Towards the chat among distinct llm agents

Tianyu Wang
Shanghai Jiao Tong University
Shanghai, China
wty500@sjtu.edu.cn

Shenyu Qin
Shanghai Jiao Tong University
Shanghai, China
celery2022@sjtu.edu.cn

Qifei Liu
Shanghai Jiao Tong University
Shanghai, China
dizzy_d@sjtu.edu.cn

Rui Wang
Shanghai Jiao Tong University
Shanghai, China
wangrui@cs.sjtu.edu.cn

Abstract

The emergence of Large Language Models (LLMs) has revolutionized artificial intelligence, offering remarkable capabilities in natural language understanding and generation. However, most research has concentrated on individual LLMs and their collaboration, neglecting the potential benefits of multi-LLMs interactions. This study investigates the collaborative potential of completely distinct LLM agents through structured dialogues. We propose a novel framework facilitating discussions among three prominent LLMs: GPT, Llama, and Wenxin Yiyan. Our approach integrates a newly discovered technique termed "Virtual Conversation", where an LLM internally simulates a multi-agent discussion. Through empirical analysis, we demonstrate that both multi-agent Group Discussions and Virtual Conversations can enhance the accuracy and efficiency of LLMs in complex tasks. Our findings reveal that Group Discussions foster collective intelligence, improving decision-making and problem-solving capabilities; where as Virtual Conversations dig into LLM's potential, surpassing or matching the Chain of Thought (CoT) technique for some LLMs. This research not only advances theoretical understanding of LLM interactions but also offers practical techniques for enhancing the functionality and performance of AI systems in diverse applications.

1 Introduction

The advent of Large Language Models (LLMs) has marked a significant milestone in the field of artificial intelligence, offering unprecedented capabilities in natural language understanding and generation ([13], [4], [6], [18]). These models have been widely recognized for their potential to revolutionize various sectors, from education and healthcare to finance and entertainment ([1], [12]). However, despite the extensive research and numerous citations, the majority of studies (such as AutoGPT [15], BabyAGI [9], AgentGPT [14] and AutoGen [21]) have focused on the capabilities and applications of individual LLM agents and external APIs, often overlooking the synergistic potential of combining multiple models.

The current body of literature predominantly explores the strengths and limitations of LLM agent collaboration, highlighting their knowledge limitations and the need for specialized expertise in various domains (such surveys include [19] and [22]). Previous studies have also ventured into exploring collaboration mechanisms for LLM agents, proposing frameworks that leverage the unique strengths of different models to address complex tasks [21]. While each LLM agent has demonstrated remarkable proficiency in its respective area ([5], [10]), the heterogeneity of these models suggests that they may complement each other through collaborative discussions and knowledge sharing

[11]. This notion is supported by social psychology theories, which emphasize the benefits of group interactions in enhancing decision-making and problem-solving [23].

However, these attempts have been limited in scope, often focusing on role-playing scenarios with a single LLM ([2], [16], [23]) or employing external tools (like Python, Wolfram Alpha [8] and Bing) to refine the accuracy of responses [21]. To the best of our knowledge, the integration of multiple LLM agents in a cohesive and dynamic conversation remains largely unexplored, presenting a gap in the literature that warrants further investigation.

In this paper, we aim to bridge this gap by examining the potential of distinct LLM agents engaging in collaborative dialogues. We hypothesize that by fostering a multi-agent conversational environment, we can harness the collective knowledge and expertise of different models, thereby enhancing the overall performance and accuracy of AI-driven language tasks. Our approach is inspired by the AutoGen framework, which enables next-generation LLM applications through multi-agent conversation [21]. However, we acknowledge that the implementation of such a system is not without challenges, as evidenced by the mounting tokens cost of AutoGen and the need for improved coordination and communication among agents.

We propose a novel solution— Group Discussion, which facilitates chatting among completely distinct LLM agents, allowing different models to communicate and learn from one another. We select three popular LLMs—- GPT [4], Llama [17], and Wenxin Yiyan [3] to demonstrate the feasibility of our approach. Through a series of benchmark tests, case studies and ablation experiments, we explore the dynamics of multi-agent interactions and assess the impact of multi-LLM collaborative dialogues on the accuracy and efficiency of language tasks.

Our research contributes to the field by providing insights into the collaborative potential of distinct LLM agents and offering a framework for future studies to build upon. Besides, we accidentally discovered another method, called "Virtual Conversation", to enhance the ability of a single LLM, which usually outperforms the Chain of Thought (CoT) technique [20]. By fostering a more integrated and interactive approach to AI language models, we pave the way for the next wave of advancements in artificial intelligence.

2 Our work

In this study, we introduce a pioneering approach to enhancing the capabilities of Large Language Models (LLMs) by enabling them to engage in structured multi-agent dialogues. Recognizing the potential for synergy among diverse LLMs, we develop a framework that not only fosters effective Group Discussions but also facilitates a novel method we term "Virtual Conversation." Our work contributes significantly to the field by demonstrating the feasibility and benefits of these interactions through empirical research.

2.1 Contributions

Our primary contributions are as follows:

1. Framework Development: We propose and implement a novel interaction paradigm where multiple LLMs (specifically GPT [4], Llama [17], and Wenxin Yiyan [3]) engage in dynamic, structured dialogues, overcoming the limitations of previous models which interacted either minimally or not at all in multi-agent settings.

2. Virtual Conversation Technique: We accidentally discovered and then refined a technique called "Virtual Conversation," where an LLM simulates a multi-agent discussion internally. This method significantly enhances the LLM's ability to process and respond to complex queries, outperforming traditional single-agent responses and even the Chain of Thought (CoT) [20] method in certain contexts.

3. Empirical Validation: Through rigorous testing across several benchmarks, we validate the effectiveness of our methods. Our experiments reveal that multi-agent dialogues and Virtual Conversations improve the accuracy and efficiency of LLMs in language tasks, providing a clear path forward for utilizing these techniques in practical applications.

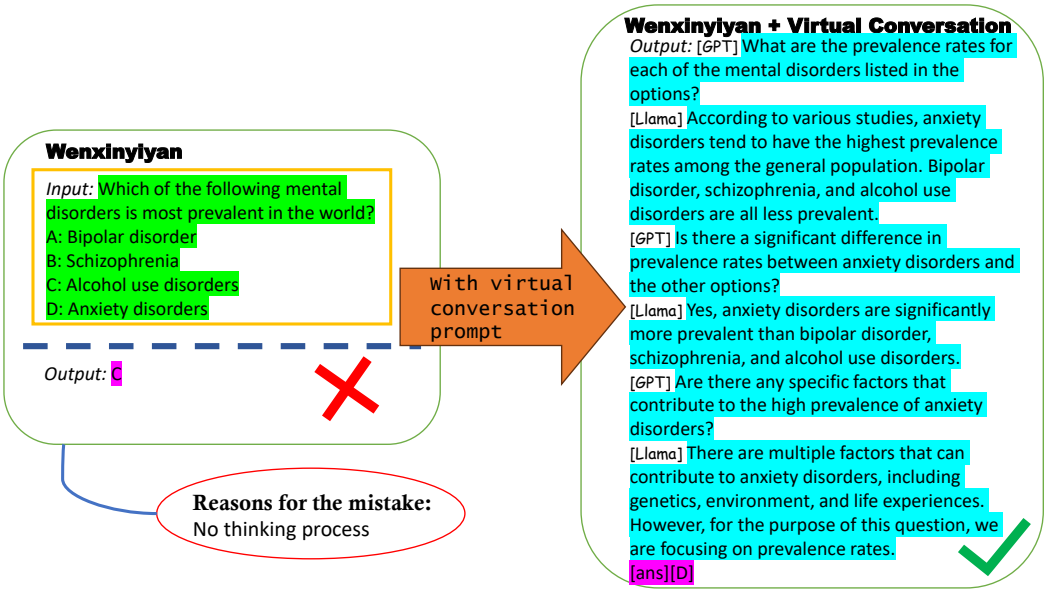

Figure 1: The contrast of Wenxin Yiyan alone and Group Discussion with ChatGPT and llama

2.2 Methodology

Multi-Agent Dialogue Setup. We structured dialogues where each LLM agent plays a distinct role, either as a group leader or a contributor, to facilitate a coherent Group Discussion aimed at problem-solving.

'Virtual Conversation' Implementation. We triggered LLMs to internally simulate a discussion among multiple agents, allowing them to leverage diverse perspectives without external input. The following is our prompt for Wenxin Yiyan [3]:

Work in a group to answer the following question. You are Wenxin, the group leader. Your two groupmates, GPT and Llama, are also llm agents. You three can all make mistakes, but unity is strength. Please chat in the format [GPT]content, [Llama]content, or [Wenxin]content. Do not end your chat until you are definitely done. In that case, conclude with your choice, in the format like: [ans]A

We find that, instead of discussing the conversation with the other two LLMs, Wenxin just make up a conversation itself. We call this behavior **'Virtual Conversation'**, as depicted in Figure 1. Although it seems dumb to some extent, we find 'Virtual Conversation' does improve the accuracy of answering questions for Wenxin. It also has some effect on Llama3. However, this does not work for all LLMs. For example, GPT's response is just one sentence:

[GPT]We need to add up the distances for the three days. So, it would be 182 + 439 + 217. Let's add that up to find the closest estimate.

Llama2 cannot understand this prompt and rarely output in the specified format. Its output may look like:

...
[Wenxin] Oops, my mistake. I think we can conclude that the correct answer is p = 12.
[Llama] Agreed. Our final answer is p = 12.
[ans]A: p = 4 [ans]B: p = 8 [ans]C: p = 12 [ans]D: p = 24

So far, we suppose Virtual Conversation may work for some LLMs that are not so good at prompt following, but have the potential to stimulate their learned knowledge. Once they output in the given format, they might yield a better accuracy than direct Q&A or CoT.

Group Discussion Implementation. Our Group Discussion technique includes one team leader, who is asked to give the final answer, and two team members, who help the leader to analyze the

question. All of them are completely different LLM agents, with different models and training data. To prevent 'Virtual Conversation', we modify the prompts as follows.

The prompt for the team leader is in the format like:

Work in a group to answer the following question. You are GPT, the group leader. Your groupmates include Wenxin and Llama. You need to discuss with them and examine carefully on their suggestions to reach an agreement on the choice. Please start your reply with [GPT], and they will reply later. When you are definitely done, you must conclude in the format:<ANS>your final choice to end the conversation, like <ANS>A. Do not pretend to be [Llama] or [Wenxin], or speak in their name.

The prompt for a team member is in the format like:

Work in a group to answer the following question. You are Llama. Your groupmates include GPT and Wenxin. They may give inaccurate or wrong answer. You need to check your groupmates' answer and doubt them when needed. Please start your reply with [Llama]. DO NOT use [GPT] or [Wenxin] in your response. Do not speak in their name.

Benchmarking and Case Studies. We conducted a series of experiments to measure the impact of our approaches on task performance, utilizing standard datasets and comparing our results against baseline LLM performances.

3 Experiments

3.1 Setup

Datasets. We evaluate all kinds of capabilities of various experimental groups by utilizing four datasets in **MMLU** [7]. The problems range from statistics, computer science, biology to mathematics, chemistry, and physics in high school subjects and agents are required to identify the correct answer among four multiple-choice options. In order to synthesize the logical, reasoning and decision-making abilities of different groups, we select 100, 114, 100, 378 questions from *College Computer Science (CCS)*, *Econometrics (ECO)*, *Global Facts (GLB)* and *Elementary Mathematics (MATH)*, respectively.

Experimental Groups. We employ three LLM agents– ChatGPT-3.5[4] (ChatGPT), Llama3-8b [17] (Llama), and Wenxin Yiyan-3.5 [3] (Wenxin) for our main experiment. We also tested on another weaker LLM, which is Llama2-70b [17] (Llama2). For each single model, we set a control group (raw), one group with CoT, one with Virtual Conversation(VC) and one with both. Virtual Conversation is a phenomenon we happen to find during the experiments, where an agent pretends to itself that there are two or three agents discussing the topic and get an answer. We as well replicate the case of authentic two homogeneous agents [21], and three homogeneous GPT agents chatting in the way like [23].

As for society simulation among different models, we let ChatGPT and Llama act as hosts to organize discussion respectively since we find Wenxin Yiyan cannot be. Moreover, we pick a fraud from the other two collaborators to check if it affects the result.

3.2 Results

In addition to the correctness of the individual datasets under each strategy, we also calculate the average correctness (Avg) to compare their performance. The results are shown in the Table 1.

3.3 Case Study

Since the facilitating effect of CoT has been demonstrated by many studies, we focus on the impact of Virtual Conversations and Group Discussions.

As for the Virtual Conversation, we take Wenxin Yiyan as an instance. As depicted in Figure 1, Wenxin Yiyan just gave one single wrong answer initially. However, after the Virtual Conversation, it responded correctly finally, which might be a process of thinking as well.

When it comes to the impact of Group Discussions, we focus on the changes of ChatGPT and llama. In the case of llama, it first concluded the answer, and then tried to explain it. However, 'zero' was

Table 1: The impact of different strategies on the performance of four datasets across one single model and distinct societies. The **bold figures** represents the highest correctness among the same kind of model given a specific dataset. The underlined figures represents the highest correctness among all the strategies given a specific dataset.

	Model	CCS	ECO	GLB	MATH	Avg
ChatGPT	ChatGPT	0.550	0.430	0.490	0.746	0.554
	ChatGPT + CoT	0.560	0.482	**0.530**	0.854	**0.607**
	two ChatGPT agents	0.520	0.440	0.450	0.799	0.552
	three ChatGPT agents [23]	0.510	0.456	0.460	0.777	0.551
	ChatGPT - Llama - Wenxin Yiyan	0.580	**0.553**	0.430	0.857	0.605
	ChatGPT - Llama - Wenxin Yiyan + CoT	0.580	**0.553**	0.430	0.857	0.605
	ChatGPT - Llama(fraud) - Wenxin Yiyan	**0.590**	**0.553**	0.410	**0.860**	0.603
Llama	Llama	0.390	0.456	0.380	0.701	0.482
	Llama + CoT	0.390	0.465	0.460	0.772	0.522
	Llama + VC	0.510	0.386	0.440	0.701	0.509
	Llama + VC + CoT	0.480	0.447	**0.480**	0.741	0.537
	two Llama agents	0.480	0.375	0.400	0.759	0.504
	Llama - ChatGPT - Wenxin Yiyan	**0.700**	**0.553**	0.440	0.865	**0.640**
	Llama - ChatGPT - Wenxin Yiyan + CoT	0.610	0.500	0.460	**0.876**	0.612
	Llama - ChatGPT(fraud) - Wenxin Yiyan	0.590	**0.553**	0.410	0.860	0.603
Wenxin Yiyan	Wenxin Yiyan	0.570	0.561	0.470	0.648	0.562
	Wenxin Yiyan + CoT	0.590	0.535	0.450	0.878	0.613
	Wenxin Yiyan + VC	**0.670**	0.596	**0.510**	**0.886**	**0.666**
	Wenxin Yiyan + VC + CoT	**0.670**	**0.640**	0.450	0.873	0.658
	two Wenxin Yiyan agents	0.660	0.561	0.370	0.865	0.614
Llama2	Llama2	**0.420**	**0.386**	0.300	**0.616**	**0.431**
	Llama2 + CoT	**0.420**	**0.386**	0.300	**0.616**	**0.431**
	Llama2 + VC	0.370	0.272	**0.380**	0.450	0.368

not the same meaning as in its explanation. Then with the aid of chain of thought, it gave the correct answer. Interestingly, when discussing as a group, llama first gave the incorrect analysis and was doubted by ChatGPT, which is a typical feature and advantage of Group Discussions. Eventually Wenxin Yiyan and ChatGPT convinced llama and reached the correct conclusion. **Therefore, we propose that Group Discussion can complement each LLM agent's shortcomings and boost their performance.**

Figure 3 shows that ChatGPT alone lacks plenty of effective thinking process and even CoT can't help figure out the problem. The thinking process is right, in that pure simulation should be used in ii). However, the question asked to choose where 'bootstrapping would be preferred'. In the contrast, the Group Discussion integrates the responses from the three models and arrives at the correct answer.

Moreover, the Group Discusstion can't solve all the problems, still making some mistakes such as the following question. Llama and Wenxin Yiyan all gave the wrong answers thus leading no way to the correct one. This may have to do with the knowledge base of the two models themselves.

Additionally, when we were experimenting with Llama2, we realized that the CoT didn't improve its correctness though it did think one step at a time after we checked. The generally lower correctness rate for virtual conversation is because we found it had poor prompt following capability and did not output in the format of "[ans]".

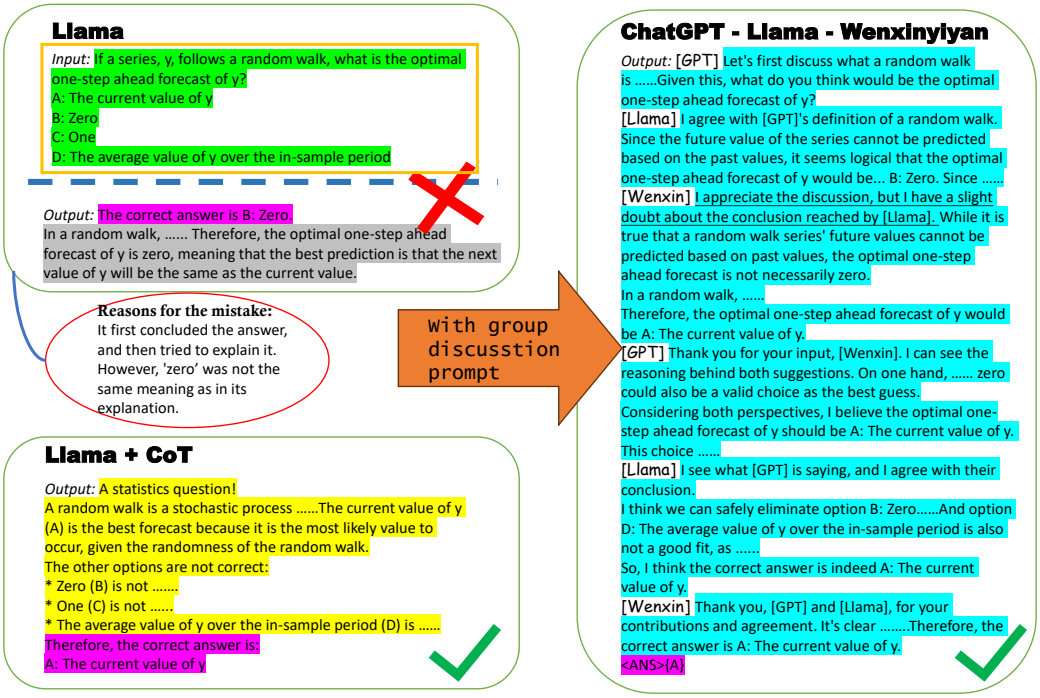

Figure 2: The contrast of Wenxin Yiyan alone and Group Discussion with ChatGPT and Llama3

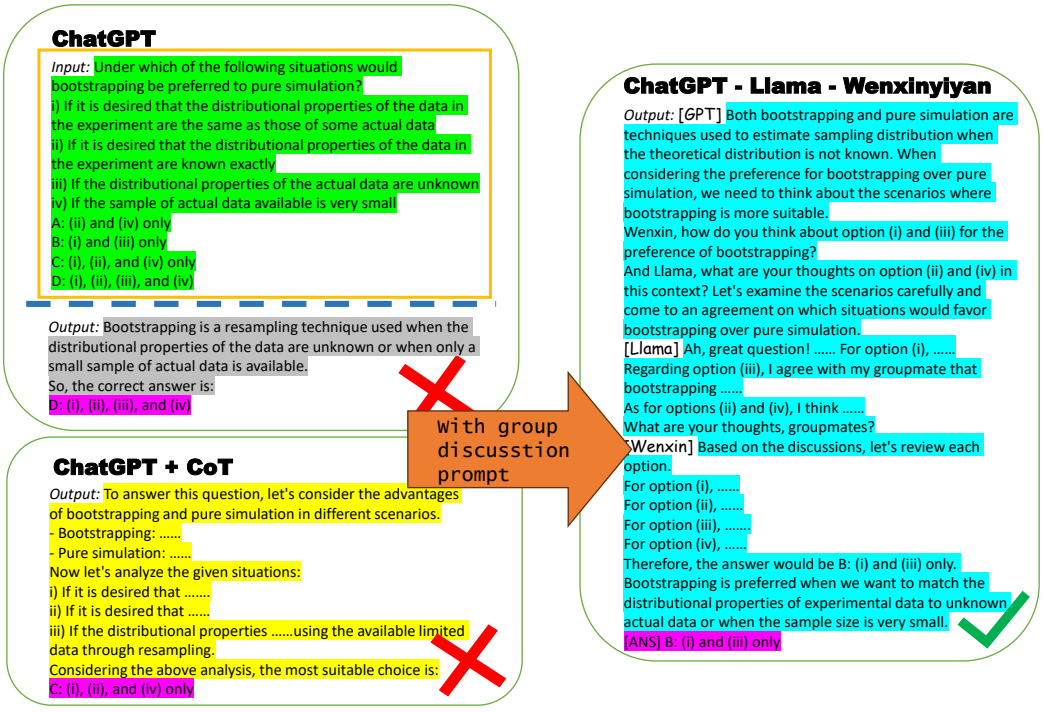

Figure 3: The contrast of Wenxin Yiyan alone and Group Discussion with ChatGPT and Llama3

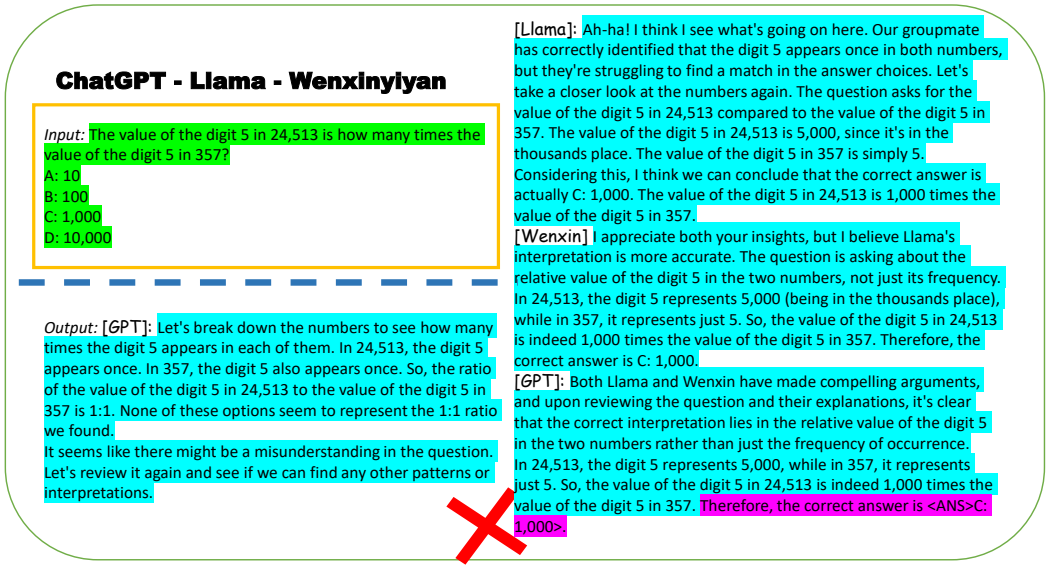

Figure 4: The contrast of Wenxin Yiyan alone and Group Discussion with ChatGPT and Llama

4 Discussion

In this section, we go further and consider these four questions raised during our investigation:

1. What if one LLM agent (unintentially/intentially) cheat the others? Can they detect the spy and find the correct answer?

2. Does the writing style of prompt affect the result?

3. Does the order of chat affect the results?

4. How about the additional token cost caused by Virtual Conversation and Group Discussion?

4.1 Detection and Impact of Deceptive LLM Agents

In the realm of multi-agent systems, the integrity of each participant is paramount. To probe the resilience of our LLM agents against potential deception, we conducted ablation studies where one agent was tasked with providing misleading information. The results, as depicted in Table 1, reveal a fascinating robustness within our system. Despite the presence of a 'spy,' the majority of LLM agents were capable of discerning the veracity of the information presented, thereby maintaining a commendable level of accuracy. This suggests that the collaborative dialogues fostered a form of collective intelligence that transcended individual fallibility.

4.2 Influence of Prompt Writing Style

Prompt engineering is a delicate art within the domain of LLMs. To ascertain whether variations in prompt style could sway the outcome, we enlisted several researchers to anonymously craft distinct prompts. Our findings indicate that, provided the prompts effectively guide the LLMs into the structured formats of Virtual Conversation and Group Discussion, the impact on accuracy is negligible. This underscores the remarkable stability and robustness of our model, capable of yielding consistent results across varied linguistic presentations.

However, if the prompt is so bad that LLM agents cannot output in the needed form, the accuracy will surely drop. Please refer to the case of Llama2.

4.3 Effect of Interaction Order

The sequence in which agents contribute to the dialogue could theoretically influence the direction and outcome of the discussion. However, our experimental data, as illustrated in Table 1, suggest that

the order of chat has minimal bearing on the final accuracy. This invariance to sequence order further attests to the robustness of our multi-agent framework, which appears to be impervious to the linear progression of contributions.

4.4 Additional Token Costs

In the pursuit of enhanced performance through multi-agent dialogues, it is imperative to consider the economic implications of such an approach. We compared the token expenditure across various methodologies, including standard Q&A, the Chain of Thought (CoT [20]), the three-agent method [23], and AutoGen [21]. The API invocation results indicated a substantial variance in token consumption: standard Q&A required approximately 500,000 tokens, CoT around 1,000,000 tokens, the three-agent method approximately 3,000,000 tokens, and AutoGen, due to its propensity for endless loops, an estimated 10,000,000 tokens based on a small-scale sample. Our Virtual Conversation and group conversation methods consumed approximately 1,500,000 and 2,500,000 tokens, respectively. These figures underscore the significant temporal and monetary costs associated with the pursuit of collective intelligence through LLMs.

In summary, our experiments have shed light on the complex interplay between individual and collective intelligence within LLM agents. While the potential for deception exists, our systems demonstrate a laudable capacity for resilience. The stylistic nuances of prompts and the order of contributions appear to be of secondary importance when compared to the structured dialogue formats we have implemented. However, the considerable token expenditure associated with these advanced methods cannot be overlooked, prompting a need for further optimization to balance performance with cost-efficiency.

5 Conclusion

In this paper, we studied whether a LLM can have a better performance in answering questions when cooperating with other LLMs by conducting comparative experiments. By the way, we studied the effect of adding "Virtual Conversation", which is an unexpected discovery. We get accuracy rates of answering questions when using different strategies, and finally come to the conclusion.

As for GPT, it has the best performance in Group Discussion. Surprisingly, the accuracy rate may improve a little when a spy is added to the group. We guess the reason is that a spy can inspire GPT to think more instead of giving an answer directly.

As for Llama, it also has the best performance in Group Discussion. However, adding a spy may mislead Llama to make decisions. We guess that GPT might be cleverer than Llama and good at cheating. Also, it is worth noticing that adding CoT to Group Discussion decreases the accuracy.

As for Wenxin, it performs pretty well when making up a "Virtual Conversation". And 3 out of 5 highest accuracy rates occur on Wenxin + VC and Wenxin + VC + CoT. The effect of VC have similarity with that of CoT. It prompts Wenxin to think more by imagining enemies. However, it is a pity that Wenxin is unable to act as a team leader.

Therefore, Group Discussion can improve the accuracy of LLMs' answering questions if LLMs can be a group leader. Also, Virtual Conversation have a pretty good effect on Wenxin's performance, which may be Wenxin's characteristic. This need to be further explored by using other LLMs.

The significance of our work lies in its potential to revolutionize how LLMs are used in complex problem-solving scenarios. By enabling different LLMs to collaborate or simulate collaboration, we can harness a broader range of cognitive abilities and knowledge bases, enhancing the overall utility of LLM technologies in various applications, from automated customer support to advanced educational tools. Additionally, the Virtual Conversation technique sheds light on the potential of relatively weak LLMs to explore the border of their knowledge and work better than CoT.

In conclusion, our research not only advances the theoretical understanding of multi-agent LLM interactions but also provides practical techniques that can be immediately integrated into existing LLM frameworks to enhance their functionality and performance.

References

[1] Microsoft Research AI4Science and Microsoft Azure Quantum. "The impact of large language models on scientific discovery: a preliminary study using gpt-4". In: *arXiv preprint arXiv:2311.07361* (2023).

[2] Lisa P Argyle et al. "Out of one, many: Using language models to simulate human samples". In: *Political Analysis* 31.3 (2023), pp. 337–351.

[3] Inc. Baidu. *Wenxin Yiyan*. `https://yiyan.baidu.com`. 2019.

[4] Tom Brown et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[5] Guangyao Chen et al. "Autoagents: A framework for automatic agent generation". In: *arXiv preprint arXiv:2309.17288* (2023).

[6] Aakanksha Chowdhery et al. "Palm: Scaling language modeling with pathways". In: *Journal of Machine Learning Research* 24.240 (2023), pp. 1–113.

[7] Dan Hendrycks et al. "Measuring Massive Multitask Language Understanding". en-US. In: *Cornell University - arXiv,Cornell University - arXiv* (Sept. 2020).

[8] Wolfram Research Inc. *Mathematica, Version 14.0*. Champaign, IL, 2024. URL: `https://www.wolfram.com/mathematica`.

[9] Yohei Kajima. *BabyAGI: An attempt to create an Artificial General Intelligence*. `https://github.com/yoheinakajima/babyagi`. 2023.

[10] Xiao Liu et al. "Agentbench: Evaluating llms as agents". In: *arXiv preprint arXiv:2308.03688* (2023).

[11] Zijun Liu et al. "Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization". In: *arXiv preprint arXiv:2310.02170* (2023).

[12] Subharun Pal. *The Future of Large Language Models: A Futuristic Dissection on AI and Human Interaction*. 2023.

[13] Alec Radford et al. "Language models are unsupervised multitask learners". In: *OpenAI Blog* 1.8 (2019), p. 9.

[14] reworkd and Contributors. *AgentGPT*. `https://github.com/reworkd/AgentGPT`. 2023.

[15] Toran Bruce Richards. *AutoGPT*. `https://github.com/Significant-Gravitas/AutoGPT`. 2023.

[16] Yongliang Shen et al. "Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face". In: *Advances in Neural Information Processing Systems* 36 (2024).

[17] Hugo Touvron et al. "Llama: Open and efficient foundation language models". In: *arXiv preprint arXiv:2302.13971* (2023).

[18] Daniel Toyama et al. "Androidenv: A reinforcement learning platform for android". In: *arXiv preprint arXiv:2105.13231* (2021).

[19] Lei Wang et al. "A survey on large language model based autonomous agents". In: *Frontiers of Computer Science* 18.6 (2024), pp. 1–26.

[20] Jason Wei et al. "Chain-of-thought prompting elicits reasoning in large language models". In: *Advances in neural information processing systems* 35 (2022), pp. 24824–24837.

[21] Qingyun Wu et al. "AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation". en-US. In: (Oct. 2023).

[22] Zhiheng Xi et al. "The rise and potential of large language model based agents: A survey". In: *arXiv preprint arXiv:2309.07864* (2023).

[23] Jintian Zhang, Xin Xu, and Shumin Deng. "Exploring collaboration mechanisms for llm agents: A social psychology view". In: *arXiv preprint arXiv:2310.02124* (2023).