BITS Pilani at SemEval-2024 Task 9: Prompt Engineering with GPT-4 for Solving Brainteasers

Dilip Venkatesh[1] and **Yashvardhan Sharma**[1]

[1]Birla Institute of Technology and Science, Pilani, Rajasthan, India
{f20201203, yash}@pilani.bits-pilani.ac.in

Abstract

Solving brainteasers is a task that requires complex reasoning prowess. The increase of research in natural language processing has lead to the development of massive large language models with billions (or trillions) of parameters that are able to solve difficult questions due to their advanced reasoning capabilities. The SemEval *BRAINTEASER* shared tasks consists of sentence and word puzzles along with options containing the answer for the puzzle. Our team uses **OpenAI's GPT-4** model along with **prompt engineering** to solve these brainteasers.

1 Introduction

There are two different types of thinking processes, vertical and lateral (Waks, 1997). Vertical thinking refers to the form of linear thinking thinking we are conditioned to. It is based on rationality and logic. Lateral thinking, or *"out-of-the-box"* thinking is a more creative way of thinking from different perspectives. This is contrary to first method.

The recent advancements of natural language processing models, more specifically large language models have achieved great progress in reasoning capabilities and therefore vertical thinking tasks (Talmor et al., 2019, Bisk et al., 2020).

This lateral, creative form of thinking has multiple use cases in the real world since rapid innovation and out of the box thinking are key functionalities of blooming institutions. Innovations are crucial to solve global scale problems like climate change and are very important to big tech companies to keep their consumers happy and engaged. Therefore an interesting part of language models are their abilities to show lateral thinking and defy default commonsense associations.

For the SemEval 2024 Task 9: *BRAINTEASER: A Novel Task Defying Common Sense* (Jiang et al., 2024) on CodaLab (Pavao et al., 2023), we aim to solve the brainteasers as a multiple-choice Question Answering (QA) tasks. Our team proposes a system for this where we use prompt engineering with GPT-4 to solve these brainteasers.

All of our code can be found on GitHub at https://github.com/dipsivenkatesh/SemEval-2024-Task-9

2 Background

2.1 Task and Data Description

The BRAINTEASER shared task[1] consists two different type of brainteasers/puzzles.

- **Sentence Puzzle**: Sentence-type brainteaser where the puzzle defying commonsense is centered on sentence snippets.
- **Word Puzzle**: Word-type brainteaser where the answer violates the default meaning of the word and focuses on the letter composition of the target question

We can find the examples of each puzzle in Table 1. In this paper we go through our team's system to solve both the sentence puzzle and word puzzle task.

The task requires us to to solve the brainteasers in the BRAINTEASER dataset (Jiang et al., 2023). The dataset was created by crawling the internet to find relevant puzzles. This is then filtered to remove irrelevant questions. The task is provided as a question-answering task in which for each puzzle we much select the correct answer from four options.

The task also consists of adversarial subsets to make sure that the approach is based on reasoning and not LLM memorization. The adversarial reconstructions are of two types.

[1]https://codalab.lisn.upsaclay.fr/competitions/15566

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 816–820
June 20-21, 2024

Question	Choices
Sentence Puzzle: A man shaves everyday, yet keeps his beard long	He is a barber. He wants to maintain his appearance. He wants his girlfriend to buy him a razor. None of the above.
Word Puzzle: What part of London is in France?	The letter N. The letter O. The letter L. None of the above.

Table 1: Sentence and Word puzzle examples.

- **Semantic Reconstruction** rephrases the original question without changing the correct answer and the distractors.
- **Context Reconstruction** keeps the original reasoning path but changes both the question and the answer to describe a new situational context.

We find instances of adversarial reconstructions in Table 2

2.2 Previous Work

The field of natural language processing has seen massive developments since the discovery of transformers (Vaswani et al., 2023). Initially used in machine translation, transformers found their way into other fields of natural language processing as well including large language models. These large language models like BERT (Devlin et al., 2019), LLaMA/Llama 2 (Touvron et al., 2023a, Touvron et al., 2023b) and OpenAI's GPT-3 (Brown et al., 2020) and GPT-4 (OpenAI, 2023) have powerful reasoning capabilities and can be applied on various tasks involving natural language.

Prompt engineering refers to structuring the input text for a large language model. Methods like prompt engineering and fine-tuning have tremendous efficacy on downstream tasks. If prompted on the role of the language model along with the input question and/or relevant data, language models do a good job on providing the correct output even in a zero-shot manner (Sanh et al., 2022).

There have been quite a few benchmarks for testing the creativity of automatic natural language systems. Identifying puns (Zou and Lu, 2019) and humour (Meaney et al., 2021) is an example of this. The shared task proposed in (Lin et al., 2021) tests the natural language understanding and creativity of it's systems by testing the systems on riddle style questions. This is pretty close to the `BRAINTEASERS` shared task that requires the system to automatically solve brainteasers. The commonsense reasoning ability of these language models are also tested with various benchmarks (Rajani et al., 2019, Ma et al., 2019, Lourie et al., 2021, Maharana and Bansal, 2022). These metrics provide a good analysis of the vertical thinking capabilities of the systems. However for the brainteaser task it is important to think in ways that go against common sense. It is also imperative for the model to understand the questions instead of just memorization as adversarial ways of forming the questions also exist in the task.

2.3 Evaluation Metrics

The systems will be evaluated on their accuracy in the question-accuracy tasks. The following two accuracy metrics are used.

- **Instance-based Accuracy**: where each question individual/adversarial are considered as a separate instance. The accuracy for the original question as well as both of the adversarial ways will be reported.
- **Group-based Accuracy**: This evaluates the accuracy of the original question along with its adversarial reconstructions combined. The value is only counted as correct if it gets all of these questions correct.

3 System Overview

3.1 GPT-4

We use the GPT-4 turbo as `gpt-4-1106-preview` model from the GPT-4 (OpenAI, 2023) family of models. We access the GPT-4 model using the OpenAI API. GPT-4 turbo has a 128,000 token context window and can solve difficult problems with

Adversarial Strategy	Question	Choice
Original	A man shaves everyday, yet keeps his beard long.	He is a barber. He wants to maintain his appearance. He wants his girlfriend to buy him a razor. None of the above.
Semantic Reconstruction	A man preserves a lengthy beard despite shaving every day.	He is a barber. He wants to maintain his appearance. He wants his girlfriend to buy him a razor. None of the above.
Context Reconstruction	Tom attends class every day but doesn't do any homework.	He is a teacher. He is a lazy person. His teacher will not let him fail. None of the above.

Table 2: Adversarial reconstructions of the brainteasers

greater accuracy than previous generation large language models. This is due to its broader general knowledge and advanced reasoning capabilities, its training data is up to the date of April 2023. We use the chat completions API in JSON mode to ensure that we get the correct option answer from the question passed to the model.

3.2 Prompts

We use prompt engineering with the roles of system prompts and user prompts to tell the model what to do and what instructions to follow.

- **Role Prompt**: You are an assistant that only responds in json. You solve riddles and brainteasers that require complex reasoning. Solve the riddle/brainteaser by selecting the correct option from the given option list. The response json should be in the format "optionindex": array index of the option selected from option list. this should be a zero-based index , "optionanswer": The answer selected from the given option list I only want the json output of this.

- **User Prompt**: Solve this brainteaser: (brainteaser question here) optionlist: (answer optionlist here)

With this we can see that we use one role prompt for the entire system, both sentences and word puzzles, and for the user prompt we specify the different questions and the options for the answer.

4 Experimental Setup

We load the BRAINTEASER test datasets (Jiang et al., 2023) provided to us by the BRAINTEASER shared task organizers using the HuggingFace datasets library (Lhoest et al., 2021). For the sentence puzzle we have 120 puzzles with 4 options corresponding to each puzzle and for the word puzzle we have 96 question s and for each question we have 4 options. The test set is unlabeled, it doesn't specify the correct option, and our systems must evaluate the correct option for each brainteaser.

We generate the prompts for each question with the methods specified above and pass them to the GPT-4 turbo chat completions API for solving the brainteasers.

5 Results

For evaluation, the organizers rank the system based on accuracy of the answers on the question-answering task. The GPT-4 with prompt engineering system that we have provided achieves 9th place on the leaderboard in the evaluation phase [2]. The performance of the system on all the different evaluation components can be found in Table 3 for the sentence puzzle and in Table 4 for the word puzzle.

[2] https://codalab.lisn.upsaclay.fr/competitions/15566#results

Team	Original	Semantic	Context	O & S	O & S & C	Overall
GPT-4 + prompt engineering	97.5	92.5	80.0	92.5	77.5	90.0
Human	90.74	90.74	94.44	90.74	88.89	91.98
ChatGPT (zero-shot)	60.77	59.33	67.94	50.72	39.71	62.68
RoBERTa-L	43.54	40.19	46.41	33.01	20.10	43.38

Table 3: Sentence puzzle result.

Team	Original	Semantic	Context	O & S	O & S & C	Overall
GPT-4 + prompt engineering	0.938	0.938	0.875	0.938	0.812	0.917
Human	91.67	91.67	91.67	91.67	89.58	91.67
ChatGPT (zero-shot)	56.10	52.44	51.83	43.90	29.27	53.46
RoBERTa-L	19.51	19.51	23.17	14.63	6.10	20.73

Table 4: Word puzzle result.

Acknowledgements

I would like to thank the organizers of the *SemEval 2024 BRAINTEASER: A Novel Task Defying Common Sense* shared task for conducting this competition and providing us with the data. I would also like to thank the faculty and research scholars at BITS Pilani for assisting me in my work.

References

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

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. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Yifan Jiang, Filip Ilievski, and Kaixin Ma. 2024. 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.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. 2021. Riddlesense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark.

Kaixin Ma, Jonathan Francis, Quanyang Lu, Eric Nyberg, and Alessandro Oltramari. 2019. Towards generalizable neuro-symbolic systems for commonsense question answering. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 22–32, Hong Kong, China. Association for Computational Linguistics.

Adyasha Maharana and Mohit Bansal. 2022. On curriculum learning for commonsense reasoning. In *Proceedings of the 2022 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 983–992, Seattle, United States. Association for Computational Linguistics.

J. A. Meaney, Steven Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 105–119, Online. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Dinh-Tuan Tran, Xavier Baro, Hugo Jair Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. 2023. Codalab competitions: An open source platform to organize scientific challenges. *Journal of Machine Learning Research*, 24(198):1–6.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

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

Yanyan Zou and Wei Lu. 2019. Joint detection and location of English puns. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2117–2123, Minneapolis, Minnesota. Association for Computational Linguistics.