yangqi at SemEval-2024 Task 9: Simulate Human Thinking by Large Language Model for Lateral Thinking Challenges

Qi Yang, Jingjie Zeng, Liang Yang*, Hongfei Lin
School of Computer Science and Technology, Dalian University of Technology, China
{2665643739, jjtail}@mail.dlut.edu.cn, {liang, hflin}@dlut.edu.cn

Abstract

This paper describes our system used in the SemEval-2024 Task 9 on two sub-tasks, BRAINTEASER: A Novel Task Defying Common Sense. In this work, we developed a system SHTL, which means simulate human thinking capabilities by Large Language Model (LLM). Our approach bifurcates into two main components: Common Sense Reasoning and Rationalize Defying Common Sense. To mitigate the hallucinations of LLM, we implemented a strategy that combines Retrieval-augmented Generation (RAG) with the the Self-Adaptive In-Context Learning (SAICL), thereby sufficiently leveraging the powerful language ability of LLM. The effectiveness of our method has been validated by its performance on the test set, with an average performance on two subtasks that is 30.1 higher than ChatGPT setting zero-shot and only 0.8 lower than that of humans.

1 Introduction

Human reasoning processes comprise two types of thinking: vertical and lateral. Vertical thinking, also known as linear, convergent, or logical thinking, is a sequential analytical process that is based on rationality, logic, and rules. Meanwhile, lateral thinking is a divergent and creative process that involves looking at a problem from a new perspective and defying preconceptions.The success of language models has inspired the natural language processing community to attend to tasks that require implicit and complex reasoning, relying on human-like Common Sense mechanisms. While such vertical thinking tasks have been relatively popular, lateral thinking puzzles have received little attention. Recently, the SemEval-2024 Task 9 BRAINTEASER: A Novel Task Defying Common Sense (Jiang et al., 2023) was proposed to bridge this gap, it was a task on a pure English dataset, testing models' ability to demonstrate lateral thinking and challenge default common sense associations. This shared task explores methods to improve models' lateral thinking capabilities.

In this paper, we introduce our entries into two BRAINTEASE subtasks. Inspired by recent research on using LLM to design Agents (Xi et al., 2023), our approach leverages an LLM to architect a system that adeptly simulates the intricacies of human divergent thinking processes. Specifically, our model capitalizes on the advanced linguistic capabilities inherent within the LLM, thereby obviating the need for supplementary training protocols. This strategy enables our system to demonstrate commendable performance across both targeted subtasks.

Furthermore, we also focus on the issue of hallucinations in LLM. LLM can sometimes generate erroneous or highly inaccurate responses. The tendency for LLM to produce these hallucinations becomes particularly noticeable when confronted with such phenomena in our investigations. To solve this problem, we draw inspiration from RAG (Lewis et al., 2020) strategies. We also discover that the performance of LLM can vary greatly depending on the specific prompt words used. As a result, we develop several sets of prompt words and ultimately select the set that achieved the highest performance on our validation tests.

Our method achieves competitive results on SemEval-2024 task 9, ranking well on all tasks, especially on more fine-grained classification tasks. Our method far exceeds the performance of ChatGPT, and on the officially provided test set, there is only a slight gap with the results of human evaluation.

2 Task Description

The BRAINTEASER QA task consists of two subtasks, sentence puzzles and word puzzles, as shown

*Corresponding author.

Question	Choice
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.
What part of London is in France?	The letter N. The letter O. The letter L. None of the above.

Table 1: Example of sentence puzzle and word puzzle.

Adversarial Strategy	Question	Choice
Orignal	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: Example of semantic reconstruction and context reconstruction.

in Table 1. It requires awareness of common sense "default values" and covering them with unconventional thinking that distinguishes these default values from hard constraints.

Sentence Puzzle: Sentence-type brain teaser where the puzzle defying common sense is centered on sentence snippets.

Word Puzzle: Word-type brain teaser where the answer violates the default meaning of the word and focuses on the letter composition of the target question

It is worth noting that both tasks include an adversarial subset, created by manually modifying the original brain teasers without changing their latent reasoning path. In order to accurately evaluate the reasoning ability of our proposed system and ensure that it truly possesses lateral thinking ability, this task constructs adversarial versions of the original data in two ways:

Semantic Reconstruction: Rephrasing the original question without changing the correct answer and the distractors, as showing in table 2.

Context Reconstruction: Keeping the original reasoning path but changing both the question and the answer to describe a new situational context.

Finally, the task also proposes two evaluation metrics to ensure the accuracy of the system in both the overall test set and each adversarial subset. These two evaluation indicators are described as follows:

Instance-based Accuracy: Consider each issue (original/adversarial) in the test set as a separate instance to test the overall accuracy of the system's output on the test set.

Group-based Accuracy: Each question and its associated adversarial instances form a group, and a system will only receive a score of 1 when it correctly solves all questions in the group.

3 Methodology

We propose a system that simulates human lateral thinking patterns, which consists of two stages. During the first stage, our system engages in a simulation of how humans typically read and interpret Brainteaser question stems. The aim here is to check the question stems meticulously, intending to pinpoint specific elements that appear to contravene established common sense norms. The second stage is to combine the parts that violate common sense with four options for thinking, find the option that can "resolve" the parts that defy common sense, and use it as the final answer. The overall architecture of the system is shown in Figure 1.

3.1 Common Sense Reason

In this stage, we use a LLM as the core to conduct common sense reasoning. We input Brainteaser's problem directly into LLM and use LLM's pow-

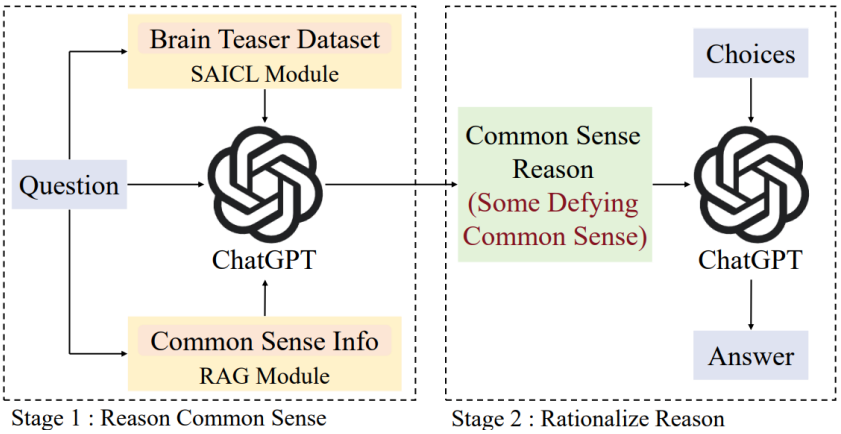

Figure 1: The overall architecture of our proposed system

erful language ability to infer the unreasonable aspects of the problem. At the same time, in order to suppress the hallucination problem of LLM, we design two modules, namely the RAG module and the SAICL module.

3.1.1 Retrieval Augmented Generation

The RAG module integrates deep learning technologies such as Retrieval and Generation. This module is designed to enhance the LLM's understanding of input questions by retrieving relevant information from a vast array of unstructured documents as well as structured knowledge graphs, specifically for Brainteaser questions, in order to produce more accurate, richer, and more relevant Defying Common Sense Reasoning. The workflow of the RAG module includes the following two steps:

Retrieval Phase: Retrieve relevant information from a large number of unstructured documents and structured knowledge graphs according to the given Brainteaser problem.

Integration Phase: The retrieved information snippets are then integrated and merged to be effectively utilized by the generation model. This includes reordering, filtering, or encoding the retrieval results to better suit the subsequent generation tasks.

3.1.2 Self-Adaptive In-Context Learning

We are inspired by Wu et al. (Wu et al., 2023) and develop a SAICL module.The SAICL module adaptively selects better In Context example data from the training set for each Brainteaser problem to improve the effectiveness of In Context Learning. The workflow of the SAICL module also consists of two phases:

Selection Phase: Using the top-K method, search for the K question and its options and answers that are closest to the Brainteaser question in the semantic space.

Sorting Phase: Using the Minimum Description Length (MDL) principle to find an organization that minimizes the compressed encoding length of the output given the input and context. This can be represented by equation (1):

$$c^* = \arg\min_{c \in \mathbf{C}} L_\theta(y \mid c, \mathbf{x}) + L(\theta), \quad (1)$$

where each c represents one possible organization of examples. $L_\theta(y \mid c, \mathbf{x})$ is the code-length required to compress and transmit testing label y given the organization c and testing input $\mathbf{x}$. $L(\theta)$ is the code-length required to describe the model, and it can be calculated in the following equation (2):

$$L_\theta(y \mid c, \mathbf{x}) \approx -\mathbb{E}_{q(y_i|Y)} \log_2 p\left(y_i \mid c, \mathbf{x}\right), \quad (2)$$

where $q\left(y_i \mid Y\right)$ is the prior of y_i among all possible labels Y. Through the above calculation, further select a suitable subset from the K examples selected in the previous phase as the context examples for Brainteaser, combine them with the output of the RAG module, and input them into LLM.

By combining these two modules with the powerful language capabilities of LLM, we can derive reasonable yet contradictory common sense reasoning from the Brainteaser problem. For example, when our question is: "How could a cowboy ride into town on Friday, stay two days, and ride out on Wednesday?" We will gain some common sense reasoning as follows:

- Cowboy rides into town on Friday.
- Cowboy stays in town for two days.
- Cowboy rides out on Wednesday.

- Sunday is two days after Friday.

In this way, we can clearly see the defying common sense part in the Brainteaser question.

3.2 Rationalize Defying Common Sense

At this stage, we combine the conflicting reasoning obtained in the previous stage with the first three options of Brainteaser. This fusion is achieved through the careful design of specific prompt words, which are crafted with the express purpose of evaluating whether any of these three preliminary options possess the capability to logically reconcile the previously identified conflicting reasoning.

For example, in the example given in section 3.1, the option: "His horse is named Wednesday" can effectively solve the Defying Common Sense part inferred from the Common Sense Reason. So it is the correct answer.

But, it is crucial to highlight that in instances where none of the first three options succeeds in producing a satisfactory rationale that effectively addresses the contradictory reasoning, our model is programmed to adopt a fallback strategy. In such scenarios, the model is designed to automatically select the fourth option, aptly labeled "None of above". This decision-making protocol ensures that our model retains the flexibility to understand situations where the presented options fail to provide a coherent resolution to the discrepancies identified, thereby maintaining the integrity of our analytical process. This strategic approach underscores the meticulousness with which our system evaluates the available options, ensuring a comprehensive and reasoned determination of the most appropriate response.

4 Experimental Setup

In this section, we introduce our system settings, and baseline model.

4.1 System Settings

In the RAG module of our system SHTL, we initially remove the stop-words from the original Brainteaser question, then use ConceptNet to retrieve the meanings and relationships of the remaining parts, followed by deduplication and sorting based on relevance to the question. Subsequently, we design appropriate prompt words to concatenate them. In the SAICL module, during the search phase, we utilize the Bert model to obtain feature vectors for each question in the training set. In the vector space, we compare these vectors with the target question's feature vector using cosine similarity, selecting the ten most similar entries. During the ranking phase, we follow the method of the original paper (Wu et al., 2023),randomly select eight entries, extracting them sixteen times, and then calculate the score for each combination obtained from these extractions according to Section 3.1.2. We then select the best combination and use appropriate prompts to link them. At the end of the first stage, we use appropriate prompts to combine the results from both the RAG and SAICL modules with the original Brainteaser question and input them into ChatGPT, obtaining Defying Common Sense. This is then combined with the options of the Brainteaser question using appropriate prompts and input into ChatGPT to derive the best answer.

4.2 Baseline

Our baseline models are categorized into three types: one consists of Large Language Models with a minimal number of prompts, another incorporates models endowed with common sense knowledge, and finally, human evaluation.

Prompted Models:

We evaluate the instruction-finetuned LLMs in few-shot setting:

• **ChatGPT** It is one of the publicly available state-of-the-art Large Language Models in the GPT series (Brown et al., 2020).

• **T0** (Sanh et al., 2022) It is an LLM trained through multi-task instruction tuning, possessing strong zero-shot generalization capabilities.

• **FlanT5** (Chung et al., 2022) It is an enhanced version of T5 (Raffel et al., 2020).

To ensure a fair comparison with human performance, when prompting ChatGPT in a zero-shot setting, we add a description indicating that the question is a brain teaser requiring creative thinking for its resolution. For the other models, we employ the same instruction templates found in their training datasets.

Common Sense Models:

To understand the impact of common sense knowledge on our task, we evaluate the following models enhanced with common sense:

• **RoBERTa-L (CSKG)** (Ma et al., 2021) It is a model fine-tuned on synthetic QA pairs generated from various Common Sense Knowledge Graphs (CSKG) (Ilievski et al., 2021).

• **CAR**(Wang et al., 2023) It is a model finetuned in a similar pipeline as (Ma et al., 2021) but with

Category	Model	Instance-based			Group-based		overall
		Original	Semantic	Context	Ori & Sem	Ori & Sem & Con	
Random		25.8	24.2	22.5	5.0	2.5	25.0
Sentence Puzzle							
Prompted Models	FlanT5(780M)	18.7	16.3	22.0	10.5	4.3	19.0
	FlanT5(3B)	26.8	25.4	35.4	20.1	12.9	29.2
	FlanT5(11B)	33.5	31.6	36.8	22.0	11.0	34.0
	T0(11B)	22.0	22.0	29.7	16.3	11.0	24.6
	T0P(11B)	23.9	22.5	34.9	17.7	12.0	27.1
	T0PP(11B)	26.3	27.3	37.8	19.1	12.0	30.5
	ChatGPT	60.8	59.3	67.9	50.7	39.7	62.7
Common Sense Models	RoBERTa-L	43.5	40.2	46.4	33.0	20.1	43.4
	RoBERTa-L(CSKG)	35.4	36.8	45.0	28.7	18.2	39.0
	CAR	10.5	10.5	11.5	5.7	2.4	10.9
Human		87.5	**90.0**	**95.0**	87.5	**87.5**	**90.8**
SHTL		**90.0**	**90.0**	87.5	**90.0**	**87.5**	89.2
Word Puzzle							
Prompted Models	FlanT5(780M)	22.6	17.7	28.7	9.1	3.7	23.0
	FlanT5(3B)	37.8	29.9	42.7	23.2	12.8	36.8
	FlanT5(11B)	42.7	32.9	43.9	28.7	20.1	39.8
	T0(11B)	17.1	14.0	23.2	9.8	6.1	18.1
	T0P(11B)	28.7	26.2	34.2	19.5	12.8	29.7
	T0PP(11B)	33.5	31.1	39.6	20.1	11.0	34.8
	ChatGPT	56.1	52.4	51.8	43.9	29.3	53.5
Common Sense Models	RoBERTa-L	19.5	19.5	23.2	14.6	6.1	20.7
	RoBERTa-L(CSKG)	18.9	16.5	30.5	12.8	6.1	22.0
	CAR	38.4	31.1	20.1	26.2	6.1	29.2
Human		84.4	87.5	**90.6**	84.4	**84.4**	**87.5**
SHTL		**90.6**	**93.8**	78.1	**90.6**	68.8	**87.5**

Table 3: Main zero-shot results over two BRAINTEASER subtasks across all models in all metrics, "Ori" is Original, "Sem" is Semantic and "Con" is Context. The best performance among all models is in bold.

enhanced negative sampling strategy and reportedly superior performance.

For reference, we also include the native RoBERTa model (Liu et al., 2019) to understand the impact of common sense knowledge.

Human Evaluation:

We recruit four volunteers who are completely unfamiliar with our task to help us test the test set, and take the average of their results as the human test result.

5 Results and Analysis

The final results of our experiments are presented in Table 3. As can be seen from Table 3, the outcomes of the majority of Prompted Models as well as Common Sense Models are essentially random, and some are even below random performance. It is noteworthy to mention the ChatGPT model, which achieves a score of 62.7 in Sentence Puzzles and 53.5 in Word Puzzles, making it the best-performing model aside from Humans and our system, SHTL. From the evaluation results of Humans, it is evident that for both Sentence Puzzles and Word Puzzles, the Human Evaluation scores for Ori & Sem and Ori & Sem & Con were identical, indicating that human lateral thinking capabilities are remarkably stable and unaffected by the Adversarial Subset. Finally, our proposed system, SHTL, can surpass Human performance in most categories, with an average score in the two subtasks that is only 0.8 lower than that of Humans. This significantly exceeds the performance achieved using ChatGPT alone, suggesting that the latent linguistic capabilities of LLMs need to be further explorer appropriately.

6 Conclusion

In this paper, we introduce a lateral thinking system named SHTL, designed to simulate human lateral thinking capabilities for solving brain teaser questions. The system is divided into two stages. The first stage focuses on common sense reasoning, primarily comprised of the RAG module and the SAICL module, which are interconnected through appropriate prompt words to generate instances of defying common sense. The second stage involves identifying the correct options to rationalize the defying common sense generated in the previous stage. This system achieves competitive results, significantly outperforming the ChatGPT setting in a zero-shot scenario, and its performance on the test set is close to that of human evaluation.

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. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Filip Ilievski, Pedro A. Szekely, and Bin Zhang. 2021. CSKG: the commonsense knowledge graph. In *The Semantic Web - 18th International Conference, ESWC 2021, Virtual Event, June 6-10, 2021, Proceedings*, volume 12731 of *Lecture Notes in Computer Science*, pages 680–696. Springer.

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, EMNLP 2023, Singapore, December 6-10, 2023*, pages 14317–14332. Association for Computational Linguistics.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. 2021. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13507–13515. AAAI Press.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. 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 Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Weiqi Wang, Tianqing Fang, Wenxuan Ding, Baixuan Xu, Xin Liu, Yangqiu Song, and Antoine Bosselut. 2023. CAR: conceptualization-augmented reasoner for zero-shot commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 13520–13545. Association for Computational Linguistics.

Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2023. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1423–1436. Association for Computational Linguistics.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huan, and Tao Gui. 2023. The rise and potential of large language model based agents: A survey. *CoRR*, abs/2309.07864.