# **Brainstorming Brings Power to Large Language Models of Knowledge Reasoning**

Zining Qin, Chenhao Wang, Huiling Qin\*, Weijia Jia Beijing Nomal University orekinana@gmail.com

# **Abstract**

Large Language Models (LLMs) have demonstrated amazing capabilities in language generation, text comprehension, and knowledge reasoning. While a single powerful model can already handle multiple tasks, relying on a single perspective can lead to biased and unstable results. Recent studies have further improved the model's reasoning ability on a wide range of tasks by introducing multi-model collaboration. However, models with different capabilities may produce conflicting answers on the same problem, and how to reasonably obtain the correct answer from multiple candidate models has become a challenging problem. In this paper, we propose the multi-model brainstorming based on prompt. It incorporates different models into a group for brainstorming, and after multiple rounds of reasoning elaboration and re-inference, a consensus answer is reached within the group. We conducted experiments on three different types of datasets, and demonstrate that the brainstorming can significantly improve the effectiveness in logical reasoning and fact extraction. Furthermore, we find that two small-parameter models can achieve accuracy approximating that of larger-parameter models through brainstorming, which provides a new solution for distributed deployment of LLMs.

# 1 Introduction

With the accumulation of data and the significant increase in computational power, large language models (LLMs) rely on their vast training data and scale to show excellent capabilities in language understanding, text generation, and knowledge reasoning[3, 4, 20]. Although existing LLMs can solve many different types of tasks, the reasoning bias caused by the limitations and instability of a single model's perspective still cannot be ignored. Recent research indicates that fusing multiple models to work collaboratively can further enhance reasoning capabilities on a wide range of tasks [5, 13, 16, 19, 25, 27, 31]. For example, one promising approach is to incorporate multiple agents generated by an LLM into a group after substituting different roles and designing a special interaction mechanism [10], and it can improve the diversity of reasoning while allowing agents to provide more stable and reliable results through debates in the interactions [6].

However, in the knowledge reasoning scenario that emphasizes the availability of definitive answers, the roles played by agents for the same type of scientific questions tend to be uniform. In such a case, the reasoning perspective of agents based on the single model becomes narrow, necessitating the involvement of models with different capabilities to improve the overall reasoning performance. When multiple models reasoning, due to the differences in the training data and training process, models with different capabilities may provide different answers to the same question. Therefore, in the face of conflicting results of reasoning, it becomes a critical and challenging problem to formulate

a rule to ensure that the group provides a uniform answer from multiple candidates. Figure 1 shows three natural interaction methods to obtain the final answer.

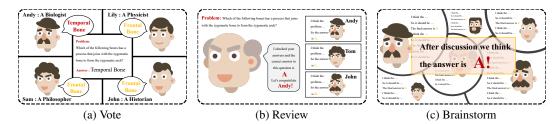


Figure 1: Differences in getting answers through voting, strong model judgments (review), and brainstorming of many different types of collaboration on large language models.

The most straightforward approach is voting, as shown in Figure 1 (a). Multiple models provide what they consider to be the correct answer, and the final answer is then determined according to the majority rule after counting the number of votes for each candidate answer [17, 33]. It is noticed that voting-based approaches tend to lack interpretability, and if the ability of a few models is significantly stronger than that of the majority, then the result obtained by voting may miss the correct answer. A further step can be taken by introducing an additional discriminative model as a reviewer [22] that evaluates multiple answers and then gives the final answer, shown as Figure 1 (b). However, the reviewer requires a higher level of knowledge and judgment than models in the group, and it is very difficult to construct such a discriminative model.

Drawing inspiration from the collective thinking and reasoning process of human beings [23], we propose a prompt-based multi-model brainstorming approach, as shown in Figure 1 (c). It allows each model to substitute a similar expert role and iteratively incorporate the reasoning process of other models into their thinking scope to update their answers, until a consensus is reached. Furthermore, in order to enhance the diversity of group thinking, we select the model with some differences in performance on different tasks for brainstorming. This ensures that each model is capable of bringing novel ways of thinking to knowledge reasoning. In this way, models enable thinking from different perspectives and capture issues that are difficult to identify from a single viewpoint.

The main contributions of our work are stated as follows:

- We propose a prompt-based multi-model brainstorming that has a more nuanced mindset than single model and the extended multi-agent debate approaches, which enhances the reasoning accuracy in both logical and factual terms.
- We utilize multi-model brainstorming instead of manual prompts in Chain of Thought (CoT)[15, 29] to reduce the cost of manual labeling. We use two small models with brainstorming to replace one large model, and exhibit approximate, even higher accuracy in knowledge reasoning in diversified datasets, providing a new solution for distributed deployment of large models.
- We find that the LLM is able to integrate different viewpoints in the brainstorm. The redundant historical dialogues or an additional summary are unnecessary to the reasoning, which avoids the problem of reduced reasoning ability due to excessive space occupation in multiple generation.

#### 2 Related Work

Prompts Enhancement in LLM Reasoning. Due to the versatile capabilities that LLMs demonstrate, researchers have leveraged LLMs as the foundation to build AI agents that can communicate with human and other LLMs and have achieved significant progress [30], including improvements via various prompts[18, 14, 26, 31, 34]. Few shot and zero shot Chain-of-thought (CoT) [15, 29] encourage LLMs to analyze the input question step by step following the procedure given by human in reasoning and generation, making the responses readable and interpretable to human and other LLM agents. Wang et al. [28] propose self-collaboration method to enable the communication of agents by assigning multi-persona to a single LLM via prompts. Zheng et al. [35] enable LLMs to "recall" and derive high-level concepts and first principles from question instances by applying Step-back prompting. With additional hints respect to the question generated by LLM agent itself,

the agent gain more chances of utilizing such principles properly and answering correctly. Our work also adopts some prompting techniques to enhance the quality of LLMs' responses.

Interaction Framework for Multiple LLMs. Compared to models that only take vectors specifically designed for them as input, LLMs bridge the gap between models and human using natural language via their tokenizers. Thanks to this property researchers can try various frameworks and strategies to combine multiple LLMs for downstream tasks. Mixture of Experts [12] trains a router network to select a subset composed of parts of LLMs (experts) to process the inner states and combine their outputs. Li et al. [17] conduct a series of experiments on majority voting among large numbers of LLMs and the performance of voting relies considerably on the great number of agents. Li et al. [16] and Chen et al. [7] propose a cooperative framework facilitating LLMs' autonomous cooperation to solve complex tasks. Chan et al.[6] build a multi-agent referee team called ChatEval to autonomously discuss and evaluate the quality of different texts, inspired by best practices of human evaluation processes. Concurrent with our work, Liang et al. [19] and Du et al. [10] also propose a similar method, making use of the multi-agent debate framework in translation, arithmetic and some optimization problems. However, Liang et al. [19] focus on the diversity that emerges from debate rather than the reasoning accuracy, while Du et al. [10] insufficiently explore the effectiveness and efficiency of the natural language interaction among LLMs.

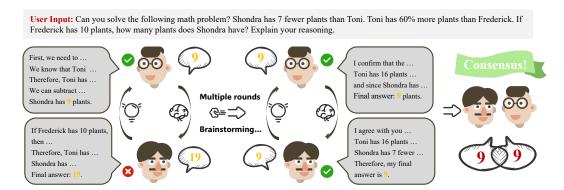


Figure 2: An illustration of LLM Brainstorming. When multiple models are involved in the inference process, they can discuss with each other and eventually reach a consensus judgment.

#### 3 Methedology

In this section, we elaborate on the principal components in LLM Brainstorming including Multimodel brainstorming, consensus in brainstorming, dialog truncating strategy, and provide a detailed overview of each component's role and functionality.

Multi-model Brainstorming. Using multiple heterogeneous LLMs for brainstorming is the key to our method. The reasoning performance of LLMs in various domains is closely related to the corpora used during training. Different organizations are likely to utilize closed-source datasets built by themselves in addition to open-source corpora for training LLMs, resulting in models' performance varying across some domains. Therefore, utilizing different models can provide more wealth of information during brainstorming. In the first round of the brainstorming, we give the question to models without additional contents in prompt. If models response with different answers, we incorporate reasoning processes as thought samples from other models into the prompt for the next round. That is, a model receives responses from the others and, in turn, sends its own responses to them in each round of the brainstorming. The brainstorming process is shown as Figure 2. With other models' reasoning output embedded, responses of models to questions can be regarded as a specific CoT (Chain-of-Thought) with more information related to the question, which can lead models to the correct answer. Note that constructing prompts for each model in brainstorming on any question does not require human intervention. The complete prompt templates are in Appendix ??.

**Consensus in Brainstorming.** The brainstorming process reveals multiple reasoning pathways and answers from LLMs' responses. How can we get the final answer with or without consensus? *Consensus* or *convergence* in LLM brainstorming occurs when all LLMs provide the same answer in

the same round of brainstorming, and the final answer is this same answer. Since there is no guarantee of consensus under all conditions, we adopt a parameter named max\_brainstorming\_round to control the brainstorming. If not all models in brainstorming provide the same answer (such as choices and numerical values), they will start another round of brainstorming until consensus is achieved or the number of brainstorming rounds reaches the max\_brainstorming\_round. As there is no confidence estimation for brainstorming, we treat all the answers generated by all the LLMs throughout the brainstorming process equally without introducing extra bias. When the brainstorming ends without consensus, we record all the answers generated by all the LLMs throughout the brainstorming process and returns the most frequent answer as the final result. This approach ensures that the final answer reflects the collective reasoning and most frequently agreed-upon responses of all the models over multiple rounds. The dialogs of LLM brainstorming are readable to human and we can easily catch up with the reasoning of brainstorming to the final answers.

Dialog Truncation Strategy. Since our method involves multiple rounds of conversation, the context will contain considerable redundancy. To ensure the efficiency and accuracy of the reasoning, we adopt dialog truncation strategies. Besides the first round of brainstorming, which includes the user's question and the reasoning and answers generated in the first round, we only keep the last n rounds of brainstorming dialogs as input for the next round. Different parameter values of n represent different dialog truncation strategies. We note that although LLMs can capture the entire brainstorming without truncation, the linear growth of the length of input tokens and the time consumption of reasoning may lead to unacceptable costs. In the experiments, we set n=1 because duplicate reasoning segments are common as the number of rounds increases, which can confuse LLMs and undermine the potential for correct reasoning. In addition, LLMs are good at summarizing previous answers, even when there is no summary order in the prompts. Truncating middle parts of dialogs can reduce input tokens and speed up brainstorming without losing the stem of previous reasoning and the question.

# 4 Experiments

We evaluate the LLM brainstorming on three datasets from different domains, which include the logical and factual reasoning question answer. For the LLMs in brainstorming, we use ChatGPT [1] and Gemini Pro [21, 24] considering their strong capability shown in the literature and practice, and we also use some smaller open-sourced models such as Qianwen1.5-7B [2], Baichuan2-7B [32] and Mistral-7B [12](Qwen, Baichuan, Mistral in short, respectively). The experimental evaluation is conducted in terms of accuracy, efficiency, and characteristic of the brainstorming.

## 4.1 Datasets

In order to evaluate the effectiveness of the brainstorming by extensive tasks, we test the ability of LLMs with respect to different factual knowledge questions. We utilize the existing MMLU dataset [11] to benchmark the accuracy of responses, which covers questions from subdomains Math, Social, Business, Humanities, Medical, etc. To further evaluate the reasoning ability in logic tasks, we consider harder mathematical reasoning tasks. Using GSM dataset [9], the models need to correctly solve grade school mathematical reasoning tasks. To evaluate the improvement of brainstorming in factual reasoning, we consider two tasks with different levels of difficulty in factual knowledge. We use the ARC-easy and ARC-challenge datasets [8], which contain only natural, grade-school science questions, to test the model's memory and ability to capture key knowledge.

## 4.2 Baselines

We evaluate the multi-model brainstorming by the following methods. By default, the multi-model brainstorming involves a pair of LLMs, and at most 8 rounds of interactions are implemented no matter whether a consensus is reached or not. As the effectiveness comparison, we compare the multi-model brainstorming with the *single-model-based* method, the *ensemble vote*, and the *multi-agent-based* method in the evaluation experiment. In the single-model-based method, we directly query an LLM to generate the response towards the evaluation. In the ensemble voter, the answer is determined by the majority rule of multiple models. In the multi-agent-based method, we employ multiple agents from the same LLM, but not multiple different LLMs, in the interaction process to obtain the answer. For all methods, the models are three open-source LLMs with diverse performance

in different domains, Qwen-7B [2], Baichuan-7B [32], Mistral-7B [12], and two larger models, GPT-3.5 [1] and Gemini Pro [21, 24].

For the evaluation of efficiency on the question answer task, we compare the brainstorming with *models using CoT prompt* and *models with double-parameter*. The CoT based models we utilize are Qianwen-7B [2], Baichuan-7B [32], and Mistral-7B [12] with 5-shot for MMLU, 8-shot for GSM, and zero-shot for ARC. In order to adapt to the 7B model size, we select Qianwen-14B [2] and Baichuan-13B [32] as the models with double parameter size. Since the Mixtral model does not have a version of size 14B, we use the MoE version (Mixtal 7×8B, with 12.9B active parameters and 46.7B total parameters) [12] instead.

#### 4.3 Results

Table 1: The accuracy of open-source models (Baichuan, Qwen, and Mistral) using brainstorming, single model, ensemble and multi-agent reasoning strategies on different datasets. The scale of all the models is 7B.  $\overline{\Delta}$  represents the average improvement compared to the related single-models.

Method	MMLU		GSM		ARC-e		ARC-c	
	Acc.	$\overline{\Delta}$	Acc.	$\overline{\Delta}$	Acc.	$\overline{\Delta}$	Acc.	$\overline{\Delta}$
Single-Model								
Baichuan	45.3%	-	29.0%	-	70.4%	-	50.8%	-
Mistral	46.2%	-	45.9%	-	70.6%	-	60.4%	-
Qwen	49.3%	-	51.1%	-	66.4%	-	55.3%	-
Ensemble Vote								
Qwen $\times$ Baichuan $\times$ Mistral	54.6%	+7.7%	54.9%	+12.9%	79.2%	+10.1%	64.9%	+9.4%
Multi-Agent								
Baichuan×2	45.9%	+0.6%	32.4%	+3.4%	72.8%	+2.4%	53.3%	+2.5%
$Mistral \times 2$	51.4%	+5.2%	48.9%	+3.0%	80.4%	+9.8%	69.5%	+9.1%
Qwen×2	54.8%	+5.5%	58.6%	+7.5%	80.1%	+13.7%	66.4%	+11.1%
Brainstorming(Ours)								
Baichuan × Mistral	52.5%	+6.8%	47.8%	+10.4%	80.3%	+9.8%	66.1%	+10.5%
Qwen × Baichuan	53.2%	+5.9%	54.1%	+14.1%	81.4%	+13.0%	64.0%	+11.0%
Qwen × Mistral	56.7%	+9.0%	56.5%	+8.0%	85.8%	+17.3%	73.7%	+15.9%

Table 2: The accuracy of commercial models (Gemini-Pro and ChatGPT) using brainstorming, single model, ensemble and multi-agent reasoning strategies on different datasets.

Method	MM	ILU	GS	M	ARC-c	
Wichiod	Acc.	$\overline{\Delta}$	Acc.	$\overline{\Delta}$	Acc.	$\overline{\Delta}$
Single-Model						_
GPT-3.5	68.8%	-	82.4%	-	84.6%	-
Gemini-Pro	68.0%	-	80.4%	-	86.3%	-
Multi-Agent						
GPT-3.5 $\times$ 2	70.9%	+2.1%	83.2%	+0.8%	86.0%	+1.4%
Gemini-Pro×2	68.5%	+0.5%	78.8%	-1.6%	86.9%	+0.6%
Brainstorming(Ours)						
GPT-3.5 $\times$ Gemini-Pro	71.0%	+2.6%	82.6%	+1.2%	88.4%	+3.0%

Accuracy Improvement for Reasoning. In Table 1, 2, we report the results of singel-model-based methods and multi-agent-based methods on MMLU, GSM, and ARC tasks. In each task, we observe that LLM brainstorming overweighs the single model. In the fine-grained logical GSM dataset and factual reasoning ARC dataset, the brainstorming even achieve an impressive 10% to 15% improvement in average accuracy, shown as the Table 1  $\overline{\triangle}$  column (Take brainstorming as an example, the  $\overline{\triangle}$  is calculated by  $Acc_{brain}(model_1, model_2) - \frac{Acc_{single}(model_1) + Acc_{single}(model_2)}{2}$ ). Although the ensemble model voting can result in an improvement in accuracy, the voting strategy still exhibits a lower reasoning accuracy than brainstorming on each task. This is due to the lack of interpretability of the voting strategy and its susceptibility to being misled by non-expert models.

Compared to the multi-agent based approach, the reasoning accuracy of brainstorming is on average 6.2%, 4.8% higher than multi-agent on logical (GSM) and factual (ARC) tasks respectively (shown as Table 1). Although the multi-agents can produce better reasoning than a single model in most tasks, restricted by its basic model, the knowledge of different agents is overlapped, which leads to lower

reasoning ability than brainstorming. Consequently, brainstorming by introducing different models and thus new knowledge can lead to correct answers on more tasks than multi-agent collaboration.

Simultaneously, in order to explore the capability gains obtained by models with different capabilities during the brainstorming, we present the accuracy distributions of 7B models on each subdomains of the MMLU dataset. Combining the results in Figure 3 and Table 1, it can be observed that the performances of Qwen and Baichuan in the math subdomain are more different, and after brainstorming, the accuracy improvement they obtained on the GSM dataset is more pronounced compared to the other pairs. Similarly, Mistral and Qwen have large differences in the distribution of subdomains that require memorization, so their brainstorming can achieve better results on the ARC dataset, which emphasizes memorization.

**Evaluation of Reasoning Efficiency.** We show the accuracy of the multi-LLM brainstorming compared to a single model enhanced with CoT in Figure.4. CoT guides the model by demonstrating a small number of examples, leading to more accurate results. Instead of manual labeling, we use the reasoning process from other model as

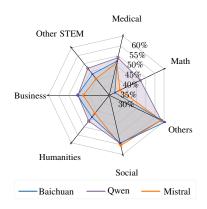


Figure 3: The accuracy distributions of the three open-source LLMs on different domains of the MMLU dataset. We divide the 57 tasks of MMLU into 7 major categories based on their fields.

CoT-like example during the brainstorming. As can be seen from the figure, brainstorming achieves comparable reasoning accuracy to the CoT-based model across the different tasks. In particular, on the factual reasoning tasks (ARC-e and ARC-c datasets), brainstorming even outperforms the single LLM with CoT, improving the average accuracy by about 14.3% compared to the CoT-based approaches. In the logical reasoning task (GSM), although Baichuan-7B as the base model (the average accuracy is around 29%) is much less effective than the other models (the average accuracy is around 50%), it still improves the reasoning accuracy to be comparable to that of the CoT-based method with the support of the brainstorming. This greatly reduces the cost of manual labeling, and provides a new solution for automation to improve the accuracy of model inference.

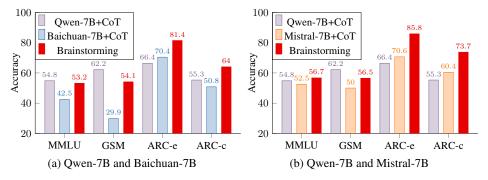


Figure 4: The accuracy comparison of brainstorming and CoT-base models on different datasets. CoT used in MMLU is 5-shot, 8-shot for GSM and 0-shot for ARC.

Table 3: The accuracy and the time consumption under saving different number of historical rounds in the brainstorming between Qwen1.5-7B and Baichuan2-7B on GSM. The time consumption is the average run time for all questions in the GSM dataset. The compute resources is in Appendix ??.

Saved Historical Rounds	1	3	5	7
Accuracy (%)	54.1	50.3	50.5	52.2
Time Consumption (s)	36.9	39.7	41.9	43.4

We also demonstrate the accuracy comparison of brainstorming based approaches with the larger parameter model. Figure 5 (a) illustrates that the reasoning accuracy of Qwen-7B and Baichuan-7B

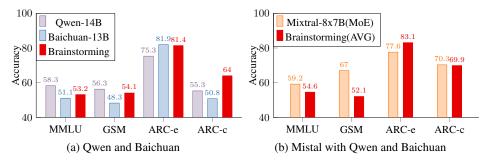


Figure 5: The accuracy comparison of the small-parameter models-based brainstorming (Qwen1.5-7B, Baichuan2-13B, Mistral-7B) with double-parameter models (Qwen1.5-14B, Baichuan2-13B) and with the MoE model (Mixtral 7 x 8B).

after brainstorming are able to reach an approximate level of the models with double-parameter (Qwen-14B and Baichuan-13B) on different tasks. Notice that, the brainstorming method is even 10% more accurate than the two-parameter model on the ARC-c dataset, which requires more common knowledge to be memorized. Since Mistral does not have a 14B version of the model, we compare the results of Mistral-7B's participation in brainstorming with its MoE version (Mixtral  $7\times8B$ ) in Figure 5 (b). We can see that the brainstorming-based approaches still possesses the ability to compete with the MoE model even when the scale of parameters is vastly different, and even obtains an outperform results of MoE in the ARC dataset. This indicates that brainstorming with models of different capabilities can dramatically improve the reasoning ability of the models through knowledge exchange. And many LLMs cannot be deployed on small devices due to the large number of parameters, while we found in our experiments that multiple small models can achieve similar capability to one large model after brainstorming, which provides a new solution for distributed deployment of large models.

Analysis of Model Properties. Considering that in a multi-round brainstorming scenario of the same problem, the reasoning process given by the model in a new round has a large portion of redundancy compared to the historical dialogs, so we evaluate whether to keep the historical dialogs or not. From Table 3, we can see that the retention of more rounds of the reasoning process does not improve the final reasoning accuracy, but rather introduces additional noise to the model, which disturbs the thinking of the model, and thus leads to the decline of the final reasoning accuracy. Moreover, retaining more rounds of inference process not only increases the burden of parsing context for the model, but also brings more memory occupation and increase in time consumption. Therefore, we believe that with the LLM already having some summa-

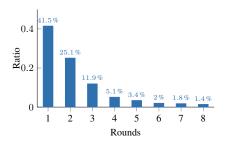


Figure 6: The proportion of tasks in the MMLU dataset where the brainstorming (Qwen1.5-7B and Baichuan2-7B) reaches consensus across rounds.

rization ability, retaining only the latest round of reasoning results in the brainstorming process can gain more benefits in terms of efficiency and reasoning accuracy.

We also explored the number of rounds required for brainstorming. As can be seen in Figure 6, the vast majority of questions can be answered consistently with about 4 rounds, with only a few questions requiring more rounds of discussion. It can also be seen in Figure 7 that the improvement in accuracy from single model reasoning to four rounds of interaction is the most significant, up to 30%. As the number of rounds rises, the gains from interaction gradually reduce, and the reasoning accuracy tends to converge when the number of rounds reaches about 4, and the accuracy increases by only 2% from 4 rounds to 8 rounds. So the brainstorming requires fewer rounds to obtain a substantial increase in reasoning accuracy.

**Case Study.** In Figure 8, we show the process of brainstorming among multiple models in a mathematical reasoning task and a biological memory task. We found that even though the models

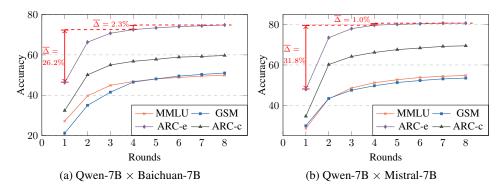


Figure 7: The accuracy of multi-model brainstorming at different number of dialog rounds, with an upper limit of 8 rounds. The red bidirectional arrows mark the change in accuracy of Brainstorming for 1 to 4 and 4 to 8 rounds with the ARC-e dataset.

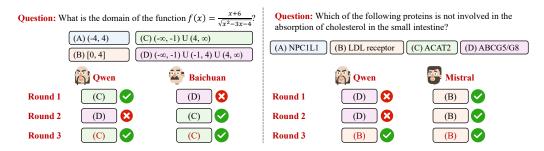


Figure 8: Changes in the reasoning results of Qwen-7B and Mistral-7B during brainstorming over logic and fact tasks.

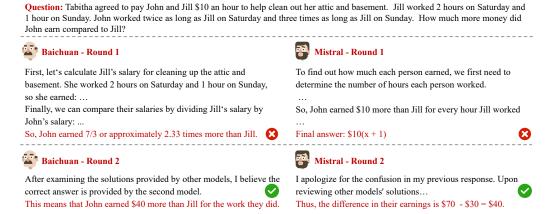


Figure 9: An illustrative example from Baichuan and Mistral brainstorming reasoning processes on a challenging mathematical task.

do not arrive at a unified answer in their initial reasoning, after multiple rounds of interaction, the initially incorrect model can correct its reasoning based on the reasoning process of other models, thus arriving at a correct answer that is consistent across multiple models.

In some more challenging tasks, as illustrated in Figure 9, neither model is able to provide the correct answer initially. Following a series of brainstorming, the models are able to reach a consensus of the correct answer. In the case of the factual task in Figure 10 that relies on memory, the models lack requisite knowledge and unable to provide the correct answer even after exchanging information with each other for several rounds. This is consistent with the typical human discussion process without requisite knowledge.

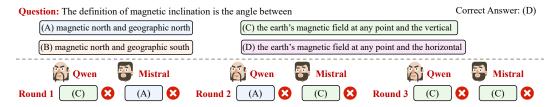


Figure 10: Owen and Mistral's brainstorming results without domain knowledge.



Figure 11: Results of debates between multiple agents with the same knowledge generated by Baichuan in logical and factual tasks.

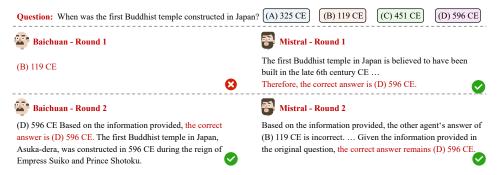


Figure 12: An illustrative of Baichuan and Mistral brainstorming when answering the same historical factual question as that in the right-hand side of Figure 11.

This phenomenon is further analyzed in the context of multi-agent debate and multi-model brain-storming. Figure 11 illustrates that, despite the single agent can not provide a correct answer in a logic task (left question of Figure 11), multi-agent is able to arrive at the correct answer through a debate. However, when a multi-agent debate is introduced for the same factual task (right question of Figure 11), it is still struggle to obtain the correct answer when each agent gives a wrong answer. This also demonstrates that the agents cannot obtain non-existent knowledge through interaction. However, by introducing another model with different capability, as shown in Figure 12, the knowledge could be introduced into the thinking process from the other model, and thus the wrong answer can be corrected after brainstorming. The detail brainstorming dialog cases are in Appendix ??.

#### 5 Discussion and Limitations

In this paper, we propose the multi-model brainstorming which contributes to improving the kownledge reasoning. We emphasize the necessity of the diverse capability and propose distinct interaction strategies as integral components within the brainstorming. Our quantitative analysis of the reasoning results conveys intuitions about knowledge exchange by brainstorming, and substantiates the effectiveness of our approach in enhancing knowledge reasoning and efficiency.

**Limitations.** In comparison to other prompting techniques, our multi-model brainstorming procedure is more computationally expensive because it requires both multiple generations and more computational space. Nevertheless, we believe that the intermediate discussion process and the generated consensus result can be distilled back to fine-tune the base model and further enhance the original knowledge reasoning ability of LLMs.

## References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- [3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.
- [4] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.
- [5] Bibo Cai, Xiao Ding, Zhouhao Sun, Bing Qin, Ting Liu, Lifeng Shang, et al. Self-supervised logic induction for explainable fuzzy temporal commonsense reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12580–12588, 2023.
- [6] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.
- [7] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, et al. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *The Twelfth International Conference on Learning Representations*, 2023.
- [8] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.
- [9] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- [10] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv* preprint *arXiv*:2305.14325, 2023.
- [11] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021.
- [12] 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, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.
- [13] Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14165–14178. Association for Computational Linguistics, 2023.
- [14] Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. Maieutic prompting: Logically consistent reasoning with recursive explanations. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1266–1279. Association for Computational Linguistics, 2022.
- [15] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

- [16] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36, 2024.
- [17] Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More agents is all you need. *arXiv preprint arXiv:2402.05120*, 2024.
- [18] Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. On the advance of making language models better reasoners. *arXiv preprint arXiv:2206.02336*, 2022.
- [19] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.
- [20] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- [21] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv* preprint arXiv:2403.05530, 2024.
- [22] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.
- [23] Winnie Street. Llm theory of mind and alignment: Opportunities and risks. *arXiv preprint* arXiv:2405.08154, 2024.
- [24] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [25] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.
- [26] Boshi Wang, Xiang Deng, and Huan Sun. Iteratively prompt pre-trained language models for chain of thought. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2714–2730. Association for Computational Linguistics, 2022.
- [27] Hongyi Wang, Felipe Maia Polo, Yuekai Sun, Souvik Kundu, Eric Xing, and Mikhail Yurochkin. Fusing models with complementary expertise. *arXiv preprint arXiv:2310.01542*, 2023.
- [28] Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *arXiv preprint arXiv:2307.05300*, 1(2):3, 2023.
- [29] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- [30] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.
- [31] Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7572–7590, 2023.
- [32] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. Baichuan 2: Open large-scale language models. *arXiv* preprint arXiv:2309.10305, 2023.

- [33] Joshua C Yang, Marcin Korecki, Damian Dailisan, Carina I Hausladen, and Dirk Helbing. Llm voting: Human choices and ai collective decision making. *arXiv preprint arXiv:2402.01766*, 2024.
- [34] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.
- [35] Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H Chi, Quoc V Le, and Denny Zhou. Take a step back: Evoking reasoning via abstraction in large language models. *arXiv preprint arXiv:2310.06117*, 2023.