Deja Vu at SemEval 2024 Task 9: A Comparative Study of Advanced Language Models for Commonsense Reasoning

Trina Chakraborty
Shahjalal University of Science & Technology
trina41@student.sust.edu

Md. Marufur Rahman
Shahjalal University of Science & Technology
marufurr701@gmail.com

Omar Faruqe Riyad
Shahjalal University of Science & Technology
riyad.omf@gmail.com

Abstract

This research systematically forms an impression of the capabilities of advanced language models in addressing the BRAINTEASER task introduced at SemEval 2024, which is specifically designed to explore the models' proficiency in lateral commonsense reasoning. The task sets forth an array of Sentence and Word Puzzles, carefully crafted to challenge the models with scenarios requiring unconventional thought processes. Our methodology encompasses a holistic approach, incorporating preprocessing of data, fine-tuning of transformer-based language models, and strategic data augmentation to explore the depth and flexibility of each model's understanding. The preliminary results of our analysis are encouraging, highlighting significant potential for advancements in the models' ability to engage in lateral reasoning. Further insights gained from post-competition evaluations suggest scopes for notable enhancements in model performance, emphasizing the continuous evolution of the models in mastering complex reasoning tasks.

1 Introduction

The reasoning ability of the human brain demonstrates a dualistic problem-solving approach, which integrates both vertical and lateral methodologies (Bala, 2014). Vertical thinking places emphasis on a methodical and logical examination, executed in a sequential fashion, guided by established norms and regulations. On the contrary, lateral thinking (De Bono, 1970) promotes innovation and fosters the ability to perceive challenges from distinct, frequently unusual observation points, thereby encouraging individuals to go beyond conventional limitations.

Over the past few years, significant progress has been made in the booming domain of Natural Language Processing (NLP), as novel technologies aim to mimic the the complicated ways in which human think (Kumar et al., 2023; Koivisto and Grassini, 2023). This undertaking overcomes conventional logical reasoning and dives into the domain of creative cognition, wherein machines are engineered to creatively navigate and interpret the complex nature of human language and thought processes. Out of these efforts, the BRAINTEASER task is particularly noteworthy for its pioneering aspect (Jiang et al., 2023). The challenge, which is part of the SemEval 2024, has been carefully constructed to evaluate a model's capacity for lateral thinking and its aptitude for questioning and redefining conventional commonplace assumptions.

BRAINTEASER (Jiang et al., 2024) takes a significant progress in the direction of bridging the divide that exists between the cognitive flexibility of humans and that of machines (Boyacı et al., 2023), exceeding the bounds of study. By selecting puzzles that require perception at both the sentence and word levels, this task highlights the significant technological advancement towards machines capable of creative thinking and logical conclusions that beat the apparent. The task encourages participants overcome the limitations of natural language processing (NLP) models by evaluating their capacity to interpret and decode language in a manner that emulates the creative reasoning of humans. Motivated by the unique characteristics of the assignment and the potential it provides to advance the domain of NLP technology, our group wholeheartedly accepted the challenge presented by BRAINTEASER. Our approach was experimental, leveraging a variety of transformer models to explore their capacity for creative and lateral thinking (Hashim et al., 2023). These models, known for their effectiveness (Nassiri and Akhloufi, 2023) in understanding and generating human language, were put to the test to see if they could indeed mimic the thought processes traditionally attributed to humans. We ranked at 20th in each of the subtasks and in average the rank was 31. The experience was rich with learning opportunities, offering

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1239–1244

us valuable insights into the capabilities and limitations of current technologies when faced with tasks that require a departure from conventional reasoning.

2 Task and Data Description

The task (Jiang et al., 2024) is making a system which is evaluated on understanding sentences and words in ways that defy usual expectations. It has two main challenges:

- Sentence Puzzle: the system must interpret sentences in unexpected ways.
- Word Puzzle: the system need to find unconventional meanings of words.

The task employs specific tests to make sure the system analyzes information deeply instead of merely remembering answers. These tests involve altering the phrasing or setting of questions without changing the basic problem, known as semantic and context reconstruction. The systems are evaluated based on two primary factors: their ability to address single questions, referred to as instance-based performance, and their consistency in answering groups of related questions, known as group-based performance. The goal of this task is to advance the system's abilities in problem-solving and creative thinking

The dataset (Jiang et al., 2023) provided for the task includes two distinct types of files, one for sentence puzzles and another for word puzzles. Each file is rich with essential elements such as the posed question, the correct answer, three alternative options (distractors), labels, a list of choices, and the sequence in which these choices are presented. During the training phase, the dataset comprises 507 sentence puzzles and 396 word puzzles, demonstrating a comprehensive range of scenarios for model training. For the testing phase, the dataset narrows down to 120 sentence puzzles and 96 word puzzles, aimed at rigorously evaluating the models' understanding and reasoning capabilities in both puzzle types

3 System Description

3.1 Data Pre-processing

Our data pre-processing for the BRAINTEASER task involved meticulous steps to prepare the dataset for effective model training. Starting with the loading and merging of two numpy arrays—'SP-train.npy' for sentence puzzles and 'WP-train.npy' for word puzzles—we created a unified dataset comprising a diverse range of puzzles. Recognizing the importance of an unbiased dataset for model training, we employed a two-step randomization process. Initially, we randomized the order of the combined dataset. Subsequently, after converting the dataset into a pandas DataFrame, we applied an additional shuffle to guarantee thorough randomness. Given the difficulties of the puzzles, converting them into a binary classification format presented unique challenges. Each puzzle was transformed into a series of question-choice pairs labeled as correct or incorrect. This binary labeling was crucial for training our models to detect the subtle differences between potential answers, thereby enhancing their reasoning capabilities and language understanding. This careful preparation, including a strategic split of 90% for training and 10% for validation, ensured that our models were ready for the BRAINTEASER challenge.

3.2 Data Augmentation

To enrich the dataset and enhance model robustness, we implemented several data augmentation techniques. These included synthesizing new puzzle questions by paraphrasing existing ones and introducing variations in the dataset to simulate a wider range of linguistic structures and puzzle formats. Such augmentation not only expanded the diversity of our training set but also provided our models with a broader linguistic context to learn from, thereby improving their generalization capabilities. This strategy was particularly beneficial in taking decisions to choose the best model.

3.3 Encoding for Models

In the next step, we take a streamlined approach to improve question-answering models through a custom class, integrating seamlessly with PyTorch's Dataset framework. This class, initialized with essential components like questions, answers, labels, a tokenizer, and a max token length, ensures comprehensive preparation of question answer pairs for training. We have tried to apply encoding each pair to produce a dictionary containing merged question-answer texts, input IDs, attention masks, and labels, all conforming to a specified maximum token length. This process, emphasizing special tokens, padding, and truncation, readies each pair for model training, significantly simplifying data handling. The class is instrumental in converting

raw data into a format conducive to learning, thus enhancing the models' ability to generate insightful responses.

3.4 Model Training

For this part, we considered different transformer models to experiment the performance. The models are :

- BERT (Do and Phan, 2022) (Bidirectional Encoder Representations from Transformers) revolutionized NLP by training on a massive corpus in a bidirectional manner, enabling it to grasp context from both directions, thus providing a deep understanding of language distinction.
- XLNet (Ghavidel et al., 2020) extends upon BERT by employing a permutation-based training method, which allows it to capture the bidirectional context more effectively, making it particularly adept at handling tasks requiring a nuanced understanding of language order and structure.
- BART (Lewis et al., 2020) (Bidirectional and Auto-Regressive Transformers) combines the best of both auto-encoding and auto-regressive approaches, excelling in text generation and comprehension tasks by reconstructing text that has been corrupted, making it highly suitable for complex comprehension and synthesis tasks.
- RoBERTa (Robustly Optimized BERT Approach) (Liu et al., 2019) iterates on BERT by modifying key hyperparameters, removing the next-sentence pretraining objective and training with much larger mini-batches and learning rates. This results in improved performance across a range of benchmark tasks.
- T5 (Text-to-Text Transfer Transformer) (Raffel et al., 2020) adopts a unified approach, treating every NLP problem as a text-to-text task, simplifying the process of applying a single model to a variety of tasks, thus streamlining the training and inference process for NLP models.

However, the training phase is structured to leverage the computational prowess of PyTorch, utilizing DataLoader for batch processing, and optimizing model performance with AdamW. We track correctness across epochs to gauge improvement, employing a stopping criterion based on minimal gains in validation accuracy to prevent overfitting. The process begins by selecting the appropriate tokenizer and model architecture based on predefined criteria. Each model is trained over several epochs, with performance on the validation set carefully monitored to ensure improvement. We adopted a learning rate between 2e-5 and 3e-5, training across 4 epochs with batches of 16 to balance efficiency and accuracy. Early stopping was implemented to halt training if validation accuracy showed minimal improvement, preventing overfitting. This process ensured each model, from BERT to T5, was precisely tuned to our dataset's details, focusing on meaningful performance gains.

3.5 Model Fine Tuning

Model fine-tuning was an important aspect of our approach, tailored to make use of the full potential of pre-trained language models. By carefully adjusting learning rates, batch sizes, and epochs, we ensured that each model was optimally adapted to the specifics of the task. Our fine-tuning process also involved a careful selection of layers to unfreeze, enabling the models to learn task-specific details without overfitting.

3.6 Prediction On Validation Set

At the very first phase, we divided the training data into 90:10 manner to get a validation set for the prediction. We utilized a specific class to assess the accuracy of transformer models like XLNet, BART, BERT, RoBERTa, T5 on validation dataset. This class predicts the correct answers by tokenizing question-choice pairs and evaluating them through the model to select the most probable answer. The effectiveness of each model is quantified by comparing predicted answers against actual labels, providing a direct measure of performance. This approach allows us to take decision for the next step.

3.7 Prediction On Test Set

By doing the previous step on validation dataset, we have got the performance analysis of each model and it makes us to observe which model has done best in this set. We choose the best performing model to conduct a prediction on the given test set for the competition for both sentence and word puzzle data.

Model	Before Data Merging	After Data Merging
Bert	91%	**87%**
RoBERTa	90%	82%
XLNet	93%	**85%**
BART	88%	79%
T5	76%	70%

Table 1: Performance on validation dataset

4 Result

Our research explored how different transformer models performed when tasked with solving two types of puzzles: sentence and word puzzles. Initially, we observed encouraging results from all models on the sentence puzzles, which were more abundant in our dataset. This success highlighted the models' proficiency in contexts where narrative clues guide the solution process. However, when we combined sentence and word puzzles into a single dataset, we noticed a significant drop in accuracy across all models (as shown in Table 1). This decline suggests that the models, while effective at processing longer, context-rich sentences, struggled with the brevity and ambiguity typical of word puzzles. This challenge was particularly evident in our competitive analysis phase, where the BERT model achieved 77% accuracy, and a subsequent re-evaluation with XLNet showed an improvement to 80% accuracy on the test set.

The better performance on sentence puzzles can be attributed to the models' inherent strengths. Both BERT and XLNet are designed to excel in understanding and processing complex narrative contexts, benefiting from extensive pre-training across diverse text types. This foundation enables them to navigate the intricate language of sentence puzzles more adeptly. On the other hand, word puzzles often rely on subtle wordplay and linguistic nuances less represented in the models' training data, posing a greater challenge..

The disparity in performance between puzzle types underscores a crucial insight: transformer models, despite their advanced capabilities, exhibit varying degrees of adaptability to different linguistic tasks. The initial high accuracy rates with sentence puzzles showcase their potential, while the subsequent drop in performance upon introducing word puzzles highlights areas for improvement, particularly in enhancing the models' versatility and ability to generalize across diverse language tasks. Our findings indicate a clear path forward—further refining these models to better capture and interpret the breadth of human language, extending their applicability beyond structured narrative contexts to include the nuanced, often unpredictable areas of word puzzles.

5 Limitation and Error Analysis

Our study had some challenges and places where error analysis showed that things could be done better. One big problem with the models is that they are skewed because they were trained on datasets that might not fully show how people use words in different situations. This might make it harder for the models to generalise to new types of data, especially when they move from sentence puzzles to word puzzles (see Tables 2 and 3). The models' different results on sentence puzzles versus word puzzles also points out a need for error analysis and better training strategies or model architectures that can handle the complex nature of both types of puzzles equally. The fact that accuracy went down when datasets were combined suggests overfitting, an important problem that needs more research to make models more reliable. These new ideas help us plan future research, like looking into bigger and more varied training datasets and making models that are specifically made to deal with the problems that come up in different language tasks.

6 Conclusion

Our research into using transformer models like BERT, XLNet, and BART for question-answering tasks shows both their strengths and weaknesses when trying to understand words like humans do. The results point to a potential way to improve system's ability to interpret, but they also show that more progress needs to be made. In the future, action should be put into improving these models so that they understand context better, are more clear, and can be used in more areas. To close the gap between what we can do now and how well we can understand complex human language,

Phase	S_ori	S_sem	S_con	S_ori_sem	S_sem_con	S_overall
Competition	0.77	0.70	0.77	0.70	0.62	0.75
Post-Competition	0.80	0.75	0.77	0.75	0.65	0.77

Table 2: Performance metrics across for Sentence Puzzle

Phase	S_ori	S_sem	S_con	S_ori_sem	S_sem_con	S_overall
Competition	0.37	0.46	0.37	0.34	0.12	0.40
Post-Competition	0.56	0.53	0.40	0.50	0.25	0.50

Table 3: Performance metrics across for Word Puzzle

we will need to work together to improve model designs, training methods, and the way we combine different types of data. Not only does this project look like it will make NLP applications smarter, but it also opens up new ways for Computers to process and come up with language-based responses.

References

Saroj Bala. 2014. Lateral thinking vs vertical thinking. *Deliberative Research*, 24(1):25.

Tamer Boyacı, Caner Canyakmaz, and Francis de Véricourt. 2023. Human and machine: The impact of machine input on decision making under cognitive limitations. *Management Science*.

Edward De Bono. 1970. Lateral thinking. *New York*, page 70.

Phuc Do and Truong HV Phan. 2022. Developing a bert based triple classification model using knowledge graph embedding for question answering system. *Applied Intelligence*, 52(1):636–651.

Hadi Abdi Ghavidel, Amal Zouaq, and Michel C Desmarais. 2020. Using bert and xlnet for the automatic short answer grading task. In *CSEDU (1)*, pages 58–67.

Muhammad Jawad Hashim, Romona Govender, Nadirah Ghenimi, Alexander Kieu, and Moien AB Khan. 2023. Lectureplus: a learner-centered teaching method to promote deep learning. *Advances in Physiology Education*, 47(2):175–180.

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.

Mika Koivisto and Simone Grassini. 2023. Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific reports*, 13(1):13601.

Kamlesh Kumar, Prince Kumar, Dipankar Deb, Mihaela-Ligia Unguresan, and Vlad Muresan. 2023. Artificial

intelligence and machine learning based intervention in medical infrastructure: a review and future trends. In *Healthcare*, volume 11, page 207. MDPI.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2020. Question and answer test-train overlap in open-domain question answering datasets. *arXiv preprint arXiv:2008.02637*.

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. *arXiv preprint arXiv:1907.11692*.

Khalid Nassiri and Moulay Akhloufi. 2023. Transformer models used for text-based question answering systems. *Applied Intelligence*, 53(9):10602–10635.

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. *Journal of machine learning research*, 21(140):1–67.