IIMAS at SemEval-2024 Task 9: A Comparative Approach for Brainteaser Solutions

Cecilia Reyes-Peña
IIMAS/ México
UPMH/ México.
ceciliareyes
@turing.iimas.unam.mx

Orlando Ramos-Flores
IIMAS / México
orlando.ramos
@aries.iimas.unam.mx

Diego Martínez-Maqueda
UPMH / México
231220009@upmh.edu.mx

Abstract

In this document, we detail our participation experience in SemEval-2024 Task 9: BRAINTEASER-A Novel Task Defying Common Sense. We tackled this challenge by applying fine-tuning techniques with pre-trained models (BERT and RoBERTa Winogrande), while also augmenting the dataset with the LLMs ChatGPT and Gemini. We achieved an accuracy of 0.93 with our best model, along with an F1 score of 0.87 for the Entailment class, 0.94 for the Contradiction class, and 0.96 for the Neutral class.

1 Introduction

The brainteasers are problems or puzzles, typically designed to be solved for amusement. To solve brainteasers is necessary the lateral and vertical think, so interpret the context itself contained in them. The SemEval 2024 BRAINTEASER: A Novel Task Defying Common Sense task poses a set of brainteasers and their answers, divided into two types: Sentence Puzzles and Word Puzzles, both in the English language and require an understanding of common sense and the ability to overwrite them through unconventional thinking that distinguishes these defaults from fixed constraints. In Sentence Puzzles, a challenge is presented that defies common sense focused on sentence fragments. In Word Puzzles, the answer challenges the predefined meaning of the word and focuses on the letter composition of the target question (Jiang et al., 2024).

Solving brainteasers requires an unconventional or out-of-the-box approach, which stimulates lateral thinking. This style of thinking is crucial for discovering ingenious solutions to complex problems and for considering situations from multiple perspectives. This type of thinking must be integrated into language models, as it enables them to provide diverse perspectives and apply them to more complex aspects of language, such as understanding metaphors, idioms, or ambiguities.

This paper documents the participation of the IIMAS team at SemEval-2024 task 9, where the resolution of brainteasers was approached using a classification framework. Our strategy relied on fine-tuning techniques applied to pre-trained models using a transformer architecture. In addition to describing our approach, we also analyze the challenges encountered during the process and discuss potential areas for improvement in future research. This paper sheds light on the application of cutting-edge techniques in natural language processing to tackle comprehension and reasoning problems, such as brainteasers, and provides valuable insight into the performance and limitations of our approach in this specific context. During the evaluation phase, the results placed us at the 33th out of 50 participants.

2 Background

We examine various methodologies for solving brain teaser challenges. In this overview, we present some of these approaches. Mitra and Baral (2015) focused on solving logic grid puzzles. Initially, they identified keywords as entities and the relationships between them. Subsequently, they constructed a pair of Answer Set Programming rules. These rules served as inputs for a logic reasoner named Logicia, equipped with a predefined set of predicates. Their model demonstrated an impressive 85.05% accuracy in classifying constituents and successfully solved 71 out of 100 test puzzles. The RIDDLESENSE challenge, introduced by Lin et al. (2021), aims to explore the task of answering riddles. This challenge presents participants with a multiple-choice question-answering scenario, where a model must select one answer from a set of five choices (one correct answer and four distractors) in response

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1121–1126

to a given riddle question. The dataset comprises 5.7k meticulously curated examples. In their experiments, researchers employed various approaches including fine-tuning pre-trained language models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ALBERT (Lan et al., 2020), alongside fine-tuning a text-to-text QA Model (Khashabi et al., 2020). Their methodology involved concatenating the question with the answer choices. During evaluation, three native English speakers achieved an average accuracy of 91.3%, with the best-performing model achieving 68.8% accuracy.

Current language models can be evaluated in what is known as vertical or convergent thinking and perform well; however, the existence of lateral or divergent thinking in the human mind leads to considering the option of evaluating these same models in this way of thinking. This idea is taken by Huang et al. (2023) to propose a way to evaluate Large Language Models (LLMs) in Lateral Thinking Puzzles, also known as situations puzzles. This type of puzzle involves a host who knows the complete truth but gives the player a story lacking certain information. The player, through questions that are only answered with Yes or No, must deduce the whole truth. The GTP-4 model from OpenAI had the best performance in this type of puzzle according to the proposed evaluation.

Tong et al. (2023) also identified the need for non-linear thinking in LLMs, so in their work, they proposed Inferential Exclusion Prompting (IEP) inspired by the method of elimination thinking. This proposal consists of, given a problem, the IEP instructs the LLMs to plan different responses and then eliminate those options that are contradictory or irrelevant. The IEP was evaluated for various problems: parajumbles, riddles, puzzles, brain teasers, and critical reasoning queries against Chain-of-Thought (CoT) prompting.

3 System overview

The data used in this task were provided by Jiang et al. (2023), comprising a set of 507 brainteasers for sentence puzzles and 396 brainteasers for word puzzles. Each of these brainteasers includes one correct answer alongside three distractors. This dataset showcases the complexity of the posed problems, suggesting that they can be effectively addressed through a natural language understanding (NLI) approach.

In this context, the BART model (Lewis et al., 2019) serves as an option for resolving Multi-Genre Natural Language Inference (MultiNLI) problems, where a model's ability to determine which of the proposed premises is true relative to a hypothesis is evaluated using a multi-choice approach. We apply zero-shot classification to the BART model, and as result, we got a low performance as we describe in Table 1.

Table 1: Multi-choice approach accuracy.

Data	Accuracy
SP-train	0.2879
WP-train	0.2449

Given the suboptimal performance of zero-shot models in multichoice tasks, the decision was made to fine-tune a model. One initially discarded proposal was to utilize the MultiNLI dataset (Williams et al., 2018)[1] for fine-tuning, as the BART model[2] is trained on this data and yielded unsatisfactory results. Therefore, the decision was made to work with data provided by the competition or data sharing of a similar nature.

To accomplish this, data transformation was necessary to operate under a classification approach, where each question serves as a value for the premise feature, and each answer is treated as a value for the hypothesis feature, these being the indicators: distractor1, distractor2, distractor(unsure), and correct answer. Each pair of data is assigned a class label. For fine-tuning bert-base-uncased, three different classes are managed. For sentence pairs containing distractor1 and distractor2, the corresponding label is Contradiction; for distractor(unsure), it is Neutral, and for the correct answer, it is Entailment (see Fig 1). For the RoBERTa Winogrande model Sakaguchi et al. (2019), it is expected that the resulting sentence from concatenating the premise with the hypothesis will have a boolean value depending on the dependencies of the hypothesis concerning the premise. Therefore, the labels used are False and True. Both models utilize the following hyperparameter values: batch_size=32, epochs=3, learning_rate=2e-5, as well as a split of the dataset with 80% for training and 20% for evaluation purposes.

In order to enhance the performance of the models, we leveraged the unique capabilities of large language models (LLMs). We employed ChatGPT

[1]https://huggingface.co/datasets/multi_nli
[2]https://huggingface.co/facebook/bart-large-mnli

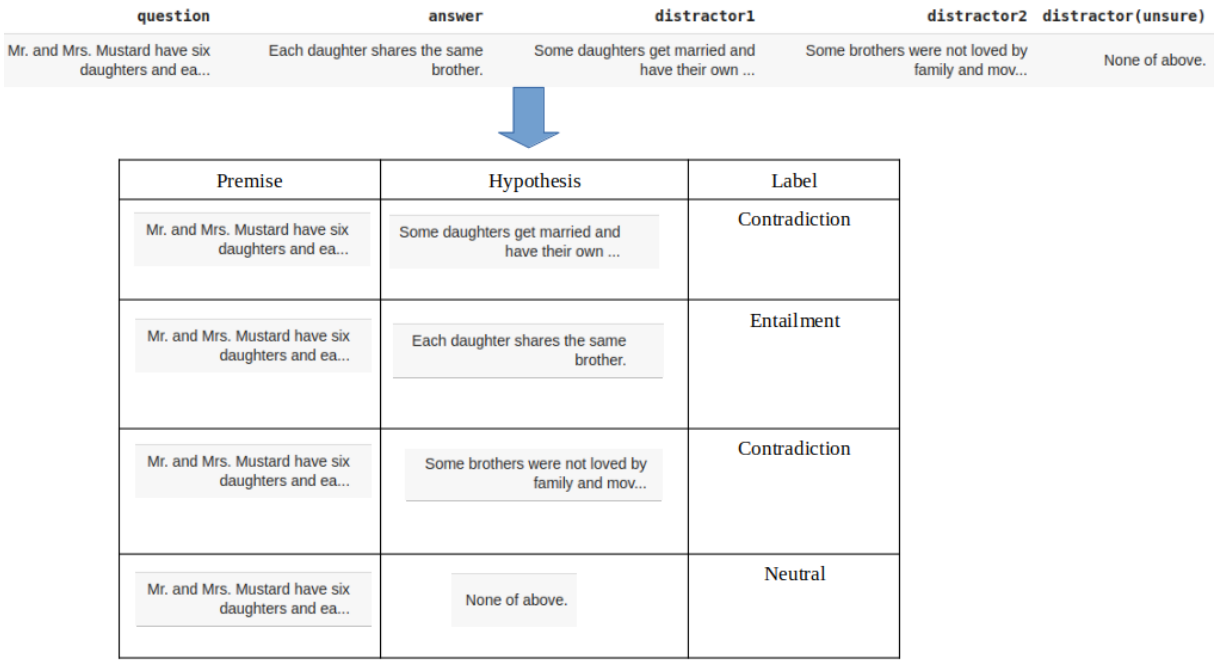

Figure 1: Data Transformation for BERT Classification Approach.

3.5[3] and Gemini[4] to generate additional brainteaser instances. These instances were then incorporated into the fine-tuning process of pre-trained models. Despite having more examples due to data transformation, additional examples were generated through language models such as ChatGPT and Gemini. The generated data underwent manual review to prevent errors regarding the correct answers to the brainteasers. With the expansion and transformation of the data, a total of 4,644 labeled pairs were obtained for fine-tuning the models with brainteasers from both tasks.

4 Experimental Setup

The evaluation results of the BERT Fine-Tuning model are presented in Table 2, revealing the model's struggle to identify the correct answer while being proficient in identifying the neutral class. Based on these findings, a decision was made to minimize the dataset size, considering the potential for model overfitting.

Consequently, the use of brainteasers generated for the Word Puzzle task was discarded, as is shown in Table 3. This decision impacted the model's performance, as evidenced in Table 4, prompting further reduction of the training dataset.

After eliminating all synthetically generated

Table 2: Evaluation Metrics of BERT Fine-Tuning Model with Original Data Train and Generated Data for Sentence Puzzle and Word Puzzle Tasks (Model 1).

Class	Precision	Recall	F1-score
Entailment	0.80	0.78	0.79
Contradiction	0.90	0.91	0.90
Neutral	**0.96**	**0.97**	**0.97**
Macro avg	0.89	0.89	0.89
Weighted avg	0.89	0.89	0.89
Accuracy			0.89

Table 3: Evaluation Metrics of BERT Fine-Tuning Model with Original Data Train and Generated Data for Sentence Puzzle task (Model 2).

Class	Precision	Recall	F1-score
Entailment	0.90	0.79	0.84
Contradiction	0.91	0.96	0.93
Neutral	**0.96**	**0.98**	**0.97**
Macro avg	0.92	0.91	0.91
Weighted avg	0.92	0.92	0.92
Accuracy			0.92

[3]https://chat.openai.com/
[4]https://gemini.google.com/

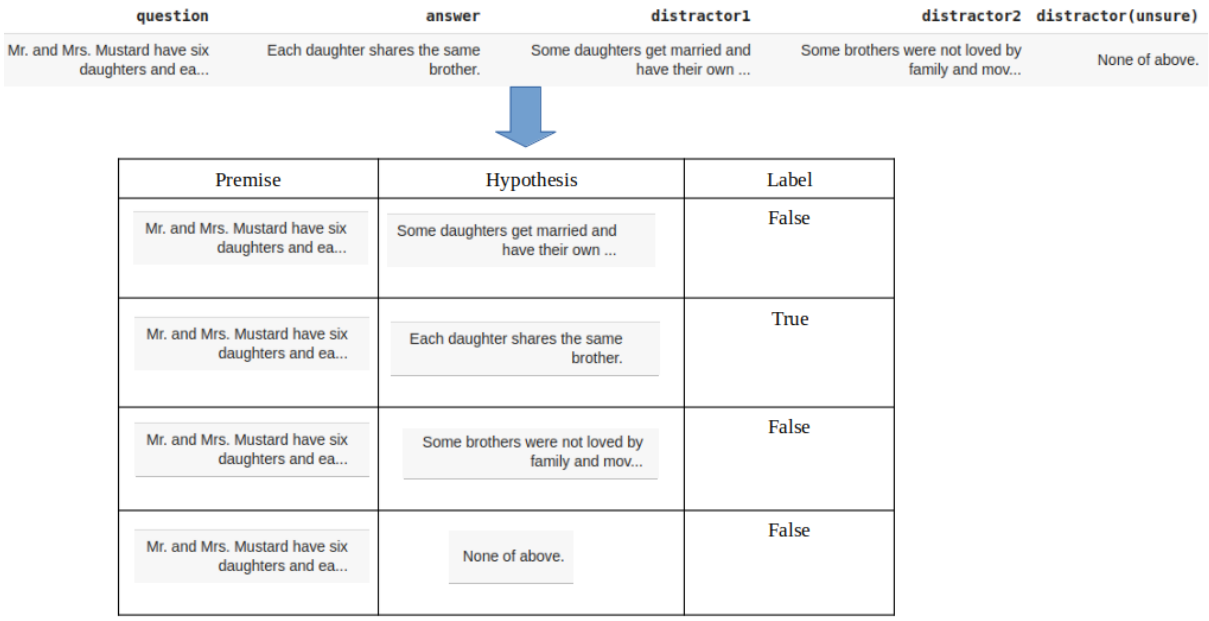

Figure 2: Data Transformation for RoBERTa Winogrande Classification Approach.

data, a noticeable improvement in the evaluation metrics for *Entailment* and *Contradiction* classes was achieved, as we present in Table 4.

Table 4: Evaluation Metrics of BERT Fine-Tuning Model with Original Data Train only for Sentence Puzzle task (Model 3).

Class	Precision	Recall	F1-score
Entailment	0.91	0.84	0.87
Contradiction	0.93	0.96	0.94
Neutral	**0.95**	**0.97**	**0.96**
Macro avg	0.93	0.92	0.92
Weighted avg	0.93	0.93	0.93
Accuracy			0.93

Finally, with the selected data in hand and the pursuit of further improvement, fine-tuning of the RoBERTa Winogrande model was carried out. However, the results were not comparable to those obtained during the fine-tuning of BERT, leading to the decision to discard this model (see Table 5).

Table 5: Evaluation model metrics.

Class	Precision	Recall	F1-score
False	0.73	1	0.84
True	0.0	0.0	0.0
Macro avg	0.36	0.50	0.42
Weighted avg	0.53	0.73	0.61
Accuracy			0.73

Table 6 displays the results obtained during the training stage using the evaluation metrics proposed for the task. For the evaluation phase, Model 3 was selected as it exhibited the best performance.

5 Result

During the evaluation phase of SemEval-2024 Task 9, administrators provided a dataset comprising 120 sentence puzzles and 96 word puzzles. The results, depicted in Table 6, demonstrate that the majority of these results surpass the baseline established by the zero-shot models. Our average final ranking, as displayed in the posted rankings table, is 33, with a score of 0.658 for Sentence Puzzle (position 23) and 0.260 for Word Puzzle (position 22), yielding an overall average score of 0.459.

5.1 Error Analysis

The primary errors of the proposed algorithm are associated with the word puzzle task, as evidenced by the imbalance of classes. Despite efforts to mitigate this imbalance by generating additional data, addressing this task has proven challenging, as the results did not exhibit improvement. One possible contributing factor to this challenge is the necessity for a deeper contextual understanding and a more figurative sense to solve these puzzles.

6 Conclusions

This work introduced a solution for SemEval-2024 Task 9: "BRAINTEASER - A Novel Task Defying Common Sense", leveraging pre-trained lan-

Table 6: SemEval2024 Task 9: BRAINTEASER train data results

	Sentence Puzzle						Word Puzzle					
Train set	**Original**	**Semantic**	**Context**	**Orig. + Sem.**	**Orig. + Sem. + Con.**	**Overall**	**Original**	**Semantic**	**Context**	**Orig. + Sem.**	**Orig. + Sem. + Con.**	**Overall**
Bard zero-shot	.284	.289	.289	.224	.13	.243	.189	.265	.28	.174	.068	.195
Model 1	.81	.828	.721	.81	.692	.77	.174	.181	.136	.09	.037	.123
Model 2	.887	.893	.846	.881	.822	.865	.272	.257	.28	.113	.03	.19
Model 3	.911	.911	.863	.911	.863	.891	.212	.174	.212	.19	.037	.145

Table 7: SemEval2024 Task 9: BRAINTEASER results table

	Sentence Puzzle						Word Puzzle					
Test set	**Original**	**Semantic**	**Context**	**Orig. + Sem.**	**Orig. + Sem. + Con.**	**Overall**	**Original**	**Semantic**	**Context**	**Orig. + Sem.**	**Orig. + Sem. + Con.**	**Overall**
Human	.907	.907	944	.907	.889	.920	.917	.917	.917	.917	.900	.917
ChatGPT	.608	.593	.679	.507	.397	.627	.561	.524	.518	.439	.292	.535
RoBERTa-L	.435	.402	.464	.330	.201	.434	.195	.195	.232	.146	.061	.207
IIMAS Team	.65	.675	.650	.600	.500	.658	.250	.250	.281	.125	.062	.260

guage models and fine-tuning them with the provided data (Jiang et al., 2023), along with additional data generated using LLMs as ChatGPT 3.5 and Gemini. Through experimentation with our pre-trained, fine-tuned models, we found that the BERT model yielded the best results compared to RoBERTa Winogrande. It is worth noting that a significant challenge in this process was defining the appropriate dataset, as certain records from the proposed set had to be discarded to enhance model performance. Ultimately, our results surpassed the task's baseline and secured a position of 33 out of 50 participants, indicating the effectiveness of our approach. However, there is room for improvement, particularly with the word puzzles, which proved to be challenging and require a deeper contextual understanding for resolution.

Acknowledgements

The authors thankfully acknowledge the computer resources, technical expertise and support provided by the Laboratorio Nacional de Supercómputo del Sureste de México, CONACYT member of the network of national laboratories. We also thank to Consejo de Ciencia, Tecnología e Innovación de Hidalgo for the support provided in the program *"Becas para Posgrados de Excelencia, Maestría y Doctorado"*.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shulin Huang, Shirong Ma, Yinghui Li, Mengzuo Huang, Wuhe Zou, Weidong Zhang, and Hai-Tao Zheng. 2023. Lateval: An interactive llms evaluation

benchmark with incomplete information from lateral thinking puzzles.

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.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

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. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1504–1515.

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*.

Arindam Mitra and Chitta Baral. 2015. Learning to automatically solve logic grid puzzles. In *Conference Proceedings - EMNLP 2015*, Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, pages 1023–1033. Association for Computational Linguistics (ACL). Publisher Copyright: © 2015 Association for Computational Linguistics.; Conference on Empirical Methods in Natural Language Processing, EMNLP 2015 ; Conference date: 17-09-2015 Through 21-09-2015.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*.

Yongqi Tong, Yifan Wang, Dawei Li, Sizhe Wang, Zi Lin, Simeng Han, and Jingbo Shang. 2023. Eliminating reasoning via inferring with planning: A new framework to guide llms' non-linear thinking.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.