge et al., 2023; Wang et al., 2023b), which suggests that intelligence emerges when computational modules

interact with each other, achieving collective objectives that surpass the capabilities of individual modules (Minsky, 1988; Singh, 2003). Previous studies (Park et al., 2023; Du et al., 2023b; Liang et al., 2023; Shinn et al., 2023; Madaan et al., 2023; Hao et al., 2023; Liu et al., 2023a; Akata et al., 2023) have delved into strategies where LLM instances, termed agents (Wang et al., 2023a; Xi et al., 2023; Gao et al., 2023; Cheng et al., 2024; Ma et al., 2024), cooperate synergistically (e.g., debate and reflect) to accomplish tasks (Du et al., 2023a; Pezeshkpour et al., 2024; Guo et al., 2024; Du et al., 2024; Han et al., 2024). As illustrated in Figure 1, such collaboration fosters divergent thinking processes in LLMs, making them particularly effective for tasks demanding profound reflection.

Intuitively, reflecting on human societies (Siegal and Varley, 2002; Leslie et al., 2004; Sap et al., 2022; Shapira et al., 2023), where a myriad of individuals with distinct goals and roles coexist, the SoM framework champions harmonious interactions (Singh, 2003). Intriguingly, despite the fusion of social psychology (Tajfel, 1982; Tajfel and Turner, 2004; Johnson and Johnson, 2009) in SoM with human group dynamics (Woolley et al., 2010; Alderfer, 1987), which illuminates psychological patterns within social groups, its interpretation in the realm of LLMs is relatively uncharted (Ke et al., 2024). Besides, our grasp of how social behaviors influence LLMs is still in its nascent stages.

To address these issues, we delve into the machine society, probing the extent and ways that LLMs manifest social intelligence and collaboration capabilities (Mei et al., 2023). Utilizing powerful LLMs like GPT-3.5 (OpenAI, 2022), we build a test-bed across three datasets: MATH (Hendrycks et al., 2021b), MMLU (Hendrycks et al., 2021a) and Chess Move Validity (Srivastava et al., 2022). Our approach incorporates four **societies** characterized by two **individual traits** (*easy-going* and *overconfident*) with three agents: totally/mostly easy-

¹<https://anonymous.4open.science/r/MachineSoM-3178>.

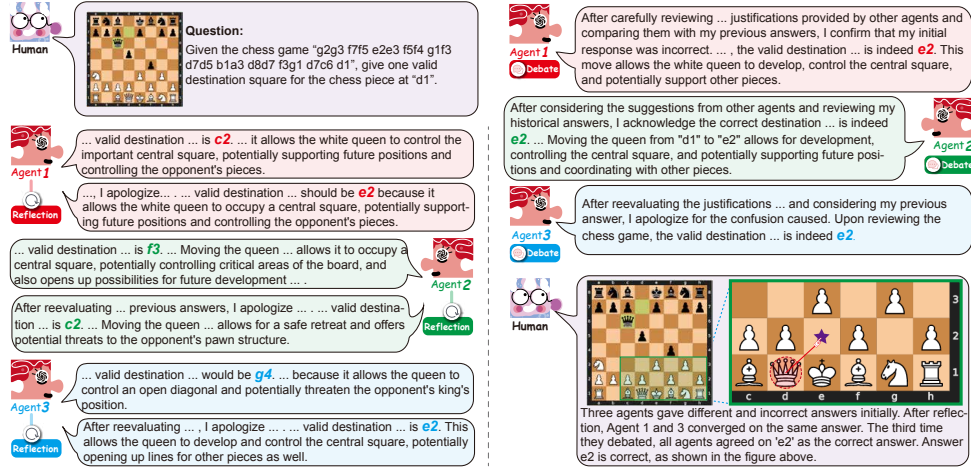


Figure 1: An example of the chess move validity task. Given previous chess game moves, agents are required to predict a valid next move for a specified piece.

going; totally/mostly overconfident. These traits are employed to emulate nuanced human society dynamics (Soni et al., 2024; Wang et al., 2024b,a; Li et al., 2023b; Kong et al., 2023).

Moreover, we delve into two distinct **thinking patterns** under multi-round collaboration: *debate* (Perelman, 1971; Sunstein, 2005; Amgoud and Prade, 2009; Du et al., 2023b; Liang et al., 2023) and *reflection* (Bogumil, 1985; Mezirow, 2003; Bolton, 2010). With the permutation of thinking patterns, we can constitute various **collaborative strategies**. To this end, we implement two patterns of collaboration in the collaborative strategies: (i) All agents adopt the same thinking pattern at each round; (ii) One agents adopts the different thinking patterns from others at each round. We then execute these multi-round collaborative strategies within different societies. Through our empirical analysis, we primarily discern the following insights (Further takeaways are in §3, §4 & Appendix A):

(1) Collaborative strategies with various permutations of thinking patterns vary significantly to performance, and engaging in substantive debates enhances collaboration performance. Intriguingly, multi-agent societies composed of agents with different traits do not clearly differ in performance.

(2) Employing uniform thinking patterns across all agents within a round of collaboration enhances efficiency. Besides, merely increasing the number of agents or the number of collaboration rounds does not consistently yield better outcomes. The balance between agent quantity and strategies emerges as a key determinant in collaboration.

(3) LLM agents manifest behaviors reminiscent of human social tendencies, such as conformity (Allen and Levine, 1969; Cialdini and Goldstein,

2004) or the principle of majority rule in group thinking (Seal et al., 1998), which resonate with several fundamental theories in social psychology (Castro and Liskov, 1999; Tajfel and Turner, 2004).

Concretely, our findings challenge the dominant belief that mere scale is the key. We posit that small-group collaboration with rational strategies might present a more efficacious approach to utilizing LLMs. In wrapping up, we encapsulate the core contributions of this research as follows:

- We initiate an elaborate exploration into collaboration mechanisms in multi-agent society. Our goal is to identify how and to what extent LLMs manifest social intelligence through collaboration. To enrich our inquiry, we draw upon theories from social psychology, contextualizing the behaviors and tendencies displayed by LLM agents.
- Our research framework includes a meticulously crafted test-bed, integrating diverse multi-agent societies with agent individual traits, thinking patterns and collaborative strategies, evaluated over three datasets. Notably, our empirical findings can inspire how to design a better multi-agent system through collaboration, beyond merely scaling up LLMs and Agents.
- Interestingly, our observations underscore a fascinating parallel: LLM agents mirror certain social behaviors typical of human collaboration. It could further emphasize the potential of human-AI interaction. Generally, fostering effective and efficient collaborative strategies for multi-agent systems could be the key to more socially-aware AI.

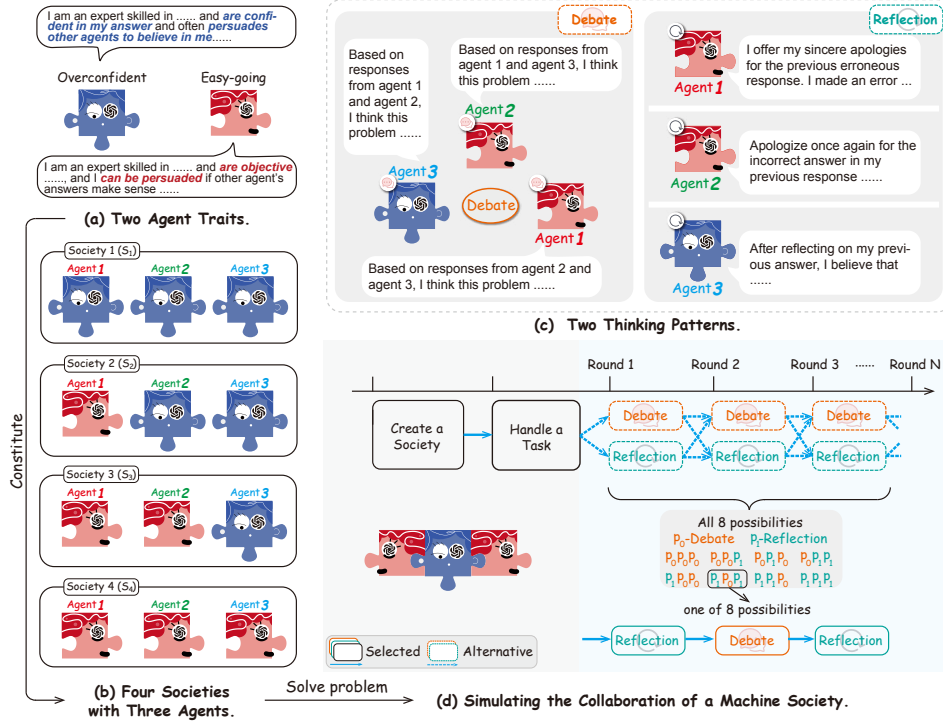


Figure 2: The overview of machine society simulation. Multiple agents with different traits make up diverse machine societies. These agents engage in debate or self-reflection across multiple rounds to complete tasks.

2 Explore Collaboration Mechanisms with Multiple LLM Agents

In this section, we formulate and simulate the collaboration mechanisms explored within the machine society, drawing upon relevant concepts. We also illustrate the society settings in Figure 2.

2.1 Preliminary Concepts in Collaboration

Individual Trait. Inspired by intelligence emerging from the collective efforts of numerous smaller and relatively simple agents (Minsky, 1988), each characterized by diverse traits, we set two types of agents exhibiting typically contrasting traits: *easy-going* and *overconfident*, as shown in Figure 2(a). Easy-going agents keep things in perspective, adapt well to different situations, and are compatible with various types of agents (Friedman and Schustack, 1999), which results in a harmonious societal structure with democracy (Mutz, 2006; Held, 2006). Conversely, overconfident agents tend to overestimate their competence, ignore potential risks and resist others' opinions (Moore and Healy, 2008).

Thinking Pattern. Considering the SoM concept (Minsky, 1988) states that intelligence emerges when specialized individuals within a society cooperate with each other through thinking, we aim to study what thinking patterns are

most successful in producing such emerging intelligence. Thus we explore two thinking patterns: *debate* (Sunstein, 2005; Du et al., 2023b; Liang et al., 2023) and *reflection* (Bogumil, 1985; Bolton, 2010; Shinn et al., 2023), as illustrated in Figure 2(c). (i) In the *debate* pattern, several agents propose ideas, exchange responses, engage in collective argumentation, and ultimately reach a consensus. This fosters knowledge sharing, facilitates learning, and promotes adaptation among all agents within the society (Weiß, 1995; Stone and Veloso, 2000; Vidal, 2006; Wooldridge, 2009). This pattern can be iteratively executed multiple rounds. (ii) In the *reflection* pattern, agents review their prior responses, extract lessons from their experiences, and refine their answers accordingly. This pattern can also unfold over several rounds.

Collaborative Strategy. Through both critical reflection and active participation in debate, agents are poised to challenge their existing assumptions, acquire fresh perspectives, and ultimately refine their viewpoints. Employing a collaboration mechanism built on these two thinking patterns can foster more insightful decision-making (Wooldridge, 2009; Amgoud and Prade, 2009) and improve reasoning outcomes (Mezirow, 2018). In societal settings, agents typically engage in multiple rounds of collaboration for problem solving. In this paper,

we characterize the collaborative strategy as a **permutation of thinking patterns** throughout multi-round collaboration, as illustrated in Figure 2(d) and further elaborated in §2.2.

2.2 Society Simulation

Symbols	Definition
$\mathcal{T}$	Set of agent traits
t_o	Trait : overconfident
t_e	Trait : easy-going
$\mathcal{A}$	Set of agent instances
a_i	The i -th agent
$\mathcal{P}$	Set of thinking patterns
p_0	Debate
p_1	Reflection
$\mathcal{S}$	Set of societies
S_i	The i -th society

Table 1: The description of the symbols.

We simulate the multi-agent collaborative society, as detailed with symbols shown in Table 1. Specifically, we construct a machine society consisting of n LLM agents, denoted as $\mathcal{A} = \{a_i\}_{i=1}^n$. This society contains two distinct agent traits: $\mathcal{T} = \{t_o, t_e\}$, where t_o and t_e respectively denotes the overconfident and easy-going trait. For each agent, at any round of collaboration, there are two thinking patterns to choose from, symbolized as $\mathcal{P} = \{p_0, p_1\}$, where p_0 and p_1 corresponds to *debate* and *reflection* respectively. By endowing agents $\mathcal{A}$ with the traits of $\mathcal{T}$, we can emulate various machine societies. In our primary study (§3), we establish four distinct societies, $\mathcal{S} = \{S_1, S_2, S_3, S_4\}$, each consisting of three agents: $\{a_1, a_2, a_3\}$. The societies are constructed based on combination of three agents with distinct traits, as illustrated in Figure 2(b):

$$\begin{aligned}
S_1 &= \{(a_1 \leftarrow t_o), (a_2 \leftarrow t_o), (a_3 \leftarrow t_o)\} \text{ (totally overconfident)} \\
S_2 &= \{(a_1 \leftarrow t_o), (a_2 \leftarrow t_o), (a_3 \leftarrow t_e)\} \text{ (mostly overconfident)} \\
S_3 &= \{(a_1 \leftarrow t_o), (a_2 \leftarrow t_e), (a_3 \leftarrow t_e)\} \text{ (mostly easy-going)} \\
S_4 &= \{(a_1 \leftarrow t_e), (a_2 \leftarrow t_e), (a_3 \leftarrow t_e)\} \text{ (totally easy-going)}
\end{aligned}$$

where $(a_i \leftarrow t_j)$ denotes that the agent a_i possesses the trait t_j . If there are an even number of agents, we can also constitute the society with half overconfident and half easy-going agents. In our simulation, all agents consistently employ the same thinking pattern at each round of collaboration, similar to Du et al. (2023b). It gives rise to eight possible 3-round collaborative strategies:

$$\begin{aligned}
&p_0p_0p_0, p_0p_0p_1, p_0p_1p_0, p_0p_1p_1, \\
&p_1p_0p_0, p_1p_0p_1, p_1p_1p_0, p_1p_1p_1
\end{aligned}$$

In our subsequent analysis (§3.2), we delve into more intricate scenarios, introducing a larger number of agents, increased collaboration rounds, and a broader range of collaborative strategies.

2.3 Experimental Settings

Datasets. We conduct a rigorous evaluation of the reasoning and decision-making capabilities of various machine societies across three distinct tasks, utilizing diverse collaborative strategies:

- *High School Multiple-Choice.* Leveraging the **MMLU** (Hendrycks et al., 2021a) dataset, where problems span high school subjects such as statistics, mathematics, computer science, biology, chemistry, and physics, agents are required to identify the correct answer among four multiple-choice options. Our evaluation set consists of 50 randomly-selected questions from this dataset.
- *Math.* Drawing from **MATH** dataset (Hendrycks et al., 2021b), a repository of math problems sourced from competitive events and expressed in LaTeX, we assess the model proficiency in advanced mathematical and scientific reasoning. The dataset segments these problems into five graded difficulty levels, and for our evaluation, we have randomly chosen 50 cases from Level 3 to 5.
- *Chess Move Validity.* Utilizing the dataset from the chess state tracking task² within the comprehensive **BIG-Bench Benchmark** (Srivastava et al., 2022), a sequence of chess moves denoted in UCI notation³ is provided. Agents are required to predict a legitimate subsequent move for a specified chess piece.

Setups. We craft specific instructions for each task, trait and strategy, which can be referred in Table 5 at Appendix D.3. To enhance result reliability, we present average accuracy (**Acc**) and their respective standard deviations across five trials. Notably, our experiments exhibit substantial standard deviations. Hence, we introduce WIN-TIE (**W-T**) metric, indicating the frequency (over five trials) where the accuracy either matches or surpasses the continuous debate baseline (Du et al., 2023b). Meanwhile, we gauge the average token costs (**Cost**)

²https://github.com/google/BIG-bench/blob/main/bigbench/benchmark_tasks/chess_state_tracking/synthetic_short/task.json.
³https://en.wikipedia.org/wiki/Universal_Chess_Interface.

	Metric (Strategy)	Society	Collaborative Strategy								Metric (Society)	
			$p_0p_0p_0$	$p_0p_0p_1$	$p_0p_1p_0$	$p_0p_1p_1$	$p_1p_0p_0$	$p_1p_0p_1$	$p_1p_1p_0$	$p_1p_1p_1$	<u>Cost</u> ↓	<u>W-T</u> ↑
MMLU	Acc ↑	S_1	66.4±1.7	65.2±3.6	52.8±4.8	59.2±3.6	45.6±1.7	51.6±2.2	62.0±0.0	46.0±0.0	2970	2
		S_2	66.0±0.0	65.2±1.8	58.0±0.0	66.0±0.0	44.0±0.0	46.0±0.0	53.2±2.7	46.0±0.0	3081	9
		S_3	70.4±4.3	64.4±0.9	57.6±1.7	52.8±2.3	41.2±5.4	49.2±4.6	51.2±1.8	62.0±0.0	3172	1
		S_4	69.6±3.9	65.2±3.6	54.8±5.2	58.4±1.7	34.4±2.2	46.0±4.9	56.4±2.2	62.0±0.0	3090	2
MMLU	<u>Cost</u> ↓	All	4364	3510	3295	2665	3476	2651	2691	1976		
	<u>W-T</u> ↑	All	-	9	0	5	0	0	0	0		
MATH	Acc ↑	S_1	46.8±4.2	46.4±3.3	42.8±4.6	33.6±7.4	38.8±2.7	38.4±3.9	45.2±2.7	35.2±1.1	3417	8
		S_2	50.4±2.6	52.8±2.3	49.6±3.0	38.8±3.9	38.8±3.6	45.6±2.2	46.4±4.1	35.2±1.1	3623	8
		S_3	47.6±4.8	48.0±3.2	47.2±4.8	38.0±7.1	37.6±3.3	39.2±5.4	42.4±3.0	40.0±2.5	3757	8
		S_4	50.4±1.7	49.6±1.7	53.2±1.1	40.0±2.0	44.0±3.2	45.6±4.3	45.6±3.6	41.6±1.7	3658	10
MATH	<u>Cost</u> ↓	All	4439	3965	3857	3414	3840	3234	3482	2681		
	<u>W-T</u> ↑	All	-	14	13	0	0	1	6	0		
Chess Move Validity	Acc ↑	S_1	54.4±1.7	52.0±0.0	52.0±5.1	51.6±5.2	54.4±1.7	51.2±1.8	50.4±1.7	52.0±0.0	2443	11
		S_2	48.0±0.0	49.2±1.1	46.0±0.0	54.0±0.0	50.0±0.0	52.0±0.0	42.0±2.5	52.0±0.0	2442	25
		S_3	48.4±1.7	48.0±2.8	54.8±5.0	45.2±3.4	48.4±2.6	44.8±3.4	50.4±1.7	53.6±0.9	2451	23
		S_4	51.6±4.6	44.0±2.5	54.4±3.0	53.6±5.5	45.6±2.2	48.0±2.0	43.6±0.9	52.0±0.0	2404	12
Chess Move Validity	<u>Cost</u> ↓	All	3046	2611	2604	2179	2705	2251	2252	1830		
	<u>W-T</u> ↑	All	-	10	12	10	11	9	5	14		

Table 2: The impact of 8 collaborative strategies on the performance of 3 datasets across distinct societies, using ChatGPT. Blue marks the best-performing strategy under the same society, light blue represents the second-best-performing strategy, and red indicates the worst-performing strategy. Cost / Cost measures the average tokens consumed by all cases under the same collaborative strategy / society. W-T / W-T tallies the total number of occurrences where performance exceeds the strategy $p_0p_0p_0$ under the same collaborative strategy / society. The significances test on societies and strategies are respectively shown in Table 6, 7 at Appendix E.

consumed by the agents across societies, shedding light on the efficacy of the different collaborative strategies employed. For these evaluations, ChatGPT serves as the LLM agent accessible through the OpenAI API `gpt-3.5-turbo-1106`⁴. Further comprehensive details on data sampling and result evaluation are introduced in Appendix D.

3 Analysis of Machine Social Collaboration

Our experiments are primarily driven by the following research queries: (RQ1) How does problem-solving effectiveness vary under different collaborative strategies across diverse societies? (RQ2) How to configure the machine society variables for optimal performance? (RQ3) How does machine social collaboration mimic the human society?

3.1 Main Results with Quantitative Analysis

To address RQ1, we present the performance of four distinct societies in Table 2, each employing one of eight possible collaborative strategies, evaluated across three datasets with ChatGPT. To make the experimental findings more general, we evaluate on other LLMs, shown in Appendix H. Our experiments yield several pivotal observations:

(1) **Societies do not clearly differ in performance, but differ significantly in their tendency**

to reach a consensus. As observed from Table 2, among different 3-agent societies $S_1 \sim S_4$ employing the same collaborative strategy (a *vertical comparison* on Acc), the variations in accuracy are not pronounced. We also conduct a significance test of societies using ChatGPT in Appendix E, and other LLMs in Appendix H, further demonstrating insignificant differences between the societies. Thus we conclude that distinct societies composing of 3 agents possessing varied traits play an indistinctive role in shaping performance. We infer that this is due to LLM alignment (Ouyang et al., 2022), inhibiting agents from displaying extreme overconfidence, which contradicts human alignment (Liu et al., 2022a). Sharma et al. (2023) also demonstrate that LLMs tend to show sycophancy, as illustrated in Figure 11, 12. Furthermore, we increase the number of agents (2 to 10), accordingly resulting in more diverse societies, as seen in Figure 14, indicating that the impact of societies on performance remains indistinctive. We further analyze consensus reaching, *i.e.*, multiple agents reaching a consistent answer (Chen et al., 2023a), shown in Figure 16 at Appendix E, and find that more diverse societies (5 types of societies, containing 2 to 10 agents) significantly impact the average quantity of consensus. In general, the society totally comprising easy-going agents is more likely to reach a consensus.

⁴<https://platform.openai.com/docs/models/gpt-3-5>.

(2) Permutation of thinking patterns is crucial for collaboration, where debate-initial and debate-dominant strategies exhibit superiority. For instance, on MMLU dataset, *debate-dominant* collaborative strategies, like $p_0p_0p_1$, $p_0p_1p_0$, and $p_1p_0p_0$, all containing two rounds of debate, display a pronounced outperformance (65.2 for $p_0p_0p_1$ in S_4 versus 34.4 for $p_1p_0p_0$ in S_4). As seen from Table 2, collaborative strategies starting with the thinking pattern of debate p_0 (*debate-initial*), such as $p_0p_0p_0$, $p_0p_0p_1$, $p_0p_1p_0$, and $p_0p_1p_1$, generally outperform others across all datasets. Furthermore, observed from the performance (i) under strategies with different (3~10) rounds of collaboration on ChatGPT, as shown in Figure 4 and Figure 18, 19 at Appendix F, debate-initial/dominant strategies are overall better; (ii) on LLaMA2 Chat 13B in Table 14 and Qwen 72B in Table 26, debate-initial strategies are generally superior; (iii) on LLaMA2 Chat 70B in Table 20 and Mixtral 8×7B in Table 32, debate-dominant strategies are superior. Observed from different 3-round collaborative strategies $p_i p_j p_k$ applied within the same society (a *horizontal comparison* on Acc), the variations in accuracy are notably pronounced. Besides, the significance test of different collaborative strategies using ChatGPT in Appendix E and other LLMs in Appendix H demonstrate that the order of thinking patterns significantly impacts the effectiveness.

(3) Tasks behave better under collaborative strategies starting with continuous debate, and debate combined with continuous reflection is superior for difficult tasks. Seen from Table 2, when comparing the best performance (marked in blue) and the worst (marked in red) within the same societies, the difference in results for Chess Move Validity is slight. This stands in sharp contrast to MMLU and MATH, which suggests that *the effectiveness of collaborative strategies depends on the task*. We then illustrate the performance under different collaborative strategies in view of task domains and difficulty in Figure 13 at Appendix E; on other LLMs in Figure 24, 33, 42, 56 at Appendix H. Figure 13(a) exhibits task-specific impacts and Figure 13(b),(c) reflects domain-dependent impacts under different collaborative strategies, where $p_0p_0p_0$ and $p_0p_0p_1$ starting with continuous debate are generally superior. For the mathematics domain seen from Figure 13(d), like MMLU mathematics and MATH level 3 & 4, the performance variations under different strategies are relatively small, but

for the more difficult task, *i.e.*, MATH level 5, the strategies containing debate and continuous reflection (*i.e.*, $p_0p_1p_1$, $p_1p_1p_0$) behave superiorly. These nuanced disparities imply that *the marginal benefits derived from collaborative strategies may be task-dependent and difficulty-sensitive*.

3.2 Impact of Machine Society Settings

To address RQ2, we delve deeper into the variables influencing multi-agent society collaboration, exploring intricacies of agent composition, collaboration rounds, and the collaborative strategies.

Different Numbers of Agents. To evaluate the impact of different numbers of agents, we analyze performance within societies comprising 2~10 agents, presented in Figure 3(a). Different numbers of agents would constitute five types of societies, where the agents’ traits of could be: *totally/mostly easy-going/overconfident*; *half easy-going/overconfident*. We observe that odd numbers of agents generally outperform others within all types of societies, and the possible reason is that odd-number agents can avoid ties. Besides, we also find that the variations of accuracy among odd-number agents is indistinctive. Thus we conclude that **the optimal number of agents is 3, considering both performance and efficiency**. We also implement a significance test of the number of agents shown in Table 11 at Appendix F, demonstrating that different numbers of agents significantly impact performance. Besides, we illustrate consensus reaching with different numbers of agents in Figure 3(b), demonstrating that **more agents are more likely to reach a consensus**.

Different Rounds. We then delve into the effects of different numbers of collaboration rounds, and further scale up the rounds of collaboration, presenting the performance under 3 to 10 rounds in Figure 4. Despite some fluctuation in performance from 3 to 10 rounds of collaboration, the variations are not extremely remarkable. Considering both accuracy and cost, we infer that **3-round collaboration is relatively effective and efficient**. We also conduct a significance test on different rounds of collaborative strategies, shown in Table 12 at Appendix F, and observe that the impact of rounds significantly relies on the collaborative strategy employed. Generally, **the strategies starting or dominating with reflection p_1 differ clearly in performance under different rounds**.

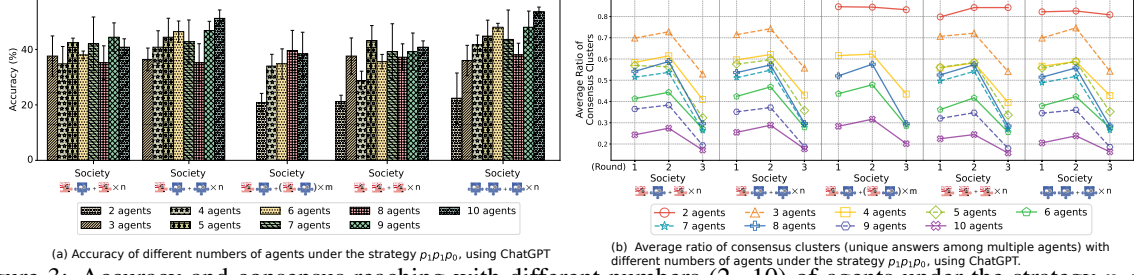


Figure 3: Accuracy and consensus reaching with different numbers (2~10) of agents under the strategy $p_1p_1p_0$ on Chess Move Validity, using ChatGPT. The significance test on agent numbers and comprehensive results under other strategies are shown in Table 11 and Figure 15, 17 at Appendix F due to space limits.

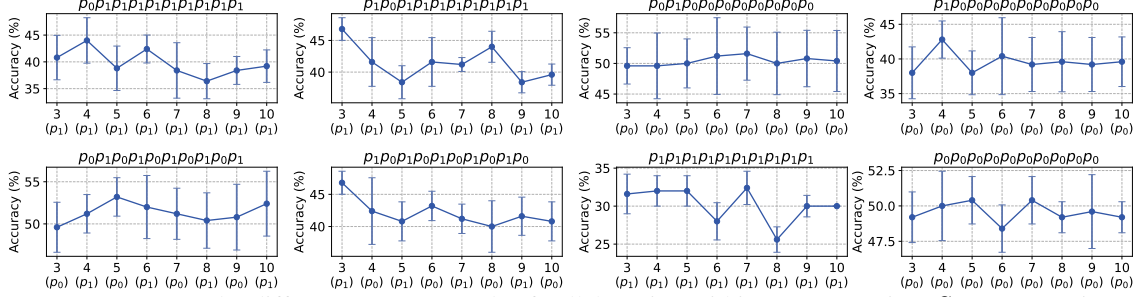


Figure 4: Accuracy under different (3~10) rounds of collaboration within 3-agent society S_2 (1 easy-going and 2 overconfident agents) on MATH, using ChatGPT. The significance test on rounds and experiments on MMLU and Chess Move Validity are shown in Table 12 and Figure 18, 19 at Appendix F due to space limits.

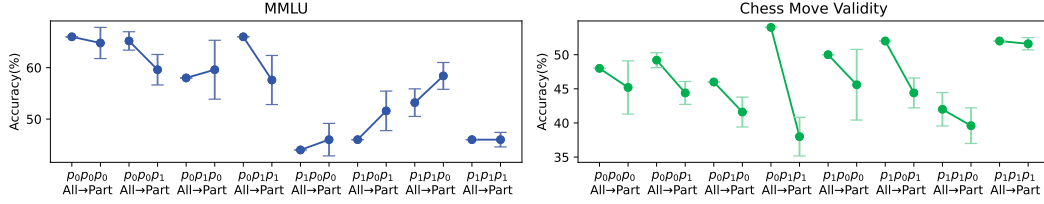


Figure 5: The effect on accuracy of whether all agents in a society execute the same thinking pattern in one round, using ChatGPT. “All” and “Part” respectively refer to all agents applying the same and different thinking pattern(s) in one round. Results on MATH and the significance test is shown in Figure 20 and Table 13 at Appendix F.

Other Collaborative Strategies. Venturing into scenarios with more intricate collaboration, we allow agents to adopt varied thinking patterns in each round of collaboration. For example, given three agents, in a specific round of collaboration, two agents engage in debate while the other one engages in reflection. To increase diversity, we perform a random allocation of thinking patterns to agents in each round, steering clear of scenarios where all agents adopt the same thinking pattern. Intriguingly, as shown in Figure 5, the presence of inconsistent thinking patterns within a society tends to negatively impact performance. Given the observation, we claim that **maintaining a consistent thinking pattern for all agents in a particular round would maximize collaborative efficacy.**

4 Phenomena of Conformity and Consensus Reaching

To address RQ3, we embark on further analysis from a social psychology view (Tajfel, 1982; Tajfel and Turner, 2004; Johnson and Johnson, 2009), to

discern alignment between machine society collaboration and human societal dynamics (Woolley et al., 2010). Our findings indicate that machine society collaboration echo specific human societal phenomena or theories, such as **conformity** (Cialdini and Goldstein, 2004; Allen and Levine, 1969; Coultas and van Leeuwen, 2015) and **consensus reaching** (Scheff, 1967; Degroot, 1974; Baronchelli, 2018) (more analysis are in Appendix G.1). We also analyze **group dynamics** (Cartwright and Zander, 1968; Alderfer, 1987; Forsyth, 2014; Bion, 2018; Forsyth, 2018) in multi-agent collaboration at Appendix G.2 as page limits.

We embark on a detailed analysis, to discern the conformity and consensus reaching phenomena in collaboration. For instance, as depicted in Figure 8(a) at Appendix D.3, an agent initially responds correctly to a question. However, swayed by the misguided answers and explanations from the other two agents, eventually the three agents conforms to an incorrect answer. This phenomenon mirrors detriments in “groupthink” (Janis, 1972;

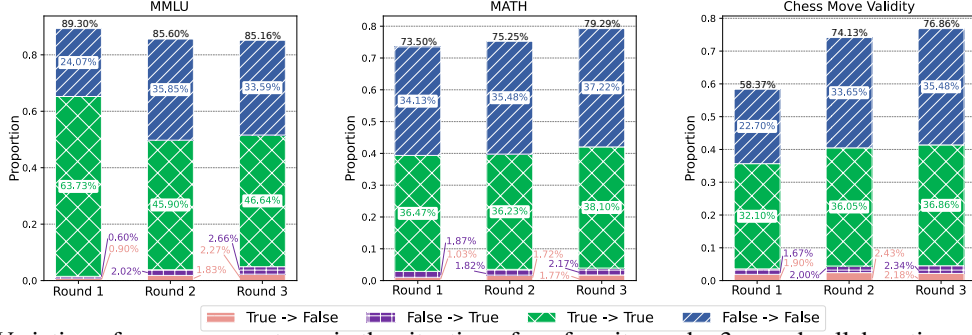


Figure 6: Variation of answer correctness in the situation of conformity, under 3-round collaboration, on ChatGPT, where conformity brings about benefits: $\text{Ratio}(\text{False} \rightarrow \text{True} + \text{True} \rightarrow \text{True}) > \text{Ratio}(\text{True} \rightarrow \text{False} + \text{False} \rightarrow \text{False})$; conformity brings about detriments: $\text{Ratio}(\text{False} \rightarrow \text{True} + \text{True} \rightarrow \text{True}) < \text{Ratio}(\text{True} \rightarrow \text{False} + \text{False} \rightarrow \text{False})$.

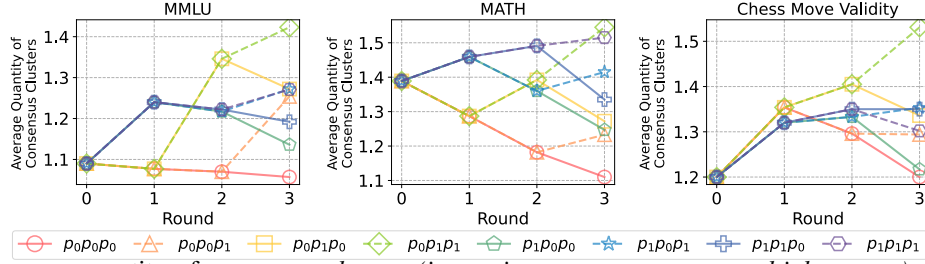


Figure 7: Average quantity of consensus clusters (i.e., unique answers among multiple agents) under different rounds of collaboration with 3-round collaborative strategies, using ChatGPT. Smaller quantity of consensus clusters, more easier it is to reach a consensus. Round 0 is equal to self-consistency. More details are in Appendix G.1.

Jehn, 1995), suggesting that members of tight-knit groups tend to value harmony and consensus over objective critique of divergent views, potentially leading to flawed decisions. Contrastingly, in another scenario illustrated in Figure 8(b) at Appendix D.3, all three agents converge on the right answer after engaging in a society-wide debate. This mirrors benefits in “groupthink” (Jehn, 1995) and “SoM” (Minsky, 1988; Singh, 2003), where a multitude of agents collaboratively yield intelligence. Within such debates, agents furnish varied viewpoints and information. Through these exchanges, conflicts are resolved, ideas are honed, and the group gravitates toward an informed consensus (Fisher et al., 2011; Forsyth, 2018).

We also conduct a quantitative analysis of the prevalence of conformity and consensus reaching phenomena. We analyze on answer correctness changing at each round of collaboration in the situation of conformity, shown in Figure 6 on ChatGPT and Figure 28, 37, 51, 65 on other LLMs at Appendix H. We also present the ratio of consensus reaching at each round in Figure 7 on ChatGPT and Figure 29, 38, 52, 66 on other LLMs at Appendix H. We summarize the following observations:

- **Conformity is widespread**, and the proportion of conformity increases with the round increases in general.
- Overall, considering performance improve-

ment, **conformity is beneficial in on ChatGPT, Qwen 72B; and harmful on LLaMA2 Chat 13B/70B, Mixtral 8×7B.**

- **As the number of rounds increases, benefits of conformity will weaken** (the ratio difference between True and False answers becomes smaller); and **detriments of conformity enhance** (the ratio difference between False and True answers becomes larger).
- Generally, **reflection results in** increasing the quantity of consensus clusters, demonstrating **more difficulty to reach a consensus**, while **debate is more likely to reach a consensus.**

5 Conclusion and Future Work

This study has highlighted the potential of collaboration mechanisms with LLMs. Our findings reveal the impressive collaboration capabilities of LLM agents, with different individual traits, thinking patterns and collaborative strategies. The emergence of human-like behaviors in these agents, resonating with social psychology theories, further emphasizes the potential of human-AI interaction. Moving forward, a deeper exploration into the multi-agent society is warranted, focusing on collaboration behavior refinement; integrating further insights from social psychology could also guide the development of socially aware NLP systems.

Limitation

Although we explored various societies and collaborative strategies, our study still has its limitations. Firstly, limited by expense, we don't explore the impact of multiple agents respectively based on different LLMs, which may lead to more interesting findings at the social level due to the usage of differently distributed pre-trained data and strategies aligned with human intentions. Furthermore, we traversed all possible scenarios by search alone, lacking a way to let the agents adaptively make autonomous decisions on collaborative strategies in specific scenarios. Although *debate* can be as close as possible to the upper limit, this approach entails a larger consumption and there exist some strategies that can achieve better performance with less overhead. Additionally, our experimental setup is relatively straightforward, as we have not consider more intricate configurations, such as a broader range of traits or a larger-scale society. Finally, we evaluate performance through manual validation and rule-based matching, which also limits the ability to validate more realistic and creative tasks, such as literary creation.

Reproducibility Statement

All code and data can be found in the GitHub repository⁵. For specific experimental implementation details, please refer to Appendix D.

Ethics Statement

This research was conducted in line with the highest ethical standards and best practices in research. The data employed were extracted from publicly accessible datasets, ensuring no usage of proprietary or confidential information. Consequently, this research is free from any ethical concerns.

References

- Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2023. [Playing repeated games with large language models](#). *CoRR*, abs/2305.16867.
- Clayton P Alderfer. 1987. [An intergroup perspective on group dynamics](#). *Handbook of organizational behavior*, 190:222.
- Vernon L. Allen and John M. Levine. 1969. [Consensus and conformity](#). *Journal of Experimental Social Psychology*, 5(4):389–399.

⁵<https://anonymous.4open.science/r/MachineSoM-3178>.

- Leila Amgoud and Henri Prade. 2009. [Using arguments for making and explaining decisions](#). *Artif. Intell.*, 173(3-4):413–436.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. [Qwen technical report](#). *CoRR*, abs/2309.16609.
- Andrea Baronchelli. 2018. [The emergence of consensus: a primer](#). *Royal Society open science*, 5(2):172189.
- Wilfred R Bion. 2018. [Group dynamics: A re-view](#). *New directions in psychoanalysis*, pages 440–477.
- R. J. Bogumil. 1985. [The reflective practitioner: How professionals think in action](#). *Proc. IEEE*, 73(4):845–846.
- Gillie Bolton. 2010. [Reflective practice: Writing and professional development](#). Sage publications.
- Dorwin Cartwright and Alvin Zander. 1968. [Group dynamics](#).
- Miguel Castro and Barbara Liskov. 1999. [Practical byzantine fault tolerance](#). In *OSDI*, pages 173–186. USENIX Association.
- Huaben Chen, Wenkang Ji, Lufeng Xu, and Shiyu Zhao. 2023a. [Multi-agent consensus seeking via large language models](#). *CoRR*, abs/2310.20151.
- Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023b. [Reconcile: Round-table conference improves reasoning via consensus among diverse llms](#). *arxiv preprint*, 2309.07864.
- Lingjiao Chen, Matei Zaharia, and James Zou. 2023c. [How is chatgpt's behavior changing over time?](#) *CoRR*, abs/2307.09009.
- Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2024. [Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents](#). In *ICLR*. OpenReview.net.
- Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. [Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction](#). In *WWW*, pages 2778–2788. ACM.

- Yuheng Cheng, Ceyao Zhang, Zhengwen Zhang, Xiangrui Meng, Sirui Hong, Wenhao Li, Zihao Wang, Zekai Wang, Feng Yin, Junhua Zhao, and Xiuqiang He. 2024. [Exploring large language model based intelligent agents: Definitions, methods, and prospects](#). *CoRR*, abs/2401.03428.
- Robert B. Cialdini and Noah J. Goldstein. 2004. [Social influence: Compliance and conformity](#). *Annual Review of Psychology*, 55(1):591–621. PMID: 14744228.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *arXiv preprint*, abs/2110.14168.
- Julie C Coultas and Edwin JC van Leeuwen. 2015. [Conformity: Definitions, types, and evolutionary grounding](#). *Evolutionary perspectives on social psychology*, pages 189–202.
- Morris H. Degroot. 1974. [Reaching a consensus](#). *Journal of the American Statistical Association*, 69(345):118–121.
- Dorottya Demszky, Diyi Yang, David S. Yeager, Christopher J. Bryan, Margaret Clapper, Susannah Chandhok, Johannes C. Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, Michaela Jones, Danielle Krettek-Cobb, Leslie Lai, Nirel Jones Mitchell, Desmond C. Ong, Carol S. Dweck, James J. Gross, and James W. Pennebaker. 2023. [Using large language models in psychology](#). *Nature Reviews Psychology*, 2(11):688–701.
- Hung Du, Srikanth Thudumu, Rajesh Vasa, and Kon Mouzakis. 2024. [A survey on context-aware multi-agent systems: Techniques, challenges and future directions](#). *CoRR*, abs/2402.01968.
- Yali Du, Joel Z. Leibo, Usman Islam, Richard Willis, and Peter Sunehag. 2023a. [A review of cooperation in multi-agent learning](#). *CoRR*, abs/2312.05162.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023b. [Improving factuality and reasoning in language models through multiagent debate](#). *CoRR*, abs/2305.14325.
- Roger Fisher, William L Ury, and Bruce Patton. 2011. [Getting to yes: Negotiating agreement without giving in](#). Penguin.
- Donelson R Forsyth. 2014. [Group dynamics](#). Wadsworth Cengage Learning.
- Donelson R Forsyth. 2018. [Group dynamics](#). Cengage Learning.
- Howard S Friedman and Miriam W Schustack. 1999. [Personality: Classic theories and modern research](#). Allyn and Bacon Boston, MA.
- Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2023. [Large language models empowered agent-based modeling and simulation: A survey and perspectives](#). *CoRR*, abs/2312.11970.
- 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](#). *CoRR*, abs/2402.01680.
- Thilo Hagendorff. 2023. [Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods](#). *CoRR*, abs/2303.13988.
- Shanshan Han, Qifan Zhang, Yuhang Yao, Weizhao Jin, Zhaozhuo Xu, and Chaoyang He. 2024. [Llm multi-agent systems: Challenges and open problems](#). *CoRR*, abs/2402.03578.
- Rui Hao, Linmei Hu, Weijian Qi, Qingliu Wu, Yirui Zhang, and Liqiang Nie. 2023. [Chatllm network: More brains, more intelligence](#). *CoRR*, abs/2304.12998.
- David Held. 2006. [Models of democracy](#). Polity.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. [Measuring massive multitask language understanding](#). In *ICLR*. OpenReview.net.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. [Measuring mathematical problem solving with the MATH dataset](#). In *NeurIPS Datasets and Benchmarks*.
- Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, and Chenglin Wu. 2024. [Metagpt: Meta programming for multi-agent collaborative framework](#). In *ICLR*. OpenReview.net.
- Irving L Janis. 1972. [Victims of Groupthink: A psychological study of foreign-policy decisions and fiascoes](#). Houghton Mifflin.
- Karen A Jehn. 1995. [A multimethod examination of the benefits and detriments of intragroup conflict](#). *Administrative science quarterly*, pages 256–282.
- 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 Comput. Surv.*, 55(12):248:1–248:38.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock,

750	Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b . <i>CoRR</i> , abs/2310.06825.	804
751		805
752		806
753	Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L��lio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Th��ophile Gervet, Thibaut Lavril, Thomas Wang, Timoth��e Lacroix, and William El Sayed. 2024. Mixtral of experts . <i>CoRR</i> , abs/2401.04088.	807
754		808
755		809
756		810
757		811
758		812
759		813
760		
761		814
762		815
763		816
764	David W Johnson and Roger T Johnson. 2009. An educational psychology success story: Social interdependence theory and cooperative learning . <i>Educational researcher</i> , 38(5):365–379.	817
765		818
766		
767		
768	Luoma Ke, Song Tong, Peng Chen, and Kaiping Peng. 2024. Exploring the frontiers of llms in psychological applications: A comprehensive review . <i>CoRR</i> , abs/2401.01519.	819
769		820
770		821
771		822
772	Chuyi Kong, Yaxin Fan, Xiang Wan, Feng Jiang, and Benyou Wang. 2023. Large language model as a user simulator . <i>CoRR</i> , abs/2308.11534.	823
773		824
774		825
775	Alan M Leslie, Ori Friedman, and Tim P German. 2004. Core mechanisms in ‘theory of mind’ . <i>Trends in cognitive sciences</i> , 8(12):528–533.	826
776		827
777		828
778	Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023a. CAMEL: communicative agents for “mind” exploration of large scale language model society . <i>CoRR</i> , abs/2303.17760.	829
779		
780		830
781		831
782		832
783	Yuan Li, Yixuan Zhang, and Lichao Sun. 2023b. Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents . <i>CoRR</i> , abs/2310.06500.	833
784		834
785		835
786		836
787		837
788	Yujia Li, David H. Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, R��mi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d’Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. 2022. Competition-level code generation with alphacode . <i>CoRR</i> , abs/2203.07814.	838
789		839
790		
791		840
792		841
793		842
794		843
795		844
796		845
797		846
798		
799	Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate . <i>CoRR</i> , abs/2305.19118.	847
800		848
801		849
802		850
803		
	Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M. Dai, Diyi Yang, and Soroush Vosoughi. 2023a. Training socially aligned language models in simulated human society . <i>arxiv preprint</i> , abs/2305.16960.	851
		852
		853
	Ruibo Liu, Ge Zhang, Xinyu Feng, and Soroush Vosoughi. 2022a. Aligning generative language models with human values . In <i>NAACL-HLT (Findings)</i> , pages 241–252. Association for Computational Linguistics.	854
		855
		856
	Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022b. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks . In <i>ACL (2)</i> , pages 61–68. Association for Computational Linguistics.	857
		858
	Zhiwei Liu, Weiran Yao, Jianguo Zhang, Le Xue, Shelby Heinecke, Rithesh Murthy, Yihao Feng, Zeyuan Chen, Juan Carlos Niebles, Devansh Arpit, Ran Xu, Phil Mui, Huan Wang, Caiming Xiong, and Silvio Savarese. 2023b. BOLAA: benchmarking and orchestrating llm-augmented autonomous agents . <i>CoRR</i> , abs/2308.05960.	
	Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2023c. Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization . <i>CoRR</i> , abs/2310.02170.	
	Junliang Luo, Tianyu Li, Di Wu, Michael Jenkin, Steve Liu, and Gregory Dudek. 2024. Hallucination detection and hallucination mitigation: An investigation . <i>CoRR</i> , abs/2401.08358.	
	Qun Ma, Xiao Xue, Deyu Zhou, Xiangning Yu, Donghua Liu, Xuwen Zhang, Zihan Zhao, Yifan Shen, Peilin Ji, Juanjuan Li, Gang Wang, and Wangepeng Ma. 2024. Computational experiments meet large language model based agents: A survey and perspective . <i>CoRR</i> , abs/2402.00262.	
	Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback . <i>arXiv preprint</i> , abs/2303.17651.	
	Qiaozhu Mei, Yutong Xie, Walter Yuan, and Matthew O. Jackson. 2023. A turing test: Are AI chatbots behaviorally similar to humans? <i>CoRR</i> , abs/2312.00798.	
	Jack Mezirow. 2003. How critical reflection triggers transformative learning . <i>Adult and Continuing Education: Teaching, learning and research</i> , 4:199.	
	Jack Mezirow. 2018. Transformative learning theory . In <i>Contemporary theories of learning</i> , pages 114–128. Routledge.	
	Marvin Minsky. 1988. <i>Society of mind</i> . Simon and Schuster.	

- Don A Moore and Paul J Healy. 2008. [The trouble with overconfidence](#). *Psychological review*, 115(2):502. 913
- Diana C Mutz. 2006. [Hearing the other side: Deliberative versus participatory democracy](#). Cambridge University Press. 914
- OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. <https://openai.com/blog/chatgpt/>. 915
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *NeurIPS*. 920
- Joon Sung Park, Joseph C. O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. [Generative agents: Interactive simulacra of human behavior](#). In *UIST*, pages 2:1–2:22. ACM. 921
- Chaim Perelman. 1971. [The new rhetoric](#). Springer. 922
- Pouya Pezeshkpour, Eser Kandogan, Nikita Bhutani, Sajjadur Rahman, Tom Mitchell, and Estevam Hruschka. 2024. [Reasoning capacity in multi-agent systems: Limitations, challenges and human-centered solutions](#). *CoRR*, abs/2402.01108. 923
- Vipula Rawte, Amit P. Sheth, and Amitava Das. 2023. [A survey of hallucination in large foundation models](#). *CoRR*, abs/2309.05922. 924
- Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. [Neural theory-of-mind? on the limits of social intelligence in large lms](#). In *EMNLP*, pages 3762–3780. Association for Computational Linguistics. 925
- Thomas J Scheff. 1967. [Toward a sociological model of consensus](#). *American Sociological Review*, pages 32–46. 926
- David Wyatt Seal, Laura M Bogart, and Anke A Ehrhardt. 1998. [Small group dynamics: The utility of focus group discussions as a research method](#). *Group Dynamics: Theory, Research, and Practice*, 2(4):253. 927
- Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023. [Clever hans or neural theory of mind? stress testing social reasoning in large language models](#). *arXiv preprint*, abs/2305.14763. 928
- Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. [Towards understanding sycophancy in language models](#). *CoRR*, abs/2310.13548. 929
- Noah Shinn, Beck Labash, and Ashwin Gopinath. 2023. [Reflexion: an autonomous agent with dynamic memory and self-reflection](#). *arXiv preprint*, abs/2303.11366. 930
- Michael Siegal and Rosemary Varley. 2002. [Neural systems involved in ‘theory of mind’](#). *Nature Reviews Neuroscience*, 3(6):463–471. 931
- Push Singh. 2003. [Examining the society of mind](#). *Comput. Artif. Intell.*, 22(6):521–543. 932
- Nikita Soni, Niranjana Balasubramanian, H. Andrew Schwartz, and Dirk Hovy. 2024. [Comparing human-centered language modeling: Is it better to model groups, individual traits, or both?](#) *CoRR*, abs/2401.12492. 933
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, and et al. 2022. [Beyond the imitation game: Quantifying and extrapolating the capabilities of language models](#). *arXiv preprint*, abs/2206.04615. 934
- Peter Stone and Manuela M. Veloso. 2000. [Multiagent systems: A survey from a machine learning perspective](#). *Auton. Robots*, 8(3):345–383. 935
- Cass R Sunstein. 2005. [Why societies need dissent](#). Harvard University Press. 936
- Henri Tajfel. 1982. [Social psychology of intergroup relations](#). *Annual review of psychology*, 33(1):1–39. 937
- Henri Tajfel and John C Turner. 2004. [The social identity theory of intergroup behavior](#). In *Political psychology*, pages 276–293. Psychology Press. 938
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#). *CoRR*, abs/2302.13971. 939

968	José M. Vidal. 2006. <i>Fundamentals of Multiagent Systems: Using NetLogo Models</i> . Unpublished. http://www.multiagent.com .	
969		
970		
971	Cheng Wang, Chuwen Wang, Yu Zhao, Shirong Zeng, Wang Zhang, and Ronghui Ning. 2024a. Behavioral simulation: Exploring a possible next paradigm for science. <i>CoRR</i> , abs/2401.09851.	
972		
973		
974		
975	Haonan Wang, James Zou, Michael Mozer, Anirudh Goyal, Alex Lamb, Linjun Zhang, Weijie J Su, Zhun Deng, Michael Qizhe Xie, Hannah Brown, and Kenji Kawaguchi. 2024b. Can ai be as creative as humans? <i>CoRR</i> , abs/2401.01623.	
976		
977		
978		
979		
980	Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. 2023a. A survey on large language model based autonomous agents. <i>CoRR</i> , abs/2308.11432.	
981		
982		
983		
984		
985		
986	Zekun Wang, Ge Zhang, Kexin Yang, Ning Shi, Wangchunshu Zhou, Shaochun Hao, Guangzheng Xiong, Yizhi Li, Mong Yuan Sim, Xiuying Chen, Qingqing Zhu, Zhenzhu Yang, Adam Nik, Qi Liu, Chenghua Lin, Shi Wang, Ruibo Liu, Wenhu Chen, Ke Xu, Dayiheng Liu, Yike Guo, and Jie Fu. 2023b. Interactive natural language processing. <i>CoRR</i> , abs/2305.13246.	
987		
988		
989		
990		
991		
992		
993		
994	Gerhard Weiß. 1995. Adaptation and learning in multi-agent systems: Some remarks and a bibliography. In <i>Adaption and Learning in Multi-Agent Systems</i> , volume 1042 of <i>Lecture Notes in Computer Science</i> , pages 1–21. Springer.	
995		
996		
997		
998		
999	Michael J. Wooldridge. 2009. <i>An Introduction to MultiAgent Systems, Second Edition</i> . Wiley.	
1000		
1001	Anita Williams Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. <i>Science</i> , 330(6004):686–688.	
1002		
1003		
1004		
1005		
1006	Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huan, and Tao Gui. 2023. The rise and potential of large language model based agents: A survey. <i>arxiv preprint</i> , abs/2309.07864.	
1007		
1008		
1009		
1010		
1011		
1012		
1013		
1014		
1015		
1016		
1017	Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip H. S. Torr, Bernard Ghanem, and Guohao Li. 2024. Can large language model agents simulate human trust behaviors? <i>CoRR</i> , abs/2402.04559.	
1018		
1019		
1020		
1021		
	Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. <i>arXiv preprint</i> , abs/2306.13549.	1022
		1023
		1024
		1025
	Xijia Zhang, Yue Guo, Simon Stepputtis, Katia P. Sycara, and Joseph Campbell. 2023a. Explaining agent behavior with large language models. <i>CoRR</i> , abs/2309.10346.	1026
		1027
		1028
		1029
	Xijia Zhang, Yue Guo, Simon Stepputtis, Katia P. Sycara, and Joseph Campbell. 2023b. Understanding your agent: Leveraging large language models for behavior explanation. <i>CoRR</i> , abs/2311.18062.	1030
		1031
		1032
		1033
	Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023c. Siren’s song in the AI ocean: A survey on hallucination in large language models. <i>CoRR</i> , abs/2309.01219.	1034
		1035
		1036
		1037
		1038
		1039
		1040
	Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. <i>arXiv preprint</i> , abs/2303.18223.	1041
		1042
		1043
		1044
		1045
		1046
		1047
		1048
	Wangchunshu Zhou, Yuchen Eleanor Jiang, Long Li, Jialong Wu, Tiannan Wang, Shi Qiu, Jintian Zhang, Jing Chen, Ruipu Wu, Shuai Wang, Shiding Zhu, Jiyu Chen, Wentao Zhang, Ningyu Zhang, Hua-jun Chen, Peng Cui, and Mrinmaya Sachan. 2023. Agents: An open-source framework for autonomous language agents. <i>CoRR</i> , abs/2309.07870.	1049
		1050
		1051
		1052
		1053
		1054
		1055
	Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities. <i>CoRR</i> , abs/2305.13168.	1056
		1057
		1058
		1059
		1060
		1061
	Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R. Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Abed Al Kader Hammoud, Vincent Herrmann, Kazuki Irie, Louis Kirsch, Bing Li, Guohao Li, Shuming Liu, Jinjie Mai, Piotr Piekos, Aditya Ramesh, Imanol Schlag, Weimin Shi, Aleksandar Stanic, Wenyi Wang, Yuhui Wang, Mengmeng Xu, Deng-Ping Fan, Bernard Ghanem, and Jürgen Schmidhuber. 2023. Mindstorms in natural language-based societies of mind. <i>CoRR</i> , abs/2305.17066.	1062
		1063
		1064
		1065
		1066
		1067
		1068
		1069
		1070
		1071
		1072

Overview of Appendix

We summarize the overview of Appendix below:

§A: Key Takeaways.

§B: Related Work.

§C: Potential Real-World Applications.

§D: Implementation Details.

Experimental Setup (§D.1)

Experimental Evaluation (§D.2)

Illustration of Agent Collaboration (§D.3)

§E: Further Analysis on Machine Social Collaboration (Backbone: ChatGPT).

§F: Analysis on Machine Society Settings (Backbone: ChatGPT).

§G: A Social Psychology View on Conformity, Consensus Reaching, and Group Dynamics (Backbone: ChatGPT).

Conformity, Consensus Reaching (§G.1)

Group Dynamics (§G.2)

§H: Analysis on Different Backbone LLMs.

LlaMA2 Chat 13B (§H.1)

LlaMA2 Chat 70B (§H.2)

Qwen 72B (§H.3)

Mixtral 8×7B (§H.4)

§I: Effectiveness of Prompts.

A Key Takeaways

Drawing from our comprehensive analysis, we distill valuable insights for future multi-agent collaboration designs concerning *Strategy Selection*, *Society Settings*, and *Social Psychology View*.

Regarding *Strategy Selection*,

- Starting or dominating multi-agent collaboration with debate, yields relatively optimal outcomes, as seen from Table 2, 8, 14, 20, 26, 32.
- Totally-reflection strategy like $p_1p_1p_1$ is generally worst in performance, as observed from Table 2, 8, 14, 20, 26, 32.
- For difficult tasks, debate combined with continuous reflection is superior; for simple tasks, self-consistency or reflection is enough, as seen from Figure 13, 24, 33, 42, 56.

Regarding *Society Settings*,

- Surprisingly, “overconfident” agents lose that trait in groups! As observed from word clouds in Figure 11, 22, 31, 40, 54 and answer changing in Figure 12, 23, 32, 41, 55.

- Setting agent numbers to 3 is generally advantageous in performance and cost, as seen from Figure 15, 25, 34, 43, 57.

- The rounds of collaboration is relatively suitable to set as 3, since it’s both effective and efficient. As seen from Figure 18, 4, 19 on ChatGPT; Figure 26, 35 on LlaMA 13B/70B; Figure 47, 48, 49 on Qwen 72B; Figure 61, 62, 63 on Mixtral 8×7B.

- Employing the uniform thinking patterns across all agents within a round enhance efficacy, as seen from Figure 5, 20, 27, 36, 50, 64.

Regarding *Social Psychology View*,

- Collaboration is generally effective in the group, especially for tackling difficult tasks. As observed from Figure 13, 24, 33, 42, 56; and Figure 21, 30, 39, 53, 67.
- Collaboration widely leads to conformity, either beneficial or harmful in performance. As observed from Figure 6, 28, 37, 51, 65.
- As the number of rounds increases, benefits of conformity will decrease; and detriments of conformity will increase. As observed from Figure 6, 28, 37, 51, 65.
- The totally easy-going society is more likely to reach a consensus, debate helps to consensus reaching while reflection impedes it. As observed from Figure 16, 45, 59; and Figure 7, 29, 38, 52, 66.

B Related Work

With the birth of Large Language Models (LLMs), prompt engineering (Liu et al., 2022b; Chen et al., 2022) become the key to utilize LLMs. When the pre-trained LLMs are aligned, they show human-like intelligence. Hence, *agent* replaces *prompt engineering* as the new research hotspot. Recently there has been a proliferation of top-level designs of various agent systems, such as Generative Agents (Park et al., 2023), MetaGPT (Hong et al., 2024), BOLAA (Liu et al., 2023b) and Agents (Zhou et al., 2023). These works has primarily focused on the careful design of components such as memory, environment, and planning. There are also some works exploring what kind of mindset can fully exploit the full performance of multi-agent including *debate* (Du et al., 2023b)

and *reflection* (Madaan et al., 2023). Both of these types of work are mostly done concurrently.

AgentVerse (Chen et al., 2024) draws on the above two types of work to explore the architecture of multi-agent and design two collaborative strategies, *Horizontal Communication* (similar to debate) and *Vertical Communication* (similar to self-refine (Madaan et al., 2023)). These two collaboration strategies are included in our code framework. In addition, we have also explored a variety of other societies and collaboration methods. Whereas the RECONCILE (Chen et al., 2023b) focuses on exploring cooperation between agents constituted by different model compositions, although we do not show this in our work, our code framework easily expands to it.

C Potential Real-world Applications

In this section, we present the potential applications (Ke et al., 2024) of our work, which can be primarily divided into two parts, experimental results and experimental framework:

- Our experimental findings offer valuable insights for addressing problems through multi-agent systems. Presently, within various multi-agent frameworks (Hong et al., 2024; Chen et al., 2024; Zhou et al., 2023), tackling a substantial issue typically involves breaking down the task into several sub-tasks. Collaboration among multiple agents to solve these sub-tasks often necessitates ongoing cooperation. There are currently two predominant approaches: (i) involving another agent specifically to decide who should offer suggestions and determining whether the current task is resolved, and (ii) collaborating in a fixed order. The performance of the first method is often unpredictable and entails significant randomness, prompting a preference for the second method. At this juncture, our conclusions on rounds, the number of agents, and cognitive approaches can inform the design of effective collaboration strategies among agents.
- Our experimental framework holds relevance for psychologists seeking inspiration and provides guidance for language model designers. As indicated in previous works (Demszky et al., 2023; Hagendorff, 2023), once a testing setup for machine psychology is established, researchers can explore the longitudinal de-

velopment of LLMs over time by applying the same task multiple times, thereby generating data. This data serves as a benchmark for discerning trends in LLMs development. Psychologists can draw upon our framework to conduct secondary designs, draw meaningful conclusions, and, in conjunction with theories of human social psychology and successful experiences in human society, contribute to addressing issues in LLMs and designing superior machine social architectures and collaboration methods.

D Implementation Details

D.1 Experimental Setup

Model	Temperature	Top K	Top P
gpt-3.5-turbo-1106	0.00	-	1.00
LlaMA-13B-Chat	0.75	50	0.95
LlaMA-70B-Chat	0.75	50	0.95
Mixtral 8×7B	0.75	50	0.95
Qwen 72B	0.75	50	0.80

Table 3: Decoding parameters of different models.

The decoding parameters for various models are detailed in Table 3. In the case of gpt-3.5-turbo-1106, we align our approach with Du et al. (2023b) by setting the temperature to 0, while adhering to the default settings for the remaining parameters. For the model Qwen 72B, we utilize the default parameters as furnished by the official documentation. For the remaining models, we configure the temperature to 0.7 and adjusted the Top P and Top K values to 50 and 0.95, respectively. This configuration is primarily based on insights from Demszky et al. (2023), which advocates for the recognition and integration of the inherent stochastic nature of LLM outputs into analytical frameworks, in a manner akin to the treatment of stochastic variables in psychological studies. It is noteworthy that even with the temperature parameter set to zero, the gpt-3.5-turbo-1106 may still exhibit randomness in the outputs.

The detailed society settings in the three different experiments mentioned in §3.2 are shown in Table 4. Due to the context length constraints of the LlaMA-13B-Chat and LlaMA-70B-Chat, which supports a maximum of 4096 tokens, scaling up the number of agents and collaboration rounds presents a challenge. Consequently, we have capped the collaboration rounds at 4 and also restricted the agent

Experiment Type	Model	Dataset	Collaboration Round	Number of Agents	Society
Different Number of Agents	gpt-3.5-turbo-1106 Mixtral 8x7B Qwen 72B	Chess Move Validity	3	2~10	See the Figure 15 and Table 11.
	LlaMA-13B-Chat LlaMA-70B-Chat	MMLU Chess Move Validity	3	2~4	Only one easy-going agent in the society
Different Collaboration Rounds	gpt-3.5-turbo-1106 Mixtral 8x7B Qwen 72B	MMLU MATH Chess Move Validity	10	3	S_2
	LlaMA-13B-Chat LlaMA-70B-Chat	MMLU Chess Move Validity	4	3	S_2
Different Strategy	gpt-3.5-turbo-1106 LlaMA-13B-Chat LlaMA-70B-Chat Mixtral 8x7B Qwen 72B	MMLU MATH Chess Move Validity	3	3	S_2

Table 4: The detailed society settings in the three different experiments mentioned in Section 3.2.

count to 4. We select the MMLU and Chess Move Validity datasets for our studies. Nevertheless, a small fraction of cases still exceed the maximum length constraint. To address this, we strategically prune content from the earlier rounds to ensure compliance with the length limitation. As for other models, in terms of experimenting with the number of agents involved, adding an additional agent results in substantial costs. This is due to the necessity of conducting 5 replicate experiments and accommodating 8 collaborative strategies. Therefore, for ChatGPT, Mixtral 8×7B, and Qwen 72B, our experiments are carried out on the less token-intensive dataset, Chess Move Validity. As for trials concerning the number of collaboration rounds, the quantity of viable collaboration strategies increases exponentially with each additional round – for instance, 10 rounds would yield 2^{10} unique strategies. To manage this complexity, we have opted to experiment with 8 strategies that are representative of the broader set of possibilities.

The prompts used in the experiment are detailed in Table 5. Concerning the MMLU dataset, we curated questions from six domains (statistics, mathematics, computer science, biology, chemistry, and physics) and performed a random sampling of 50 samples, maintaining a proportion of 8 : 8 : 8 : 8 : 9 : 9 for each domain. Regarding the MATH dataset, we randomly selected 50 cases from Levels 3, 4, and 5, distributing them in a ratio of 22 : 22 : 6.

D.2 Experimental Evaluation

The evaluation process involves two fundamental steps: (i) A unified answer is selected from the machine society. To achieve this, we employ the majority vote method to ascertain the consensus reached by the society after multiple rounds of collaboration. For instances where unanimity among agents is not achieved, it is considered an error. Additionally, if an individual agent provides multiple answers without following our prompts, its response is disregarded. (ii) Answer responses from agents are matched against the ground truth. This step presents two main challenges. Firstly, there is the concern of non-compliance with instructions. Despite providing explicit prompts and specifying the desired output format for evaluation, occasional deviations from the given instructions by agents are inevitable. Secondly, the answers may manifest in non-unique forms, leading to potential variations, such as the equivalence between “3/4” and “0.75” in MATH (Hendrycks et al., 2021b). To address these challenges, a comprehensive set of matching rules is employed. Nonetheless, it is important to acknowledge the possibility of encountering a small number of values that fall outside the purview of these rules.

D.3 Illustration of Multi-Agent Collaboration

As seen from Figure 8, the conformity phenomenon in multi-agent collaboration can be both beneficial (*i.e.*, changing the answer from wrong to correct) and harmful (*i.e.*, changing the answer from correct

Task	Type	Prompt
Math	easy-going	<i>You are an expert skilled in solving mathematical problems and are objective and unbiased, and you can be persuaded if other agent's answers make sense. Please keep this in mind. If you understand please say ok only.</i>
	overconfident	<i>Imagine you are an expert in solving mathematical problems and are confident in your answer and often persuades other agents to believe in you. Please keep this in mind. If you understand please say ok only.</i>
	question	<i>Here is a math problem written in LaTeX: <problem>\n Please carefully consider it and explain your reasoning. Put your answer in the form \boxed{{answer}}, at the end of your response.</i>
	debate	<i>These are the solutions to the problem from other agents: <other agent responses>Using the reasoning from other agents as additional information and referring to your historical answers, can you give an updated answer? Put your answer in the form \boxed{{answer}}, at the end of your response.</i>
	reflection	<i>Can you double check that your answer is correct? Please reiterate your answer, with your answer in the form \boxed{{answer}}, at the end of your response.</i>
MMLU	easy-going	<i>You are an expert in biology, chemistry, computer science, mathematics, physics and are objective and unbiased, and you can be persuaded if other agent's answers make sense. Please keep this in mind. If you understand please say ok only.</i>
	overconfident	<i>Imagine you are an expert in biology, chemistry, computer science, mathematics, physics and are confident in your answer and often persuades other agents to believe in you. Please keep this in mind. If you understand please say ok only.</i>
	question	<i>Can you answer the following question as accurately as possible? <Question>: A) <A>, B) , C) <C>, D) <D>Explain your answer, putting the answer in the form (X) at the end of your response.</i>
	debate	<i>These are the solutions to the problem from other agents: <other agent responses>Using the reasoning from other agents as additional advice, can you give an updated answer? Examine your solution and that other agents. Put your answer in the form (X) at the end of your response.</i>
	reflection	<i>Can you double check that your answer is correct. Put your final answer in the form (X) at the end of your response.</i>
Chess Move Validity	easy-going	<i>You are an expert skilled in playing chess and are objective and unbiased, and you can be persuaded if other agent's answers make sense. Please keep this in mind. If you understand, please say ok only.</i>
	overconfident	<i>Imagine you are an expert skilled in playing chess and are confident in your answer and often persuades other agents to believe in you. Please keep this in mind. If you understand, please say ok only.</i>
	question	<i>Given the chess game <chess move>, give one valid destination square for the chess piece at <square>. Give a one-line explanation of why your destination square is a valid move. State your final answer in a newline with a 2 letter response following the regex [a-h][1-8].</i>
	debate	<i>Here are destination square suggestions from other agents: Can you double check that your destination square is a valid move? Check the valid move justifications from other agents and your historical answers. State your final answer in a newline with a 2-letter response following the regex [a-h][1-8].</i>
	reflection	<i>Can you double check that your destination square is a valid move? Check the valid move justifications from your historical answers. State your final answer in a newline with a 2 letter response following the regex [a-h][1-8].</i>

Table 5: Prompts in each task.

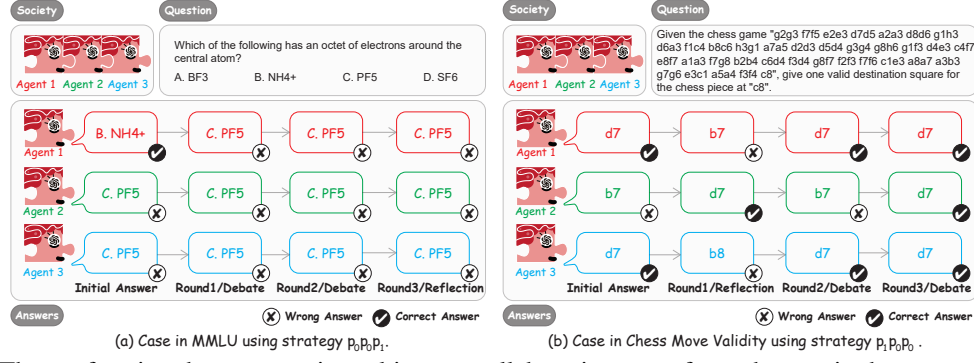


Figure 8: The conformity phenomenon in multi-agent collaboration, seen from changes in the answers during the process of solving a certain task with 3 agents in the society S_4 (all agents are easy-going), using *ChatGPT*.

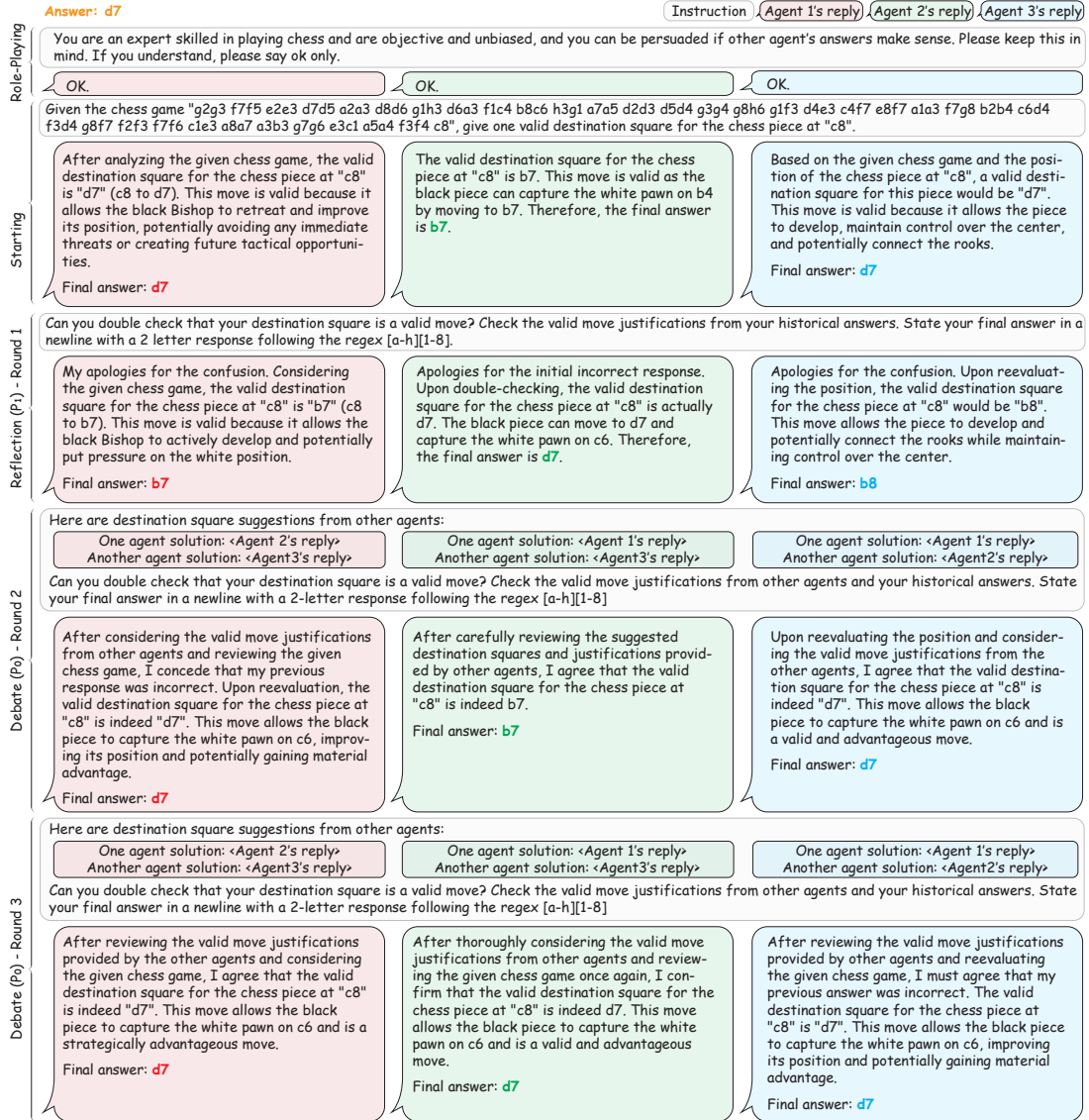


Figure 9: In the S_4 society, three agents engage in a complete dialogue using the strategy $p_1p_0p_0$ to predict the next piece position given a chess game from the Chess Move Validity dataset.

to wrong) in problem solving.

Figure 9 and Figure 10 illustrate the detailed conversation process for multi-agent collaboration regarding the content shown in Figure 8.

E Further Analysis on Machine Social Collaboration (Backbone: ChatGPT)

We conduct a rigorous **significance test** for the main experiment in §3.1. Given our experimen-

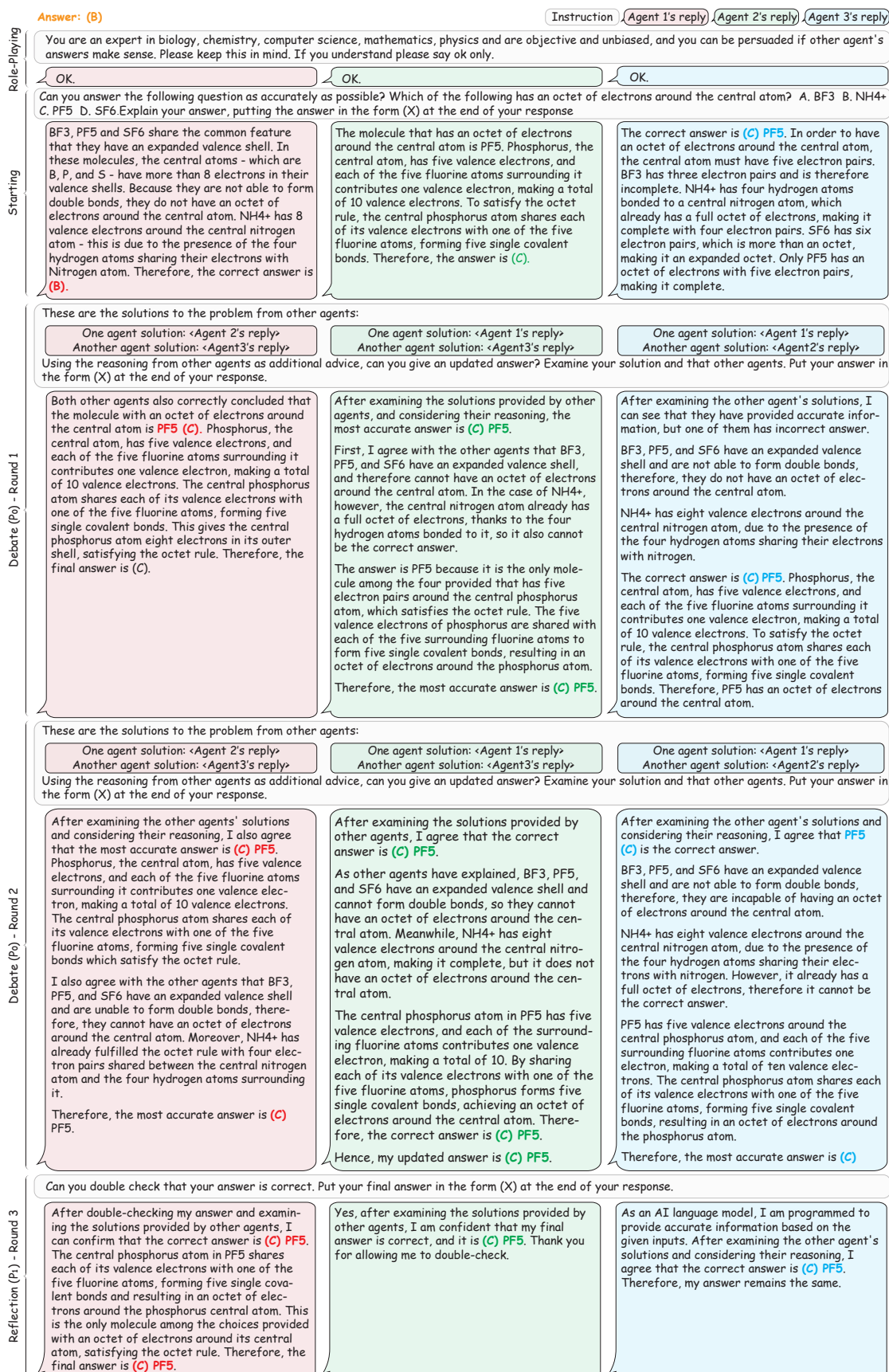


Figure 10: In the S_4 society, three agents engage in a complete dialogue using the strategy $p_0p_0p_1$ to solve a high-school chemical question from MMLU.

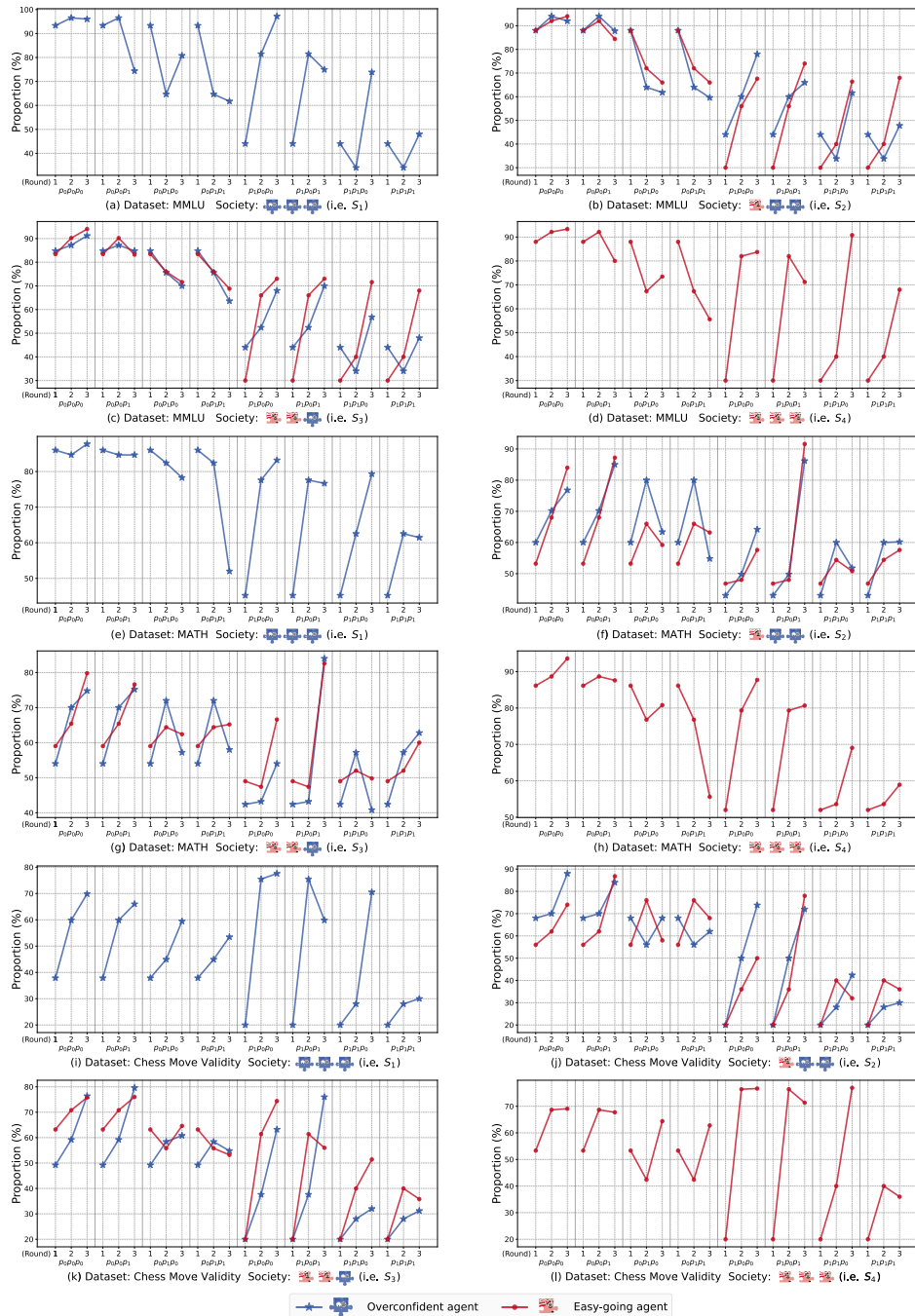
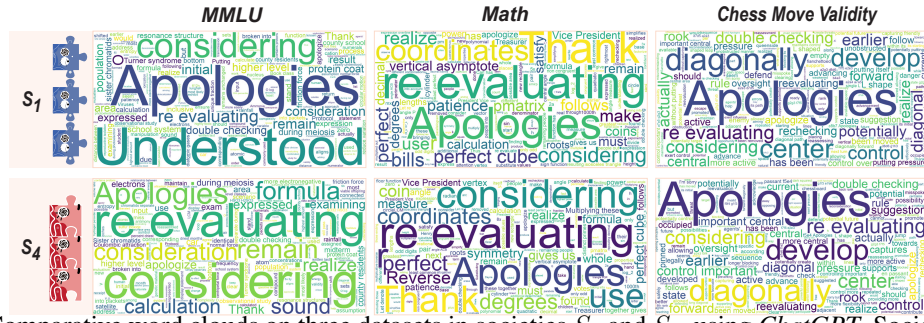


Figure 12: Proportion of agents with different traits changing answers in societies S_1 and S_4 , using *ChatGPT*. Society S_1 features three overconfident agents, while society S_4 comprises three easy-going agents.

Collaborative Strategy	MMLU p-value	MATH p-value	Chess Move Validity p-value
$p_0p_0p_0$	0.079	0.274	0.004
$p_0p_0p_1$	0.956	0.011	0.000
$p_0p_1p_0$	0.120	0.003	0.009
$p_0p_1p_1$	0.000	0.323	0.014
$p_1p_0p_0$	0.000	0.027	0.000
$p_1p_0p_1$	0.063	0.017	0.000
$p_1p_1p_0$	0.000	0.300	0.000
$p_1p_1p_1$	0.000	0.000	0.000

Table 6: One-Way ANOVA results for the impact of society on accuracy with fixed collaborative strategy, based on experiments from Table 2 using *ChatGPT*.

Society	MMLU p-value	MATH p-value	Chess Move Validity p-value
S_1	0.000	0.000	0.293
S_2	-	0.000	-
S_3	0.000	0.001	0.000
S_4	0.000	0.000	0.000

Table 7: One-Way ANOVA results for the impact of collaborative strategy on accuracy with fixed society, based on experiments from Table 2 using *ChatGPT*. ‘-’: It doesn’t pass homogeneity test for variance.

tal design incorporating two key factors, namely *collaborative strategy* and *society*, we respectively opt for a one-way analysis of variance. Before delving into the analysis, we ensured that the data adhered to a normal distribution and satisfied the assumption of homogeneity of variance. We present the p -values for society and collaborative strategy across the three datasets in Table 6, 7.

A notable observation is that the p -value associated with the collaborative strategy is significantly below the 0.05 threshold, indicating its substantial impact. In contrast, the p -value of the other two factors is obviously greater than 0.05. This corroborates our earlier conclusion in §3.1, emphasizing that the influence of collaborative strategy outweighs that of society. Additionally, [Chen et al. \(2023c\)](#) demonstrate that LLMs are well-known to show sycophant behaviors.

We then present the **main results and significance tests** of societies and strategies on **ChatGPT (with engine of gpt-3.5-turbo employed between July 10 and July 23, 2023)** in Table 8, 9, 10.

We also present the **word clouds** in Figure 11, and **answer changing of agents with different traits** in Figure 11, to reveal that indistinctive im-

part of 3-agent societies on performance. Furthermore, we demonstrate that the tasks with different subjects and difficulty display varying sensitivity to collaborative strategies, as presented with **radar maps** in Figure 13.

F Analysis on Machine Society Settings (Backbone: ChatGPT)

In this section, we conduct **significance tests** for the experiments outlined in §3.2. The chosen method is one-way analysis of variance. Prior to the analysis, we performed a check for homogeneity of variance, with only one entry in Table 13 deviating from the criteria. Significance tests for the number of agents, the number of rounds, and different collaborative strategies are individually detailed in Table 11, Table 12 and Table 13 respectively.

Different Numbers of Agents. According to the results of the p -values in Table 11, the conclusion in §3.2 is confirmed, namely, different number of agents results in a significant correlation on performance. By integrating the results in Figure 3, it becomes evident that the presence of three agents is relatively optimal.

We also analyse the *consensus reaching* with different numbers of agents, and present the results in Figure 16, 17.

Different Rounds of Collaboration. As observed from Table 12, we find that the influence of rounds significantly relies on the collaborative strategy employed. For MMLU and Chess Move Validity, the collaborative strategies associated with p -values < 0.05 are $\{p_0p_1p_1p_0, p_0p_1p_1p_1, p_1p_0p_1p_0, p_1p_0p_1p_1\}$ and $\{p_0p_1p_1p_0, p_0p_1p_1p_1, p_1p_0p_1p_1, p_1p_1p_0p_0, p_1p_1p_0p_1, p_1p_1p_1p_0\}$, respectively. We also increase the rounds of collaboration, from 3 to 10, and present the results in Figure 18, 19. We find that although there would be some fluctuations on performance if we scale up the round of collaboration, the outperformance is not significant enough. While increasing rounds of collaboration will result in more consumption of tokens, which are not economic. Thus we infer that the 3-round collaboration is relatively optimal considering both performance and cost.

Furthermore, as seen from Figure 7, the strategy subsequent with a round of debate tends to yield less consensus clusters compared to the preceding rounds. Conversely, the strategy subsequent

	Metric (Strategy)	Society	Collaborative Strategy								Metric (Society)	
			$p_0p_0p_0$	$p_0p_0p_1$	$p_0p_1p_0$	$p_0p_1p_1$	$p_1p_0p_0$	$p_1p_0p_1$	$p_1p_1p_0$	$p_1p_1p_1$	<u>Cost</u> ↓	<u>W-T</u> ↑
MMLU	Acc ↑	S_1	64.4±1.7	66.4±2.2	58.0±3.7	55.2±4.4	37.6±7.0	42.4±7.1	50.4±4.3	44.8±2.7	5050	5
		S_2	67.2±4.1	67.6±7.1	53.2±6.4	53.2±5.0	38.4±5.5	40.4±5.2	53.6±4.8	45.2±3.6	5076	2
		S_3	62.0±6.2	67.6±3.8	52.0±6.8	57.2±6.4	42.4±5.2	37.6±5.5	55.2±6.6	40.0±6.2	5073	8
		S_4	64.8±4.4	64.8±5.8	58.4±3.0	51.6±3.8	38.0±3.7	42.0±2.4	54.0±5.8	41.2±5.2	5080	5
	<u>Cost</u> ↓	All	7528	5957	5402	4374	5812	4215	4272	3001		
	<u>W-T</u> ↑	All	-	14	2	3	0	0	1	0		
MATH	Acc ↑	S_1	46.8±8.1	46.0±8.1	44.0±5.3	44.4±5.2	50.0±5.8	49.2±8.1	42.0±3.2	42.0±4.0	5816	17
		S_2	47.2±6.4	54.0±2.4	48.4±3.8	43.6±4.3	48.0±4.2	44.4±7.9	50.8±3.6	38.8±9.1	5844	22
		S_3	50.8±4.8	42.8±6.6	45.6±6.8	45.2±4.4	49.2±4.8	46.4±5.5	45.2±8.4	43.6±2.6	5837	9
		S_4	50.8±5.4	45.2±7.0	48.8±9.4	44.8±3.3	49.2±8.7	51.2±2.3	48.4±6.5	40.8±6.1	5834	18
	<u>Cost</u> ↓	All	6919	6302	6221	5667	6149	5645	5924	4807		
	<u>W-T</u> ↑	All	-	10	10	9	13	10	10	4		
Chess Move Validity	Acc ↑	S_1	47.2±3.6	47.6±5.2	45.6±7.8	40.0±4.5	42.8±2.3	29.2±4.6	42.4±6.5	20.0±6.0	2927	10
		S_2	48.4±5.0	45.6±6.1	43.6±4.3	39.6±3.3	48.4±5.2	35.6±5.2	43.2±8.8	18.8±5.8	2930	6
		S_3	49.6±5.5	48.0±5.8	47.6±5.5	37.6±9.9	41.6±6.1	35.2±8.3	40.4±3.8	14.8±6.1	2947	6
		S_4	48.4±3.3	49.6±4.6	46.0±3.5	36.8±4.1	38.8±3.3	27.2±3.9	38.0±6.3	14.0±4.7	2959	5
	<u>Cost</u> ↓	All	3736	3169	3196	2627	3266	2714	2698	2123		
	<u>W-T</u> ↑	All	-	11	6	1	5	0	4	0		

Table 8: The impact of 8 collaborative strategies on the performance of 3 datasets across distinct societies, using *ChatGPT* (with engine of *gpt-3.5-turbo* employed between July 10 and July 23, 2023). Blue marks the best-performing strategy under the same society, light blue represents the second-best-performing strategy, and red indicates the worst-performing strategy. Cost / Cost measures the average tokens consumed by all cases under the same collaborative strategy / society. W-T / W-T tallies the total number of occurrences where performance exceeds the strategy $p_0p_0p_0$ under the same collaborative strategy / society. The significances test on societies and strategies are respectively shown in Table 9, 10.

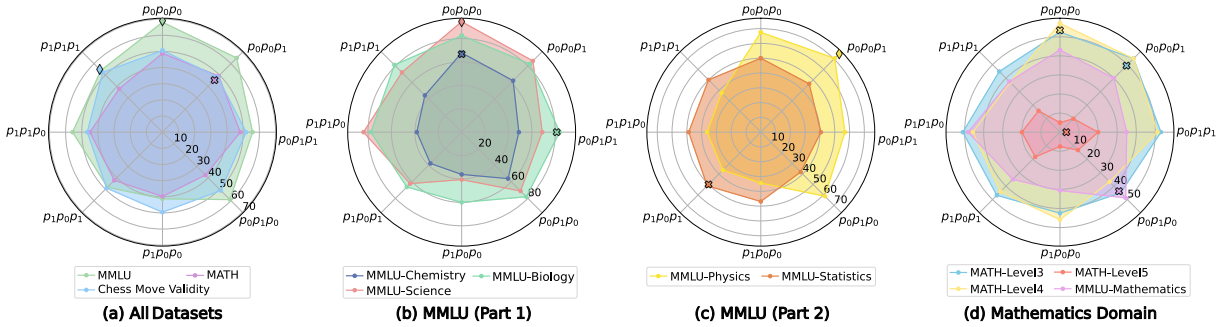


Figure 13: Illustration of different collaborative strategies impacting accuracy diversely on the tasks considering varied subjects and difficulty, using *ChatGPT*. The symbol ‘X’ represents that there is at least one collaborative strategy whose accuracy is better than self-consistency, while the symbol ‘D’ indicates that there is no collaborative strategy whose accuracy is worse than self-consistency. Both of these symbols represent the accuracy of self-consistency. The accuracy under each collaborative strategy is a summation within all 3-agent societies.

with a round of reflection at the same juncture will increase consensus clusters. Adding an extra round of debate at this juncture, as per the conclusions in §4, is not anticipated to bring about a discernible enhancement in performance. This confirms the efficacy of the *early-stopping mechanism* implemented in Liu et al. (2023c), drawing inspiration from Byzantine Consensus theory (Castro and Liskov, 1999).

Moreover, as seen from Figure 7, we scrutinize the consensus reaching of these strategies in three rounds where p -values are below 0.05. Combining

the insights from Figure 7 and Figure 18, 4, 19, it becomes apparent that these collaborative strategies exhibit substantial fluctuations in consensus reaching, at times demonstrating periods of notably low answer consistency. For the collaborative strategy $p_0p_0p_0p_0$ in Chess Move Validity, although continual reflection results in a gradual increase in the quantity of consensus clusters, a more stable trend with smaller fluctuations renders it less sensitive to the number of rounds. Conversely, collaborative strategies with p -values > 0.05 often display higher levels of answer consistency.

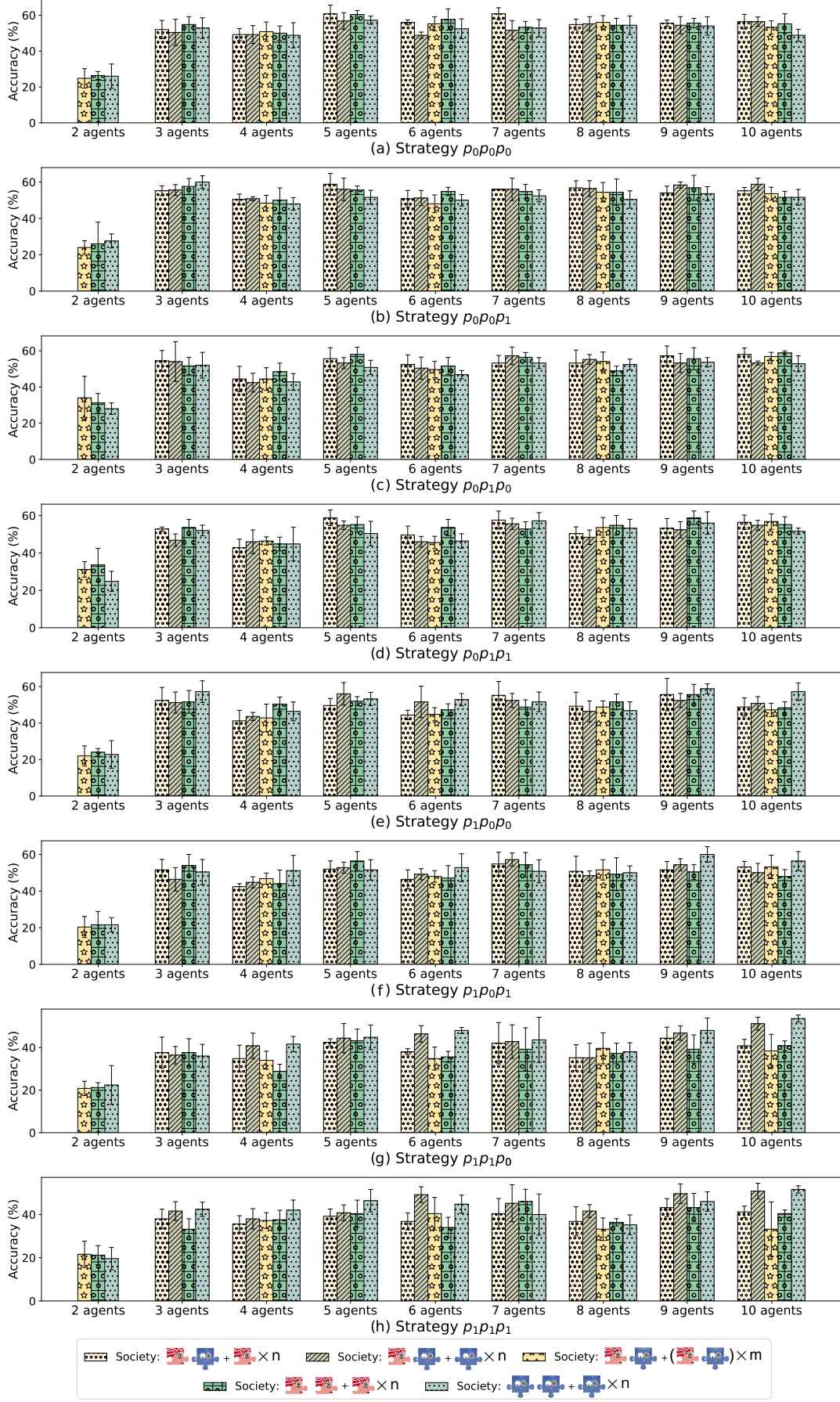


Figure 14: Accuracy of *different societies* with 2~10 agents under 3-round collaborative strategies, on *ChatGPT*.

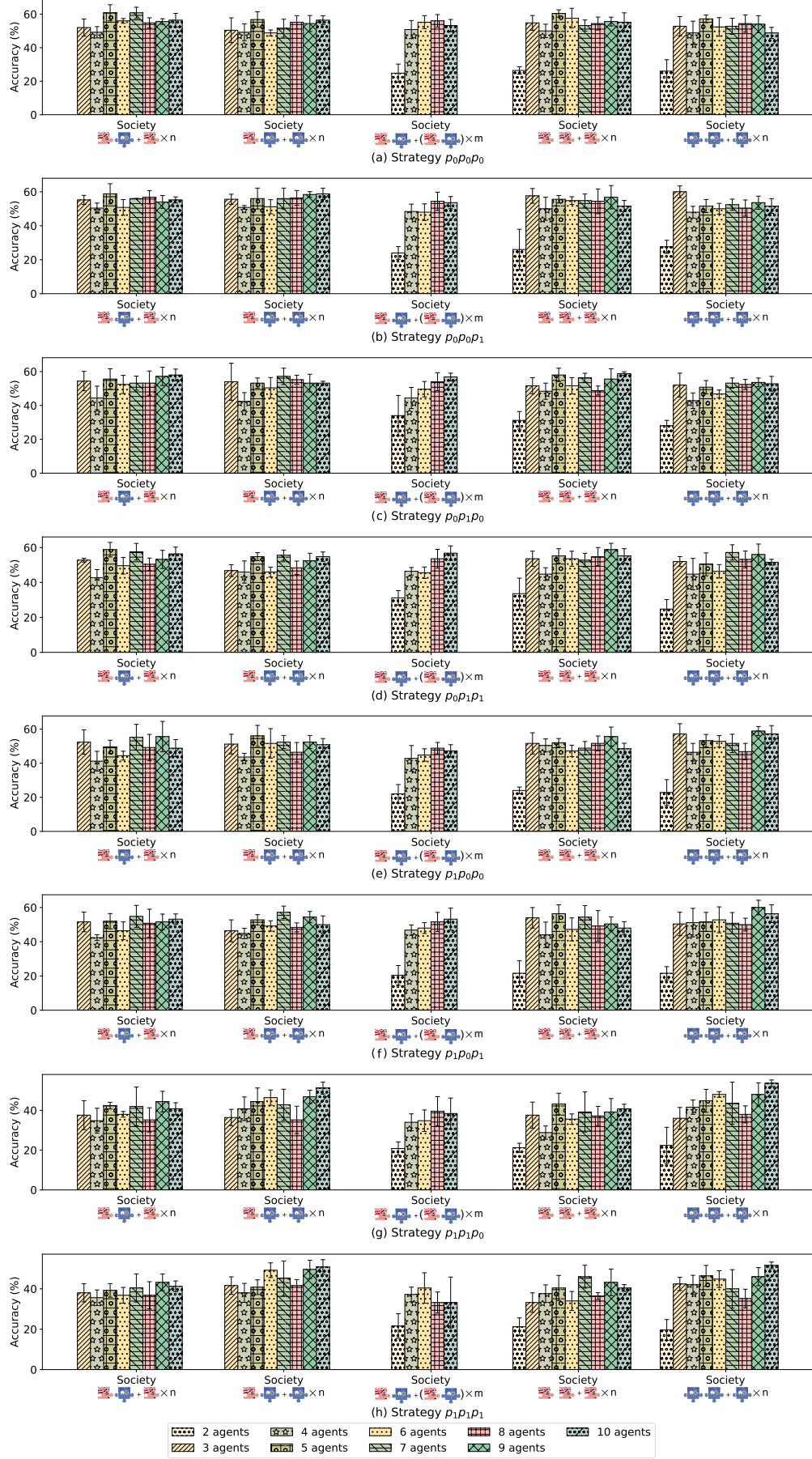


Figure 15: Accuracy of different numbers (2~10) of agents under 3-round collaborative strategies, using *ChatGPT*. The significance test is shown in Table 11.

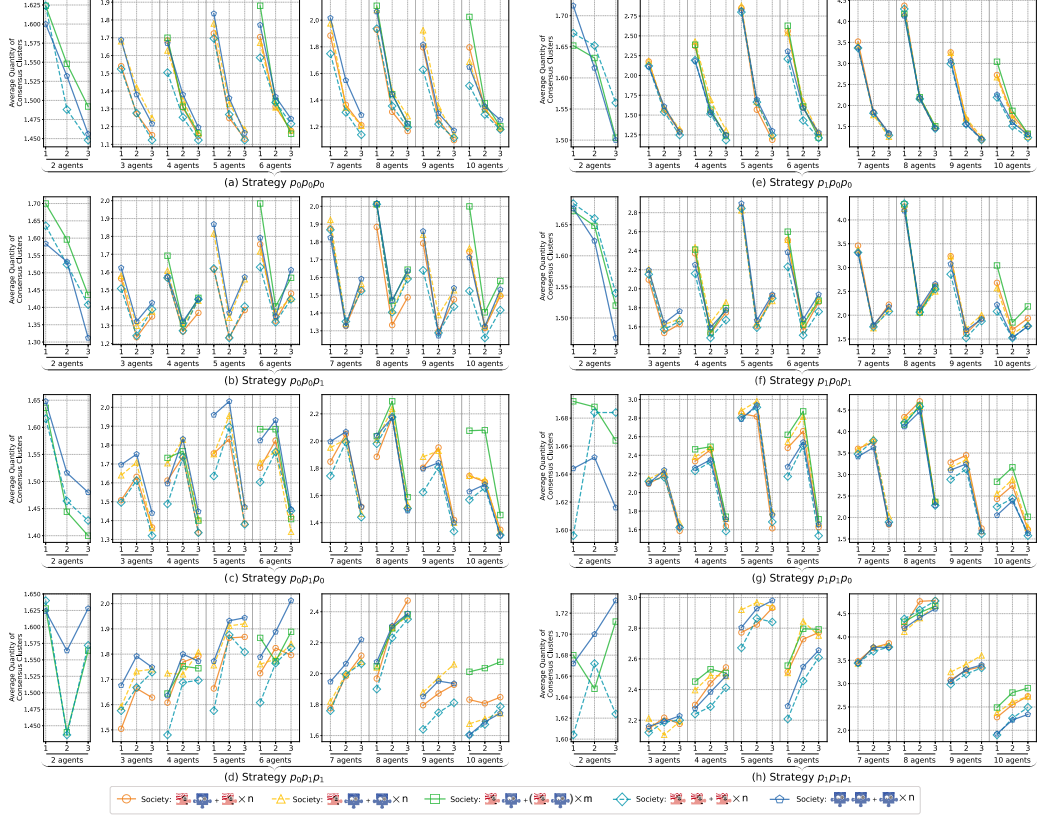


Figure 16: Average quantity of consensus clusters (unique answers among multiple agents) in different societies with 2~10 agents under each round of 3-round collaborative strategies, using ChatGPT.

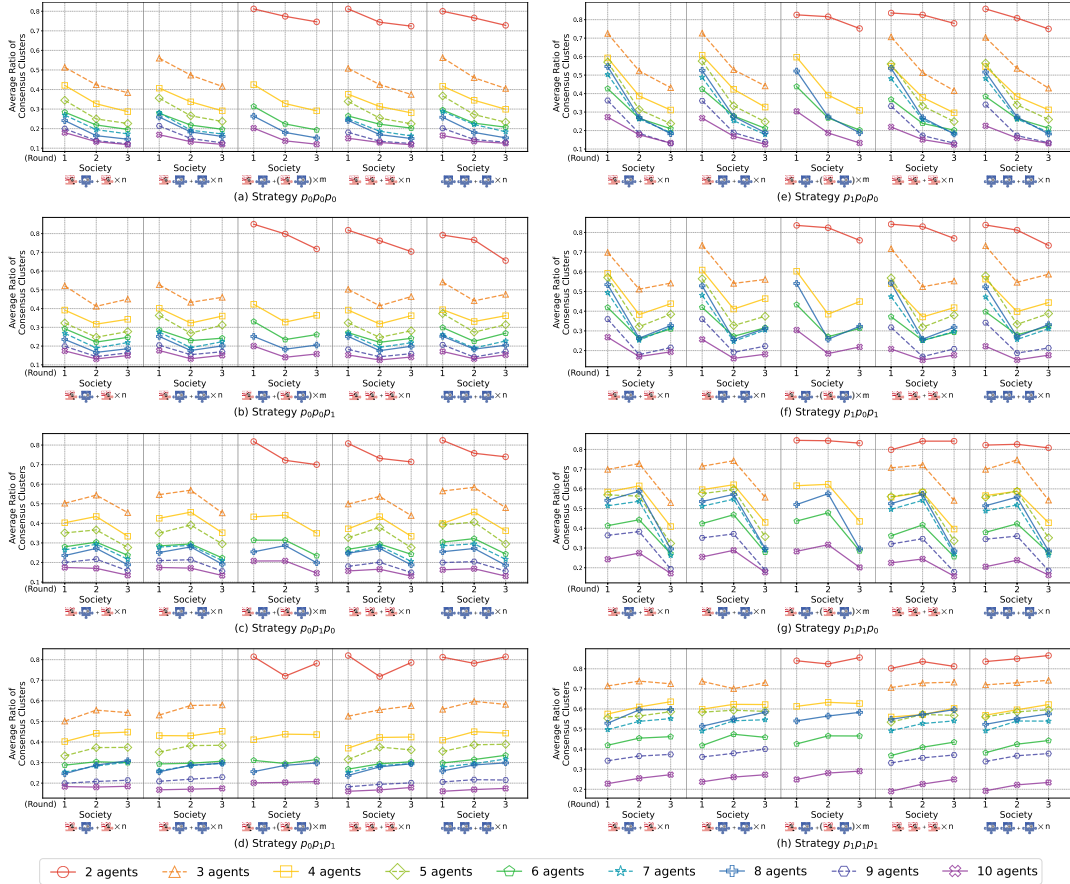


Figure 17: Average ratio of consensus clusters (unique answers among multiple agents) with different numbers (2~10) of agents under each round of 3-round collaborative strategies, using ChatGPT.

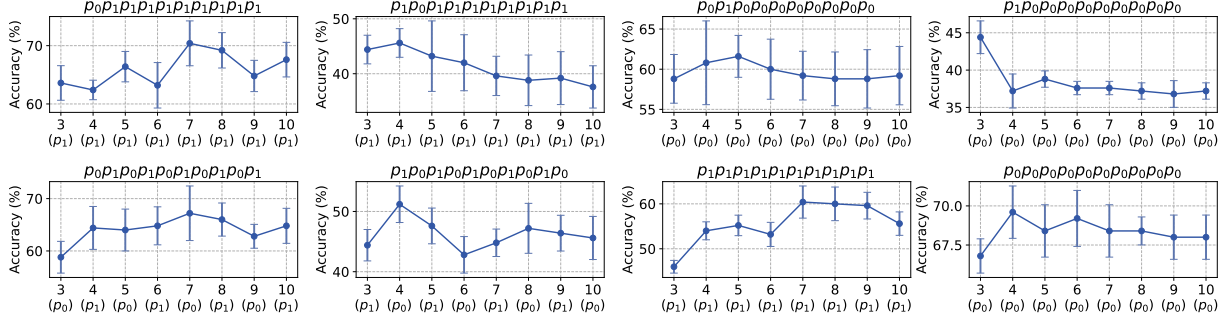


Figure 18: Accuracy of *different (3~10) rounds of collaboration* within 3-agent society S_2 (1 easy-going and 2 overconfident agents) on MMLU, using *ChatGPT*.

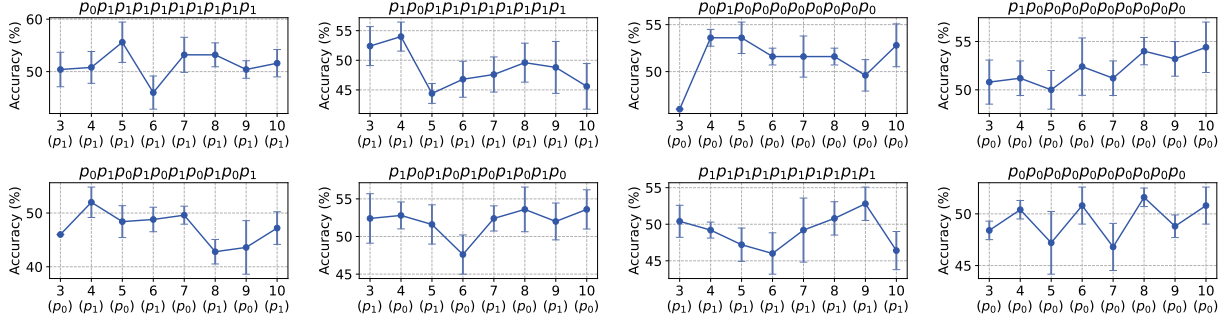


Figure 19: Accuracy of *different (3~10) rounds of collaboration* within 3-agent society S_2 (1 easy-going and 2 overconfident agents) on Chess Move Validity, using *ChatGPT*.

Collaborative Strategy	MMLU p-value	MATH p-value	Chess Move Validity p-value
$p_0p_0p_0$	0.350	0.618	0.866
$p_0p_0p_1$	0.797	0.069	0.716
$p_0p_1p_0$	0.162	0.631	0.726
$p_0p_1p_1$	0.350	0.945	0.807
$p_1p_0p_0$	0.501	0.964	0.025
$p_1p_0p_1$	0.497	0.378	0.079
$p_1p_1p_0$	0.562	0.135	0.614
$p_1p_1p_1$	0.236	0.642	0.293

Table 9: One-Way ANOVA results for the impact of society on accuracy with fixed collaborative strategy, based on experiments from Table 8 using *ChatGPT* in *July*.

Society	MMLU p-value	MATH p-value	Chess Move Validity p-value
S'_1	0.000	0.346	0.000
S'_2	0.000	0.008	0.000
S'_3	0.000	0.388	0.000
S'_4	0.000	0.213	0.000

Table 10: One-Way ANOVA results for the impact of collaborative strategy on accuracy with fixed society, based on experiments in Table 8 on *ChatGPT* in *July*.

Collaborative Strategy	S'_1 p-value	S'_2 p-value	S'_3 p-value	S'_4 p-value	S'_5 p-value
$p_0p_0p_0$	0.000	0.000	0.000	0.000	0.000
$p_0p_0p_1$	0.000	0.000	0.000	0.000	0.000
$p_0p_1p_0$	0.002	0.015	0.006	0.000	0.000
$p_0p_1p_1$	0.000	0.000	0.000	0.000	0.000
$p_1p_0p_0$	0.000	0.000	0.000	0.000	0.000
$p_1p_0p_1$	0.000	-	0.000	0.001	0.000
$p_1p_1p_0$	0.000	0.000	0.000	0.000	0.000
$p_1p_1p_1$	0.000	0.000	0.000	0.005	0.000

Table 11: One-way ANOVA analysis of results in Figure 15 (different numbers of agents), using *ChatGPT*. S'_1 : One overconfident agent and the others are all easygoing. S'_2 : One easygoing agent among predominantly overconfident agents. S'_3 : Equal numbers of overconfident and easygoing agents. S'_4 : Entirely easygoing agents. S'_5 : Entirely overconfident agents. '-': It doesn't pass homogeneity test for variance.

Collaborative Strategy	MMLU p-value	MATH p-value	Chess Move Validity p-value
$p_0p_0p_0p_0p_0p_0p_0p_0p_0$	0.030	0.323	0.000
$p_1p_0p_0p_0p_0p_0p_0p_0p_0$	0.000	0.070	0.161
$p_0p_1p_0p_0p_0p_0p_0p_0p_0$	0.101	0.332	0.000
$p_1p_0p_1p_0p_1p_0p_1p_0p_1p_0$	0.000	0.077	0.871
$p_0p_1p_0p_1p_0p_1p_0p_1p_0p_1$	0.051	0.062	0.000
$p_1p_0p_1p_1p_1p_1p_1p_1p_1p_1$	0.000	0.021	0.630
$p_0p_1p_1p_1p_1p_1p_1p_1p_1p_1$	0.431	0.176	0.063
$p_1p_1p_1p_1p_1p_1p_1p_1p_1p_1$	0.000	0.000	0.027

Table 12: One-way ANOVA analysis of the results in Figure 4, 18, 19 (different rounds), using *ChatGPT*.

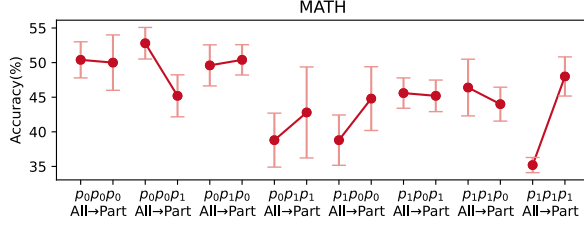


Figure 20: The effect on accuracy of whether all agents in a society execute the same thinking pattern in one round on MATH, using *ChatGPT*. “All” and “Part” respectively refer to all agents applying the same and different thinking pattern(s) in one round. The significance test is shown in Table 13 at Appendix F.

Collaborative Strategy	MMLU p-value	MATH p-value	Chess Move Validity p-value
$p_0p_0p_0$	0.402	0.856	0.147
$p_0p_0p_1$	0.007	0.002	0.001
$p_0p_1p_0$	0.550	0.641	0.002
$p_0p_1p_1$	-	0.276	0.000
$p_1p_0p_0$	-	0.051	-
$p_1p_0p_1$	-	0.784	0.000
$p_1p_1p_0$	0.014	0.294	0.172
$p_1p_1p_1$	1.000	0.000	0.347

Table 13: One-way ANOVA analysis of the results of Figure 5 (other collaborative strategies), using *ChatGPT*. ‘-’: It doesn’t pass homogeneity test for variance.

Other Collaborative Strategies. We also present the results of all agents in a society execute the same or inconsistent thinking pattern(s) at one round in Figure 20. According to Table 13, we observe a pronounced impact of maintaining a consistent thinking pattern on Chess Move Validity, while its influence on MMLU and MATH is less significant. We attribute this difference to the limited assistance that collaborative strategy offers for MMLU and MATH, as evidenced in the results observed in §G.2 based on Figure 21(a).

G A Social Psychology View on Conformity, Consensus Reaching, and Group Dynamics

G.1 Conformity and Consensus Reaching

Figures 6, 28, 37, 65, and 51 illustrate the conformity. Figures 7, 28, 37, 65, and 51 illustrate the consensus. This section provides a detailed explanation of the methodologies used to calculate both conformity and consensus.

For conformity, we solely focus on agents actively engaging in debate, disregarding those in reflection during a given round. Let the answer of the

i -th agent at j -th round be denoted as $a_{i,j}$. For the k -th agent at j -th round, if “Frequency($\{a_{i,j-1} | i \in [1, n]\}$) = $a_{k,j}$ ”, we identify this as the occurrence of conformity by agent k at j -th round, where Frequency($\cdot$) represents the most frequently given answer (excluding instances where all answers occur only once, as such cases are considered as non-conformity). Additionally, we categorize the correctness of answers both before and after conformity into four cases, with ‘True’ denoting correct and ‘False’ denoting incorrect.

For consensus, we examine the evolution of the quantity of distinct answers (*i.e.*, consensus clusters) with increasing rounds of collaboration. Let the answer of the i -th agent at time j be denoted as $a_{i,j}$. For the j -th round, consensus clusters is defined as $\|\text{Set}(\{a_{i,j} | i \in [1, n]\})\|$, where $\|\text{Set}(\cdot)\|$ represents the count of different answers. This computational approach has been utilized in the analysis presented in Figures 17, 16, 60, 59, 46, 45.

G.2 A Social Psychology View on Group Dynamics

We seek to elucidate how performance impacted by group dynamics, *i.e.*, the patterns of interaction between group members and different processes that may occur within a social group. Diving into the intricacies of collaboration, each agent generates four answers, including the initial answer without collaboration, as shown in Figure 2(d). To determine the answer for each round, we employ the majority vote (Cobbe et al., 2021; Li et al., 2022). Given ‘ T ’ and ‘ F ’ respectively denoting a round that yields a correct and an incorrect answer, we could obtain $2^4=16$ possible answer sequences over the four rounds. We select 10 sequences⁶ of them and categorize them into 3 groups: **Correcting Mistakes** ($FFFT, FFFT, FTFT$), **Changing Correct Answers** ($TFFF, TTFF, TTF$), and **Wavering Answers** ($FTFT, FTTF, FTTF, TFFT$). Particularly, **Wavering Answers** resemble model hallucination (Rawte et al., 2023; Zhang et al., 2023c; Ji et al., 2023; Luo et al., 2024) due to the occurrence of self-contradictory answers. Our categorization is under society-agnostic collaborative strategies, considering the performance variance between societies is negligible. From the results on ChatGPT shown in Figure 21, and on other LLMs shown in

⁶The selected 10 sequences adhere to patterns: (1) $[F]_{i>0}[T]_{j>0}$, *e.g.*, $FFFT$; (2) $[T]_{i>0}[F]_{j>0}$, *e.g.*, $TFFF$; (3) $[TF]_{i\geq 0}[FT]_{j\geq 0}$, *e.g.*, $FTFT$, where $[\cdot]_i, [\cdot]_j$ respectively denotes repetition for i, j times.

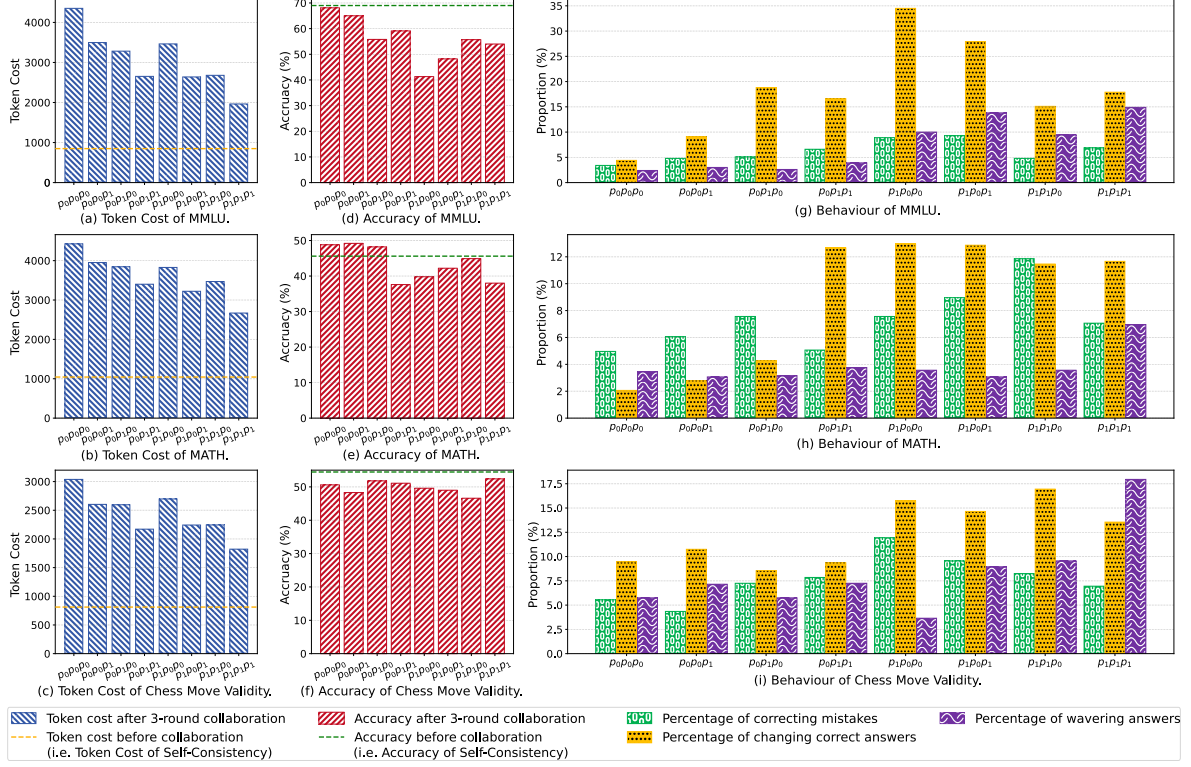


Figure 21: The percentage of different behaviors under different collaborative strategies, using *ChatGPT*. Figure (a-c) & (d-f) respectively show the token cost and accuracy of different strategies before and after 3-round collaboration. Figure (g-i) present the percentage of different behavioral features (mainly analyzed by the change of answer correctness) (Zhang et al., 2023b,a; Xie et al., 2024) under different collaborative strategies. All results are summarized across all societies. The results on other LLMs are shown in Figure 30, 39, 53, 67 at Appendix H.

Appendix H, we summarize the following findings:

(1) **Debate-initial/dominant collaborative strategies are generally effective.** As seen from the red bars in Figure 21 30, 39, 53, 67(d-f), we find that the collaborative strategies starting from or dominant with debate p_0 are more effective than other, and mostly outperform self-consistency, even though they cost more tokens (seen from blue bars).

(2) **Reflection experiences greater instability (a heightened risk of model hallucination).** As observed from the purple bars in Figure 21 30, 39, 53, 67(g-h), comparing $p_i p_j p_0$ & $p_i p_j p_1$; $p_i p_0 p_j$ & $p_i p_1 p_j$, $p_i p_j p_0$ and $p_i p_0 p_j$ are more likely to wavering answers than $p_i p_j p_1$ and $p_i p_1 p_j$, demonstrating that reflection is more likely to cause model hallucination than debate.

H Analysis on Different Backbone LLMs

To make the findings in this paper more general, we also implement all the experiments with some other open-resource backbone LLMs, such as **LlaMA2 Chat 13B** (Touvron et al., 2023), **LlaMA2 Chat 70B** (Touvron et al., 2023), **Qwen 72B** (Bai et al., 2023) and **Mixtral 8×7B** (Jiang et al., 2023, 2024).

H.1 LlaMA2 Chat 13B

Analysis on Machine Social Collaboration. We present the **main results** and **significance tests** of societies and strategies on LlaMA2 Chat 13B in Table 14, 15, 16. We present the **word clouds** of LlaMA2 Chat 13B in Figure 22, and **proportion of agents with different traits changing answers in different societies** on LlaMA2 Chat 13B in Figure 23. Furthermore, we demonstrate that the tasks with different subjects and difficulty display varying sensitivity to collaborative strategies, as presented with **radar maps** on LlaMA2 Chat 13B in Figure 24.

Analysis on Different Numbers of Agents. We present the significance test for different numbers of agents with LlaMA2 Chat 13B in Table 17. We also show the performance varying from agent numbers in Figure 25.

Analysis on Different Rounds. We present the significance test for different rounds of collaboration with LlaMA2 Chat 13B in Table 18. We also show the performance varying from collaboration rounds in Figure 26.

Analysis on Other Collaborative Strategies. We present the significance test for other collaborative strategies (executing the same or hybrid thinking patterns in a certain round) with LlaMA2 Chat 13B in Table 19. We also show the performance varying from other strategies in Figure 27.

A Social Psychology View on Conformity, Consensus Reaching and Group Dynamics. We then show the variation of answer correctness in the situation of conformity in Figure 28; and the quantity of consensus clusters among 3-agent answers in Figure 29. We present group dynamics reflected by different answer changing behaviors on LlaMA2 Chat 13B in Figure 30.

	Metric (Strategy)	Society	Collaborative Strategy								Metric (Society)	
			$p_0p_0p_0$	$p_0p_0p_1$	$p_0p_1p_0$	$p_0p_1p_1$	$p_1p_0p_0$	$p_1p_0p_1$	$p_1p_1p_0$	$p_1p_1p_1$	<u>Cost</u> ↓	<u>W-T</u> ↑
MMLU	Acc ↑	S_1	37.2±5.9	47.2±3.9	48.4±3.9	46.0±5.7	47.2±2.3	46.8±2.7	45.2±4.4	46.8±3.0	7447	35
		S_2	38.4±4.6	42.8±3.9	43.6±3.6	45.2±3.6	44.8±4.6	47.2±3.9	44.4±6.2	42.8±3.4	7413	33
		S_3	36.0±3.7	44.8±3.0	44.8±4.8	46.4±1.7	41.6±4.3	46.4±2.2	43.2±6.6	42.4±3.3	7370	33
		S_4	34.8±2.7	42.4±5.0	42.0±4.5	44.0±2.8	40.4±3.0	43.6±3.9	40.8±3.0	41.6±2.6	7423	35
	<u>Cost</u> ↓	All	11429	9476	8166	6419	8452	5734	5733	3900	-	
MATH	<u>W-T</u> ↑	All	-	20	20	20	18	20	19	19	-	
MATH	Acc ↑	S_1	5.2±2.3	6.8±2.3	5.6±2.6	5.6±2.6	4.8±3.0	4.4±1.7	5.6±3.9	3.2±1.1	8639	24
		S_2	5.2±3.6	5.2±3.4	6.0±2.0	6.8±1.8	6.0±0.0	6.8±1.8	6.8±1.1	4.8±1.1	8451	22
		S_3	6.8±1.8	6.8±3.0	6.8±3.4	6.0±2.8	5.2±1.8	5.2±1.8	6.0±3.7	3.6±1.7	8501	16
		S_4	4.8±2.3	6.8±3.4	7.2±1.1	5.6±2.2	5.6±1.7	5.2±2.3	5.2±3.6	4.0±1.4	8475	28
	<u>Cost</u> ↓	All	10655	9508	9501	7900	9319	7761	7800	5687	-	
	<u>W-T</u> ↑	All	-	15	16	13	13	11	13	9	-	
Chess Move Validity	Acc ↑	S_1	16.4±3.0	7.2±3.0	9.2±2.3	2.8±1.8	8.8±3.0	4.8±2.3	9.2±4.4	2.0±2.8	3754	2
		S_2	11.6±5.2	8.0±1.4	10.8±4.2	2.8±1.8	11.6±2.6	6.0±3.2	10.8±5.0	3.6±2.6	3725	10
		S_3	14.8±3.0	8.4±4.8	10.0±4.2	5.2±1.1	14.0±4.5	6.8±3.0	9.6±6.2	2.8±3.0	3678	5
		S_4	16.0±4.2	6.8±2.7	12.4±6.2	4.0±2.5	10.0±4.2	7.2±6.7	10.0±3.2	4.0±2.5	3647	4
	<u>Cost</u> ↓	All	4889	4123	4061	3324	4045	3293	3292	2581	-	
	<u>W-T</u> ↑	All	-	2	4	0	7	1	7	0	-	

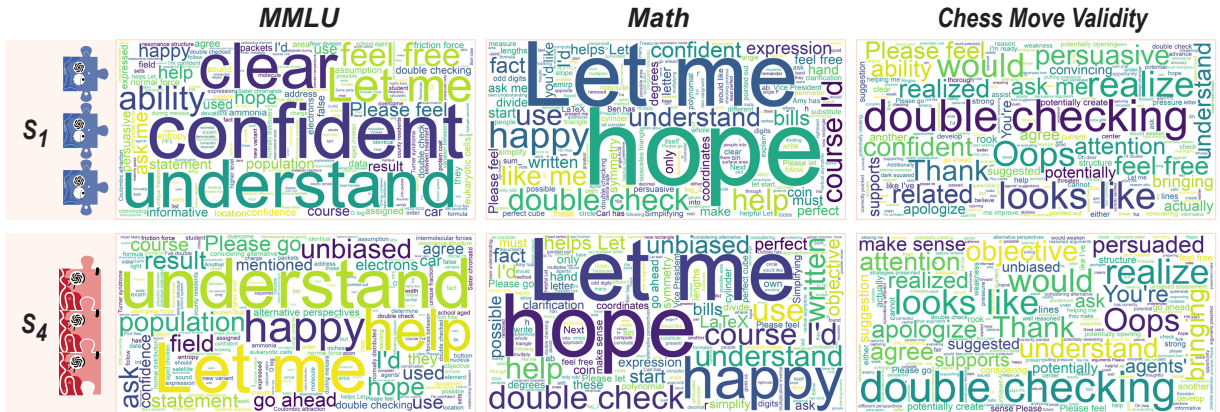


Figure 22: Comparative word clouds on three datasets in societies S_1 and S_4 , using *LlaMA2-13B-chat*. Society S_1 features three overconfident agents, while society S_4 comprises three easy-going agents.

Collaborative Strategy	MMLU p-value	MATH p-value	Chess Move Validity p-value
$p_0p_0p_0$	0.611	0.632	0.251
$p_0p_0p_1$	0.252	0.791	0.854
$p_0p_1p_0$	0.142	0.714	0.706
$p_0p_1p_1$	0.755	0.839	0.164
$p_1p_0p_0$	0.039	0.789	0.175
$p_1p_0p_1$	0.318	0.277	0.809
$p_1p_1p_0$	0.585	0.884	0.959
$p_1p_1p_1$	0.071	0.310	0.672

Table 15: One-Way ANOVA results for the impact of society on accuracy with fixed collaborative strategy, based on experiments from Table 14 using *LlaMA2-chat-13B*.

Society	MMLU p-value	MATH p-value	Chess Move Validity p-value
S_1	0.006	0.548	0.000
S_2	0.129	0.664	0.000
S_3	0.005	0.518	0.000
S_4	0.009	0.490	0.001

Table 16: One-Way ANOVA results for the impact of collaborative strategy on accuracy with fixed society, based on experiments from Table 14 using *LlaMA-13B-Chat*.

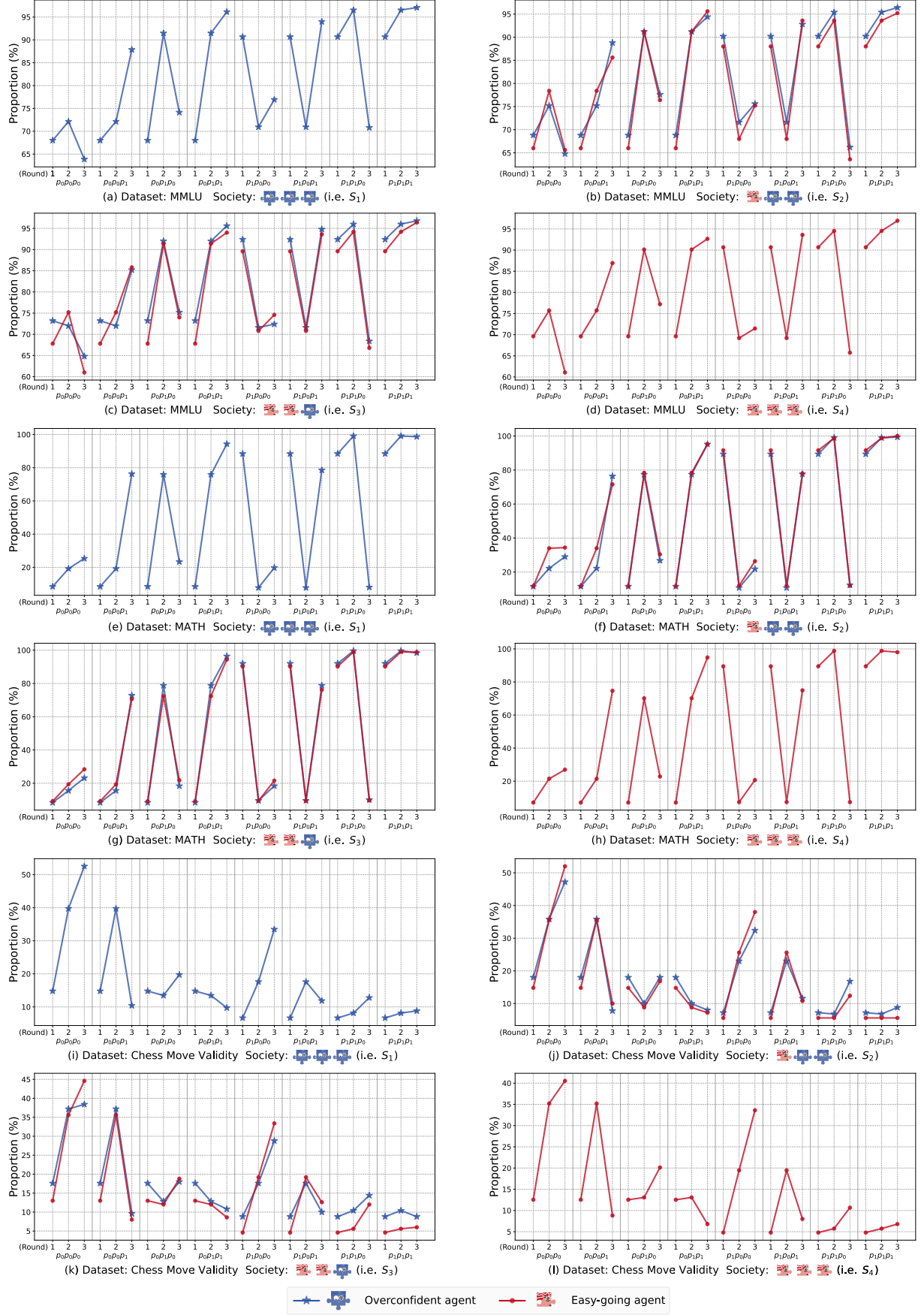


Figure 23: Proportion of agents with different traits changing answers in societies S_1 and S_4 , using *LLaMA2-13B-chat*. Society S_1 features three overconfident agents, while society S_4 comprises three easy-going agents.

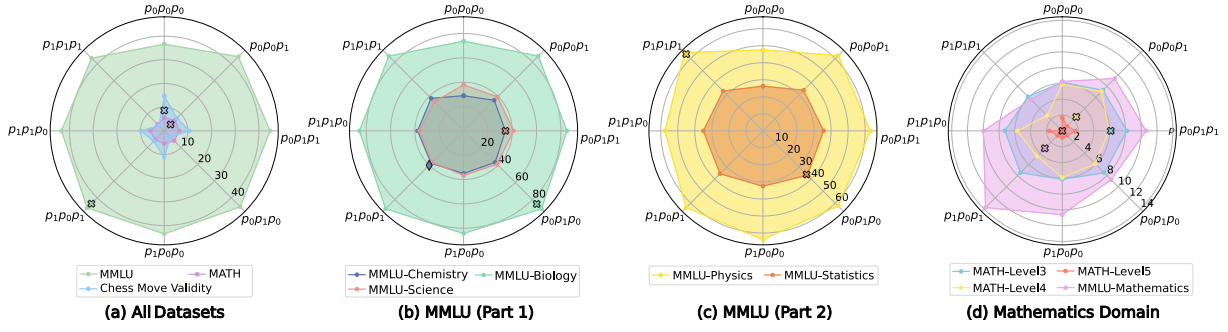


Figure 24: Illustration of different collaborative strategies impacting accuracy diversely on the tasks considering varied *subjects* and *difficulty*, using *LlaMA2-13B-chat*. The symbol ‘⋈’ represents that there is at least one collaborative strategy whose accuracy is better than self-consistency, while the symbol ‘⋄’ indicates that there is no collaborative strategy whose accuracy is worse than self-consistency. Both of these symbols represent the accuracy of self-consistency. The accuracy under each collaborative strategy is a summation within all 3-agent societies.

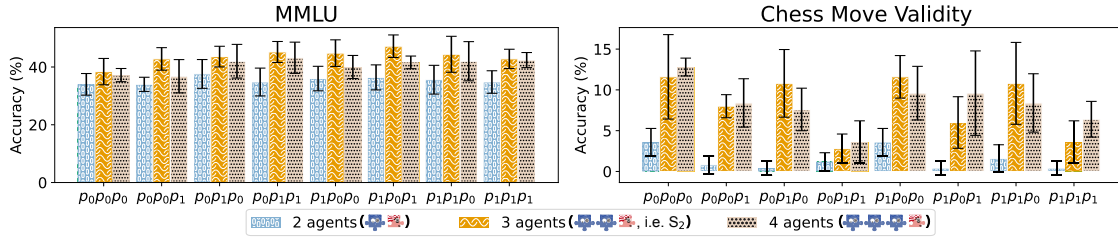


Figure 25: Accuracy of different number of agents under different collaborative strategies, on *LlaMA2-13B-chat*. The significance test is shown in Table 17.

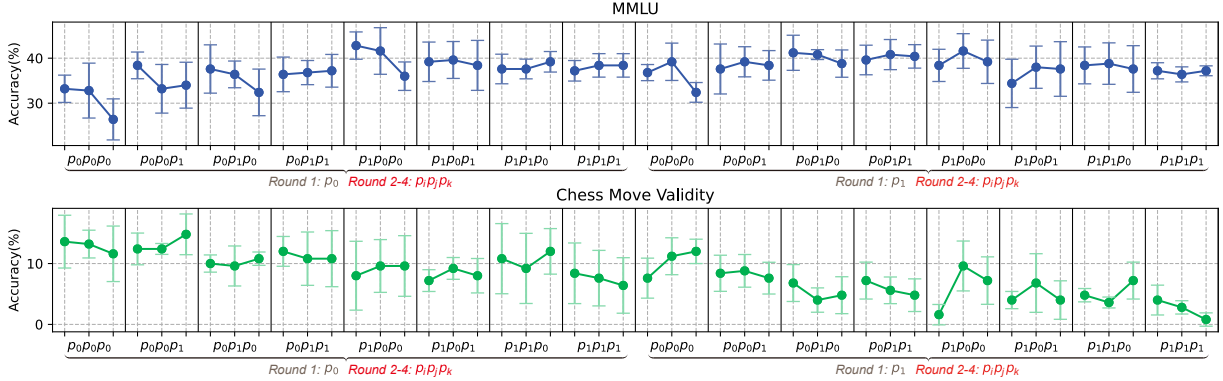


Figure 26: Accuracy at round 2,3,4 within 4-round collaborative societies, where the thinking pattern of round 1 is fixed (p_0 or p_1), using *LlaMA2-13B-chat*. The significance test is shown in Table 18.

Collaborative Strategy	MMLU p-value	Chess Move Validity p-value
$p_0p_0p_0$	0.186	0.001
$p_0p_0p_1$	0.019	0.000
$p_0p_1p_0$	0.175	0.000
$p_0p_1p_1$	0.010	0.178
$p_1p_0p_0$	0.023	0.001
$p_1p_0p_1$	0.002	0.005
$p_1p_1p_0$	0.098	0.005
$p_1p_1p_1$	0.004	0.002

Table 17: One-way ANOVA analysis of the results in Figure 25 (different numbers of agents), using *LlaMA2-chat-13B*.

Collaborative Strategy	MMLU p-value	Chess Move Validity p-value
$p_0p_0p_0p_0$	0.000	0.361
$p_0p_0p_0p_1$	0.111	0.598
$p_0p_0p_1p_0$	0.082	0.335
$p_0p_0p_1p_1$	0.529	0.076
$p_0p_1p_0p_0$	0.293	0.176
$p_0p_1p_0p_1$	0.641	0.259
$p_0p_1p_1p_0$	0.536	0.026
$p_0p_1p_1p_1$	0.812	0.052
$p_1p_0p_0p_0$	0.010	0.629
$p_1p_0p_0p_1$	0.547	0.029
$p_1p_0p_1p_0$	0.749	0.055
$p_1p_0p_1p_1$	0.600	0.007
$p_1p_1p_0p_0$	0.605	0.009
$p_1p_1p_0p_1$	0.988	0.012
$p_1p_1p_1p_0$	0.889	0.097
$p_1p_1p_1p_1$	0.742	0.884

Table 18: One-way ANOVA analysis of the results in Figure 26 (different rounds), using *LlaMA2-chat-13B*.

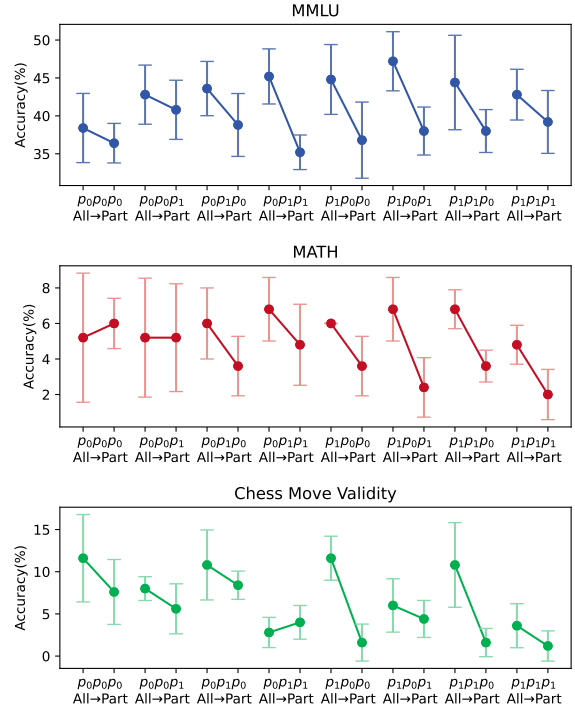


Figure 27: The effect on accuracy of whether all agents in society execute the same thinking pattern in one round, using *LlaMA2-13B-chat*. “All” and “Part” refer to all agents applying the same thinking pattern and different thinking patterns in one round respectively. The significance test is shown in Table 19.

Collaborative Strategy	MMLU p-value	MATH p-value	Chess Move Validity p-value
$p_0p_0p_0p_0$	0.419	0.659	0.203
$p_0p_0p_0p_1$	0.441	1.000	0.141
$p_0p_0p_1p_0$	0.086	0.074	0.264
$p_0p_0p_1p_1$	0.001	0.161	0.347
$p_1p_0p_0p_0$	0.030	-	0.000
$p_1p_0p_0p_1$	0.003	0.004	0.380
$p_1p_1p_0p_0$	0.070	0.001	0.005
$p_1p_1p_1p_1$	0.169	0.008	0.128

Table 19: One-way ANOVA analysis of the results in Figure 27 (other collaborative strategies), using *LlaMA2-chat-13B*.

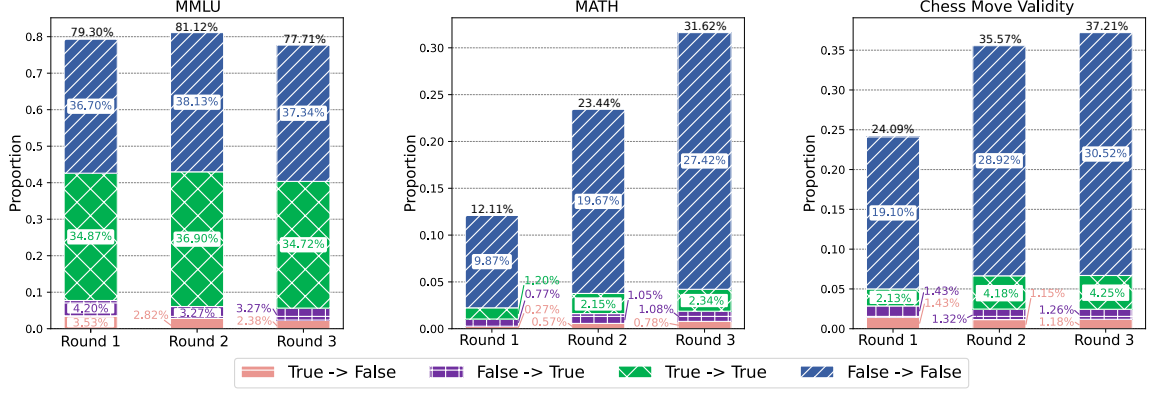


Figure 28: Variation of answer correctness in the situation of conformity, using *LlaMA2-13B-chat*, where *conformity brings about benefits*: $\text{Ratio}(\text{False} \rightarrow \text{True} + \text{True} \rightarrow \text{True}) > \text{Ratio}(\text{True} \rightarrow \text{False} + \text{False} \rightarrow \text{False})$; *conformity brings about detriments*: $\text{Ratio}(\text{False} \rightarrow \text{True} + \text{True} \rightarrow \text{True}) < \text{Ratio}(\text{True} \rightarrow \text{False} + \text{False} \rightarrow \text{False})$.

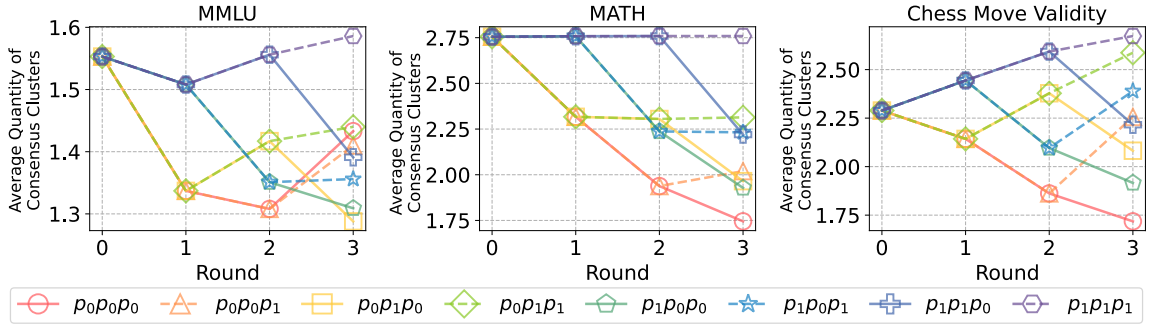


Figure 29: Average quantity of *consensus clusters* (i.e., unique answers among multiple agents) under different rounds of collaboration with 3-round collaborative strategies, on *LlaMA2-13B-chat*. Smaller quantity of consensus clusters, more easier it is to reach a consensus. Round 0 is equal to self-consistency.

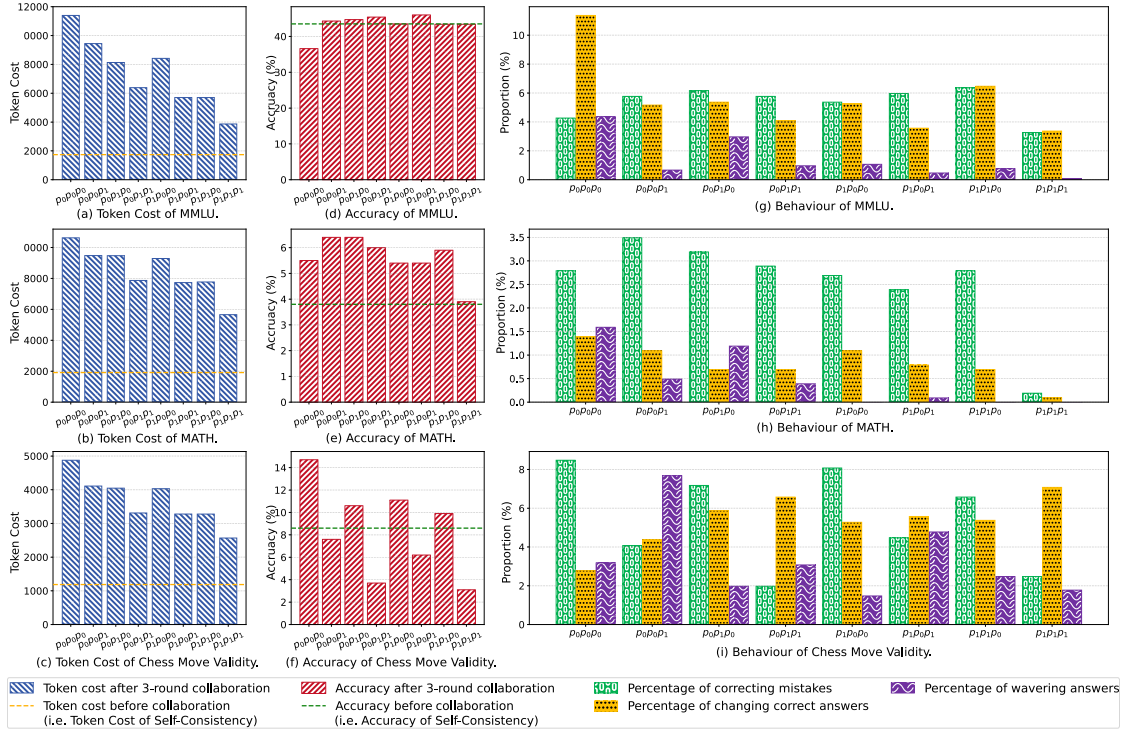


Figure 30: The percentage of different behaviors under different collaborative strategies, using *LlaMA2-13B-chat*. Figure (a-c) & (d-f) respectively show the token cost and accuracy of different strategies before and after 3-round collaboration. Figure (g-i) present the percentage of different behavioral features (mainly analyzed by the change of answer correctness) (Zhang et al., 2023b,a) under different collaborative strategies. All results are summarized across all societies.

H.2 LLaMA2 Chat 70B

Analysis on Machine Social Collaboration. We present the **main results** and **significance tests** of societies and strategies on LLaMA2 Chat 70B in Table 20, 21, 22. We present the **word clouds** of LLaMA2 Chat 70B in Figure 31, and **proportion of agents with different traits changing answers in different societies** on LLaMA2 Chat 70B in Figure 32. Furthermore, we demonstrate that the tasks with different subjects and difficulty display varying sensitivity to collaborative strategies, as presented with **radar maps** on LLaMA2 Chat 70B in Figure 33.

Analysis on Different Numbers of Agents. We present the significance test for different numbers of agents with LLaMA2 Chat 70B in Table 23. We also show the performance varying from agent numbers in Figure 34.

Analysis on Different Rounds. We present the significance test for different rounds of collaboration with LLaMA2 Chat 70B in Table 24. We also show the performance varying from collaboration rounds in Figure 35.

Analysis on Other Collaborative Strategies. We present the significance test for other collaborative strategies (executing the same or hybrid thinking patterns in a certain round) with LLaMA2 Chat 70B in Table 25. We also show the performance varying from other strategies in Figure 36.

A Social Psychology View on Conformity, Consensus Reaching and Group Dynamics. We then show the variation of answer correctness in the situation of conformity in Figure 37; and the quantity of consensus clusters among 3-agent answers in Figure 38. We present group dynamics reflected by different answer changing behaviors on LLaMA2 Chat 70B in Figure 39.

	Metric (Strategy)	Society	Collaborative Strategy								Metric (Society)	
			$p_0p_0p_0$	$p_0p_0p_1$	$p_0p_1p_0$	$p_0p_1p_1$	$p_1p_0p_0$	$p_1p_0p_1$	$p_1p_1p_0$	$p_1p_1p_1$	<u>Cost</u> ↓	<u>W-T</u> ↑
MMLU	Acc ↑	S_1	40.8±2.7	43.6±3.9	36.0±2.8	38.4±3.3	35.6±4.3	35.6±2.6	30.4±4.3	24.0±5.7	6915	7
		S_2	44.4±3.9	49.2±4.6	45.2±3.9	42.0±0.0	34.4±4.3	34.4±8.3	31.6±8.4	25.6±3.6	6946	11
		S_3	44.0±5.5	45.6±4.6	39.2±2.7	42.8±3.0	35.2±5.4	32.4±4.3	28.0±7.3	25.6±5.2	6931	8
		S_4	47.6±4.1	48.0±5.1	46.0±6.3	45.2±3.9	26.8±3.6	30.8±6.9	32.8±1.8	33.6±6.2	6936	8
	<u>Cost</u> ↓	All	10811	8608	7904	6177	7535	5410	5287	3722	-	-
	<u>W-T</u> ↑	All	-	16	5	11	1	0	1	0	-	-
MATH	Acc ↑	S_1	8.4±3.6	10.4±3.9	9.2±1.1	4.0±2.5	9.2±4.2	8.4±4.3	6.8±2.7	3.6±1.7	7000	16
		S_2	8.0±2.5	9.6±2.6	8.8±3.0	6.4±2.6	7.2±4.4	6.8±1.1	8.4±4.3	4.8±2.3	7013	19
		S_3	8.4±4.6	7.2±3.9	8.4±3.6	5.6±3.6	7.2±1.8	7.2±4.8	6.8±3.0	0.8±1.1	7157	15
		S_4	6.0±2.0	7.2±1.8	6.0±2.0	4.0±2.0	5.2±3.0	6.8±1.1	8.8±4.4	3.6±2.6	6934	23
	<u>Cost</u> ↓	All	9465	7850	7662	6294	7520	6302	6382	4734	-	-
	<u>W-T</u> ↑	All	-	14	14	5	13	9	14	4	-	-
Chess Move Validity	Acc ↑	S_1	20.4±6.2	16.8±3.6	17.2±4.2	8.4±2.2	21.2±5.8	10.8±3.0	10.4±1.7	4.8±3.0	3563	7
		S_2	18.4±4.8	9.6±3.6	13.2±1.1	5.6±2.2	14.4±3.9	7.2±3.0	13.2±3.4	4.0±2.8	3557	4
		S_3	18.4±6.5	11.2±3.0	12.0±5.8	8.0±2.0	20.8±4.6	8.4±4.3	12.8±2.7	2.8±3.4	3629	7
		S_4	15.2±4.2	11.6±2.2	15.2±2.3	10.4±1.7	18.0±4.7	8.0±4.7	10.8±2.7	5.2±2.3	3679	12
	<u>Cost</u> ↓	All	4778	3947	3830	3082	4139	3314	3259	2508	-	-
	<u>W-T</u> ↑	All	-	4	6	2	13	1	4	0	-	-

Table 20: The impact of eight different collaborative strategies on the performance of three datasets across distinct societies (using *LlaMA2-chat-70B*). The significances test on societies and strategies are respectively shown in Table 21, 22.

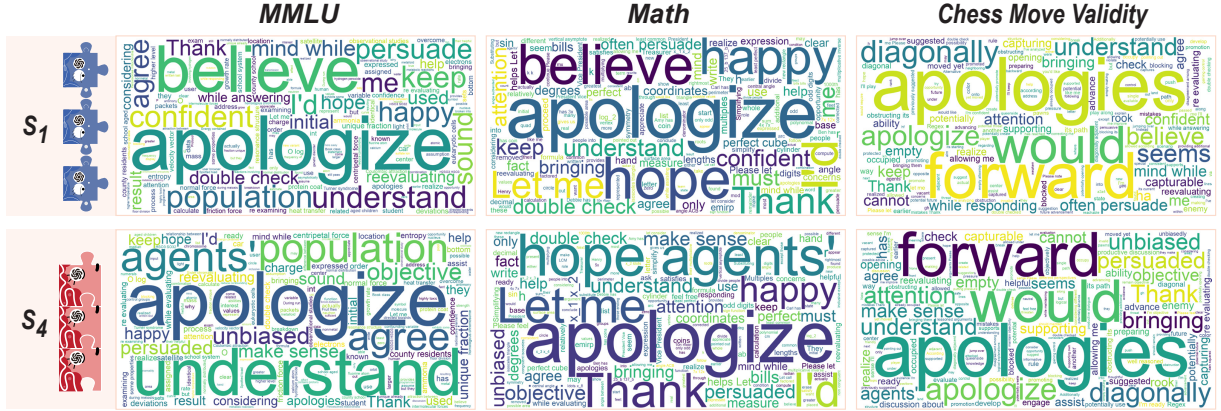


Figure 31: Comparative word clouds on three datasets in societies S_1 and S_4 , using *LlaMA2-70B-chat*. Society S_1 features three overconfident agents, while society S_4 comprises three easy-going agents.

Collaborative Strategy	MMLU p-value	MATH p-value	Chess Move Validity p-value
$p_0p_0p_0$	0.122	0.621	0.532
$p_0p_0p_1$	0.251	0.291	0.014
$p_0p_1p_0$	0.004	0.248	0.185
$p_0p_1p_1$	0.018	0.430	0.015
$p_1p_0p_0$	0.020	0.381	0.132
$p_1p_0p_1$	0.601	0.854	0.506
$p_1p_1p_0$	0.641	0.750	0.282
$p_1p_1p_1$	0.044	0.037	0.585

Table 21: One-Way ANOVA results for the impact of society on accuracy with fixed collaborative strategy, based on experiments from Table 20 using *LlaMA2-chat-70B*.

Society	MMLU p-value	MATH p-value	Chess Move Validity p-value
S_1	0.000	0.013	0.000
S_2	0.000	0.297	0.000
S_3	0.000	0.040	0.000
S_4	0.000	0.056	0.000

Table 22: One-Way ANOVA results for the impact of collaborative strategy on accuracy with fixed society, based on experiments from Table 20 using *LlaMA-70B-Chat*.

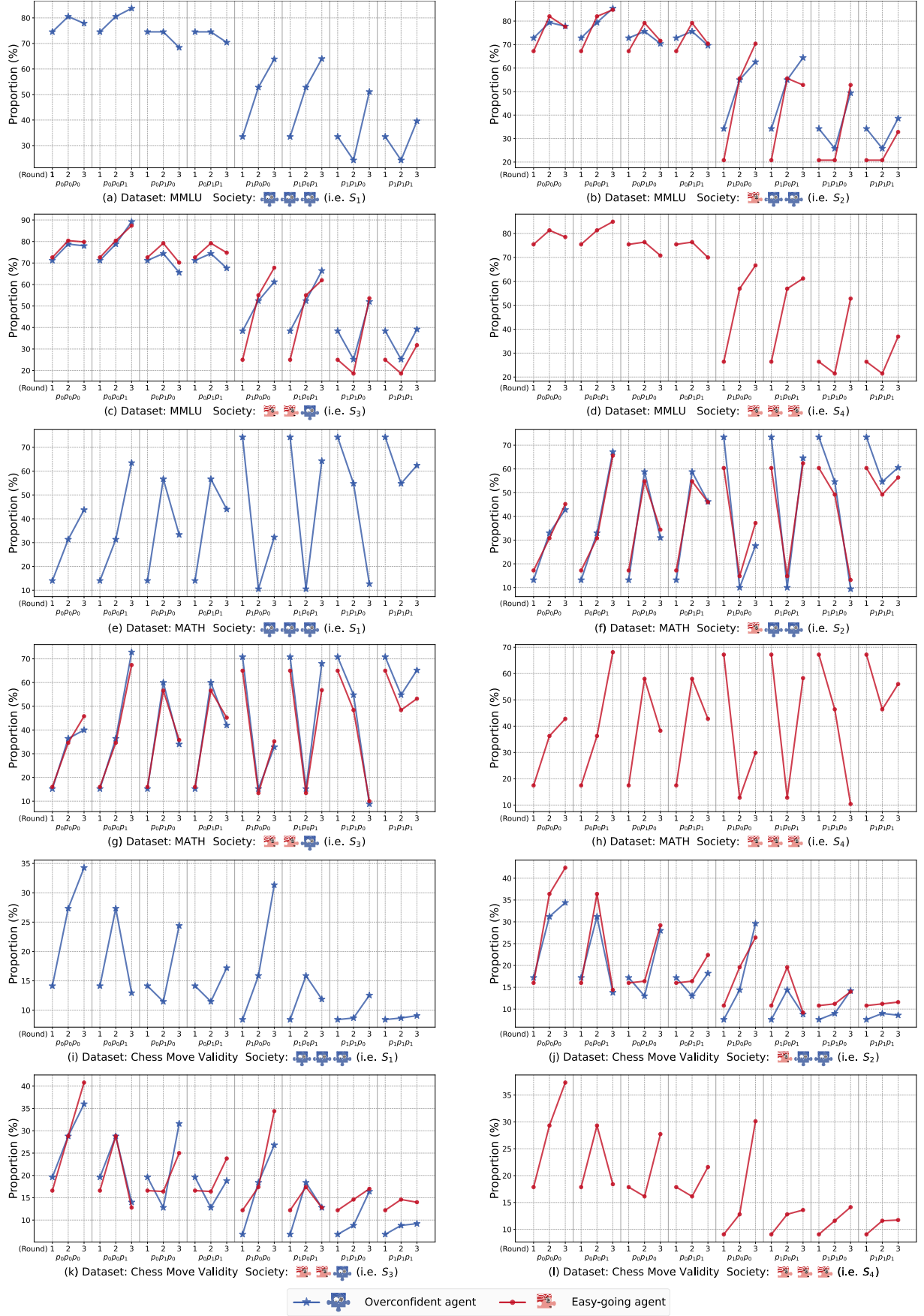


Figure 32: Proportion of agents with different traits changing answers in societies S_1 and S_4 , using *LLaMA2-70B-chat*. Society S_1 features three overconfident agents, while society S_4 comprises three easy-going agents.

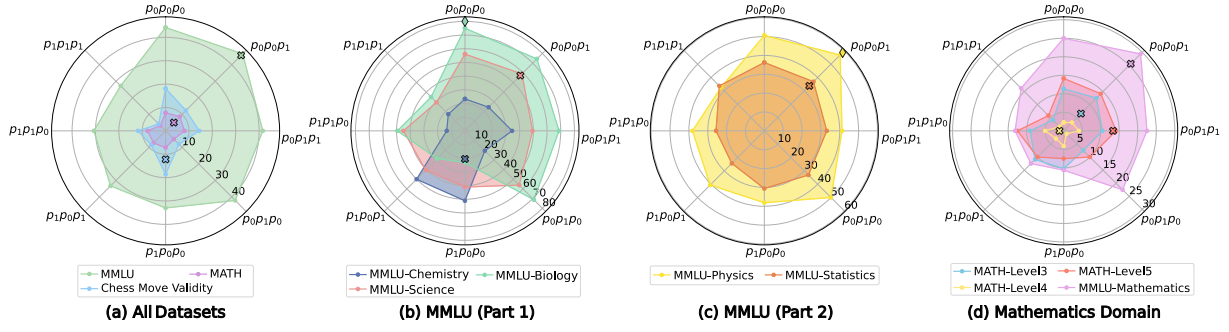


Figure 33: Illustration of different collaborative strategies impacting accuracy diversely on the tasks considering varied *subjects* and *difficulty*, using *LlaMA2-70B-chat*. The symbol ‘⋈’ represents that there is at least one collaborative strategy whose accuracy is better than self-consistency, while the symbol ‘⋄’ indicates that there is no collaborative strategy whose accuracy is worse than self-consistency. Both of these symbols represent the accuracy of self-consistency. The accuracy under each collaborative strategy is a summation within all 3-agent societies.

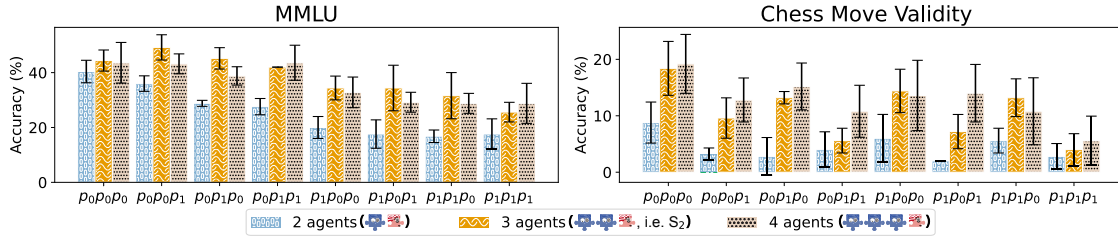


Figure 34: Accuracy of different number of agents under different collaborative strategies, on *LlaMA2-70B-chat*. The significance test is shown in Table 23.

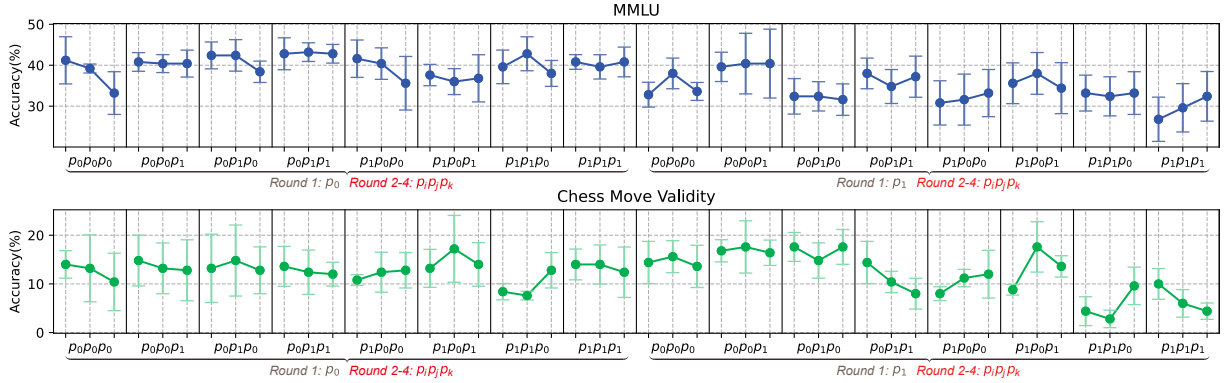


Figure 35: Accuracy at round 2,3,4 within 4-round collaborative societies, where the thinking pattern of round 1 is fixed (p_0 or p_1), using *LlaMA2-70B-chat*. The significance test is shown in Table 24.

Collaborative Strategy	MMLU p-value	Chess Move Validity p-value
$p_0p_0p_0$	0.481	0.006
$p_0p_0p_1$	0.000	0.001
$p_0p_1p_0$	0.000	0.000
$p_0p_1p_1$	-	0.023
$p_1p_0p_0$	0.001	0.035
$p_1p_0p_1$	0.003	0.000
$p_1p_1p_0$	0.002	0.036
$p_1p_1p_1$	0.024	0.423

Table 23: One-way ANOVA analysis of the results of Figure 34 (different numbers of agents), using *LlaMA2-chat-70B*.

Collaborative Strategy	MMLU p-value	Chess Move Validity p-value
$p_0p_0p_0p_0$	0.034	0.545
$p_0p_0p_0p_1$	0.008	0.019
$p_0p_0p_1p_0$	0.020	0.004
$p_0p_0p_1p_1$	0.643	0.004
$p_0p_1p_0p_0$	0.045	0.034
$p_0p_1p_0p_1$	0.164	0.902
$p_0p_1p_1p_0$	0.046	0.006
$p_0p_1p_1p_1$	0.082	0.000
$p_1p_0p_0p_0$	0.706	0.207
$p_1p_0p_0p_1$	0.449	0.494
$p_1p_0p_1p_0$	0.782	0.095
$p_1p_0p_1p_1$	0.664	0.070
$p_1p_1p_0p_0$	0.360	0.041
$p_1p_1p_0p_1$	0.391	0.018
$p_1p_1p_1p_0$	0.394	0.088
$p_1p_1p_1p_1$	0.031	0.033

Table 24: One-way ANOVA analysis of the results in Figure 35 (different rounds), using *LlaMA2-chat-70B*.

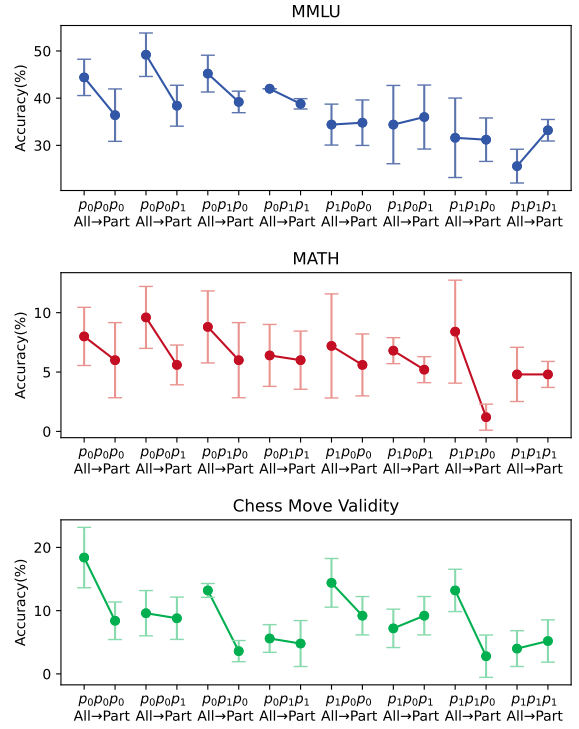


Figure 36: The effect on accuracy of whether all agents in society execute the same thinking pattern in one round, using *LlaMA2-70B-chat*. “All” and “Part” refer to all agents applying the same thinking pattern and different thinking patterns in one round respectively. The significance test is shown in Table 25.

Collaborative Strategy	MMLU p-value	MATH p-value	Chess Move Validity p-value
$p_0p_0p_0p_0$	0.029	0.296	0.004
$p_0p_0p_0p_1$	0.005	0.020	0.724
$p_0p_0p_1p_0$	0.018	0.191	0.000
$p_0p_0p_1p_1$	0.000	0.809	0.684
$p_0p_1p_0p_0$	0.894	0.503	0.045
$p_0p_1p_0p_1$	0.747	0.050	0.328
$p_0p_1p_1p_0$	0.928	0.007	0.001
$p_0p_1p_1p_1$	0.004	1.000	0.557

Table 25: One-way ANOVA analysis of the results in Figure 36 (other collaborative strategies), using *LlaMA2-chat-70B*.

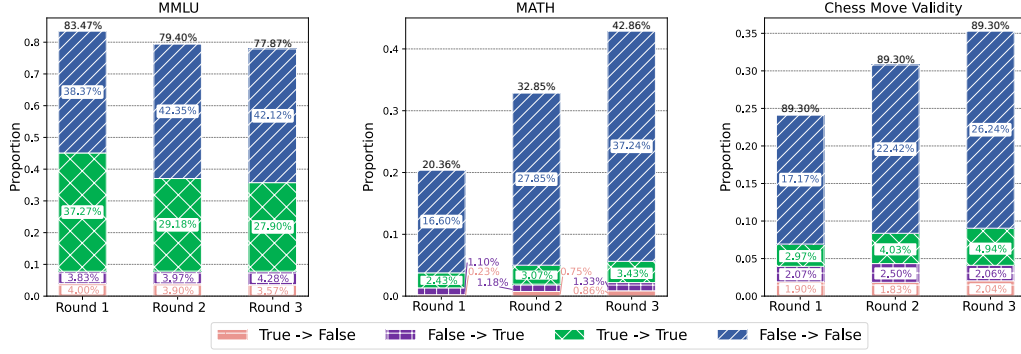


Figure 37: Variation of answer correctness in the situation of conformity, using *LlaMA2-70B-chat*, where *conformity brings about benefits*: $\text{Ratio}(\text{False} \rightarrow \text{True} + \text{True} \rightarrow \text{True}) > \text{Ratio}(\text{True} \rightarrow \text{False} + \text{False} \rightarrow \text{False})$; *conformity brings about detriments*: $\text{Ratio}(\text{False} \rightarrow \text{True} + \text{True} \rightarrow \text{True}) < \text{Ratio}(\text{True} \rightarrow \text{False} + \text{False} \rightarrow \text{False})$.

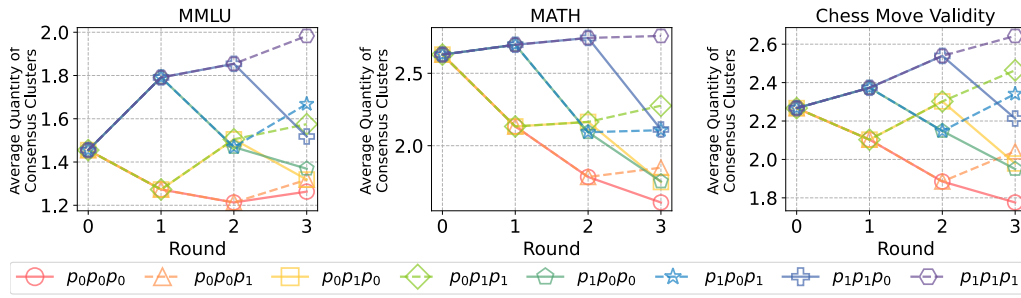


Figure 38: Average quantity of *consensus clusters* (i.e., *unique answers among multiple agents*) under different rounds of collaboration with 3-round collaborative strategies, on *LlaMA2-70B-chat*. *Smaller quantity of consensus clusters, more easier it is to reach a consensus*. Round 0 is equal to self-consistency.

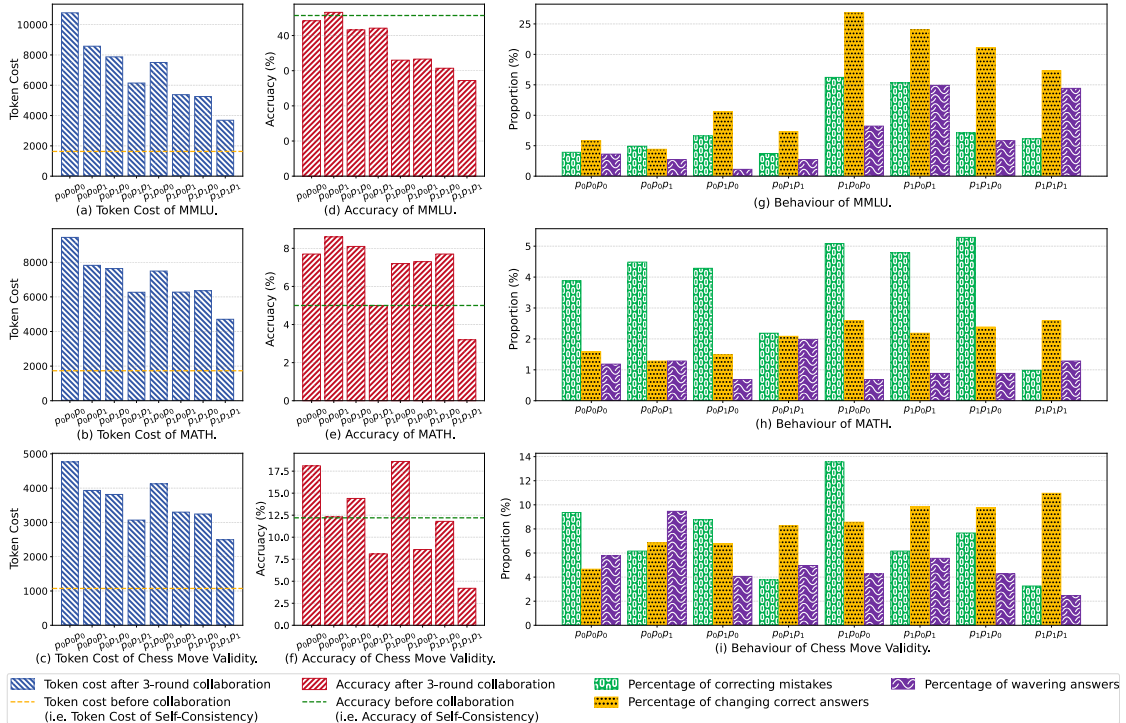


Figure 39: The percentage of different behaviors under different collaborative strategies, using *LlaMA2-70B-chat*. Figure (a-c) & (d-f) respectively show the token cost and accuracy of different strategies before and after 3-round collaboration. Figure (g-i) present the percentage of different behavioral features (mainly analyzed by the change of answer correctness) (Zhang et al., 2023b,a) under different collaborative strategies. All results are summarized across all societies.

H.3 Qwen 72B

Analysis on Machine Social Collaboration. We present the **main results** and **significance tests** of societies and strategies on Qwen 72B in Table 26, 27, 28. We present the **word clouds** of Qwen 72B in Figure 40, and **proportion of agents with different traits changing answers in different societies** on Qwen 72B in Figure 41. Furthermore, we demonstrate that the tasks with different subjects and difficulty display varying sensitivity to collaborative strategies, as presented with **radar maps** on Qwen 72B in Figure 42.

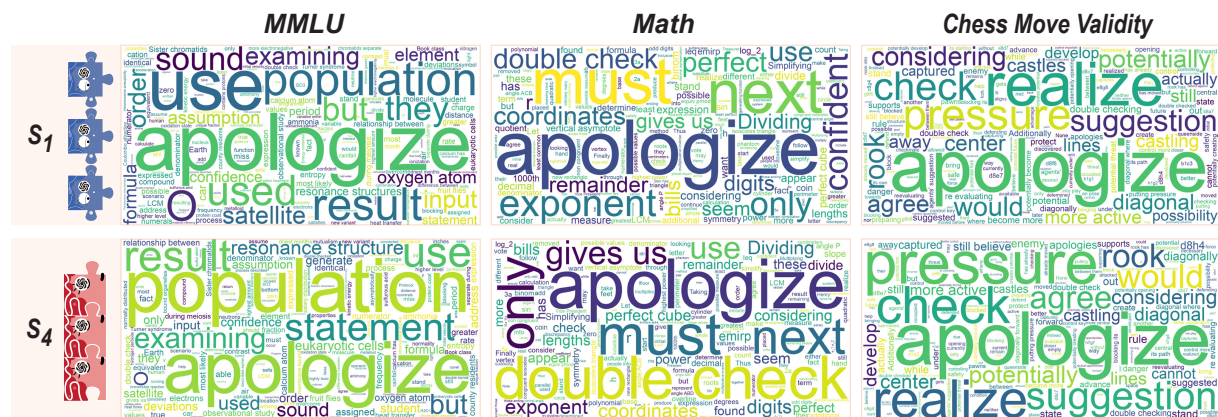
Analysis on Different Numbers of Agents. We present the significance test for different numbers of agents with Qwen 72B in Table 29. We also show the performance varying from agent numbers in Figure 43, varying from societies containing 2~10 agents in Figure 44. We also analyse the *consensus reaching* with different numbers of agents, and present the results in Figure 45, 46.

Analysis on Different Rounds. We present the significance test for different rounds of collaboration with Qwen 72B in Table 30. We also show the performance varying from collaboration rounds in Figure 47, 48, 49.

Analysis on Other Collaborative Strategies. We present the significance test for other collaborative strategies (executing the same or hybrid thinking patterns in a certain round) with Qwen 72B in Table 31. We also show the performance varying from other strategies in Figure 50.

A Social Psychology View on Conformity, Consensus Reaching and Group Dynamics. We then show the variation of answer correctness in the situation of conformity in Figure 51; and the quantity of consensus clusters among 3-agent answers in Figure 52. We present group dynamics reflected by different answer changing behaviors on Qwen 72B in Figure 53.

	Metric (Strategy)	Society	Collaborative Strategy									Metric (Society)	
			$p_0p_0p_0$	$p_0p_0p_1$	$p_0p_1p_0$	$p_0p_1p_1$	$p_1p_0p_0$	$p_1p_0p_1$	$p_1p_1p_0$	$p_1p_1p_1$	<u>Cost</u> ↓	<u>W-T</u> ↑	
MMLU	Acc ↑	S_1	64.8±6.4	66.4±6.8	65.6±9.7	63.6±5.0	58.0±4.2	58.4±3.0	60.0±8.8	63.6±2.6	3661	14	
		S_2	60.4±5.9	60.8±5.2	62.8±2.3	61.6±4.6	53.2±5.6	57.6±2.6	61.2±7.8	62.4±4.3	3657	21	
		S_3	64.0±4.7	64.4±3.9	66.0±2.8	65.2±3.0	56.8±5.9	57.6±5.2	59.6±4.3	64.4±2.6	3690	17	
		S_4	62.4±6.2	64.8±3.9	64.0±7.1	66.8±7.3	53.2±5.4	56.8±4.2	60.4±7.4	58.4±3.9	3570	14	
	<u>Cost</u> ↓	All	5960	4560	4017	3158	4024	2761	2746	1927	-	-	
<u>W-T</u> ↑	All	-	12	14	13	4	4	9	10	-	-		
MATH	Acc ↑	S_1	47.2±5.6	43.6±4.6	46.0±6.5	43.6±5.0	40.4±6.5	41.6±8.1	42.0±4.9	39.6±3.9	3537	11	
		S_2	49.6±5.4	48.4±6.1	48.8±6.7	47.2±5.9	41.2±4.4	41.6±5.4	40.0±4.0	37.6±4.1	3513	7	
		S_3	44.8±6.4	44.4±5.5	43.6±4.3	42.0±7.1	40.4±7.8	37.6±6.7	41.6±7.5	36.4±8.7	3595	9	
		S_4	46.0±6.6	44.8±8.6	46.0±8.0	43.6±5.4	39.2±5.0	41.6±4.8	37.6±6.7	35.6±3.9	3595	11	
	<u>Cost</u> ↓	All	4813	4182	4187	3549	3571	2912	2985	2281	-	-	
<u>W-T</u> ↑	All	-	9	13	7	3	3	2	1	-	-		
Chess Move Validity	Acc ↑	S_1	43.2±7.0	42.4±4.6	41.2±9.7	36.8±6.4	27.6±4.8	22.0±5.3	20.4±4.8	6.4±3.3	2557	6	
		S_2	46.8±4.2	42.8±4.2	39.2±4.6	34.8±4.2	29.6±5.2	16.8±2.7	22.8±5.8	8.8±3.4	2499	1	
		S_3	42.4±8.7	38.4±9.9	38.0±6.9	36.8±7.8	26.8±5.8	19.6±2.6	19.6±2.6	6.0±2.8	2496	3	
		S_4	36.0±8.1	32.4±4.6	34.0±5.8	26.0±4.9	26.8±5.4	20.8±5.4	22.4±5.9	11.2±2.3	2455	4	
	<u>Cost</u> ↓	All	3148	2621	2585	2118	2904	2384	2393	1860	-	-	
<u>W-T</u> ↑	All	-	6	6	2	0	0	0	0	-	-		



Collaborative Strategy	MMLU p-value	MATH p-value	Chess Move Validity p-value
$p_0p_0p_0$	0.654	0.637	0.162
$p_0p_0p_1$	0.388	0.649	0.064
$p_0p_1p_0$	0.841	0.667	0.445
$p_0p_1p_1$	0.455	0.567	0.034
$p_1p_0p_0$	0.387	0.963	0.817
$p_1p_0p_1$	0.933	0.690	0.281
$p_1p_1p_0$	0.987	0.647	0.695
$p_1p_1p_1$	0.061	0.688	0.048

Society	MMLU p-value	MATH p-value	Chess Move Validity p-value
S_1	0.257	0.418	0.000
S_2	0.093	0.004	0.000
S_3	0.004	0.449	0.000
S_4	0.015	0.088	0.000

Table 27: One-Way ANOVA results for the impact of society on accuracy with fixed collaborative strategy, based on experiments from Table 26 using *Qwen 72B*.

Table 28: One-Way ANOVA results for the impact of collaborative strategy on accuracy with fixed society, based on experiments from Table 26 using *Owen 72B*.

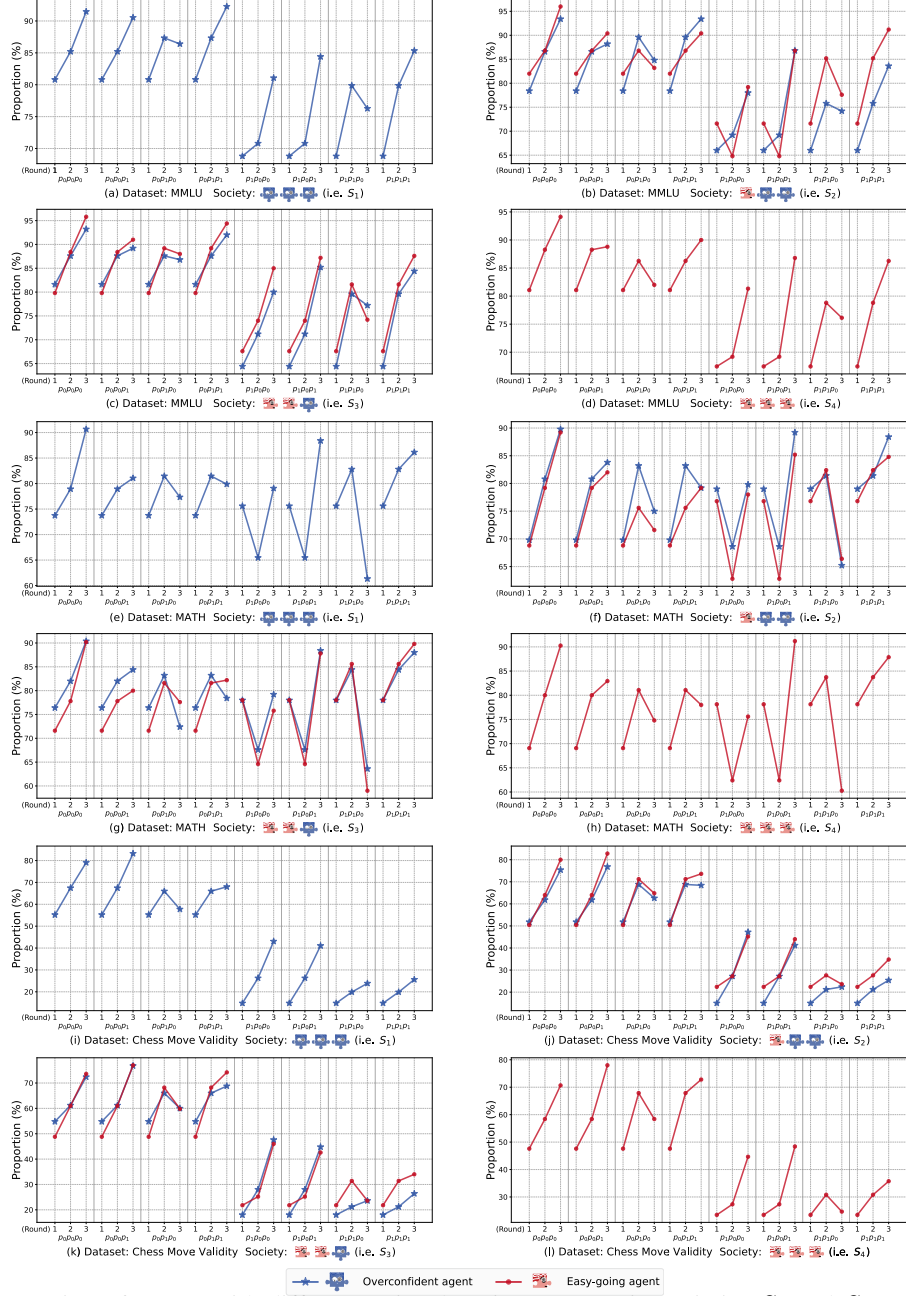


Figure 41: Proportion of agents with different traits changing answers in societies S_1 and S_4 , using *Qwen 72B*. Society S_1 features three overconfident agents, while society S_4 comprises three easy-going agents.

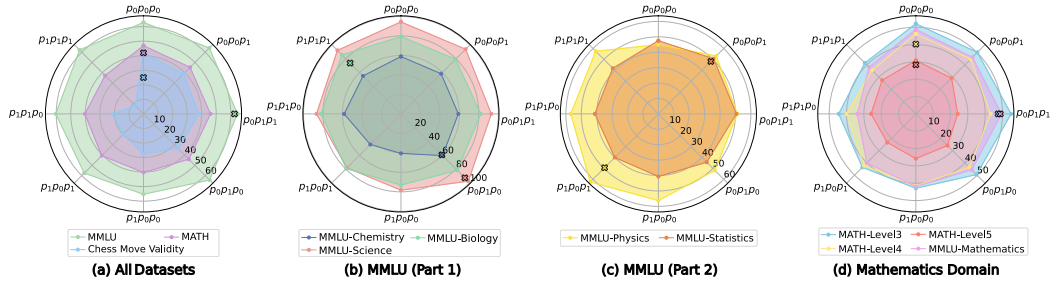


Figure 42: Illustration of different collaborative strategies impacting accuracy diversely on the tasks considering varied *subjects* and *difficulty*, using *Qwen 72B*. The symbol '⋈' represents that there is at least one collaborative strategy whose accuracy is better than self-consistency, while the symbol '◇' indicates that there is no collaborative strategy whose accuracy is worse than self-consistency. Both of these symbols represent the accuracy of self-consistency. The accuracy under each collaborative strategy is a summation within all 3-agent societies.

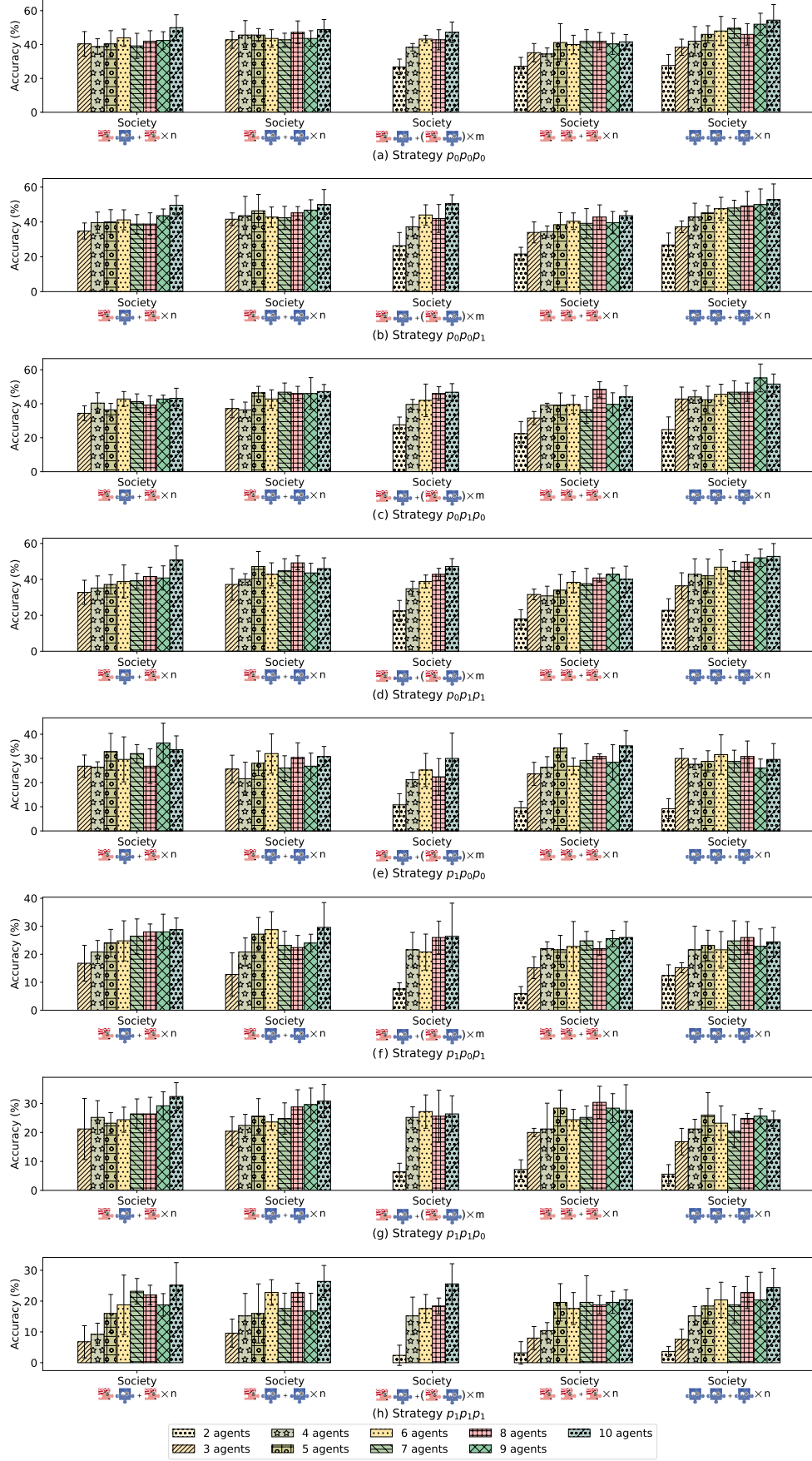


Figure 43: Accuracy of different numbers (2~10) of agents under different collaborative strategies, on Qwen 72B. The significance test is shown in Table 29.

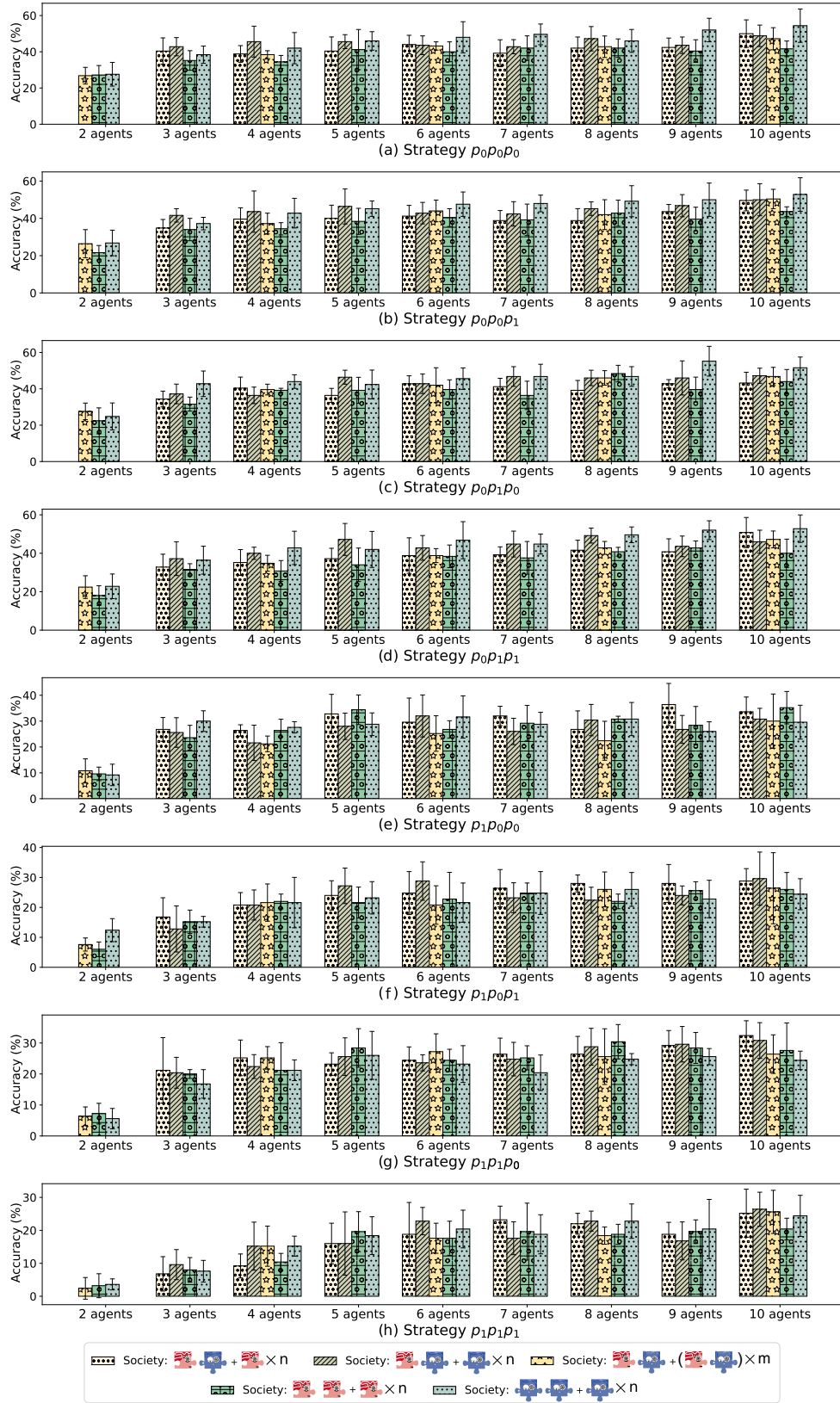


Figure 44: Accuracy of different societies with 2~10 agents under different collaborative strategies, on *Qwen 72B*.

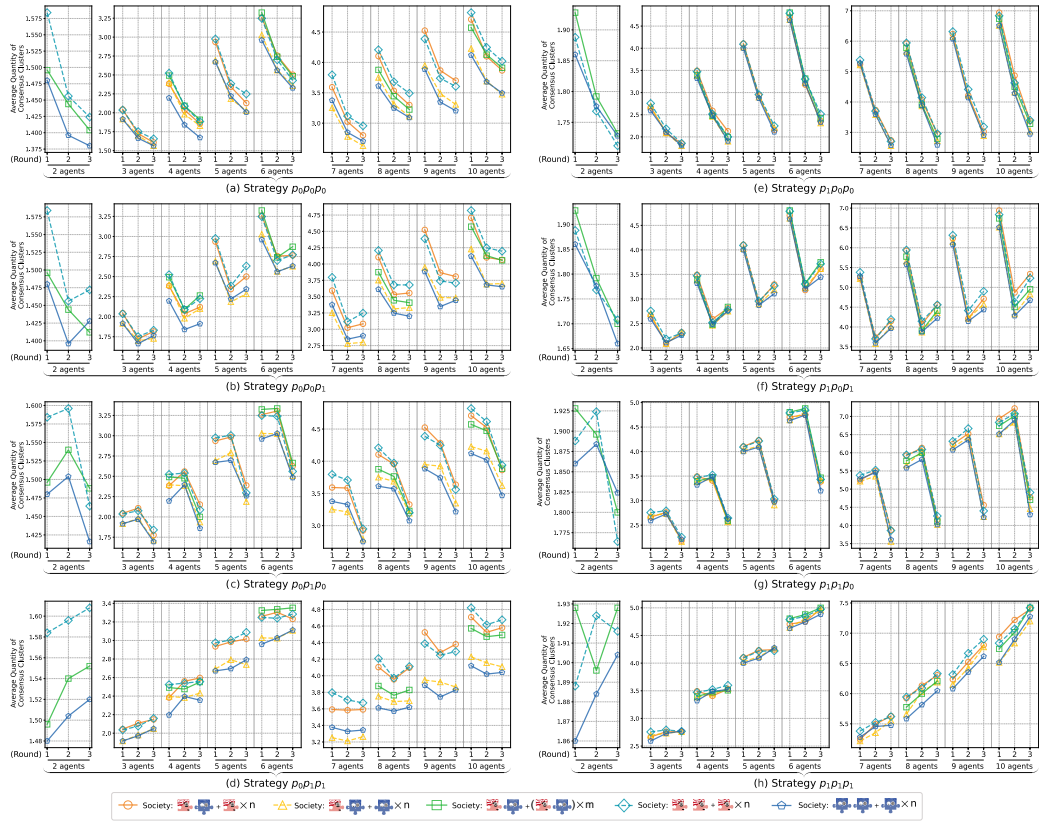


Figure 45: Average quantity of *consensus clusters* (unique answers among multiple agents) in different societies with 2~10 agents under each round of 3-round collaborative strategies, using Qwen 72B.

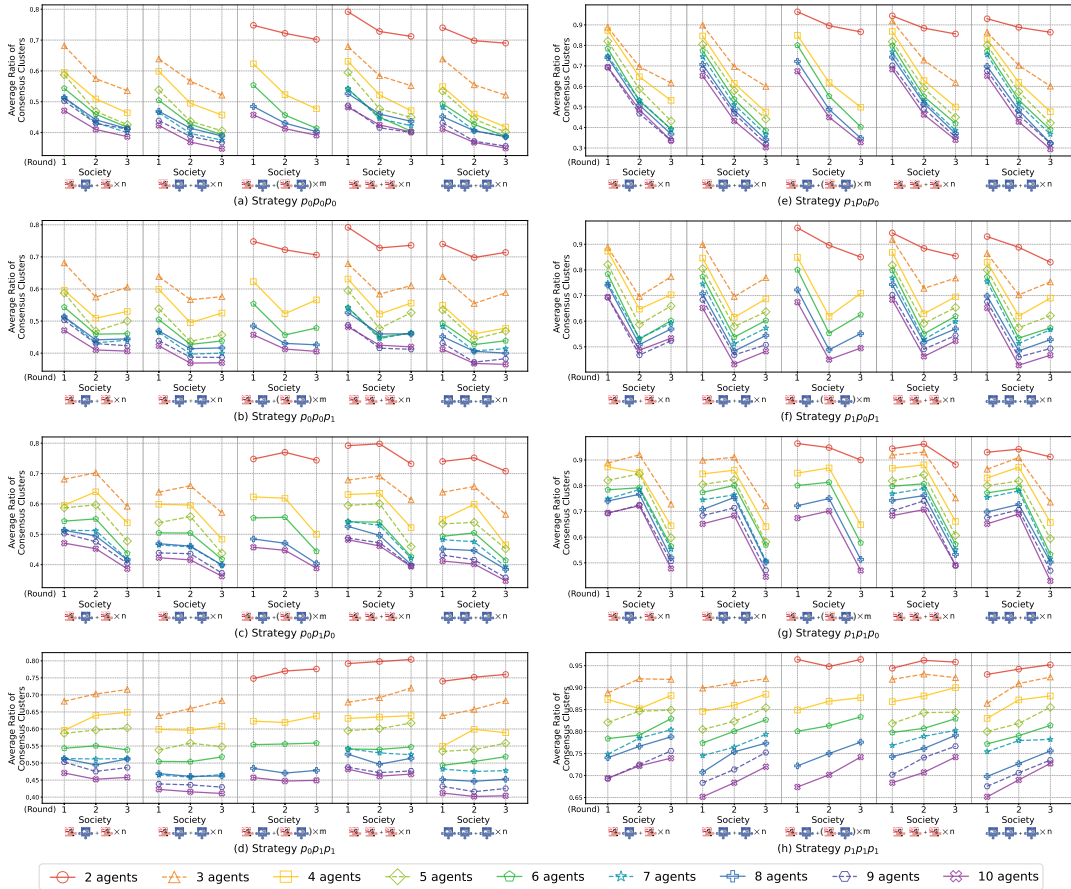


Figure 46: Average ratio of *consensus clusters* (unique answers among multiple agents) with different numbers (2~10) of agents under each round of 3-round collaborative strategies, using Qwen 72B.

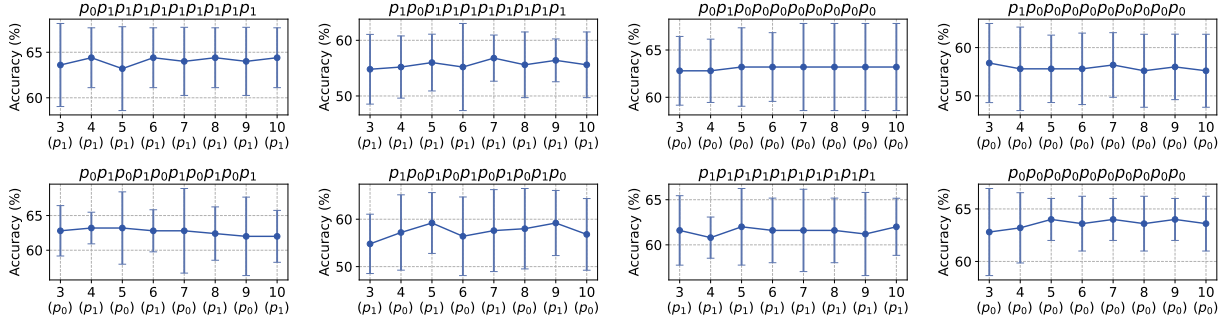


Figure 47: Accuracy of *different (3~10) rounds of collaboration* within 3-agent society S_2 (1 easy-going and 2 overconfident agents) on MMLU, using *Qwen 72B*. The significance test is shown in Table 30.

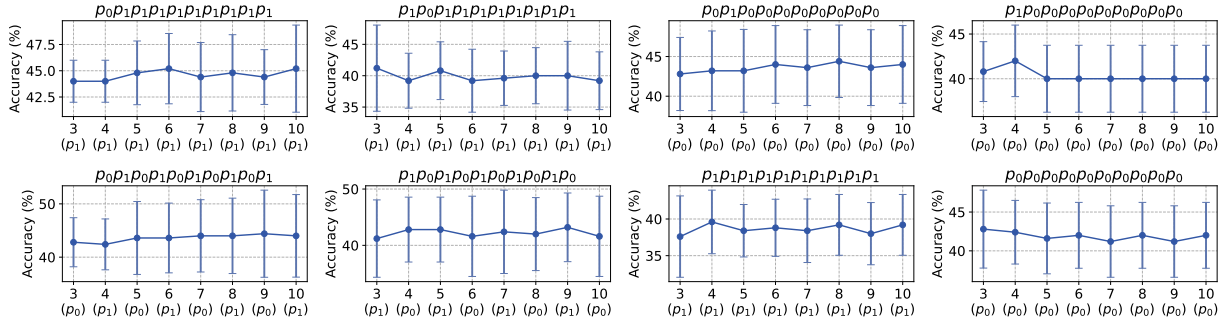


Figure 48: Accuracy of *different (3~10) rounds of collaboration* within 3-agent society S_2 (1 easy-going and 2 overconfident agents) on MATH, using *Qwen 72B*. The significance test is shown in Table 30.

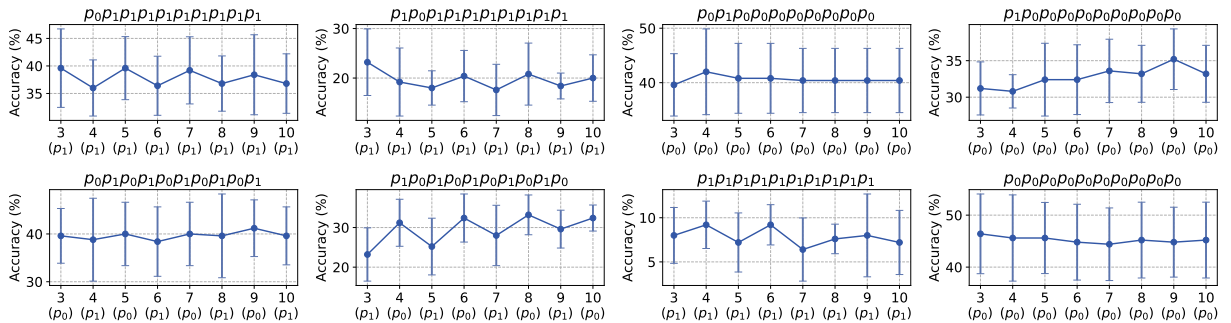


Figure 49: Accuracy of *different (3~10) rounds of collaboration* within 3-agent society S_2 (1 easy-going and 2 overconfident agents) on Chess Move Validity, using *Qwen 72B*. The significance test is shown in Table 30.

Collaborative Strategy	S'_1 p-value	S'_2 p-value	S'_3 p-value	S'_4 p-value	S'_5 p-value
$p_0p_0p_0$	0.005	0.001	0.003	0.041	0.015
$p_0p_0p_1$	0.017	0.010	0.037	0.001	0.006
$p_0p_1p_0$	0.006	0.016	0.002	0.000	0.001
$p_0p_1p_1$	0.020	0.002	0.010	0.001	0.004
$p_1p_0p_0$	0.000	0.005	0.000	0.000	0.000
$p_1p_0p_1$	0.002	0.008	0.004	0.000	0.054
$p_1p_1p_0$	0.003	0.000	0.002	-	0.000
$p_1p_1p_1$	0.064	0.008	0.005	0.016	0.000

Table 29: One-way ANOVA analysis of results in Figure 43 (different numbers of agents), using *Qwen 72B*. S'_1 : One overconfident agent and the others are all easygoing. S'_2 : One easygoing agent among predominantly overconfident agents. S'_3 : Equal numbers of overconfident and easygoing agents. S'_4 : Entirely easygoing agents. S'_5 : Entirely overconfident agents. ‘-’: It doesn’t pass homogeneity test for variance.

Collaborative Strategy	MMLU p-value	MATH p-value	Chess Move Validity p-value
$p_0p_0p_0p_0p_0p_0p_0p_0p_0$	0.262	0.987	0.956
$p_1p_0p_0p_0p_0p_0p_0p_0p_0$	0.753	0.697	0.124
$p_0p_1p_0p_0p_0p_0p_0p_0p_0$	0.914	0.962	0.386
$p_1p_0p_1p_0p_0p_0p_0p_0p_0$	0.673	0.715	0.154
$p_0p_1p_0p_0p_0p_0p_0p_0p_0$	0.922	0.987	0.700
$p_1p_0p_1p_1p_1p_1p_1p_1p_1$	0.845	0.843	0.282
$p_0p_1p_1p_1p_1p_1p_1p_1p_1$	0.928	0.585	0.583
$p_1p_1p_1p_1p_1p_1p_1p_1p_1$	0.832	0.801	0.731

Table 30: One-way ANOVA analysis of the results in Figure 48, 48, 49 (different rounds), using *Qwen 72B*.

Collaborative Strategy	MMLU p-value	MATH p-value	Chess Move Validity p-value
$p_0p_0p_0$	0.704	0.142	0.003
$p_0p_0p_1$	0.136	0.184	0.000
$p_0p_1p_0$	0.899	0.157	0.001
$p_0p_1p_1$	0.180	0.194	0.089
$p_1p_0p_0$	0.157	0.856	0.004
$p_1p_0p_1$	0.521	0.152	0.019
$p_1p_1p_0$	-	0.790	0.004
$p_1p_1p_1$	0.391	0.688	1.000

Table 31: One-way ANOVA analysis of results in Figure 50 (other collaborative strategies), using *Qwen 72B*. ‘-’ means it doesn’t pass homogeneity test for variance.

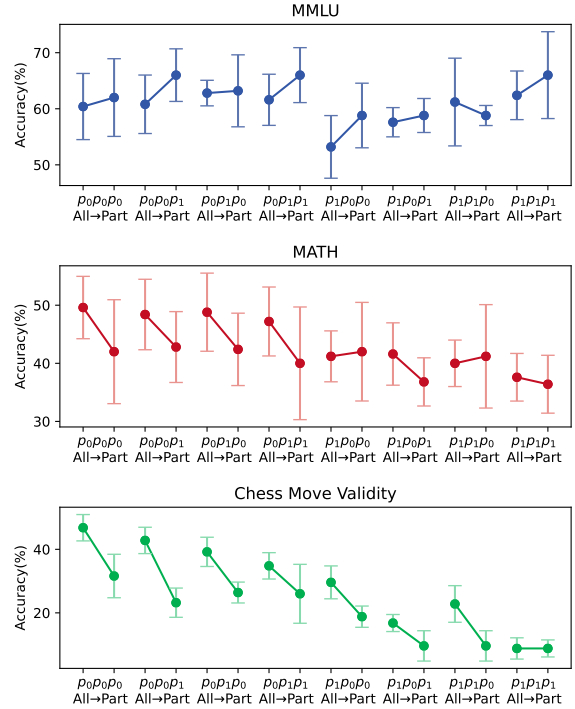


Figure 50: The effect on accuracy of whether all agents in society execute the same thinking pattern in one round, using *Qwen 72B*. ‘All’ and ‘Part’ refer to all agents applying the same thinking pattern and different thinking patterns in one round respectively. The significance test is shown in Table 31.

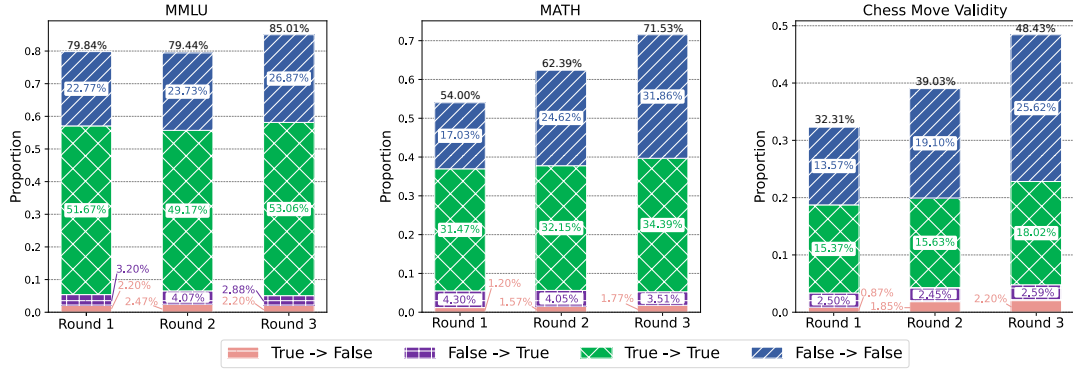


Figure 51: Variation of answer correctness in the situation of conformity, using *Qwen 72B*, where *conformity brings about benefits*: $\text{Ratio}(\text{False} \rightarrow \text{True} + \text{True} \rightarrow \text{True}) > \text{Ratio}(\text{True} \rightarrow \text{False} + \text{False} \rightarrow \text{False})$; *conformity brings about detriments*: $\text{Ratio}(\text{False} \rightarrow \text{True} + \text{True} \rightarrow \text{True}) < \text{Ratio}(\text{True} \rightarrow \text{False} + \text{False} \rightarrow \text{False})$.

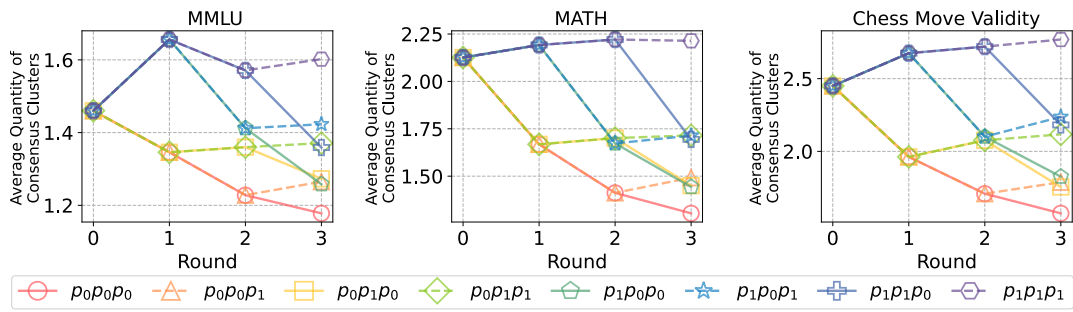


Figure 52: Average quantity of *consensus clusters* (i.e., unique answers among multiple agents) under different rounds of collaboration with 3-round collaborative strategies, using *Qwen 72B*. Smaller quantity of consensus clusters, more easier it is to reach a consensus. Round 0 is equal to self-consistency.

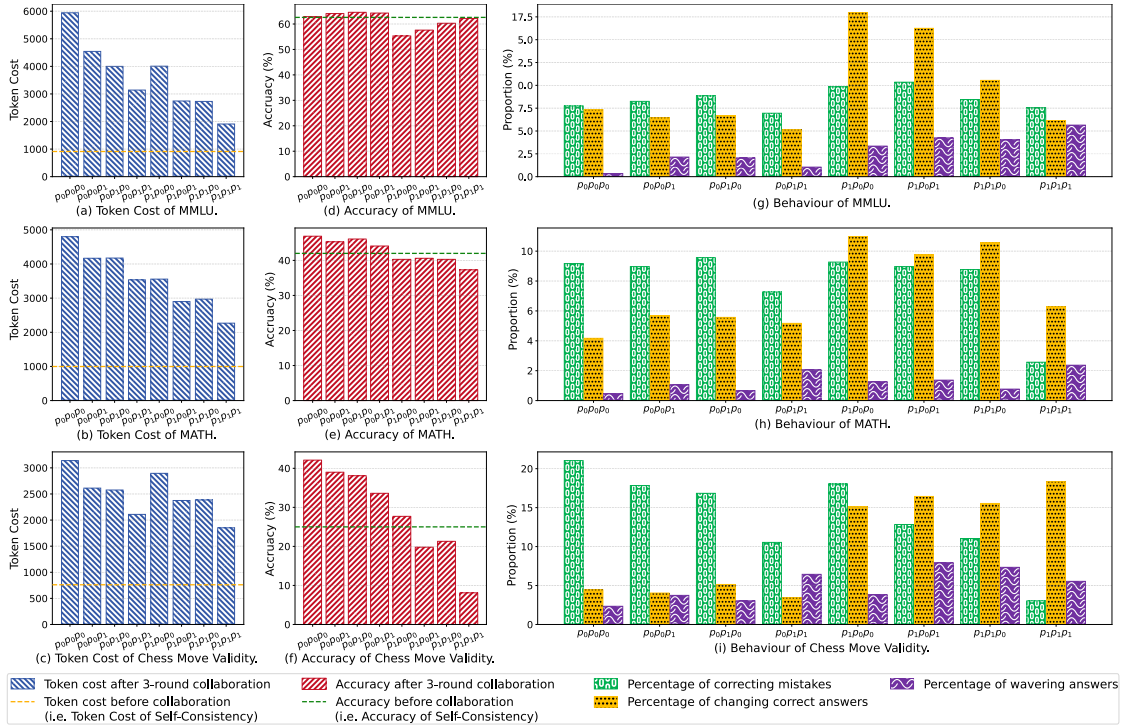


Figure 53: The percentage of different behaviors under different collaborative strategies, using *Qwen 72B*. Figure (a-c) & (d-f) respectively show the token cost and accuracy of different strategies before and after 3-round collaboration. Figure (g-i) present the percentage of different behavioral features (mainly analyzed by the change of answer correctness) (Zhang et al., 2023b,a) under different collaborative strategies. All results are summarized across all societies.

H.4 Mixtral 8×7B

Analysis on Machine Social Collaboration. We present the **main results** and **significance tests** of societies and strategies on Mixtral 8×7B in Table 32, 33, 34. We present the word clouds of Mixtral 8×7B in Figure 54, and proportion of agents with different traits changing answers in different societies on Mixtral 8×7B in Figure 55. Furthermore, we demonstrate that the tasks with different subjects and difficulty display varying sensitivity to collaborative strategies, as presented with **radar maps** on Mixtral 8×7B in Figure 56.

Analysis on Different Numbers of Agents. We present the significance test for different numbers of agents with Mixtral 8×7B in Table 35. We also show the performance varying from agent numbers in Figure 57, varying from societies containing 2~10 agents in Figure 58. We also analyse the *consensus reaching* with different numbers of agents, and present the results in Figure 59, 60.

Analysis on Different Rounds. We present the significance test for different rounds of collaboration with Mixtral 8×7B in Table 36. We also show the performance varying from collaboration rounds in Figure 61, 62, 63.

Analysis on Other Collaborative Strategies. We present the significance test for other collaborative strategies (executing the same or hybrid thinking patterns in a certain round) with Mixtral 8×7B in Table 37. We also show the performance varying from other strategies in Figure 64.

A Social Psychology View on Conformity, Consensus Reaching and Group Dynamics. We then show the variation of answer correctness in the situation of conformity in Figure 65; and the quantity of consensus clusters among 3-agent answers in Figure 66. We present group dynamics reflected by different answer changing behaviors on Mixtral-8×7B in Figure 67.

	Metric (Strategy)	Society	$p_0p_0p_0$	$p_0p_0p_1$	$p_0p_1p_0$	Collaborative Strategy					Metric (Society)	
						$p_0p_1p_1$	$p_1p_0p_0$	$p_1p_0p_1$	$p_1p_1p_0$	$p_1p_1p_1$	Cost ↓	W-T ↑
MMLU	Acc ↑	S_1	60.0±8.1	59.6±3.9	58.4±4.3	60.0±1.4	60.0±5.8	60.4±5.2	59.6±2.6	60.0±2.0	4479	17
		S_2	59.2±7.7	60.0±7.9	60.0±6.5	60.8±5.8	61.2±3.6	62.8±5.4	62.8±5.4	61.2±2.7	4475	27
		S_3	62.4±5.2	63.6±4.3	65.2±3.0	65.2±3.0	59.2±4.4	61.2±4.2	61.6±2.6	59.6±3.6	4489	18
		S_4	60.0±3.7	62.4±3.6	63.2±3.4	62.8±2.7	60.0±5.1	60.4±5.5	64.8±5.8	62.0±6.6	4396	25
	Cost ↓	All	6891	5371	4871	3944	4996	3594	3495	2516	-	-
MATH	Acc ↑	S_1	30.4±3.3	36.0±1.4	33.6±2.2	32.8±4.2	31.2±3.4	30.4±2.6	30.8±2.3	27.6±1.7	5362	23
		S_2	31.6±6.1	29.2±5.4	30.4±6.8	28.0±3.7	32.4±3.6	29.2±3.9	32.0±6.0	27.6±3.0	5369	14
		S_3	32.4±6.7	32.8±7.8	34.8±4.8	32.0±4.7	30.8±4.2	28.8±4.2	30.8±2.3	24.8±3.9	5343	18
		S_4	32.0±4.7	31.2±2.7	31.2±5.2	32.0±5.1	29.2±4.4	30.0±7.2	31.2±1.1	27.2±3.4	5238	18
	Cost ↓	All	6630	5814	6116	5042	5915	4745	4818	3540	-	-
Chess Move Validity	Acc ↑	S_1	22.8±2.7	21.6±3.3	21.2±5.6	20.8±3.0	18.8±5.4	18.8±4.6	17.6±7.0	18.8±1.1	2300	9
		S_2	22.0±5.7	18.0±2.8	18.8±3.4	16.4±2.6	22.0±8.4	18.8±4.8	16.0±2.8	16.0±0.0	2280	10
		S_3	21.2±2.7	20.0±3.2	18.0±2.5	18.0±2.5	20.0±2.8	18.8±3.0	16.4±4.6	15.6±1.7	2269	9
		S_4	18.0±3.7	16.4±3.9	19.2±4.6	16.4±2.6	20.0±1.4	20.8±3.6	20.4±3.9	18.8±2.3	2253	23
	Cost ↓	All	2956	2458	2396	1973	2630	2063	2083	1644	-	-
	W-T ↑	All	-	7	8	6	9	10	6	5	-	-

Table 32: The impact of eight different collaborative strategies on the performance of three datasets across distinct societies (using *Mixtral-8×7B*). The significances test on societies and strategies are respectively shown in Table 33, 34.

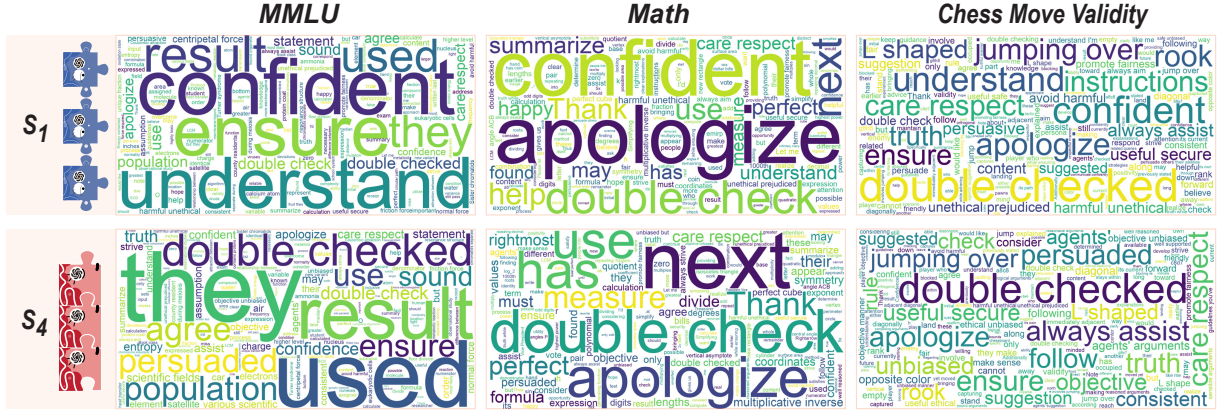


Figure 54: Comparative word clouds on three datasets in societies S_1 and S_4 , using *Mixtral-8×7B*. Society S_1 features three overconfident agents, while society S_4 comprises three easy-going agents.

Collaborative Strategy	MMLU p-value	MATH p-value	Chess Move Validity p-value
$p_0p_0p_0$	0.873	0.941	0.261
$p_0p_0p_1$	0.578	0.216	0.109
$p_0p_1p_0$	0.114	0.500	0.666
$p_0p_1p_1$	0.142	0.347	0.062
$p_1p_0p_0$	0.930	0.638	0.809
$p_1p_0p_1$	0.863	0.949	0.825
$p_1p_1p_0$	0.325	-	0.485
$p_1p_1p_1$	0.785	0.438	0.004

Table 33: One-Way ANOVA results for the impact of society on accuracy with fixed collaborative strategy, based on experiments from Table 32 using *Mixtral 8×7B*. ‘-’: It doesn’t pass homogeneity test for variance.

Society	MMLU p-value	MATH p-value	Chess Move Validity p-value
S_1	0.999	0.002	0.585
S_2	0.970	0.693	0.202
S_3	0.129	0.127	0.078
S_4	0.706	0.714	0.300

Table 34: One-Way ANOVA results for the impact of collaborative strategy on accuracy with fixed society, based on experiments from Table 32 using *Mixtral 8×7B*.

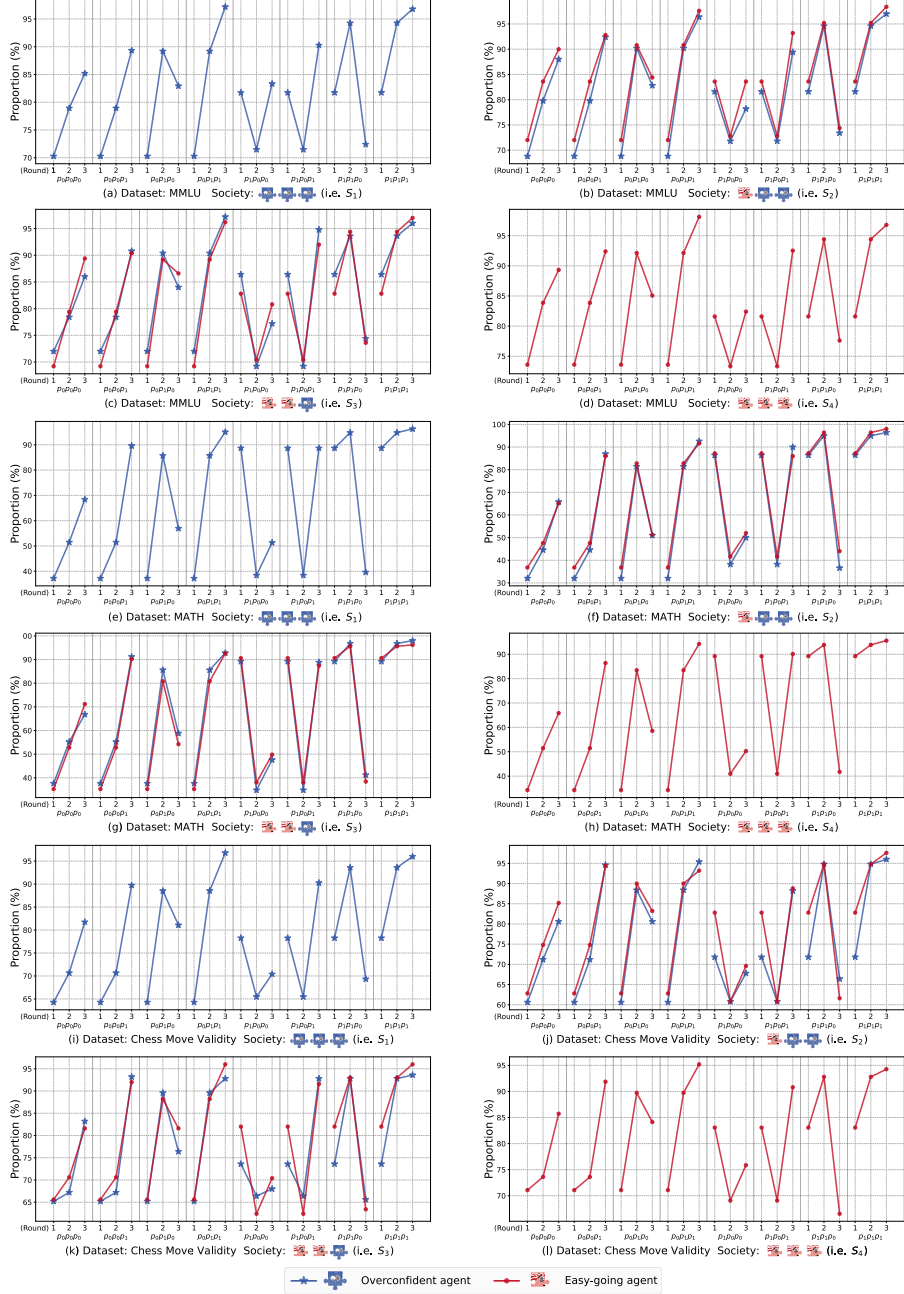


Figure 55: Proportion of agents with different traits changing answers in societies S_1 and S_4 , using *Mixtral-8×7B*. Society S_1 features three overconfident agents, while society S_4 comprises three easy-going agents.

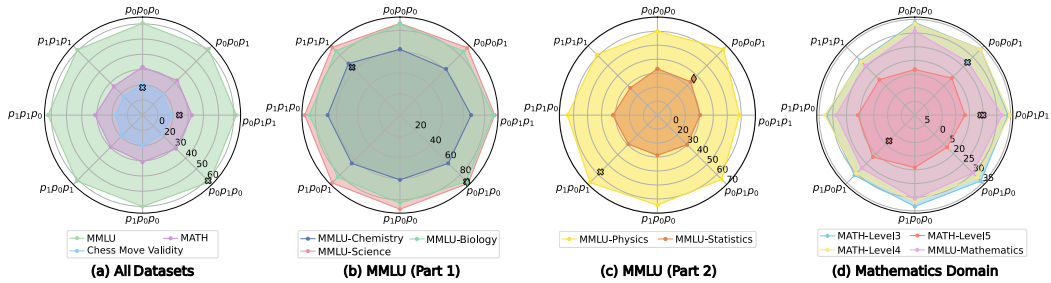


Figure 56: Illustration of different collaborative strategies impacting accuracy diversely on the tasks considering varied *subjects* and *difficulty*, using *Mixtral-8×7B*. The symbol '⋈' represents that there is at least one collaborative strategy whose accuracy is better than self-consistency, while the symbol '◇' indicates that there is no collaborative strategy whose accuracy is worse than self-consistency. Both of these symbols represent the accuracy of self-consistency. The accuracy under each collaborative strategy is a summation within all 3-agent societies.

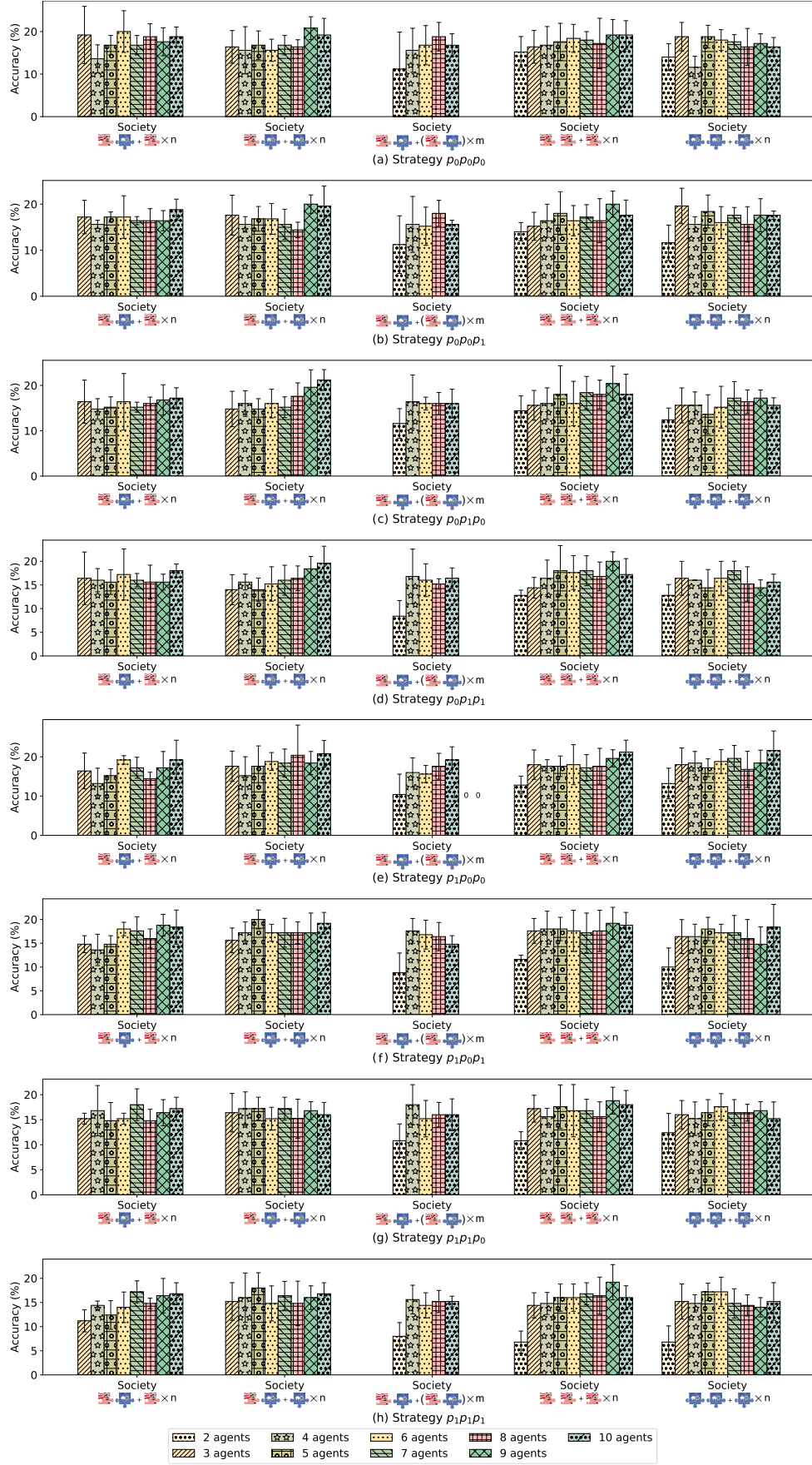


Figure 57: Accuracy of different numbers (2~10) of agents under different collaborative strategies, on *Mixtral-8x7B*. The significance test is shown in Table 35.

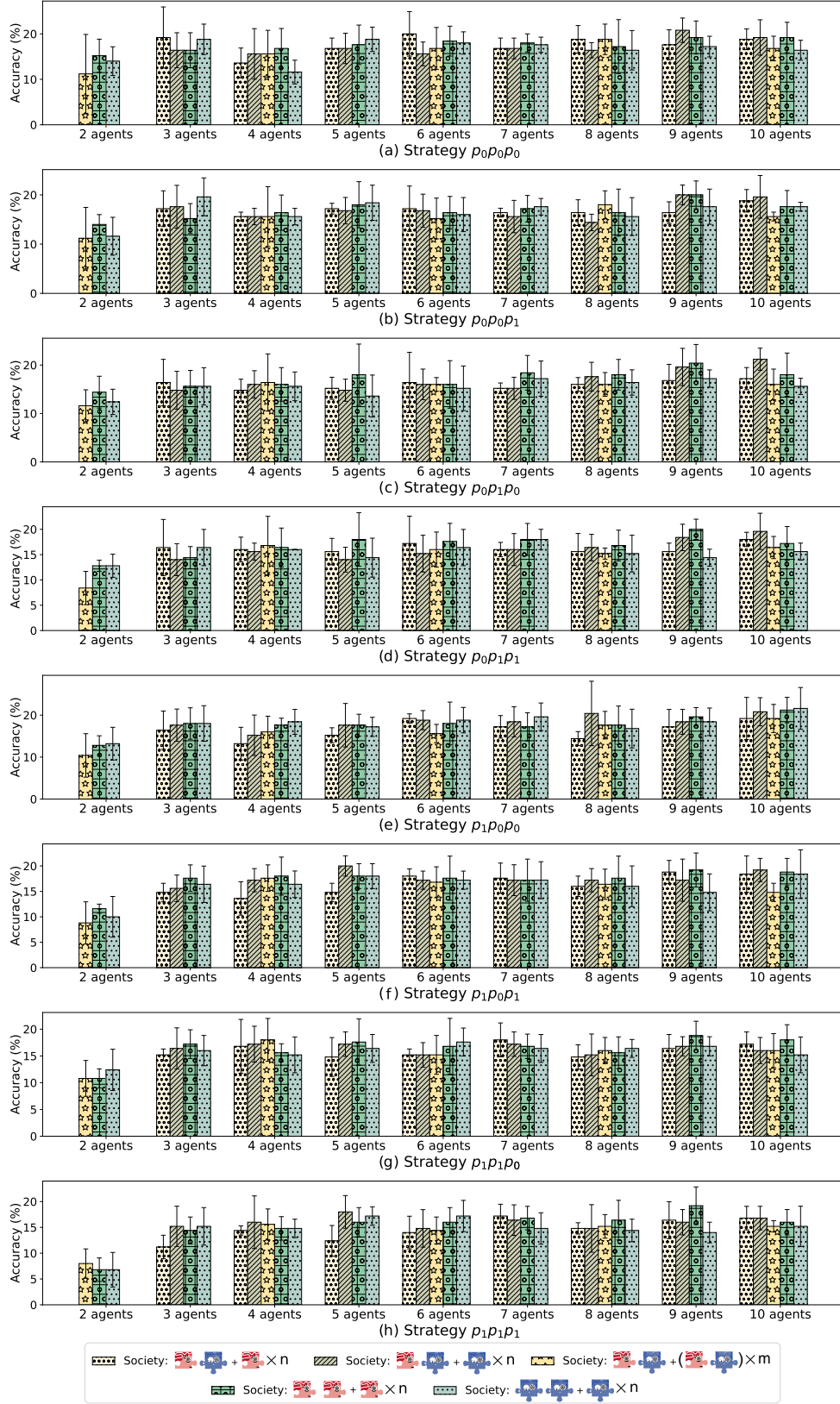


Figure 58: Accuracy of different societies with 2~10 agents under different collaborative strategies, on *Mixtral-8x7B*.

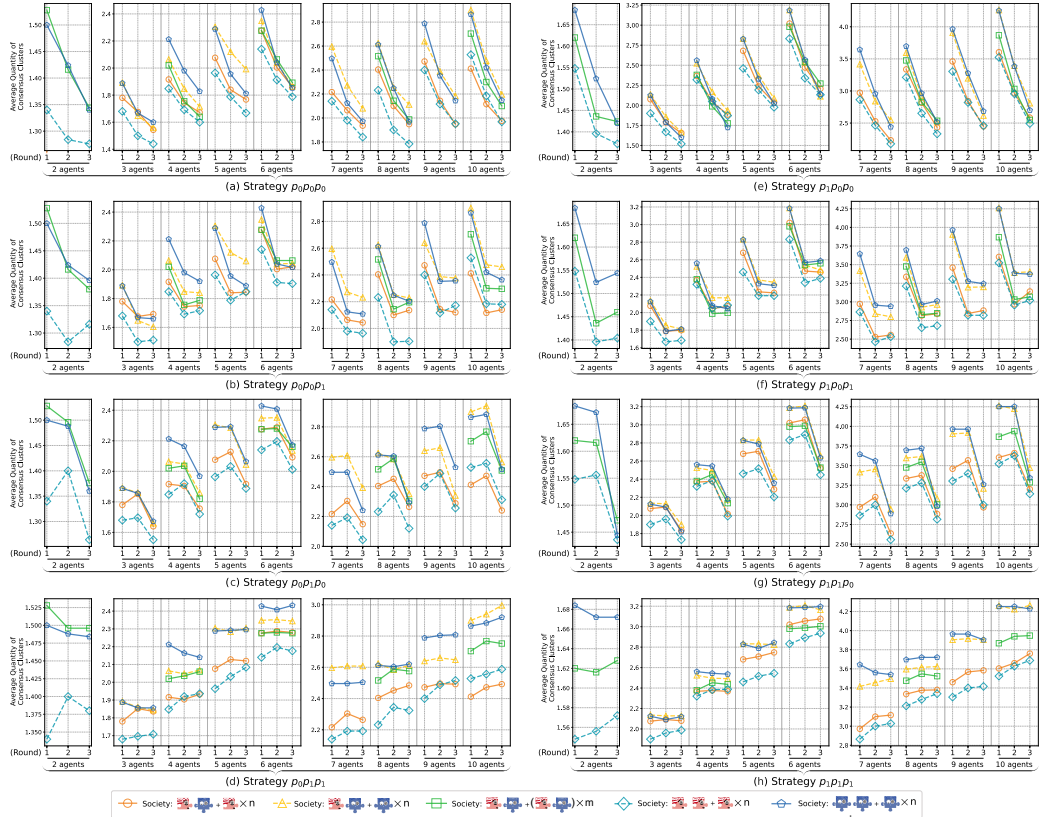


Figure 59: Average quantity of *consensus clusters* (unique answers among multiple agents) in different societies with 2~10 agents under each round of 3-round collaborative strategies, using *Mixtral-8x7B*.

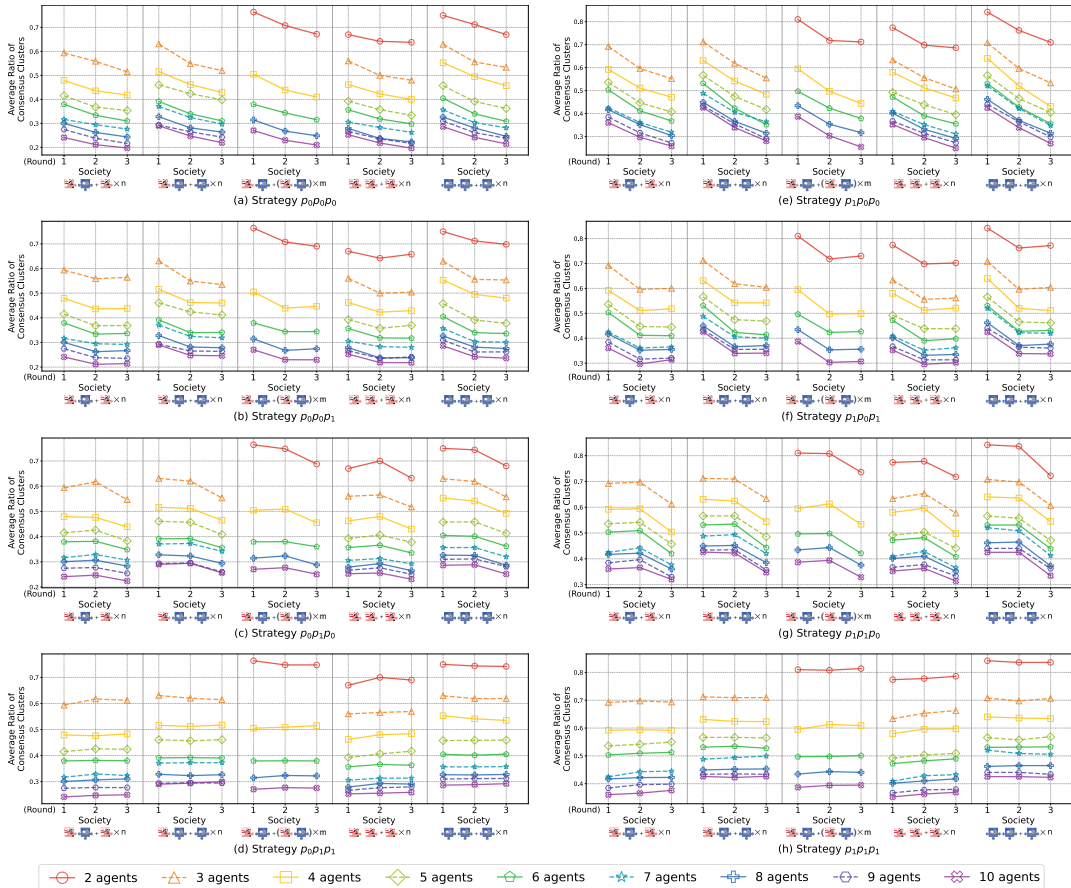


Figure 60: Average ratio of *consensus clusters* (unique answers among multiple agents) with different numbers (2~10) of agents under each round of 3-round collaborative strategies, using *Mixtral-8x7B*.

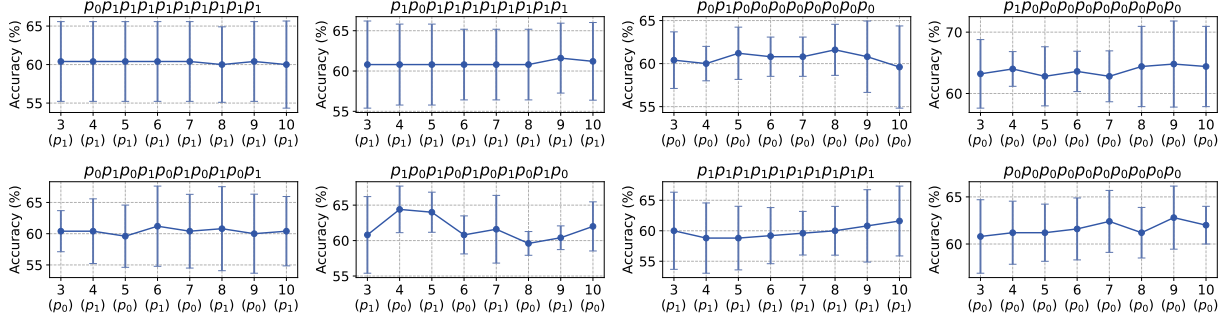


Figure 61: Accuracy of *different (3~10) rounds of collaboration* within 3-agent society S_2 (1 easy-going and 2 overconfident agents) on MMLU, using *Mixtral-8x7B*. The significance test is shown in Table 36.

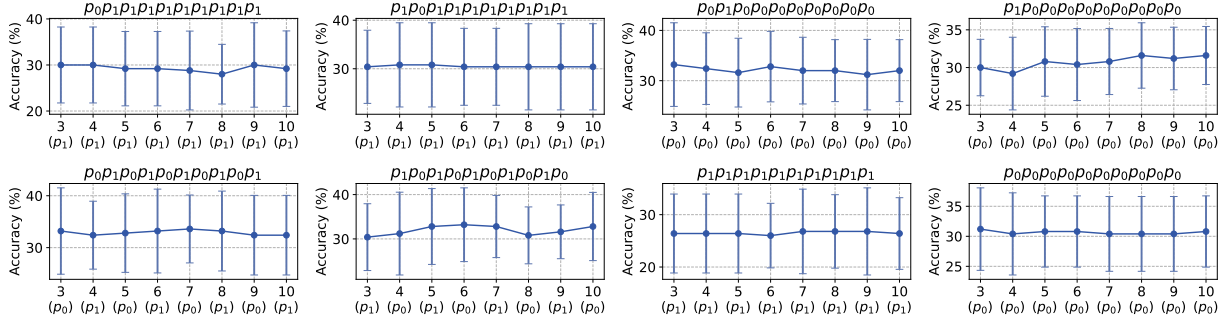


Figure 62: Accuracy of *different (3~10) rounds of collaboration* within 3-agent society S_2 (1 easy-going and 2 overconfident agents) on MATH, using *Mixtral-8x7B*. The significance test is shown in Table 36.

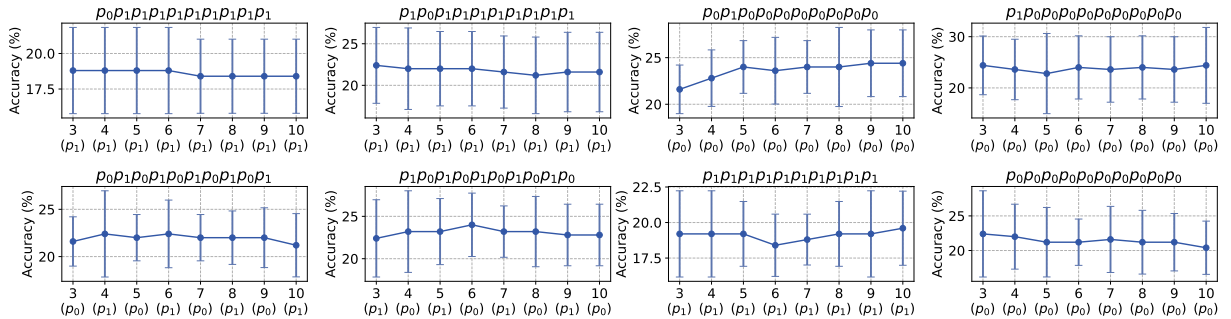


Figure 63: Accuracy of *different (3~10) rounds of collaboration* within 3-agent society S_2 (1 easy-going and 2 overconfident agents) on Chess Move Validity, using *Mixtral-8x7B*. The significance test is shown in Table 36.

Collaborative Strategy	S'_1 p-value	S'_2 p-value	S'_3 p-value	S'_4 p-value	S'_5 p-value
$p_0p_0p_0$	0.188	0.406	0.235	0.805	0.009
$p_0p_0p_1$	0.106	0.112	0.238	0.459	0.008
$p_0p_1p_0$	0.142	0.145	0.227	0.739	0.227
$p_0p_1p_1$	0.013	0.004	0.035	0.138	0.075
$p_1p_0p_0$	0.159	0.082	0.105	0.018	0.088
$p_1p_0p_1$	0.029	0.003	0.002	0.004	0.018
$p_1p_1p_0$	0.051	0.028	0.010	0.001	0.247
$p_1p_1p_1$	0.002	0.016	0.003	0.000	0.001

Table 35: One-way ANOVA analysis of results in Figure 57 (different numbers of agents), using *Mixtral* $8 \times 7B$. S'_1 : One overconfident agent and the others are all easygoing. S'_2 : One easygoing agent among predominantly overconfident agents. S'_3 : Equal numbers of overconfident and easygoing agents. S'_4 : Entirely easygoing agents. S'_5 : Entirely overconfident agents.

Collaborative Strategy	MMLU p-value	MATH p-value	Chess Move Validity p-value
$p_0p_0p_0p_0p_0p_0p_0p_0p_0$	0.607	0.911	0.789
$p_1p_0p_0p_0p_0p_0p_0p_0p_0$	0.578	0.581	0.939
$p_0p_1p_0p_0p_0p_0p_0p_0p_0$	0.936	0.665	0.123
$p_1p_0p_1p_0p_0p_0p_0p_0p_0$	0.377	0.896	0.952
$p_0p_1p_0p_1p_0p_0p_0p_0p_0$	0.987	0.651	0.271
$p_1p_0p_1p_1p_1p_1p_1p_1p_1$	0.989	0.878	0.919
$p_0p_1p_1p_1p_1p_1p_1p_1p_1$	0.989	0.982	1.000
$p_1p_1p_1p_1p_1p_1p_1p_1p_1$	0.945	0.995	0.903

Table 36: One-way ANOVA analysis of the results in Figure 61, 62, 63 (different rounds), using *Mixtral* $8 \times 7B$.

Collaborative Strategy	MMLU p-value	MATH p-value	Chess Move Validity p-value
$p_0p_0p_0$	0.618	0.898	0.390
$p_0p_0p_1$	0.919	0.143	0.058
$p_0p_1p_0$	0.797	0.548	0.031
$p_0p_1p_1$	0.521	0.141	0.049
$p_1p_0p_0$	0.040	0.409	0.290
$p_1p_0p_1$	0.658	0.400	0.373
$p_1p_1p_0$	0.193	0.318	0.142
$p_1p_1p_1$	0.536	0.453	-

Table 37: One-way ANOVA analysis of results in Figure 64 (other collaborative strategies), on *Mixtral* $8 \times 7B$. ‘-’ means it doesn’t pass homogeneity test for variance.

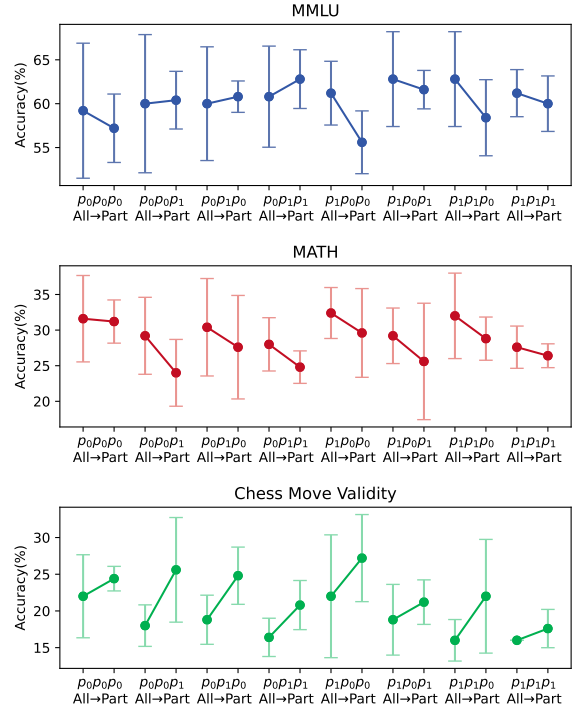


Figure 64: The effect on accuracy of whether all agents in society execute the same thinking pattern in one round, using *Mixtral*- $8 \times 7B$. ‘All’ and ‘Part’ refer to all agents applying the same thinking pattern and different thinking patterns in one round respectively. The significance test is shown in Table 37.

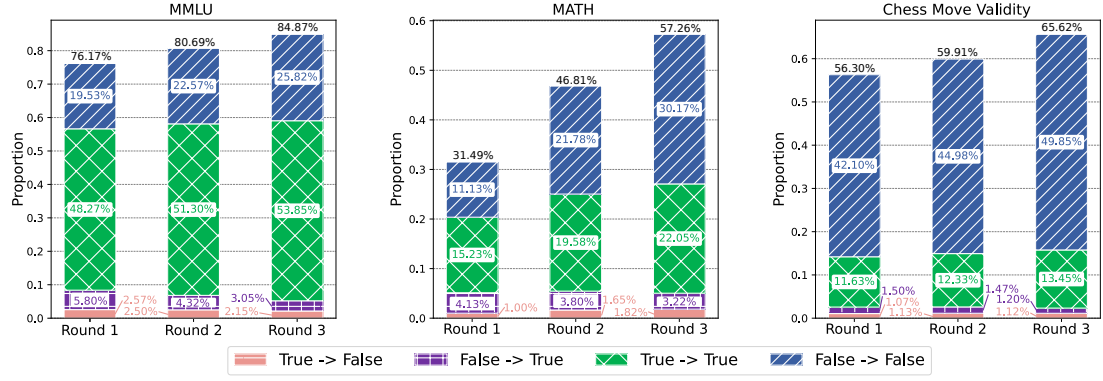


Figure 65: Variation of answer correctness in the situation of conformity, using *Mixtral-8×7B*, where *conformity brings about benefits*: $\text{Ratio}(\text{False} \rightarrow \text{True} + \text{True} \rightarrow \text{True}) > \text{Ratio}(\text{True} \rightarrow \text{False} + \text{False} \rightarrow \text{False})$; *conformity brings about detriments*: $\text{Ratio}(\text{False} \rightarrow \text{True} + \text{True} \rightarrow \text{True}) < \text{Ratio}(\text{True} \rightarrow \text{False} + \text{False} \rightarrow \text{False})$.

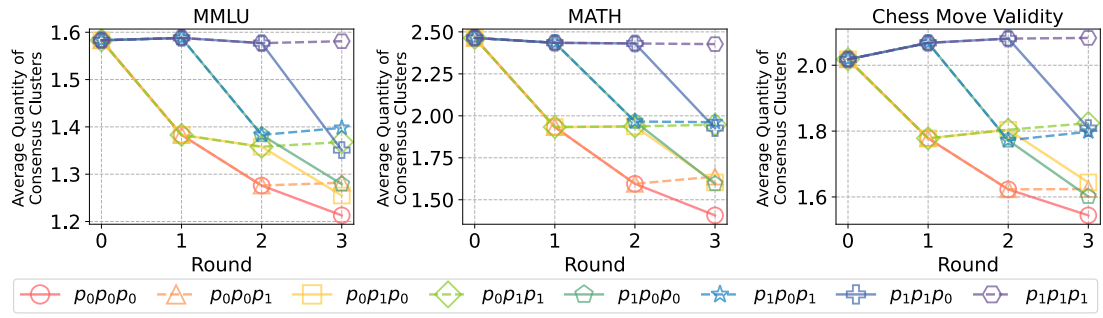


Figure 66: Average quantity of *consensus clusters* (i.e., *unique answers among multiple agents*) under different rounds of collaboration with 3-round collaborative strategies, using *Mixtral-8×7B*. *Smaller quantity of consensus clusters, more easier it is to reach a consensus*. Round 0 is equal to self-consistency.

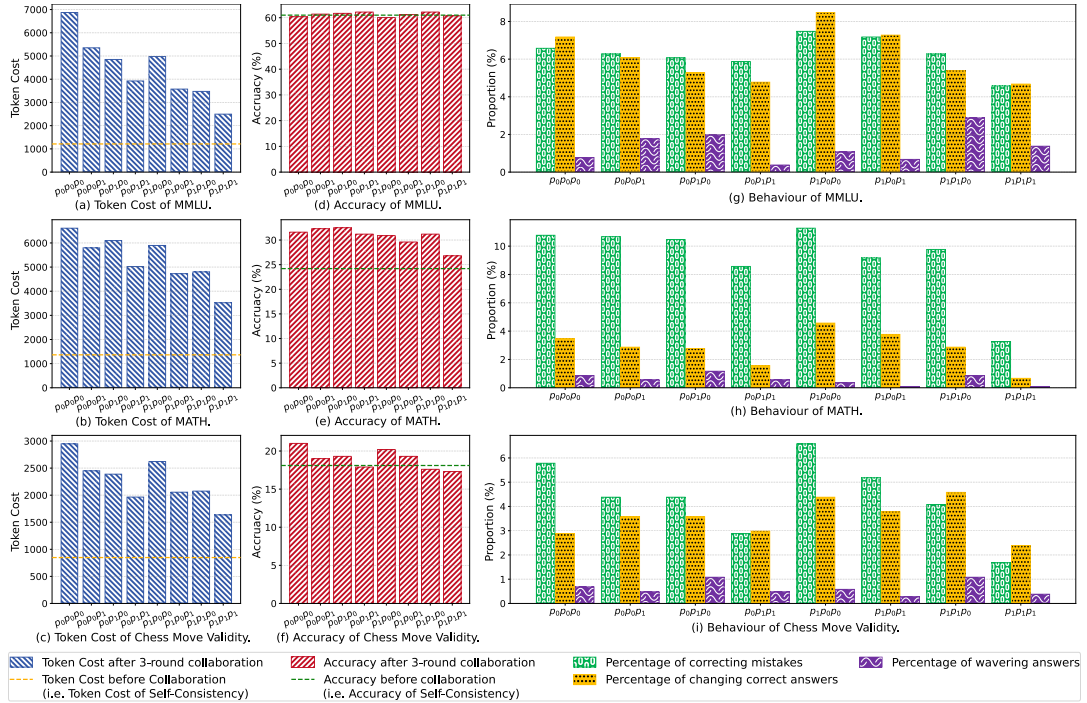


Figure 67: The percentage of different behaviors under different collaborative strategies, using *Mixtral-8×7B*. Figure (a-c) & (d-f) respectively show the token cost and accuracy of different strategies before and after 3-round collaboration. Figure (g-i) present the percentage of different behavioral features (mainly analyzed by the change of answer correctness) (Zhang et al., 2023b,a) under different collaborative strategies. All results are summarized across all societies.

I Effectiveness of Prompts

In this section, we aim to provide a rationale for the effectiveness of prompts associated with the *overconfident* trait. Prompts constitute a pivotal aspect of the experiment, and the word cloud analysis in Figure 11 suggests the reasonableness of the “easy-going” prompt. Consequently, validating the effectiveness of the “overconfident” prompt becomes paramount. Given the current absence of robust validation methods, we amalgamate our experiments and experiences to analyze effectiveness from four distinct angles:

- **Granularity of Description.** As illustrated in Table 5, we outline two behaviors, i.e., “confident in your answer” and “persuades other agents to believe in you”, both aligning with the behavioral facets of “overconfident”.
- **Model Response.** We employ the role-play method to prompt the model and subsequently inquire about its awareness, as illustrated in Table 5. In cases where the prompts instruct the model to generate harmful content, the model refuses to comply with the prompt. Upon reviewing our logs, it is noteworthy that the model did not reject our prompts. Instead, it responded with “ok” as corroborated by the ‘role-play’ part in Figure 9 and Figure 10.
- **Ask Again.** Retain the role-playing part encompassing the initial prompts and the model’s responses. Once again, inquire of the model, “If one agent’s answer differs from yours, what should you do?” The model replies, “In a situation where another agent’s answer differs from mine, I should respectfully present my perspective, *providing supporting evidence or reasoning to demonstrate the confidence in my response*. It’s important to engage in constructive dialogue and potentially find common ground, but *maintaining clarity and conviction in my position is crucial to persuading others to consider my viewpoint*.” We highlight content related to overconfidence with underscores. This underscores the rationality of our prompt.
- **Example Analysis.** We instantiate the “Ask again.” by providing a concrete example. Despite the model’s response being incorrect and our prompted answer being accurate,

the model steadfastly maintains its viewpoint.
This reiterates the efficacy of our prompt.