

OUNLP at SemEval-2024 Task 9: Retrieval-Augmented Generation for Solving Brain Teasers with LLMs

Vineet Saravanan
Cranbrook Schools
Bloomfield Hills, MI
vineetsaravanan@gmail.com

Steven Wilson
Oakland University
Rochester, MI
stevenwilson@oakland.edu

Abstract

The advancement of natural language processing has given rise to a variety of large language models (LLMs) with capabilities extending into the realm of complex problem-solving, including brainteasers that challenge not only linguistic fluency but also logical reasoning. This paper documents our submission to the SemEval 2024 Brainteaser task, in which we investigate the performance of state-of-the-art LLMs, such as GPT-3.5, GPT-4, and the Gemini model, on a diverse set of brainteasers using prompt engineering as a tool to enhance the models' problem-solving abilities. We experimented with a series of structured prompts ranging from basic to those integrating task descriptions and explanations. Through a comparative analysis, we sought to determine which combinations of model and prompt yielded the highest accuracy in solving these puzzles. Our findings provide a snapshot of the current landscape of AI problem-solving and highlight the nuanced nature of LLM performance, influenced by both the complexity of the tasks and the sophistication of the prompts employed. All the code, along with the data used, is available on our GitHub¹

1 Introduction

The pursuit of creating artificial intelligence models with advanced reasoning and problem-solving capabilities has led researchers down the path of deploying brainteasers as a benchmark for AI systems' linguistic and reasoning prowess. These brainteasers are more than trivial or recreational challenges; they are testaments to the complexity of human cognition, embedding layers of semantics, pragmatics, and world knowledge that remain elusive to AI systems. The gulf between the operational logic of current AI models and the intricate understanding displayed by the human mind

is significant, particularly in domains necessitating advanced reasoning and a robust common sense foundation. This disparity is not only observed but keenly felt in the context of AI systems' interaction with human language and thought (Mahowald et al., 2023).

The limitations of pattern recognition as the mainstay of AI systems' learning mechanisms have been critically examined, sparking a discourse that emphasizes the imperative for AI systems to transcend these confines. Rigorous benchmarks that challenge AI systems to demonstrate inferential reasoning are essential to catalyze this evolution (Sawada et al., 2023). Brainteasers emerge as one medium through which AI systems' competencies can be evaluated. They are not simply puzzles to be solved but are reflective of the complex, often ambiguous nature of human communication and problem-solving.

The BRAINTEASER task introduced at SemEval 2024 (Jiang et al., 2024) is part of this evolution of AI system assessment, standing at the center of linguistic analysis and computational intelligence. It is designed to evaluate what machines can understand and how they can apply this understanding in a manner similar to human thought processes. Language models, such as GPT-3.5² and GPT-4 (OpenAI, 2023), are increasingly being subjected to these tests to gauge their mastery over language and logic, as demonstrated in recent comparative analyses (Espejel et al., 2023). The BRAINTEASER task's format, which intertwines linguistic cues with logical conundrums, requires systems to not only comprehend the text at a superficial level but to delve into the implied, the inferred, and the intuitive aspects that are second nature to human beings.

By benchmarking language models against brainteasers within the framework of the BRAIN-

¹<https://github.com/VSPuzzler/OUNLP-at-SemEval-2024-Task-9>

²<https://openai.com/blog/chatgpt>

TEASER task, we are able to learn more about the current capabilities of popular LLMs. This work can help to provide a direction for future research by pinpointing where current models fall short and where the next wave of innovation is urgently needed.

Our approach involved testing different LLM models. We web-scraped example riddles and used them as an example for the model. Additionally, we tested with the closest riddle and the most different riddle and found that GPT-4.0 oneshot with a similar riddle worked best for the Sentence Puzzle and GPT-4.0 oneshot with a different riddle worked best for the Word Puzzle. The Word Puzzle turned out to be a significantly harder task than the Sentence Puzzle.

2 Related Work

The exploration of reasoning abilities in large language models (LLMs) has been the focus of several studies in recent years. Notably, work by OpenAI provides foundational insights into the capabilities of GPT-3, especially highlighting its potential in solving reasoning tasks through few-shot learning (Brown et al., 2020). This work is particularly relevant as it demonstrates how providing a few examples can significantly improve an LLM’s ability to solve reasoning problems, akin to the one-shot and few-shot techniques examined in our study.

Furthering the discussion on reasoning, work has been done that discusses the ‘chain-of-thought’ (CoT) prompting method, where models are guided to articulate intermediate steps when solving complex tasks (Wei et al., 2022). This process is similar to the explanation method in solving the brainteasers, which encourages models to elaborate on their reasoning, leading to improved performance.

The brainteasers in the training data provided often require making analogies and similarities in reasoning. Work has been done that offers an analysis of how word embeddings capture semantic relationships, which can be fundamental in retrieving similar examples to aid reasoning (Allen and Hospedales, 2019). This is directly linked to our one-shot similar and few-shot in-context learning approaches, where the ability of an LLM to use analogous examples influences its problem-solving effectiveness.

Moreover, the strategies for solving brainteasers with AI systems have been enriched by incorporating external knowledge bases. An investigation

has been conducted into the inherent knowledge within language models and their ability to function as knowledge bases (Petroni et al., 2019). The integration of external knowledge is particularly pertinent to tasks requiring common sense and real-world information, underscoring the importance of knowledge retrieval in the context of a brainteaser. Lastly, a pivotal study has been done that introduces a dataset designed to probe AI systems’ common sense reasoning capabilities (Talmor et al., 2018). This study aligns with our aim in solving brainteasers to evaluate the capacity of LLMs to handle questions that necessitate an understanding of the world as humans perceive it.

3 Methodology

Our experimental design relies on prompt engineering to explore the effectiveness of language models in solving brainteasers. In this study, we experimented with different prompt structures to determine their impact on the model’s performance. The primary prompt format tested was structured as follows: “Please pick the best choice for the brain teaser. Each brain teaser has only one possible solution, including the choice ‘none of the above.’ The answer should only provide the choice text.” This directive was chosen to explicitly instruct the model to select a single, most appropriate answer from a set of given options. To ensure a controlled variable, we explicitly presented the model with the choices, observing how it navigates the selection process when options are directly provided.

An interesting observation was made regarding the specification of the type of brainteaser. Initially, it was hypothesized that indicating whether the puzzle was a ‘word puzzle’ or a ‘sentence puzzle’ would aid the models in narrowing down their reasoning scope, thereby improving accuracy. However, the results indicated that such specifications did not significantly affect the models’ performance. This finding suggests that the models possess a level of task generalization, wherein they apply similar reasoning processes to both types of puzzles without the need for explicit differentiation.

Furthermore, we explored the effect of including choices within the prompt. By contrasting scenarios with and without provided options, we aimed to assess whether the presence of choices would guide the model to a correct answer more efficiently. Prompts structured to request the model

to pick from provided choices explicitly did not significantly alter the success rate compared to when no choices were given. This aspect of the study aimed to discern the degree to which the models rely on contextual clues versus intrinsic problem-solving capabilities. For example, in a test with a bad prompt on GPT-3.5 without choice, a 27.60% accuracy was achieved, while with the same prompt with choices, a 28.80% accuracy was achieved. Due to the slight increase in accuracy, we decided to include choices for the rest of the prompts used.

We also implemented explanation and chain of thought (CoT) reasoning to guide the language model toward a more structured and reasoned approach when tackling brain teasers. Explanation prompts encouraged the model to articulate the rationale behind its chosen answers. Similarly, CoT prompts aimed to simulate a step-by-step reasoning process, mirroring how humans might approach problem-solving.

To further enhance the accuracy of our language model in solving brain teasers, we adopted a one-shot in-context learning approach, leveraging a large dataset of riddles as context for the model. We extracted a comprehensive collection of 3,899 riddles by downloading texts from the riddles.com website, including both the questions and their corresponding answers. This dataset was a reference for the model to draw upon when presented with new puzzles.

We employed two distinct strategies for selecting a relevant riddle from this dataset to use as an example in our one-shot method. The first strategy aimed to identify the riddle most *similar* to the brain teaser in question, believing that a similar context might prime the model more effectively for the task at hand. Conversely, the second strategy sought out the riddle most *dissimilar* to the brain teaser, hypothesizing that a contrasting example could stimulate a broader range of the model’s reasoning capabilities.

To facilitate the rapid identification of the most similar or dissimilar riddle, we encoded each riddle question into a vector representation using the all-MiniLM-L6-v2 model (Reimers and Gurevych, 2019) in the SentenceTransformers³ Python package. This allowed us to compute the cosine similarity between the vector representation of the new brain teaser and those of the riddles in

³<https://sbert.net>



Figure 1: Example prompts tested

our dataset. Using this similarity score, we could efficiently identify the riddle that was either closest or farthest in semantic space from the brain teaser presented to the model. The highest cosine similarity is used as the similar riddle, while the lowest cosine similarity is used as the dissimilar riddle.

This approach significantly improved the efficiency of selecting an appropriate example riddle, enabling a more streamlined integration of the one-shot learning method into our experimental setup. The rationale behind using cosine similarity was to leverage the high-dimensional space in which language representations reside, making it possible to quantify the semantic proximity between different textual inputs effectively. A depiction of the different elements of different prompts is depicted in Figure 1.

In an effort to enhance the accuracy of solutions provided by the language model for word puzzles, we also experimented with an innovative approach modeled after collective human problem-solving dynamics. This method involved simulating a "council" of three hypothetical individuals engaged in a discussion about a puzzle, with the aim of reaching a consensus on the answer.⁴ The intent was to emulate the collaborative approach often used in human group problem-solving, where different perspectives and thought processes can lead to more accurate solutions. The implementation of this method required the model to generate three distinct responses, each purportedly from a different "council member," who would then "discuss" their reasoning and perspectives on the puzzle. Following this simulated deliberation, the model was prompted to synthesize the viewpoints into a single, collective answer (see example in Appendix A). Despite the creative nature of this technique, the results were not as promising as anticipated. The accuracy of the word puzzles did not show significant improvement using the council-based discussion method. This outcome suggests that while the approach mirrors human group interactions, it may not translate effectively within the constraints of a single AI model's processing capabilities.

4 Results

We ran and tested the LLMs on the training set (Jiang et al., 2023) with each combination of model and prompt. This allows us to get a comprehensive view of the performance of the LLMs across different prompts. It is important to increase the general accuracy across all LLMs by adding more information about the question to the prompt, along with examples. The models we evaluated included GPT-3.5, GPT-4, Gemini Pro (Team et al., 2023), and a suite of memory-efficient language models from the languagemodels repository.⁵ The models used from this package included neural-chat-7b-v3-1, flan-alpaca-xl, flan-alpaca-gpt4-xl, flan-t5-xl, fastchat-t5-3b-v1.0, LaMini-Flan-T5-783M, flan-t5-large, LaMini-Flan-T5-248M, flan-alpaca-base, flan-t5-base, dialogstudio-t5-base-v1.0, LaMini-Flan-T5-77M, flan-t5-small, phi-1_5, LaMini-GPT-774M, and LaMini-GPT-124M. This set explores sixteen models, yet they never outperformed the

⁴<https://github.com/dave1010/tree-of-thought-prompting>

⁵<https://github.com/jncraton/languagemodels>

other LLMs (GPT and Gemini series). Therefore, we only report the results from the best of these models for each prompt and task in the final result tables, Table 1 and Table 2.

Prompt	GPT-3.5	GPT-4.0	Gemini	languagemodels
basic	0.288	0.649	0.803	0.359
task desc.	0.477	0.645	0.753	0.383
+ CoT	0.722	0.692	0.671	0.314
+one-shot sim.	0.650	0.809	0.753	0.633
+one-shot diff.	0.680	0.825	0.759	0.345
+one-shot sim. + CoT	0.710	0.686	0.637	0.686
+one-shot diff. + CoT	0.670	0.704	0.655	0.347

Table 1: Accuracy of LMs using different prompts on the **Sentence Puzzle** task. **Bold** indicates the best model for a given prompting strategy, and **underlined** indicates the best overall approach for the task. The languagemodels column shows the best score achieved by any model from the languagemodels library.

Prompt	GPT-3.5	GPT-4.0	Gemini	languagemodels
basic	0.346	0.508	0.531	0.341
task desc.	0.341	0.487	0.494	0.354
+ CoT	0.520	0.641	0.351	0.323
+one-shot sim.	0.485	0.649	0.530	0.553
+one-shot diff.	0.470	0.621	0.505	0.356
+one-shot sim. + CoT	0.553	0.540	0.384	0.242
+one-shot diff. + CoT	0.513	0.586	0.354	0.333

Table 2: Accuracy of LMs using different prompts on the **Word Puzzle** task. **Bold** indicates the best model for a given prompting strategy, and **underlined** indicates the best overall approach for the task. The languagemodels column shows the best score achieved by any model from the languagemodels library.

It was found that the one-shot method consistently had the top 2 accuracy in the prompts studied, proving the efficiency of one-shot methods for LLMs in general. Gemini's accuracy when only using the basic prompt was very high for the Sentence Puzzle task, which shows Gemini's versatility and adaptability to different questions with high accuracy without needing examples to perform well. The GPT-4 system with the basic prompt with a chain-of-thought method also proved to be highly accurate.

The Chain of Thought approach has been shown

to improve accuracies for LLMs. Despite these efforts, we observed that this strategy did not lead to a measurable increase in accuracy. This outcome suggests that while such prompts can often lead to more interpretable answers, they do not necessarily enhance the model’s ability to deduce the correct solution in the context of brain teasers. Further research may explore whether the complexity of the puzzles or the inherent limitations of the models’ understanding contributed to this result.

Despite the overall increase in accuracy observed with more informative prompts, GPT-4 did not always outperform GPT-3.5 with prompts like the task description + explain. This highlights that while advancements in model architecture contribute to enhanced performance, they do not guarantee superior outcomes in every scenario, particularly in specialized tasks like puzzle-solving. Since GPT-4 is trained on a larger range of data sources to improve general performance across a broad range of tasks, the generalized training approach may lead GPT-4 not performing as well in this specific task.

Furthermore, we found that employing chain-of-thought and explanation methodologies did not significantly improve performance in this context. This deviation from expected outcomes may indicate that for certain types of puzzles, these approaches do not align with the models’ strengths or the nature of the problem-solving process required.

The performance of open-source packages like `language-models` was notably lower compared to their commercial counterparts. This gap underscores the developmental distance that open-source models need to traverse to reach the sophistication level of models like GPT-4 or Gemini, suggesting that access to extensive datasets, computing resources, and proprietary algorithms plays a significant role in model performance. Typically the models from this set that worked best were the `LaMini-Flan-T5` class, which was always the case for the one-shot setting. The main exceptions to this were in the zero-shot scenario, specifically with the basic prompt (neuralchat worked best for Sentence Puzzle and `dialogstudio` worked best for Word Puzzle) and the task description prompt (phi-1.5 for SP and neuralchat for WP). These cases provide positive examples of situations in which lightweight, open models are more competitive with proprietary, closed models.

Additionally, the increased difficulty of word puzzles presents a notable challenge, potentially

due to their reliance on nuanced understanding, cultural context, and semantic associations that can be challenging even for human solvers. This complexity is reflected in the lower accuracy rates across all models for word puzzles when compared to sentence puzzles, implying that word puzzles may represent a closer analog to human-level problem-solving and, as such, provide a more stringent test of AI reasoning and language capabilities.

The three submission prompts submitted are GPT-4 one-shot different for Sentence Puzzle and GPT-4 one-shot similar for word puzzles, GPT-4 one-shot similar for Sentence Puzzle and GPT-4 one-shot different for word puzzles, and Gemini basic for Sentence Puzzle and GPT-4 basic + CoT for word puzzle. The first submission received an accuracy score of 0.925 for the sentence puzzle and 0.9375 for the word puzzle, the second a score of 0.95 for the sentence puzzle and 0.78125 for the word puzzle, and the third a score of 0.625 for the sentence puzzle and 0.46875 for the word puzzle. Note that these accuracies are from the “Original” riddles. The second submission was the highest and ranked us 14th in terms of average score, 11th on the Word Puzzle task, and 10th on the Sentence Puzzle task under the name `vspuzzler`. Our submitted system performed exceptionally well on the “Original” version of the brainteasers (ranking 3rd overall for Sentence Puzzle and 7th for Word Puzzle within this subcategory) but underperformed on the “Context Reconstruction” variations of the brainteasers in which the original reasoning path was used within a new situational context.

5 Conclusion

This study’s evaluation of LLMs across a range of prompt types provides insights into the strengths and limitations of current AI systems in solving brain teasers. Our findings revealed that in-context learning methods are highly effective, particularly for the Gemini model and GPT-4, when solving sentence puzzles. However, GPT-4 did not consistently outperform GPT-3.5 across all prompt types, which suggests that the latest models do not always guarantee an improvement in task-specific performance. The chain-of-thought and explanation strategies, while enhancing interpretability, did not necessarily translate into higher accuracy, indicating the need for further research into how these models process complex language tasks. The performance gap between proprietary and open-source

models highlights the significant role of resources and proprietary technology in developing LLMs. The increased difficulty of word puzzles suggests that tasks requiring a nuanced understanding and cultural context remain challenging for AI systems, closely mirroring the complexity of human cognition. This study underscores the importance of tailored prompting strategies to leverage the capabilities of LLMs and the potential for future advancements in AI-based problem-solving.

References

- Carl Allen and Timothy Hospedales. 2019. Analogies explained: Towards understanding word embeddings. In *International Conference on Machine Learning*, pages 223–231. PMLR.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Jessica López Espejel, El Hassane Ettifouri, Mahaman Sanoussi Yahaya Alassan, El Mehdi Chouham, and Walid Dahhane. 2023. Gpt-3.5, gpt-4, or bard? evaluating llms reasoning ability in zero-shot setting and performance boosting through prompts. *Natural Language Processing Journal*, 5:100032.
- 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.
- Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*.
- R OpenAI. 2023. Gpt-4 technical report. *arXiv*, pages 2303–08774.
- Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Tomohiro Sawada, Daniel Paleka, Alexander Havrilla, Pranav Tadepalli, Paula Vidas, Alexander Kranias, John J Nay, Kshitij Gupta, and Aran Komatsuzaki. 2023. Arb: Advanced reasoning benchmark for large language models. *arXiv preprint arXiv:2307.13692*.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

A Appendix

Chain of thought was used as one of the prompts. Below is an example of it in use:

Prompt:

Please pick the best choice for the sentence puzzle brain teaser (Sentence-type brain teaser where the puzzle defying commonsense is centered on sentence snippets). Each brain teaser has only one possible solution including the choice none of above, answer should only provide the choice:

Question: Mr. and Mrs. Mustard have six daughters and each daughter has one brother. But there are only 9 people in the family, how is that possible? Choice:

- (A) Some daughters get married and have their own family.
- (B) Each daughter shares the same brother.
- (C) Some brothers were not loved by family and moved away.
- (D) None of above.

Answer:

```
{"response": "B",  
"explanation": "Each daughter has one  
brother, which means there is only one  
brother shared among all six daughters.  
If Mr. and Mrs. Mustard have six daugh-  
ters and one son, together with the two  
of them, that makes a total of nine family  
members. Therefore, the correct answer  
is choice B." }
```

A council method was also used as one of the prompts. Below is an example of it in use:

Prompt:

Please pick the best choice for the sentence puzzle brain teaser (Sentence-type brain teaser where the puzzle-defying commonsense is centered on sentence snippets). Each brain teaser has only one possible solution including the choice none of above, answer should only provide the choice. Imagine three different experts are answering this question. All experts will write down 1 step of their thinking, then share it with the group. Then all experts will go on to the next step, etc. If any expert realises they're wrong at any point then they leave.

Question:Mr. and Mrs. Mustard have six daughters and each daughter has one brother. But there are only 9 people in the family, how is that possible? Choice:

- (A) Some daughters get married and have their own family.
- (B) Each daughter shares the same brother.
- (C) Some brothers were not loved by family and moved away.
- (D) None of above.

Answer:

```
{"response": "B",  
"Expert1Thinking": "If there are six  
daughters and each has one brother, it  
would initially seem like there should be  
six brothers. However, the question is  
phrased in a way that suggests that each  
daughter has the same single brother.",  
"Expert2Thinking": "Taking into ac-  
count that there are Mr. and Mrs. Mus-  
tard, the six daughters can collectively
```

have only one brother to make the total count nine. This brother would be shared among all sisters.",

```
"Expert3Thinking": "It's a puzzle meant  
to trick the reader into thinking there  
are more people than there actually are.  
The family must consist of Mr. and Mrs.  
Mustard, their six daughters, and one son  
who is the brother of all daughters. So,  
there's only one brother, not six." }
```