GeminiPro at SemEval-2024 Task 9: BrainTeaser on Gemini

Kyu-Hyun Choi and **eung-Hoon Na**
Division of Computer Science and Engineering, Jeonbuk National University, South Korea
chlrbgus321@naver.com and nash@jbnu.ac.kr

Abstract

It is known that human thought can be distinguished into lateral and vertical thinking. The development of language models has thus far been focused on evaluating and advancing vertical thinking, while lateral thinking has been somewhat neglected. To foster progress in this area, SemEval has created and distributed a brainteaser dataset based on lateral thinking consist of sentence puzzles and word puzzle QA. In this paper, we test and discuss the performance of the currently known best model, Gemini, on this dataset.

1 Introduction

Human thought is known to be distinguished into lateral and vertical thinking. (Jiang et al., 2023) cites (Waks, 1997) in mentioning, based on modern neuroscience, that vertical thinking is associated with the left hemisphere of the brain, while lateral thinking is associated with the right hemisphere. Moreover, this paper notes that during the development of language models, there has been a focus on problem-solving abilities in vertical thinking, neglecting the capabilities based on lateral thinking. This paper anticipates that lateral thinking puzzles may not be easily solved with just additional adaptations and extensions of the LLM (Large Language Model) approach, yet this paper is prepared to counter that expectation. It evaluates the performance of Google's ambitious model, Gemini(Team et al., 2023), which is said to surpass GPT-4, by measuring performance solely through changes in demonstration, as it was not possible to fine-tune Gemini.

Gemini is anticipated to show increased performance due to the scaling law mentioned in (Kaplan et al., 2020), as it utilizes significantly more parameters than GPT-4. Being a multimodal model trained with additional learning resources such as visual and auditory inputs, it is speculated that these characteristics might give rise to unique emergent abilities. (Wei et al., 2022) These two aspects are expected to contribute to performance improvements in lateral thinking.

Our approach is straightforward. First, we formalize Gemini's responses by adding demonstrations, following the same few-shot provision method as used by SemEval, and second, we provide only the relevant task few-shot examples for the two brainteaser tasks: sentence puzzles and word puzzles. Through the second method, we investigate whether providing clear few-shot examples for tasks alone can aid in performance improvement.

2 Background

2.1 Vertical thinking

As illustrated in Figure 1 of the (Jiang et al., 2023), vertical thinking is generally considered a logical form of thought. The first example of vertical thinking in Figure 1 of the paper, regarding the question "How do you flood a room?", involved associating the meaning of the word "flood" with the span "Cover with water". This association led to the selection of a similar meaning, "Fill it with water", as the answer. The second example of vertical thinking was in response to the question "I have five fingers, but I am not alive. What am I?". Here, the span "five fingers" led to the association of a similar span "Five separate parts", and "Not alive" led to the association of "item like a hand", which, despite "not alive" having a broader meaning, was contextually restricted by the span "five fingers". The only option that simultaneously had the properties of "Five separate parts" and "item like a hand" was "Glove". This problem, even though it is a riddle as mentioned in brainteasers, could be solved through vertical thinking. Such an ability to associate a specific word span with another span of similar meaning could be implemented in transformer

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1602–1606

models, as mentioned in (Dai et al., 2021), where the feed-forward network contains knowledge, and the context patterns created in the attention layer act as a key, enabling the association of a particular part of the input with another similar span.

2.2 Lateral thinking

Let's look at the first example of lateral thinking from Figure 1 of the overview paper. It is common sense to associate "Man shaves everyday" with "His beard gets clean everyday". However, the condition "yet keeps his beard long" blocks this inference path. Therefore, the model must use a different reasoning path, and to solve the problem, it must break away from the common sense that the man shaves himself and instead think of the possibility that he shaves someone else. This example forces the most commonsensical reasoning path to be blocked and requires navigating an alternative reasoning path.

The second example asks, "What type of cheese is made backwards?" This question is not commonsensical in itself. However, if "made" is not considered as a verb but as a sequence of letters, the problem is solved. Reversing "made" spells "edam," which is a type of cheese.

2.3 Brain-Teaser Benchmark

Brain teaser tasks (Jiang et al., 2024) are designed to explore whether language models are capable of lateral thinking, diverging from traditional methods. These tasks involve reading a question and providing an answer in a QA format, structured as a multiple-choice question with options (A), (B), (C), (D) to ensure clear output.

Sentence puzzles involve semantic exercises that break conventional thinking, while word puzzles use arrangements of alphabets in words to provide answers that play on words, challenging common sense.

There are two variations of both sentence and word puzzles. One is semantic reconstruction, where the question is paraphrased to measure if the problem can still be solved effectively while the answer and options remain unchanged. The other is context reconstruction, where the thought process to solve the problem remains the same, but the question and options are changed.

Two methods are used to measure performance: instance-based accuracy, which measures the accuracy of original, semantic, and context reconstructions separately, and group-based accuracy, which increases accuracy if the original and semantic reconstructions are answered correctly together or if correct answers are provided for original, semantic, and context reconstructions all at once.

Approximately 1,000 training examples were provided by Semeval, but this study measures the intrinsic ability of Gemini without using the training dataset.

3 System overview

In this study, we used Gemini, an ambitious model released by Google, known to surpass ChatGPT. We utilized the Gemini-Pro API and followed the ChatGPT evaluation method provided by SemEval.

3.1 Add Demonstration

Gemini tends to include explanations in its responses, resulting in varying output styles for each question. For example, it can be feel like this:

The answer is (A), because [explanation.....]

[explanation.....] so, the answer is (A)

(B) is [explanation.....]
(C) is [explanation.....]
(D) is [explanation.....]
so, the answer is (A)

To use the brain teaser score calculator, the answers must be clear in the form of (A), (B), (C), or (D). The output style described above is not suitable for input into the answer calculator, especially in the last example where all options (A), (B), (C), and (D) are included in the output. Implementing an algorithm to post-process this and select a clear single answer, like (A), from such outputs is complex. Therefore, to avoid these difficulties, we structured the demonstration to include the following feel.

[demonstration...]
question
option (A)
option (B)
option (C)
option (D)

3.2 Use only relevant few-shot examples

To determine if providing only sentence puzzle examples for sentence puzzles or only word puzzle

examples for word puzzles helps resolve confusion between examples and aids in problem-solving, we conducted 1-shot, 2-shot, and 4-shot evaluations using the same set of examples.

In this case, we did not add a demonstration because the few-shot examples clearly provide the style of output. When using only relevant few-shot examples, we follow this format:

For sentence puzzles:

N examples
[sentence puzzle question
option (A)
option (B)
option (C)
option (D)
Answer: (A) or (B) or (C) or (D)]

problem
[sentence puzzle question
option (A)
option (B)
option (C)
option (D)
Answer:]

For word puzzles:

N examples
[word puzzle question
option (A)
option (B)
option (C)
option (D)
Answer: (A) or (B) or (C) or (D)]

problem
[word puzzle question
option (A)
option (B)
option (C)
option (D)
Answer:]

4 Experimental setup

Although SemEval provided aproxymately 1,000 training examples, this study did not use the training data as it did not involve fine-tuning. Instead, we directly used the brain teaser test data to measure the intrinsic capabilities of Gemini-Pro. Gemini-Pro occasionally does not output an answer. In such cases, we considered (D) as the answer. If it does not output a response in the structured form of (A), (B), (C), (D), we also treated it as (D). For all other cases, we followed the ChatGPT methodology as outlined by SemEval.

5 Results

5.1 With Demonstration

As shown in table 1, for zero-shot, sentence puzzle performance was generally superior to chatGPT, except it showed exceptionally lower performance in context reconstruction. In few-shot, when two examples were provided, it only showed superiority in original, and tied with four-shot in ori&sem&con, while four-shot generally showed superior performance elsewhere, and performance actually decreased in eight-shot.

For word puzzles, zero-shot performance was superior to chatGPT in original, semantic, and context, but uniquely showed lower performance in Ori&sem and ori&sem&con. In few-shot, original showed overwhelming performance in two-shot, semantic was superior in eight-shot, and context had the best performance in four-shot. Overall, the best performance was seen in eight-shot.

5.2 Without Demonstration and Use only relevant few-shot examples

When only sentence puzzle examples were provided for sentence puzzles, the performance in two-shot and four-shot was comparable to the original method. In two-shot, original performance dropped by 10 points, semantic increased by 8 points, context increased by 3 points, ori&sem dropped by 3 points, and ori&sem&con dropped by 5 points, with overall scores remaining the same. In four-shot, original remained unchanged, semantic dropped by 3 points, context dropped by 10 points, Ori&sem increased by 8 points, and ori&sem&con remained the same, with overall dropping by 5 points.

For word puzzles in two-shot, original dropped by 10 points, semantic increased by 6 points, context increased by 4 points, but uniquely, performance remained the same in ori&sem and ori&sem&con, with overall performance unchanged. In four-shot, original performance increased by 10 points, semantic by 19 points, context dropped by 3 points, ori&sem increased by 16 points, and ori&sem&con by 16 points, with an

	Instance-based			Group-based		Overall
	Original	**Semantic**	**Context**	**Ori & Sem**	**Ori & Sem & Con**	
With Demonstration						
Sentence Puzzle						
Zero-shot	0.67	0.62	0.62	0.52	0.4	0.64
Two-shot	0.75	0.67	0.72	0.6	0.57	0.71
Four-shot	0.72	0.7	0.75	0.67	0.57	0.72
Eight-shot	0.7	0.67	0.67	0.57	0.45	0.68
Word Puzzle						
Zero-shot	0.65	0.53	0.53	0.43	0.25	0.57
Two-shot	0.78	0.78	0.71	0.5	0.40	0.69
Four-shot	0.69	0.56	0.87	0.43	0.40	0.69
Eight-shot	0.71	0.71	0.84	0.59	0.53	0.76
Without Demonstration and Use only relevant few-shot examples						
Sentence Puzzle						
One-shot	0.72	0.72	0.62	0.65	0.52	0.69
Two-shot	0.65	0.75	0.75	0.57	0.57	0.71
Four-shot	0.72	0.67	0.65	0.65	0.57	0.67
Word Puzzle						
One-shot	0.75	0.59	0.78	0.78	0.56	0.70
Two-shot	0.68	0.65	0.75	0.5	0.40	0.69
Four-shot	0.75	0.75	0.84	0.59	0.56	0.76

Table 1: Result of evaluation on Gemini-Pro

overall increase of 7 points.

Sentence puzzles showed a tendency for scores to drop, regardless of how the examples were organized, making it unclear whether the scores dropped randomly. Word puzzles showed a tendency for significant performance increases, but with only 96 test examples for word puzzles and no clear direction in the fluctuations of scores, it is uncertain whether the performance increase was due to providing only word puzzle examples or if the performance randomly improved.

6 Conclusion

As observed in Figure 2 of (Jiang et al., 2023), increasing the number of examples in sentence puzzles did not consistently improve performance, and while an overall upward trend in performance for word puzzles was noted, it did not improve regularly. Similarly, in the experiments of this paper, performance fluctuations with the number of examples were erratic, but it is clear that performance is generally higher compared to chatGPT. The leaderboard for brainteasers often shows many cases scoring over 90, which is likely due to the use of fine-tuning methods on the brainteaser training set. Without training specialized for brainteasers, the effect of using the method of demonstration appears to be minimal or almost nonexistent in a pure model state, and it has been found that larger models exhibit more pronounced performance improvements. Particularly, Gemini, despite being a multimodal model trained with both visual and auditory inputs, significantly underperforms compared to human capabilities. Contrary to the original paper's expectation, it was observed that merely increasing the model size could spontaneously develop problem-solving abilities for lateral thinking tasks, suggesting that even the capability for lateral thinking falls within the range of emergent abilities. According to (Jawahar et al., 2019), as the training of transformer models progresses, layers specialized for tasks are formed, with it being speculated that lateral associations are made in highly differentiated semantic layers in layers closer to the end. Perhaps the improvement in performance in LLMs, as mentioned in the context of brainteasers, might simply be due to memorizing content from the corpus.(Carlini et al., 2022) It remains to be seen whether probing layers specialized for semantic tasks in the future could unveil the mechanism behind lateral thinking.

7 Acknowledgement

This work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub)

References

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. 2021. Knowledge neurons in pretrained transformers. *CoRR*, abs/2104.08696.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

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.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

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*.

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

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.