HW-TSC at SemEval-2024 Task 9: Exploring Prompt Engineering Strategies for Brain Teaser Puzzles Through LLMs

Yinglu Li, Yanqing Zhao, Min Zhang, Yadong Deng, Aiju Geng, Xiaoqin Liu, Mengxin Ren, Yuang Li, Chang Su, Xiaofeng Zhao, Xiaosong Qiao, Ming Zhu, Yilun Liu, Mengyao Piao,Feiyu Yao, Shimin Tao, Hao Yang, Yanfei Jiang
Huawei Translation Services Center, Beijing, China
{liyinglu, zhaoyanqing, zhangmin186, yanghao30}@huawei.com

Abstract

Large Language Models (LLMs) have demonstrated impressive performance on many Natural Language Processing (NLP) tasks. However, their ability to solve more creative, lateral thinking puzzles remains relatively unexplored. In this work, we develop methods to enhance the lateral thinking and puzzle-solving capabilities of LLMs. We curate a dataset of word-type and sentence-type brain teasers requiring creative problem-solving abilities beyond commonsense reasoning. We first evaluate the zero-shot performance of models like GPT-3.5 and GPT-4 on this dataset. To improve their puzzle-solving skills, we employ prompting techniques like providing reasoning clues and chaining multiple examples to demonstrate the desired thinking process. We also fine-tune the state-of-the-art Mixtral 7x8b LLM on our dataset. Our methods enable the models to achieve strong results, securing 2nd and 3rd places in the brain teaser task. Our work highlights the potential of LLMs in acquiring complex reasoning abilities with the appropriate training. The efficacy of our approaches opens up new research avenues into advancing lateral thinking and creative problem-solving with AI systems.

1 Introduction

In recent years, the advent of advanced language models has revolutionized the field of NLP, steering research towards challenges that necessitate intricate and implicit reasoning processes akin to human commonsense reasoning. Such tasks often require vertical thinking, an analytical and methodical approach to problem-solving. This paradigm has enjoyed substantial popularity and success within the NLP community. However, lateral thinking puzzles, which demand creative reasoning and the ability to perceive indirect or non-obvious solutions, have not been equally explored. Lateral thinking involves breaking away from conventional patterns to reveal novel insights, a feat that models based on rigid commonsense associations often struggle with.

Task	Type	Train size	Eval size	Test size
Subtask 1	Word Puzzle	396	120	96
Subtask 2	Sentence Puzzle	507	120	120

Table 1: Task dataset description

Recognizing this disparity, we introduce LLMs in "BRAINTEASER," a meticulously curated multiple-choice Question Answering (QA) task in order to evaluate LLMs' capabilities for lateral thinking. The dataset(Jiang et al., 2023b, 2024b) contains two subtasks: word and sentence brain teasers. Word puzzles are word-type brain teasers where the answer deviates from the typical meaning of the word and instead focuses on the letter composition. Sentence puzzles are sentence-type brain teasers centered around nonsensical or illogical snippets of text. The key characteristics of the dataset are described in Table 1.

In our approach, we employ the formidable GPT-4 language model to address BRAINTEASER's questions under both zero-shot and few-shot conditions, thereby assessing its inherent reasoning capabilities without and with limited context. Additionally, we leverage prompt engineering strategies and incorporate a Chain of Thought (CoT) prompting technique to enhance GPT-4's comprehension of the task requirements. This innovative methodology not only facilitates clearer demonstration of the problem-solving process but also aligns the model's reasoning with human-like thought patterns.

Examples of the word and sentence puzzle samples are provided in Tables 2 and 3, respectively. Semantic Reconstruction (SR) rephrases the original question without altering the correct answer or distractor. Context reconstruction (CR) maintains

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1646–1651

ID	Question	Choice List
WP-0	*How do you spell COW in thirteen letters?*	SEE OH DEREFORD **SEE O DOUBLE YOU.** COWCOWCOWCOWW. None of above.
WP-0_SR	*In thirteen letters, how do you spell COW?*	SEE OH DEREFORD **SEE O DOUBLE YOU.** COWCOWCOWCOWW. None of above.
WP-0_CR	*How do you spell COB in seven letters?*	COBCOBB COBBLER **SEE O BEE.** None of above.

Table 2: Dataset samples for subtask 1: word puzzles. Each choice list has four choices. The ground truth is bold.

ID	Question	Choice List
SP-48	*Why is it so cold on Christmas?*	**Because it's in December.** Because people are waiting for the New Year. Because people are celebrating. None of above.
SP-48_SR	*Why is Christmas Day so chilly?*	**Because it's in December.** Because people are waiting for the New Year. Because people are celebrating. None of above.
SP-48_CR	*Why is Independence Day so hot?*	Because people are enjoying the firework. Because people are celebrating. **Because it's in July.** None of above.

Table 3: Dataset samples for subtask 2: sentence puzzles. Each choice list has four choices. The ground truth is bold.

the reasoning path but changes both the question and answer to reflect a new situational context.

The results of our experiments are both promising and insightful. Our model achieved commendable rankings, securing 2nd and 3rd places in the task, which underscores the potential of LLMs in mastering complex, creative problem-solving tasks that extend beyond the scope of traditional commonsense reasoning. These outcomes not only validate the efficacy of our methods but also pave the way for further explorations into the untapped potential of lateral thinking in AI-driven language understanding.

2 Related work

2.1 LLM

Language is a uniquely human ability that allows us to communicate, express ourselves, and record information. In AI research, language models refer to models that can predict the next word or token in a sequence given the previous words or context. Early language models are based on statistical techniques that calculate the probability of each possible next word. These statistical language models are later superseded by neural network-based models, which can more accurately estimate the probability of the next token using deep-learning methods. The development of neural language models marks a major advance in NLP capabilities. By utilizing neural networks to model the complexities of language, today's state-of-the-art language models can generate surprisingly human-like text and show impressive language understanding abilities.

Subsequently, pretrained language models (PLM) like BERT(Devlin et al., 2018), BART(Lewis et al., 2019), and GPT2(Radford et al., 2019) are proposed. These models represent milestones in the development of language models, as they are based on the classical transformer architecture(Vaswani et al., 2023) and significantly increase the text generation capabilities of models. Initially, most of these models have relatively small sizes.

Research has shown that even by solely increasing model size while keeping model architecture similar, abilities on difficult tasks can substantially improve(Brown et al., 2020). This phenomenon of emerging abilities with scale is referred to as

emergent behavior(Wei et al., 2022). This has led to the development of LLMs which have profoundly impacted research and society. For example, the release of LLM has created much interest due to its strong text generation abilities like abstract writing and logical reasoning. This has catalyzed further research into LLMs, with models like LLaMA(Touvron et al., 2023a), LLaMA 2(Touvron et al., 2023b), Mistral 7B(Jiang et al., 2023a), GPT 4(OpenAI et al., 2023), and Mixtral 8x7B(Jiang et al., 2024a) demonstrating impressive performance on various tasks.

2.2 Prompt Engineering

Template-based prompts are among the early attempts at single-stage prompting (Paranjape et al., 2021).

However, the Chain of Thought (CoT) technique leads to more significant improvements in model capabilities (Wei et al., 2023) and attracts substantial interest. By providing a few reasoning demonstrations or "exemplars" in the prompt, CoT yields impressive performance gains. CoT also reveals LLMs' innate zero-shot reasoning abilities — simply prompting the model with "Let's think step-by-step!" enables complex inferential reasoning.

Additionally, prompt quality factors like reasoning complexity in exemplars, number of reasoning steps, and diversity of exemplars impact performance of LLM.

Since single-stage prompting may enable end-to-end reasoning, (Press et al., 2023) also explores constructing multi-stage prompts with follow-up questions and answers to provide detailed reasoning. (Jung et al., 2022) propose prompts based on trees of explanations generated abductively and recursively, e.g. X is true, because Y; Y is true, because...

(Zhou et al., 2023) find that decomposing complex questions into a series of simpler subquestions was beneficial for constructing effective prompts.

3 Method

3.1 GPT-4: From Zero-Shot to Few-Shot

Since GPT-3.5 and GPT-4 demonstrate strong performance on tasks like QA and text generation, we utilize these models to directly answer the training questions by providing the question and choice list.

For the zero-shot stage, we first explain what a word or sentence puzzle is in the prompt, presenting the question and options simultaneously. Then we use GPT-3.5 to predict answers one by one.

During this stage, we observe precision of only 17% for word puzzles on the training set. Errors frequently occur because many questions defy common knowledge, leading models to be overconfident in the "None of the above" choice. Therefore, we modify the prompt by appending "please don't choose 'None of the above', because in most cases, it is not the correct answer", increasing the precision to 66%.

We also notice some questions are too difficult for the model, such as "How many days are there in a month?". We think that providing the model with reasoning clues or demonstration may be beneficial. For this challenging sample, we guide the model to not only simply count days, but also approach the question from a new perspective — identifying which words on a calendar contain "day", like Monday and Tuesday, rather than numerals like January 1st.

A similar puzzle is "How many seconds are there in one year?". GPT-4 cannot find a correct answer if it counts the actual number of seconds in a year. We should tell it this is not to count the actual number of seconds and it should try to answer the question in another way, that is, to count the number of dates that contain second (2nd) in a year. For the hard sample "What is in front of a woman and at the end of a cow?", as an explanation, we tell GPT-4 this is a word game, and it should interpret the questions in two parts and find which letter is at the start/beginning of one word and at the end of the other word. For the question "What is at the end of a cow and in front of a woman?", we remind the model that the word woman starts with the letter "w", and the word cow ends with the letter "w". The correct answer is the letter "W". "What is at the beginning of eternity and the end of time?" For this question, the word "eternity" starts with the letter "e", and the word "time" ends with the letter "e". The correct answer is the letter "E". In this way, GPT-4 can think in the way we expect and correctly answer similar categories of brain teaser puzzles.

To address incorrectly answered examples, we identify and categorize over 20 challenging training instances to include in an extended prompt, as shown in 2. This prompt is designed to guide the model towards lateral thinking. Each illustrative example comprises the original question, choice

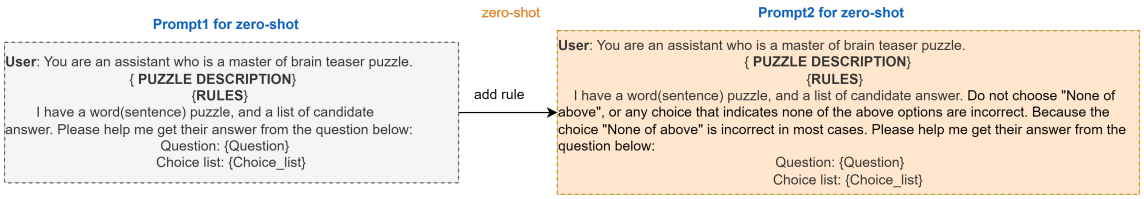

Figure 1: The left figure illustrates Prompt 1, which only provides the definition of a word or sentence puzzle before concatenating the question and choice list from the dataset. As the model tended to select 'None of the above', Prompt 2 adds a rule to avoid this answer. Both prompts are used in a zero-shot setting without examples.

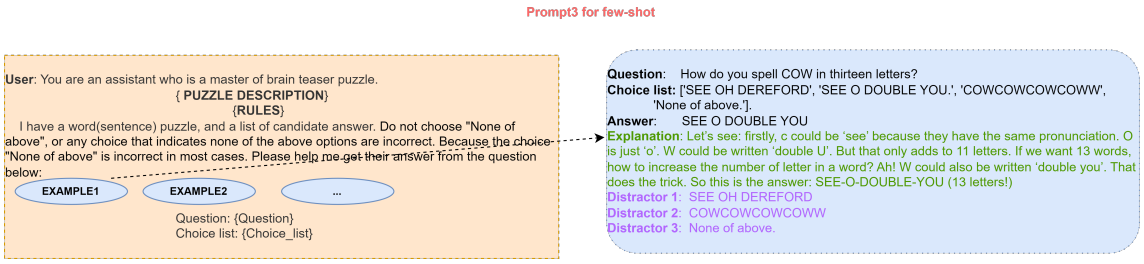

Figure 2: For the third strategy, we concatenate explanation of each example as well as distractors in the choice list. This is a few-shot strategy.

list, correct answer, and an explanatory reasoning clue extracted from the training data. Additionally, we find that supplementing each example with the three distractor options further improves GPT performance. Therefore, the full set of multiple-choice options is appended to each illustrated case. As depicted, these elements are combined to demonstrate the desired thought process.

3.2 Mixstral Fine-tuning

We also experiment with fine-tuning the Mixtral 7x8b model to predict solutions for these brain teaser puzzles. Mixtral 7x8b is a leading open-source LLM, comprised of a Mixture-of-Experts (MOE) architecture with approximately 45 billion parameters. It is regarded as state-of-the-art, outperforming models such as LLaMA 270B and GPT-3.5 on many benchmarks. Mixtral 7x8b offers both a base model and an instruct model, with the latter fine-tuned for enhanced performance on conversational tasks. Therefore, we select Mixtral-7x8b-instruct-v0.1 for fine-tuning on our dataset of around 1000 puzzle examples.

4 Experiment and Result

4.1 Experiment

Experiments are conducted on a test set to evaluate the three prompt designs introduced previously. Initially, GPT-3.5 was used to test Prompts 1 and 2 for subtask 1 (word puzzles). As the evaluation deadline approached, we switched to GPT-4 for greater efficiency. We evaluated Prompt 3 five times, and an ensemble voting strategy was adopted. Besides, we proceeded with only Prompt 3 (GPT4, with ensemble) for subtask 2's test set, omitting Prompts 1 and 2.

4.2 Result and Analysis

Experiment results on the training set are shown in Table 5, and our final results are shown in Table 6. As shown in Table 5, there is a substantial performance increase from Prompt 2 (GPT-3.5, zero-shot) to Prompt 3 (GPT-4, few-shot, with ensemble). This demonstrates the efficacy of our strategy utilizing Prompt 3 with GPT-4 in a few-shot learning setting. Besides, for the same question, GPT-4 would sometimes generate inconsistent answers or refuse to answer. To mitigate this, we ensemble the answers from 5 evaluations of each prompt by a voting strategy. This ensemble approach improves performance compared to single evaluations. Ultimately, we achieve an accuracy of 0.980 on the training subset.

	ft_mixtral_instruct
WP training set	0.21
SP training set	0.26

Table 4: Result of ft_mixtral_instruct

WP Training set (random 100 data samples)	Prompt 1 zero-shot	Prompt 2 zero-shot	Prompt 3 few-shot, with Ensemble
GPT-3.5	0.170	0.660	-
GPT-4	-	-	0.980

Table 5: Result on subtask 1: word puzzle. We use three kinds of prompt strategies on the training dataset for this subtask. We try GPT-3.5 to verify Prompt 1 and Prompt 2, and then use GPT-4 for Prompt 3. The latter strategy shows a much better performance.

SP Test Set	S_ori	S_sem	S_con	S_ori_sem	S_ori_sem_con	S_overall
	1.000	0.975	0.925	0.975	0.900	0.967
WP Test Set	W_ori	W_sem	W_con	W_ori_sem	W_ori_sem_con	W_overall
	0.969	0.938	1.000	0.938	0.938	0.969

Table 6: Final result on subtask 1 and subtask 2. We use three kinds of prompt strategies on the training dataset for the subtask. We try GPT-3.5 to verify Prompt 1 and Prompt 2, and then use GPT-4 for Prompt 3. The latter strategy shows a much better performance.

5 Conclusion

In conclusion, we demonstrate our prompt design method to enhance creative problem-solving in LLMs, enabling strong performance on brain teaser puzzles. Through prompting strategies and model fine-tuning, our methods attain 2nd and 3rd place rankings on this lateral thinking task. These results validate our techniques and highlight the potential for developing multifaceted reasoning skills in AI. Our work provides promising pathways toward more human-like language understanding and flexible thinking in natural language models. In summary, we take steps toward training AI systems capable of creative problem-solving.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b.

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. 2023b. 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.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, and ... 2023. Gpt-4 technical report.

Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021. Prompting contrastive explanations for commonsense reasoning tasks.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models.