MasonTigers at SemEval-2024 Task 9: Solving Puzzles with an Ensemble of Chain-of-Thought Prompts

Md Nishat Raihan, Dhiman Goswami, Al Nahian Bin Emran,
Sadiya Sayara Chowdhury Puspo, Amrita Ganguly, Marcos Zampieri
George Mason University, USA
mraihan2@gmu.edu

Abstract

This paper presents the *MasonTigers*' submission to the SemEval-2024 Task 9 which provides a dataset of puzzles for testing natural language understanding. We employ large language models (LLMs) to solve this task through several prompting techniques. We show that zero-shot and few-shot prompting with proprietary LLMs outperform open-source models. Results are further improved with chain-of-thought prompting. We obtain our best results by utilizing an ensemble of chain-of thought prompts, ranking 2^{nd} in the word puzzle subtask and 13^{th} in the sentence puzzle sub-task.

1 Introduction

In recent years, LLMs have achieved impressive performance on several question answering and language understanding tasks when provided with appropriate prompting (Brown et al., 2020). However, complex reasoning abilities often present a challenge for these models. SemEval-2024 Task 9 (Jiang et al., 2024b) introduces a novel dataset called *BrainTeaser* (Jiang et al., 2023) which includes a set of complex puzzles and brainteasers. Such tasks involve solving word and sentence puzzles, which require multi-step inference and deduction. The dataset covers a diverse range of puzzle types including sequences, analogies, classification, mathematical reasoning, inferences about implicit relationships, and more. Solutions frequently demand a chained application of knowledge and logic across multiple steps to uncover insights or concepts not directly stated in the problem description.

Solving these elaborate reasoning problems is a challenging scenario for NLP systems. We explore whether and how LLMs can succeed on this task. We employ proprietary models such as GPT-4 (OpenAI, 2023) and Claude 2.1 (Anthropic, 2023) through APIs. These models have shown promising few-shot reasoning ability. We also use Mixtral (Jiang et al., 2024a), an open-source LLM that shows state-of-the-results in several language reasoning tasks. The prompting paradigm involves providing models with natural language descriptions that encode the reasoning process step-by-step (Liu et al., 2021). We implement various prompting approaches for mapping puzzles to conditional text and systematically transforming reasoning into explanation chains. Our core method, chain-of-thought prompting (Wei et al., 2022), iteratively breaks down the deduction into simplified logical steps.

Experiments reveal that while zero-shot performance lags due to a lack of grounding, multi-step prompts can unlock substantial reasoning ability in models. Performance improves with more steps and more specificity in the prompt. While introducing few-shot prompting generates good results, we observed that models do significantly better with chain-of-thought prompting. We experiment with several chains of thought and achieve mostly similar results with each attempt. To make a more empirically confident guess towards solving the puzzles we adopt an ensemble of these chains based on majority voting. Our approach achieves competitive performance, ranking *2nd* on the word puzzle subtask and *13th* on sentence puzzles.

2 Related Work

LLMs have been widely used for complex and challenging language processing tasks recently (Raihan et al., 2023a,b; Goswami et al., 2023). They have shown good reasoning abilities in several tasks. The task of solving puzzles and the BrainTeaser dataset (Jiang et al., 2023) represent both a novel task and a novel dataset respectively. Similarly to their multiple choice questions (MCQs) approach, a few datasets like MathQA (Austin et al., 2021), have been compiled. However, these are intended for specific tasks in which domain knowledge is usually enough thus they not requiring deep reason-

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1358–1363

ing. A similar work is done by Saeedi et al. (2020) where they investigate a task that combines natural language understanding and commonsense reasoning. They present deep learning architectures for distinguishing between sensible and nonsensical statements.

Pun detection by (Zou and Lu, 2019) is a puzzle-like activity that is similar to BrainTeaser. It presents a method for joint pun detection and localization utilizing a sequence labeling perspective. This highlights the complexity of language comprehension, especially in detecting subtle wordplay. Another dataset, LatEval is curated by Huang et al. (2023) that delves further into lateral thinking and commonsense reasoning, highlighting the challenges faced by language models in tasks requiring unconventional thinking and creativity. Zhou et al. (2023) presents ROME, a dataset designed to assess vision-language models' capacity to reason beyond intuitive understanding, highlighting the shortcomings of existing models in understanding events that defy common sense.

In the field of reasoning task, a *chain-of-thought* (Wei et al., 2022) implies a logical sequence of connected ideas, fostering coherence and depth in responses. On the other hand, a *tree-of-thought* suggests branching out into various related ideas, offering a more comprehensive exploration of a topic. While few-shot prompting is effective for some tasks by providing examples to guide the model, it may have limitations in capturing the complexity of nuanced conversations. The optimal choice may involve a hybrid approach, where a few-shot prompt sets the initial (Yao et al., 2023) context, and the model subsequently follows a chain or tree of thought to generate more contextually rich and coherent responses useful for reasoning tasks.

Tan (2023) shows the performance of LLM's on the reasoning of arithmetic word problems. It states that higher degrees of realization are associated with better overall accuracy on arithmetic problems. And chain-of-thought is really helpful in this aspect as it covers a variety of prompts to strengthen the reasoning. Similarly, Mo and Xin (2023) presents a new reasoning framework for large language models by addressing a gap in prior tree-based reasoning methods which overlooked inherent uncertainties in intermediate decision points made by models. Overall, the key innovation is leveraging uncertainty estimation locally within the models during tree reasoning to enable more precise problem-solving and reasoning.

3 The BrainTeaser Dataset

The BrainTeaser dataset (Jiang et al., 2023), introduced with the task (Jiang et al., 2024b) is a question-answering benchmark designed to evaluate models' ability for lateral thinking, i.e., to defy default commonsense associations and reason creatively. The dataset contains 1,100 multiple-choice questions divided into two sub-tasks - 627 sentence-based puzzles relying on narrative context and common phrases and 492 puzzles focused on the literal form and letters of words.

For a fair comparison with human performance, the dataset also provides a separate human evaluation set with 102 randomly sampled questions. Each question in BrainTeaser has one correct answer and three distractor choices, including the option "none of the above". To prevent memorization of training data, the dataset also contains semantically and contextually reconstructed variants for every question while preserving the original reasoning process and answers. The key statistics of the dataset are shown in Table 1.

	Sentence	**Word**
Number of puzzles	627	492
Avg. tokens (prompt)	34.88	10.65
Avg. tokens (choices)	9.11	3.0

Table 1: Key statistics of the BrainTeaser dataset in the sentence and word puzzle sub-task.

During the SemEval-2024 Task 9 development phase, a total of 240 prompts (120 for both sentence and word puzzles) are provided. During the test phase, a total of 216 prompts (120 for sentence and 96 for word puzzles) are provided.

4 Experiments

In our experiments, we focus on several prompting strategies by employing three state-of-the-art models including proprietary models like GPT-4 (OpenAI, 2023) and Claude 2.1 (Anthropic, 2023) (accessed via API key) and one open-source model - Mixtral (Jiang et al., 2024a).

4.1 Zero-Shot Prompting

We start with zero-shot prompting by assigning the AI a role, describing the task, and giving it one puzzle at a time, as shown in Figure 1.

Role:
You are a helpful AI assistant. You are given the task of solving a sentence/word puzzle.

+

Definition:
It's a Multiple Choice Question paired with 4 potential answers. Choose the most suitable one.

+

Question: What part of London is in France?
Choices:
1. The letter N.
2. The letter O.
3. The letter L.
4. None of the above.

Choose the most suitable answer. Thanks.

Figure 1: Sample structure for Zero-Shot Prompting.

4.2 Few-Shot Prompting

In order to give the LLMs more context we integrate more examples and design prompts for few-shot prompting. We include 4 solved puzzles from the train set and then attach one puzzle from the test set each time we prompt the models. We also use some tags for better extracting the generated answers, as shown in Figure 2.

4.3 Chain-of-Thought

To guide the models toward better reasoning - we experiment with chain-of-thought prompting. We give the model the puzzle, and the potential answers and work with every example one-by-one in order to choose the most reasonable one. Like the original CoT approach (Wei et al., 2022), we do not assign any role or explain the task - just pose the question, the CoT, and the answer (see Figure 3). We do this as 2-shot, 4-shot, and 8-shot for all three models.

4.4 Ensemble of Chain-of-Thought Prompts

To assess model performance, an ensemble approach is utilized with chain-of-thought prompting to make more confident guesses regarding the correct answers. Specifically, majority voting is done across an ensemble of models prompted by different question groups. For each prompt, 8 different random questions are selected from the BrainTeaser training set - repeated 5 times in total. Finally, the predictions are aggregated through voting to output the overall ensemble prediction.

This ensemble methodology with chain-of-thought prompting helps improve robustness to outlier examples and noise compared to using a single model. By prompting the ensemble components on different random question subsets, diversity is promoted to capture a more holistic representation of the overall data distribution. The voting also helps cancel out issues with single models latching onto spurious patterns. Overall, the ensemble approach with multiple chain-of-thought prompt sets provides a robust assessment strategy suited for the open-ended nature and diversity of the BrainTeaser puzzles.

Role:
You are a helpful AI assistant. You are given the task of solving a sentence/word puzzle.

+

Definition:
It's a Multiple Choice Question paired with 4 potential answers. Choose the most suitable one.

+

Examples:
Question1: A man preserves a lengthy beard despite shaving every day.
1. He is a barber.
2. He wants to maintain his appearance.
3. He wants his girlfriend to buy him a razor.
4. None of the above.
Correct Answer: <ans1> 1 **</ans1>**
.....
Question4: Tom attends class every day but doesn't do any homework.
1. He is a teacher.
2. He is a lazy person.
3. His teacher will not let him fail.
4. None of the above.
Correct Answer: <ans4> 1 **</ans4>**

+

Question5: The brother of a beggar passed away, but the deceased had no brothers. How is that possible?
Choices:
1. The beggar was the man's sister.
2. The man is angry for his brother being a beggar and cut ties with him.
3. The bagger's brother is a murderer.
4. None of the above.
Correct Answer: <ans5> ? **</ans5>**
Choose the most suitable answer. Thanks.

Figure 2: Sample structure for Few-Shot Prompting.

5 Results

We analyze the performance of the three models - including GPT4 Turbo, Claude 2.1, and Mixtral.

Question 1: How do you spell COW in thirteen letters?

–

Choices 1:

1. SEE OH DEREFORD.
2. SEE O DOUBLE YOU.
3. COWCOWCOWCOWW.
4. None of above.

–

Chain-of-Thought 1:

1. SEE OH DEREFORD: Doesn't seem to spell out "COW" in any conventional or playful manner.
2. SEE O DOUBLE YOU: Spells out "COW" in a creative way, matching the letter count required.
3. COWCOWCOWCOWW: Simply repeats the word "COW" without cleverly meeting the thirteen-letter criteria.
4. None of the above: Not applicable since there is a viable option.

–

Decision 1: The answer "SEE O DOUBLE YOU" creatively spells "COW" using thirteen letters, making it the correct choice.

–

Answer 1: 2.

.....

.....

.....

.....

Question 8: –

Choices 8: –

Chain-of-Thought 8: –

Decision 8: –

Answer 8: –

———

———

Question 9: How do you spell COB in seven letters?

–

Choices 9:

1. COBCOBB
2. COBBLER
3. SEE O BEE.
4. None of the above.

–

Figure 3: Sample structure for Chain-of-Thought Prompting (8-shot).

These models are tested with different types of prompts - regular and chain-of-thought, and with a varying number of examples, known as shots, ranging from zero to eight. Additionally, an ensemble method is applied to the eight-shot chain-of-thought prompting to see if it can further improve the models' performance. The results, shown in Table 2, reveal how the models performed under each condition. A human baseline with scores of 0.91 for both Sentence and Word puzzles in the test set is provided by the task organizers for comparison purposes.

GPT4 Turbo shows the best performance, especially with chain-of-thought prompting and an increasing number of shots. The model performs best with the eight-shot chain-of-thought prompting combined with the ensemble method ([E]), reaching the highest Sentence and Word scores of 0.93 and 0.95 in the test set, respectively. This shows that chain-of-thought prompting and the ensemble method significantly improve the model's understanding and output. Claude 2.1 also improves with chain-of-thought prompting and more shots. Its best scores were with the eight-shot chain-of-thought with the ensemble, achieving Sentence and Word scores of 0.86 and 0.95 in the test set, respectively. The asterisk (*) mark in Table 2 denotes our submission during the test phase. Even though Mixtral's performance is inferior to the performance of the other two models, it consistently gets better with more shots and chain-of-thought prompting. Mixtral delivered best results with the eight-shot chain-of-thought and the ensemble technique, with Sentence and Word scores of 0.88 and 0.82 in the test set, respectively.

Finally, the results highlight the effectiveness of chain-of-thought prompting in boosting the performance of LLMs. This approach, especially when combined with more examples and the ensemble method, greatly improves models' abilities to process and generate more accurate responses. GPT4 Turbo's top performance is likely due to its advanced design, which makes the most of these strategies. On the other hand, Claude 2.1's results point to the importance of model-specific adjustments.

6 Conclusion and Future Work

In this paper, we presented MasonTigers' approach to SemEval-2024 Task 9 on solving puzzles using LLMs. We explored various prompting strategies to guide the models, including zero-shot, few-shot, and chain-of-thought prompting. Our key method involved iteratively breaking down reasoning into simplified logical steps to decompose the complex

Model	Prompting	# of Shot	Sen_Dev	Sen_Test	Word_Dev	Word_Test
Human Baseline	–	–	–	0.91	–	0.91
GPT4 Turbo	Regular	Zero Shot	0.79	0.76	0.81	0.79
GPT4 Turbo	Regular	4 Shot	0.90	0.91	0.87	0.86
GPT4 Turbo	CoT	2 Shot	0.87	0.88	0.85	0.89
GPT4 Turbo	CoT	4 Shot	0.89	0.90	0.92	0.91
GPT4 Turbo	CoT	8 Shot	0.93	0.92	0.94	0.94
GPT4 Turbo	CoT **[E]**	8 Shot	**0.94**	**0.93**	**0.96**	**0.95**
Claude 2.1	Regular	Zero Shot	0.76	0.77	0.71	0.62
Claude 2.1	Regular	4 Shot	0.87	0.84	0.87	0.85
Claude 2.1	CoT	2 Shot	0.84	0.81	0.83	0.84
Claude 2.1	CoT	4 Shot	0.91	0.84	0.90	0.94
Claude 2.1	CoT	8 Shot	0.90	0.84	0.90	0.94
Claude 2.1	CoT **[E] [*]**	8 Shot	***0.91***	***0.86***	***0.91***	***0.95***
Mixtral	Regular	Zero Shot	0.71	0.66	0.45	0.51
Mixtral	Regular	4 Shot	0.81	0.82	0.79	0.75
Mixtral	CoT	2 Shot	0.79	0.75	0.63	0.70
Mixtral	CoT	4 Shot	0.84	0.86	0.77	0.76
Mixtral	CoT	8 Shot	0.89	0.86	0.80	0.81
Mixtral	CoT **[E]**	8 Shot	0.89	0.88	0.81	0.82

Table 2: Comparing the results generated by the models with different prompting strategies. **[CoT]** - denotes chain-of-thought. **[E]** - denotes Ensemble (as described in 4.4). **[*]** - denotes submission during the test phase on the Leaderboard.

deduction process.

Our experiments revealed promising results. While zero-shot performance was limited, providing explanatory prompts substantially improved the models' reasoning abilities. Performance increased with more prompt specificity and steps. Our best results came from an ensemble approach applying majority voting across multiple chain-of-thought prompts.

Ultimately, our system achieved competitive rankings on the leaderboard, placing 2^{nd} in the word puzzle sub-task and 13^{th} on sentence puzzles. The strong capability unlocked through guided prompting highlights these models' latent reasoning potential when given a structured thought process. Our work sheds light on how explanatory chains can elicit more of the knowledge within large language model parameters.

A few key limitations remain to be addressed in future work. First, constructing effective prompts requires extensive human effort and insight - automating this prompting process could improve scalability. Additionally, performance still lags behind human levels, indicating that there is room for advancement. Architectural constraints related to long-term memory and reasoning likely need to be overcome. Finally, our approach focused narrowly on the given puzzles rather than teaching broader inferential skills - developing more generalizable reasoning through prompts is an open challenge.

Acknowledgments

We would like to thank the shared task organizers for proposing this interesting competition and for providing participants with the BrainTeaser dataset.

References

Anthropic. 2023. Claude 2.1: Updates and improvements. https://www.anthropic.com/news/claude-2-1.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*.

Dhiman Goswami, Md Nishat Raihan, Antara Mahmud, Antonios Anastasopoulos, and Marcos Zampieri. 2023. OffMix-3L: A novel code-mixed test dataset in bangla-english-hindi for offensive language identification. In *Proceedings of SocialNLP (AACL)*.

Shulin Huang, Shirong Ma, Yinghui Li, Mengzuo Huang, Wuhe Zou, Weidong Zhang, and Hai-Tao Zheng. 2023. Lateval: An interactive llms evaluation benchmark with incomplete information from lateral thinking puzzles. *arXiv preprint arXiv:2308.10855*.

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. 2024a. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

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 EMNLP*.

Peng Liu, Nicholas Lourie, Sean Welleck, and Julia Hockenmaier. 2021. What makes good in-context examples for gpt-3? In *Proceedings of EMNLP*.

Shentong Mo and Miao Xin. 2023. Tree of uncertain thoughts reasoning for large language models. *arXiv preprint arXiv:2309.07694*.

OpenAI. 2023. Gpt-4 technical report. https://arxiv.org/abs/2303.08774.

Md Nishat Raihan, Dhiman Goswami, Antara Mahmud, Antonios Anastasopoulos, and Marcos Zampieri. 2023a. SentMix-3L: A novel code-mixed test dataset in bangla-english-hindi for sentiment analysis. In *Proceedings of SEALP (AACL)*.

Md Nishat Raihan, Umma Hani Tanmoy, Anika Binte Islam, et al. 2023b. Offensive language identification in transliterated and code-mixed bangla. In *Proceedings of BLP (EMNLP)*.

Sirwe Saeedi, Aliakbar Panahi, Seyran Saeedi, and Alvis C Fong. 2020. Cs-nlp team at semeval-2020 task 4: Evaluation of state-of-the-art nlp deep learning architectures on commonsense reasoning task. *arXiv preprint arXiv:2006.01205*.

Juanhe TJ Tan. 2023. Causal abstraction for chain-of-thought reasoning in arithmetic word problems. In *Proceedings of BlackboxNLP (EMNLP)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, et al. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Kankan Zhou, Eason Lai, Wei Bin Au Yeong, Kyriakos Mouratidis, and Jing Jiang. 2023. Rome: Evaluating pre-trained vision-language models on reasoning beyond visual common sense. *arXiv preprint arXiv:2310.19301*.

Yanyan Zou and Wei Lu. 2019. Joint detection and location of English puns. In *Proceedings of NAACL*.