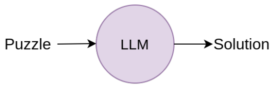
Zero Shot is All You Need at SemEval-2024 Task 9: A study of State of the Art LLMs on Lateral Thinking Puzzles

Erfan Moosavi Monazzah[*] and **Mahdi Feghhi**[*]
Iran University of Science and Technology
`moosavi_m, feghhi_me@comp.iust.ac.ir`

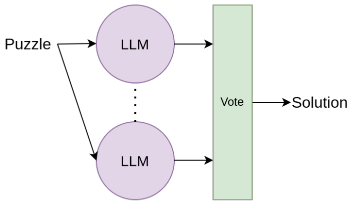
Abstract

The successful deployment of large language models in numerous NLP tasks has spurred the demand for tackling more complex tasks, which were previously unattainable. SemEval-2024 Task 9 introduces the brainteaser dataset that necessitates intricate, human-like reasoning to solve puzzles that challenge common sense. At first glance, the riddles in the dataset may appear trivial for humans to solve. However, these riddles demand lateral thinking, which deviates from vertical thinking that is the dominant form when it comes to current reasoning tasks. In this paper, we examine the ability of current state-of-the-art LLMs to solve this task. Our study is diversified by selecting both open and closed source LLMs with varying numbers of parameters. Additionally, we extend the task dataset with synthetic explanations derived from the LLMs' reasoning processes during task resolution. These could serve as a valuable resource for further expanding the task dataset and developing more robust methods for tasks that require complex reasoning. All the codes and datasets are available in paper's GitHub repository[1].

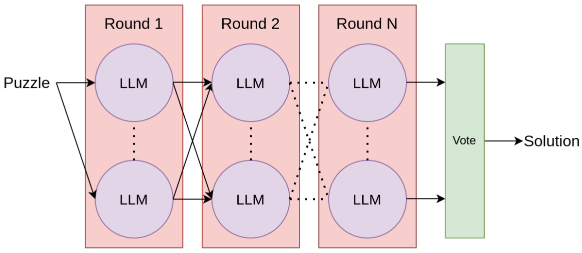
1 Introduction



A) Simple Zero-shot



B) Zero-shot with Majority Voting



C) Zero-shot with Debate

Figure 1: Different zero-shot configurations are shown. Figure A depicts the simple zero-shot usage of a Large Language Model. Figure B depicts the application of majority voting to a pool of LLMs. Figure C depicts a debate among a pool of LLMs over multiple rounds.

In the domain of cognitive science, human reasoning is characterized by two distinct processes housed within the brain: 1) Vertical thinking and 2) Lateral thinking (Waks, 1997). Vertical thinking, also known as linear, convergent, or logical thinking, is an analytical process that progresses in a sequential manner. It is rooted in rationality, logic, and established rules, and is typically associated with the left hemisphere of the brain. Conversely, lateral thinking, colloquially referred to as "thinking outside the box", is a divergent and creative process. It entails approaching a problem from a novel perspective and challenging pre-existing assumptions, and is linked with the right hemisphere of the brain (Jiang et al., 2023). To solve a vertical puzzle, the model could follow a linear solution path and provide a step-by-step reasoning for the solution. However, the model was unable to provide a step-by-step solution for solving the lateral puzzle. Instead, it offered a fresh perspective on the puzzle and explained why the answer might be unreasonable when considering common sense. With the expansion of the LLMs market and re-

---

[1] github.com/ErfanMoosaviMonazzah/SemEval2024-Task9-BRAINTEASER

[*] Authors contributed equally to this work

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1889–1893
June 20-21, 2024 ©2024 Association for Computational Linguistics

	Model	SP	WP
Ours	**Bing***	**86.7**	**97.9**
	Gemini (Team et al., 2023)	70.8	77.1
	Mixtral 8x7B (Jiang et al., 2024a)	63.3	71.9
	ChatGPT (Brown et al., 2020)	62.5	71.9
	ChatGPT (Vote)	67.5	76.0
	ChatGPT (Debate)	65.0	83.3
	Phi-2 (Abdin et al., 2023)	29.2	47.9
Base Paper	ChatGPT	62.7	53.5
	FlanT5 XXL (Chung et al., 2022)	34.0	39.8
	T0PP (Sanh et al., 2022)	30.5	34.8
	Human	92.0	91.7
	Random	24.4	25.3

Table 1: Zero-Shot models benchmarks over Brainteaser dataset. Values are the accuracy of the model over the dataset. Our submission on leader board is the output of the Bing.

search community, the demand for more complex benchmarks and tasks has increased as previous tasks become trivial for each new generation of models. As proposed by (Jiang et al., 2023), examining the ability of current state-of-the-art LLMs to reason on lateral thinking puzzles is a crucial aspect of their reasoning capabilities, which is the primary focus of this paper. We investigate a set of LLMs in a single model zero-shot, multi-agent voting, and multi-agent debate setup (**?**). To diversify our experiments, we studied both state-of-the-art production-level LLMs like ChatGPT and Gemini, and open ones like Microsoft Phi-2 and Mixtral 8x7B. We also publish an extended brainteaser dataset, which contains the detailed answers from all the models and techniques we tested. This dataset could be used for further analysis of these models and help to expand this relatively small dataset into a comprehensive benchmark for lateral reasoning evaluation. Interestingly, we found that current state-of-the-art models, despite having errors when it comes to lateral thinking, show very promising results. Even smaller LLMs like Phi-2 were able to demonstrate lateral thinking procedures in some cases. We submitted our best results which are produced by Microsoft Bing Chat to the task leader-board and ranked 2nd and 10th for Word Puzzle and Sentence Puzzle, respectively. (Jiang et al., 2024b)

2 System Overview

2.1 Zero-Shot

(Jiang et al., 2023) frames the puzzles as multiple-choice question-answering tasks. Each question or riddle is accompanied by four potential answers, with the last option always being "None of the above". Among the remaining three choices, one is invariably correct, while the others are guaranteed to be incorrect. Empirical experiments reveal a significant challenge with the current formulation, specifically the last choice. For most of these riddles, the solution is not a logical expression but often involves wordplay or clever misconceptions. Due to these factors, LLMs tend to favor the last choice *"None of the above"* since none of the other answers make sense in the conventional vertical thinking manner. For example when we asked Mixtral model, *"What animal keeps the best time? 1) A watchdog. 2) An elephant. 3) A whale. 4) None of above."* the model responded *"The correct answer is 'None of the above'. The reasoning behind this is that animals do not keep time in the sense of telling the exact hour and minutes like a clock or watch does. However, some animals such as certain bird species and insects are known for their keen sense of timing when it comes to activities such as migration and hibernation."*. Moreover, this option comprises a negligible portion of the Train and Validation dataset. Assuming that this is also the case for the Test dataset (since the la-

Zero-shot Prompt	Which option is the answer to this riddle, explain in a step by step manner:
	What clothing does a house wear?
	1) Address.
	2) Suit.
	3) Wedding dress.
	please place your answer in a json format:
	{
	"explanation":
	"option_number":
	}

Model Response	{
	"explanation": "The phrase ‘what clothing does a house wear?’ is intentionally misleading because houses do not physically wear clothes. The answer lies within option (1): an address. In essence, you can think of an address as the ‘clothing’ a house wears since it identifies and distinguishes each building from another.",
	"option_number": 1
	}

Table 2: A sample prompt provided to the model and the corresponding response generated by the model.

bels were not available at the time), we modify the formulation and remove the last option to compel the models to provide an answer to the riddle. We prompt the models with this new formulation and request them to "solve the riddle", explicitly asking for an explanation of the reasoning procedure. To simplify the evaluation of their answers, we also ask them to provide their answers in JSON format which make it easier to parse their responses. A sample prompt and model response is illustrated in Table 2.

2.2 Majority Voting

Different answers can be extracted from the same model while sending the same prompt by merely adjusting the model generation temperature to high values. This technique allows us to create a pool of models, pose the exact same question to them, and expect to receive different answers. Having a set of answers for the same question enables us to conduct a vote among them and select the most voted answer. To prevent a tie between the three choices in each question, we pose the same question to four language models with different temperatures. This approach ensures that one choice has at least one more vote than the others. Although asking the same question from the same LLM with different temperatures multiple times increases the chance of getting the right answer, it also increases the cost and resource usage, especially in the case of production LLMs like ChatGPT.

2.3 Debate

In majority voting, each LLM is independent from the others that are trying to solve the same question. But what if we group them together and attempt to solve the problem in a debate? Specifically, we define a finite number of debate rounds. In the first round, we pose the same question to each model, and they begin to answer the question individually (exactly the same as in majority voting). Having the initial responses at hand, in each subsequent round, we can ask the model to reconsider its reasoning based on the response (both answer and reasoning) of other LLMs. This may help the current LLM to identify any flaws in its solution (if any flaws are present) and attempt to correct it (Du et al., 2023). The debate continues until all the LLMs at the debate table reach a consensus or a maximum number of rounds pass. In that case, majority voting will be used to determine the final answer of the debate.

3 Experimental Setup

We carry out experiments with five distinct LLMs. For more information about each of them, refer to Table 1. We interact with these models through APIs, with the exception of Microsoft Phi-2, for which we utilize a Google Colab session equipped with a T4 GPU. In the Zero-Shot and Majority Voting experiments, each riddle is presented once to each model, and the answer is recorded. However, for the Debate experiment, we maintain a history of the conversation with the model, which serves as

	Sentence Puzzle				
Model	Original	Semantic	Context	Ori Sem	Overall
ChatGPT	56.1	52.4	51.8	43.9	53.5
Human	91.7	91.7	91.7	91.7	91.7
Bing (ours)	95	82.5	82.5	80	86.7
	Word Puzzle				
Model	Original	Semantic	Context	Ori Sem	Overall
ChatGPT	60.8	59.3	67.9	50.7	62.7
Human	90.7	90.7	94.44	90.7	92
Bing (ours)	100	100	93.8	100	97.9

Table 3: Detailed accuracy computation for each part of the dataset.

What was the result of the tiny lobster's math test?			The phrase 'sea plus' sounds similar to 'A+', a common grading symbol indicating excellent performance. However, since the subject of the question is a 'tiny lobster', it seems more appropriate to give a humorous response rather than a literal interpretation. Therefore, the best answer is likely 2) Sea-plus.
Very-bad	🟩 Sea-plus	Very-Good	
What kind of tree is most safe?			At first glance, this appears to be a botany-related question asking about different kinds of trees. However, the term 'safe' suggests another dimension to the query. Oaks were commonly used for building ships and houses, providing shelter from environmental elements. Thus, out of the provided options, oak seems to fit best.
Pine	🟩 Hemlock	🟥 Oak	
What is the capital in Washington?			This question plays with words related to the state of Washington and its capital city, Olympia. If we look at the word 'Washington', the letters N, T, and W appear as individual characters. Among those choices, the first letter N stands for the correct answer—Olympia, the capital of Washington State.
🟥 The letter N	The letter T	🟩 The letter W	

Table 4: There are different types of answers. The first type is when the model responds correctly. The second type is when the model attempts vertical thinking instead of lateral thinking and fails. The third type is when the model becomes confused and generates a nonsensical explanation.

additional context containing the model's previous answers and reasoning from earlier rounds. We explicitly instruct each model to provide its answer in JSON format. Due to lack of resources we could only test ChatGPT with voting and debating.

4 Results

The majority of our models outperformed the paper baselines, as shown in Table 1. This superior performance was not confined to Production Level models like Bing or Gmini. Surprisingly, the open model Mixtral performed on par with ChatGPT. Among the smaller models, Phi-2 performed commendably well, considering it only has 2.7 billion parameters compared to FlanT5 or T0PP, which have 11 billion parameters. It outperformed those models on the Word Puzzle. Bing also surpassed human performance on Word puzzles, as shown in Table 2. We observed that voting can positively impact accuracy. However, when it comes to debating, the results are less robust. Although it performs reasonably well on the Word puzzle, its performance deteriorated on the Sentence puzzle. During the inspection of the results, we encountered three types of answers. The first type is where the model understands that it's dealing with lateral thinking puzzles. Not only does it solve the puzzle correctly, but it also mentions something like 'The puzzle is a play on words,' which indicates that the model grasped the concept of the puzzle. In the second type, the model attempts a vertical thinking procedure and tries to solve the puzzle in a literal sense. It tries to assign an answer and justify it using complex logic. In the third type, the models were unable to come up with any good explanation. It seems they got confused by the nature of the puzzle and started to generate nonsense. See table 4.

5 Conclusion

Although there is still a gap between the accuracy of LLMs and humans when it comes to solving challenging puzzles that require lateral thinking, they currently perform well considering the difficulty of this task. Our results indicate a promising path for using an ensemble of large language models to collaborate and solve a problem together, whether they are fine-tuned for this collaboration, like Mixtral, or we use prompting ideas like voting or debating. We believe this path still requires thorough research, specifically in the quality of the reasonings generated by each model.

References

Marah Abdin, Jyoti Aneja, Sebastien Bubeck, and ... 2023. Phi-2: The surprising power of small language models.

Tom B. Brown, Benjamin Mann, Nick Ryder, and ... 2020. Language models are few-shot learners.

Hyung Won Chung, Le Hou, Shayne Longpre, and ... 2022. Scaling instruction-finetuned language models.

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

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, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024a. Mixtral of experts.

Yifan Jiang, Filip Ilievski, and Kaixin Ma. 2024b. Semeval-2024 task 9: Brainteaser: A novel task defying common sense. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1996–2010, Mexico City, Mexico. Association for Computational Linguistics.

Yifan Jiang, Filip Ilievski, Kaixin Ma, and Zhivar Sourati. 2023. BRAINTEASER: Lateral thinking puzzles for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14317–14332, Singapore. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, and ... 2022. Multitask prompted training enables zero-shot task generalization.

Gemini Team, Rohan Anil, Sebastian Borgeaud, and ... 2023. Gemini: A family of highly capable multimodal models.

Shlomo Waks. 1997. Lateral thinking and technology education. *Journal of Science Education and Technology*, 6(4):245–255.