ROSHA at SemEval-2024 Task 9: BRAINTEASER A Novel Task Defying Common Sense

Mohammadmostafa Rostamkhani, Shayan Mousavinia, Sauleh Eetemadi
Iran University of Science and Technology
{mo_rostamkhani97, sh_mousavinia}@comp.iust.ac.ir, sauleh@iust.ac.ir

Abstract

In our exploration of SemEval 2024 Task 9, specifically the challenging BRAINTEASER: A Novel Task Defying Common Sense, we employed various strategies for the BRAINTEASER QA task, which encompasses both sentence and word puzzles. In the initial approach, we applied the XLM-RoBERTa model both to the original training dataset and concurrently to the original dataset alongside the BiRdQA dataset and the original dataset alongside RiddleSense for comprehensive model training. Another strategy involved expanding each word within our BiRdQA dataset into a full sentence. This unique perspective aimed to enhance the semantic impact of individual words in our training regimen for word puzzle (WP) riddles. Utilizing ChatGPT-3.5, we extended each word into an extensive sentence, applying this process to all options within each riddle. Furthermore, we explored the implementation of RECONCILE (Round-table conference) using three prominent large language models—ChatGPT, Gemini, and the Mixtral-8x7B Large Language Model (LLM). As a final approach, we leveraged GPT-4 results. Remarkably, our most successful experiment yielded noteworthy results, achieving a score of 0.900 for sentence puzzles (S_ori) and 0.906 for word puzzles (W_ori).

1 Introduction

Human reasoning involves two primary types of thinking: vertical and lateral. Vertical thinking, synonymous with linear, convergent, or logical thinking, follows a sequential analytical process based on rationality and rules. Conversely, lateral thinking, often referred to as "thinking outside the box," is a divergent and creative process that challenges preconceptions by approaching problems from new perspectives. Despite the success of language models in tasks requiring implicit and complex reasoning, there is a notable lack of attention to lateral thinking puzzles within the NLP community. To address this gap, the BRAINTEASER Question Answering task (Jiang et al., 2023) has been introduced, designed to evaluate a model's ability to exhibit lateral thinking and challenge default commonsense associations. SemEval 2024 Task 9, BRAINTEASER (Jiang et al., 2024b) comprises two subtasks, Sentence Puzzle and Word Puzzle, which require unconventional thinking to overcome commonsense "defaults" without violating hard constraints. An adversarial subset is included in both tasks, created by manually modifying original brain teasers without altering their underlying reasoning paths. In our initial series of experiments, our focus is on fine-tuning XLM-RoBERTa (Conneau et al., 2020) in three variations: once on the original training data, once alongside the BiRdQA dataset (Zhang and Wan, 2022), and once alongside the RiddleSense dataset (Lin et al., 2021). Additionally, we introduced an innovative approach involving the extension of each word in the BiRdQA dataset into a complete sentence. This method aims to enhance the contextual meaning of individual words during the training process for word puzzle (WP) riddles. To achieve this, we utilized ChatGPT-3.5 to expand each word into a comprehensive sentence, applying this transformation to all options within each riddle. Subsequently, our exploration extends to the application of RECONCILE (Round-table conference) (Chen et al., 2023), incorporating three substantial language models: GPT 3.5, Gemini, and the Mixtral-8x7B (Jiang et al., 2024a) Large Language Model (LLM), a pre-trained generative Sparse Mixture of Experts. Noteworthy is the superior performance of the Mixtral-8x7B model compared to Llama 2 70B across various benchmarks. In the third set of experiments, we assess the zero-shot performance of GPT-4 using the Copilot GUI. Our observations highlight a significant superiority of GPT-4 over alternative models and methods. Furthermore, our

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1038–1042

findings underscore the collaborative utilization of Large Language Models (LLMs) in a round-table format, showcasing substantial enhancements in overall performance. Evaluation metrics are based on two accuracy measures: Instance-based accuracy, treating each question (original/adversarial) as a distinct instance, and group-based accuracy, where each question and its associated adversarial instances form a group, and a system is awarded a score of 1 only if it correctly solves all questions within the group. Our submission to the evaluation phase comprised XLM-RoBERTa fine-tuned on the original training dataset and BiRdQA dataset. The resulting method ranked 25 out of 31 in sentence puzzles and 20 out of 23 in word puzzles. For a detailed implementation of our method, refer to our GitHub repository.

2 Background

The model's inputs consisted of the puzzle and its corresponding choices, provided as input to XLM-RoBERTa. For alternative methods, we employed a prompt, feeding both the puzzle and choices to the model. All puzzles were written in English. To enhance the training of XLM-RoBERTa, we augmented the primary training dataset with additional datasets, namely BiRdQA and RiddleSense. In the context of word puzzles, we further enriched each choice by transforming it into a complete sentence using ChatGPT. The output from all models and methods was expressed as a numerical representation, denoting the correct choice in a zero-based format.

3 System overview

3.1 Preprocessing

In the preprocessing stage, we employ the following steps for the XLM-RoBERTa model: Each choice is concatenated with the corresponding question and subsequently tokenized. In the case of the BiRdQA and RiddleSense datasets, each riddle initially contains 5 options. However, the standard format, based on data validation, necessitates 4 options. To handle this, we transform each riddle into two separate riddles. The approach involves first removing the correct answer from the set of 5 options, resulting in 4 shuffled options. We then create two new riddles from this set by selecting 3 options for each. Finally, we add the correct answer back to the list of labels for each of the new riddles. In an alternative approach, we endeavored to transform each word into a sentence for every option within the BiRdQA dataset. This strategy aimed to enhance the robustness of our model, facilitating a more comprehensive understanding of each option. The rationale behind this was rooted in the notion that comprehending a sentence is generally more straightforward than understanding an isolated word. To execute this transformation, we presented each option to ChatGPT-3.5 with the prompt: "What is the definition of "text"? Write in a sentence." This process generated an extensive file resembling a dictionary. Throughout our training procedure, instead of utilizing individual words, we incorporated the respective definitions created by ChatGPT into our model. For methods utilizing Large Language Models (LLMs), no specific preprocessing is applied. Instead, we use the data in the format of our prompt without any additional preprocessing steps.

Hyperparameter	Value
Optimizer	AdamW
Learning rate	1×10^{-5}
Epochs	10
Batch size	4
Scheduler	Cosine Annealing
Loss Function	Categorical Cross Entropy

Table 1: Values of hyperparameters

3.2 Dataset

To construct the dataset for the XLM-RoBERTa model, we store tokenized sentences for each choice, concatenated with the corresponding question, and include the corresponding label indicating the correct answer to the riddle. Additionally, we incorporate the BiRdQA dataset, designed for bilingual question answering on challenging riddles, and the RiddleSense dataset, alongside the original training dataset. The creation of new datasets from these sources is detailed in the preprocessing section. The original train and test datasets for sentence puzzles comprise 507 and 120 instances, respectively. For word puzzles, the train and test datasets consist of 396 and 96 instances, respectively. The original RiddleSense and BiRdQA datasets initially contain 3510 and 4093 instances, and after applying the transformations outlined in the preprocessing section, they expand to 7020 and 8186 instances, respectively.

DataSet	W_ori	W_sem	W_con	W_ori_sem	W_ori_sem_con	W_overall
Original Dataset	0.438	0.469	0.438	0.344	0.188	0.448
Original + BiRdQA	**0.625**	0.469	**0.469**	0.468	0.281	**0.521**
Original + RiddleSense	0.531	**0.562**	0.438	**0.5**	**0.375**	0.51
Original + BiRdQA (Word Extender)	0.468	0.468	0.25	0.406	0.125	0.375

Table 2: Results of fine-tuned models

Round	Model	S_ori	S_sem	S_con	S_ori_sem	S_ori_sem_con	S_overall
Round 1	**ChatGPT**	0.575	0.700	0.475	0.525	0.300	0.583
	Gemini	0.750	0.750	0.675	0.675	0.575	0.725
	Mixtral-8x7B	0.725	0.625	0.600	0.600	0.450	0.650
Round 2	**ChatGPT**	0.625	0.725	0.700	0.525	0.450	0.683
	Gemini	0.750	0.775	**0.725**	**0.700**	**0.600**	0.750
	Mixtral-8x7B	0.700	0.725	0.600	0.625	0.450	0.675
Round 3	**ChatGPT**	0.700	0.725	0.650	0.625	0.550	0.692
	Gemini	**0.775**	**0.800**	0.700	**0.700**	0.550	**0.758**
	Mixtral-8x7B	0.725	0.650	0.525	0.625	0.375	0.633
Round 4	**ChatGPT**	0.650	0.750	0.675	0.600	0.525	0.692
	Gemini	0.725	**0.800**	0.650	0.650	0.525	0.725
	Mixtral-8x7B	0.675	0.725	0.575	0.625	0.450	0.658

Table 3: Results of Round-Table on sentence puzzle

3.3 Model

We opted for XLM-RoBERTa as our model for this problem due to its pre-training on 100 different languages, indicating a robust understanding of language. Our fine-tuning process involved updating all the model weights using gradient descent on datasets we created. The architecture includes a multiple-choice head with 4 choices over the XLM-RoBERTa model, and we apply Categorical Cross-Entropy loss. For implementing the RECONCILE method, we leverage GPT-3.5, Gemini, and the Mixtral-8x7B Large Language Model (LLM), with certain adaptations to the original method designed for binary classification. We modified it to suit multiple-choice questions and incorporated 4 rounds for our specific application. In each round, the model is prompted to think step by step (Zhou et al., 2023), generating the correct answer and providing a confidence level (0 to 100) along with a reasoning for the selected choice. The original authors suggested that 4 rounds are sufficient for convergence. The output of all models from the previous round serves as input for the next round, where the model evaluates its logical consistency. No fine-tuning is applied to this method. When utilizing GPT-4 with the Copilot interface, we prompt the model to think step by step and generate the correct option. The model provides the corresponding confidence level and a rationale for choosing that particular choice.

4 Experimental setup

We allocated 20% of the original dataset for our validation set, resulting in 80 samples for validation in the word puzzle (WP) domain and 102 samples for validation in the sentence puzzle (SP) domain. Notably, when incorporating additional datasets into our training data, we maintained consistency by retaining the original validation dataset throughout the training process. This decision was driven by the recognition that the supplementary data introduced distinct variations compared to the original training and testing data. Preserving the originality of the validation data aimed to uphold the quality and uniqueness of the final model.
For fine-tuning using XLM-RoBERTa, we utilized the Hugging Face platform and implemented a cosine annealing scheduler.

5 Results

Leveraging the BiRdQA and RiddleSense datasets led to enhancements across all the metrics utilized for evaluating our model, surpassing the performance observed with the original dataset.

The findings presented in table 3 and table

Round	Model	W_ori	W_sem	W_con	W_ori_sem	W_ori_sem_con	W_overall
Round 1	**ChatGPT**	0.375	0.313	0.438	0.219	0.125	0.375
	Gemini	**0.719**	0.594	**0.813**	0.500	**0.438**	**0.708**
	Mixtral-8x7B	0.688	**0.625**	0.469	0.500	0.281	0.594
Round 2	**ChatGPT**	0.5	0.469	0.469	0.406	0.219	0.479
	Gemini	0.656	0.594	0.594	**0.531**	0.375	0.615
	Mixtral-8x7B	0.594	0.563	0.469	0.406	0.188	0.542
Round 3	**ChatGPT**	0.500	0.344	0.469	0.313	0.156	0.438
	Gemini	0.625	0.563	0.625	0.438	0.313	0.604
	Mixtral-8x7B	0.594	0.500	0.531	0.406	0.219	0.542
Round 4	**ChatGPT**	0.500	0.406	0.438	0.375	0.156	0.448
	Gemini	0.594	0.531	0.594	0.438	0.281	0.573
	Mixtral-8x7B	0.500	0.406	0.469	0.344	0.156	0.458

Table 4: Results of Round-Table on word puzzle

Model	S_ori	S_sem	S_con	S_ori_sem	S_ori_sem_con	S_overall	W_ori	W_sem	W_con	W_ori_sem	W_ori_sem_con	W_overall
XLM-RoBERTa (fine-tuned on original dataset)	0.525	0.550	0.625	0.500	0.400	0.567	0.438	0.469	0.438	0.344	0.188	0.448
GPT-4 (Copilot)	**0.900**	**0.875**	**0.825**	**0.875**	**0.775**	**0.867**	**0.906**	**0.875**	**0.875**	**0.844**	**0.719**	**0.885**

Table 5: Comparison between copilot and XLM-RoBERTa results

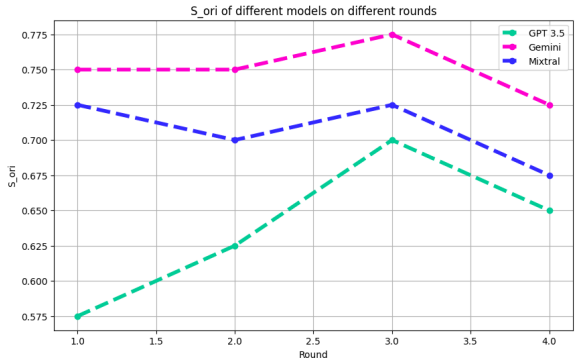

Figure 1: Visualization of Round-Table results for sentence puzzle

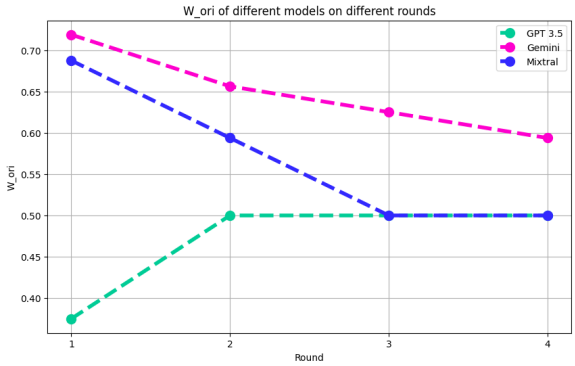

Figure 2: Visualization of Round-Table results for word puzzle

4 indicate that the incorporation of round-table discussions can enhance model performance in sentence puzzles, but conversely, it leads to a decrease in performance for word puzzles. This discrepancy may stem from the fact that, in solving sentence puzzles, some models can provide correct reasoning and influence others positively, whereas the complexity of reasoning in word puzzles may result in incorrect reasoning leading other models astray. Optimal results suggest that employing 3 rounds is most effective for sentence puzzles, while 1 round is preferable for word puzzles. Notably, Gemini consistently outperforms all other models across all rounds. Furthermore, this approach demonstrates its efficacy in boosting the performance of GPT 3.5 in both sentence and word puzzles.

GPT-4 consistently outperformed other models by a significant margin, demonstrating superior results across all metrics.

6 Conclusion

This study explores various methodologies for tackling SemEval 2024 Task 9: "BRAINTEASER: A Novel Task Defying Common Sense." To enhance our model's performance in word puzzles, we incorporate additional datasets for fine-tuning. Additionally, we introduce a modified round-table approach implemented over four rounds. We also evaluate the zero-shot performance of GPT-4 on this task,

Question	Options	BiRdQA	Orginal	RiddleSense	BiRdQA Word Extender	Correct
What kind of stock doesn't have shares?	Small-cap stock, Livestock, Growth stock, None of above	0	0	**1**	2	**1**
What kind of birds always make noise?	Humming bird, Hawk, Owl, None of above	**0**	2	2	1	**0**
What type of chase never involves running?	Escape chase, Paperchase, Risky chase, None of above	0	2	**1**	0	**1**
What kind of tree can you hold in your hands?	Oak, Pine, Palm, None of above	0	0	1	**2**	**2**
What species of geese engages in snake-fighting?	Canada goose, Snow goose, Mongoose, None of above	1	1	1	0	**2**

Table 6: Examples of predictions from different models

which demonstrates superior results across all metrics.

References

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023. Reconcile: Round-table conference improves reasoning via consensus among diverse llms.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

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. 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.

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.

Yunxiang Zhang and Xiaojun Wan. 2022. Birdqa: A bilingual dataset for question answering on tricky riddles.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers.