ALF at SemEval-2024 Task 9: Exploring Lateral Thinking Capabilities of LMs through Multi-task Fine-tuning

Seyed Ali Farokh
Department of Computer Engineering
Amirkabir University of Technology
alifarokh@aut.ac.ir

Hossein Zeinali
Department of Computer Engineering
Amirkabir University of Technology
hzeinali@aut.ac.ir

Abstract

Recent advancements in natural language processing (NLP) have prompted the development of sophisticated reasoning benchmarks. This paper presents our system for the SemEval 2024 Task 9 competition and also investigates the efficacy of fine-tuning language models (LMs) on BrainTeaser—a benchmark designed to evaluate NLP models' lateral thinking and creative reasoning abilities. Our experiments focus on two prominent families of pre-trained models, BERT and T5. Additionally, we explore the potential benefits of multi-task fine-tuning on commonsense reasoning datasets to enhance performance. Our top-performing model, DeBERTa-v3-large, achieves an impressive overall accuracy of 93.33%, surpassing human performance. The code and models associated with this study are publicly available at https://github.com/alifarrokh/SemEval2024-Task9.

1 Introduction

The SemEval 2024 Task 9, BrainTeaser, is a multiple-choice question-answering task, organized by (Jiang et al., 2024) and based on the BrainTeaser benchmark (Jiang et al., 2023) that aims to test the ability of NLP models to exhibit lateral thinking, a creative type of human reasoning process that often requires looking at problems from a new perspective. Unlike similar benchmarks for computational creativity, such as RiddleSense (Lin et al., 2021), which focus on problems resolvable through commonsense associations, the BrainTeaser benchmark comprises questions that challenge models to defy default commonsense associations and linear inference chains (Jiang et al., 2023).

The task includes two subtasks: Sentence Puzzle and Word Puzzle. While the puzzles in the first subtask focus on the meaning of sentences, the word puzzles concentrate on the letter composition of questions and their choices. The following are examples of questions in each subtask.

- **Example Sentence Puzzle**
  Question: A man shaves everyday, yet keeps his beard long. How is that possible? (A) He is a barber. (B) He wants to maintain his appearance. (C) He wants his girlfriend to buy him a razor. (D) None of above.
  Answer: A

- **Example Word Puzzle**
  Question: What part of London is in France? (A) The letter O. (B) The letter N. (C) The letter L. (D) None of above.
  Answer: B

(Lin et al., 2021) discusses three types of popular methods for commonsense question answering: 1) Fine-tuning pre-trained language models such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019), 2) Fine-tuning text-to-text question answering models such as T5 (Raffel et al., 2020), 3) Incorporating knowledge graphs for graph-based language reasoning similar to KagNet (Lin et al., 2019) and MHGRN (Feng et al., 2020). An advantage of using graph-based reasoners is the interpretability of their results due to the symbolic structures of knowledge graphs. Motivated by the superior performance achieved by fine-tuning language models or text-to-text models in achieving the best results on the RiddleSense benchmark, our study investigates the vertical thinking capabilities of these models. We accomplish this by fine-tuning them on the BrainTeaser dataset.

We solely engage in the first subtask of BrainTeaser (Sentence Puzzles) and, due to resource constraints, confine our experiments to models with fewer than one billion parameters. In the subsequent section (Section 2), we provide a brief discussion of the models we fine-tuned. Subsequently, we offer a more detailed introduction to the task in

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1523–1528
June 20-21, 2024 ©2024 Association for Computational Linguistics

Section 3. Section 4 delves into the specifics of our experiments and their outcomes, while Section 5 presents our results in the competition alongside a concise error analysis.

2 System Overview

Inspired by the recent progress in pre-trained language models, our work investigates the performance of fine-tuned language models on the BrainTeaser task. Specifically, we fine-tuned two groups of models, i.e., BERT-based and T5-based models.

2.1 BERT-based Models

The models included in this group are ALBERT v2 (Lan et al., 2019)[1], RoBERTa (Liu et al., 2019), and DeBERTa v3 (He et al., 2023). We refer to this group as BERT-based models because all of them are inspired by BERT, a pre-trained bidirectional transformer encoder (Vaswani et al., 2017), with slight improvements in their pre-training objectives or architectures. The overall process of fine-tuning BERT-based models for multiple choice question answering is illustrated in Figure 1.

Note that for the experiments in which multiple datasets with different numbers of choices are used during fine-tuning, we have to normalize the questions so they consist of the same number of choices, and the model can be fine-tuned with a shared linear projection layer. This is simply achieved by either randomly removing extraneous options from questions with too many choices or by adding dummy options to other ones. Since dummy options are constant in all the questions, the model can easily learn to ignore them and assign a zero probability to them.

As a side note, we also fine-tuned BERT in a sequence classification format where all options are fed into the model so it can infer the correct one by looking at the others. However, the performance was suboptimal in this case, so we did not include the results in the paper.

2.2 T5-based Models

This group includes Flan T5 (Chung et al., 2022) and Unified-QA v2 (Khashabi et al., 2022), pre-trained encoder-decoder transformers that convert all NLP problems into a text-to-text format. These models are fine-tuned to generate the correct choice conditioned on the input question (Figure 2).

[1]ALBERT v2 was introduced in their GitHub repository at https://github.com/google-research/albert

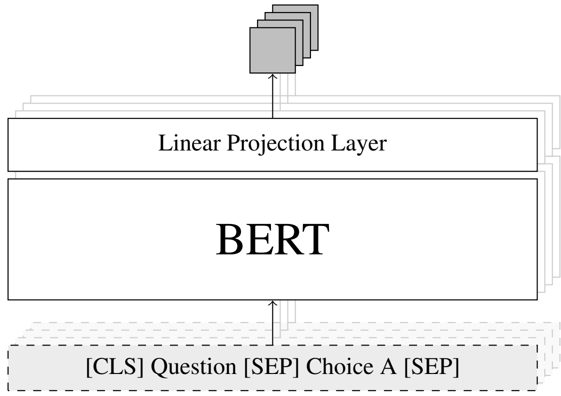

Figure 1: Fine-tuning BERT for multiple-choice question answering involves computing n forward passes simultaneously for questions with n choices. The output embeddings are then projected into a vector of size n, which is fed into a SoftMax function to compute the Cross-Entropy Loss. This optimization process aims to maximize the score of the correct choice.

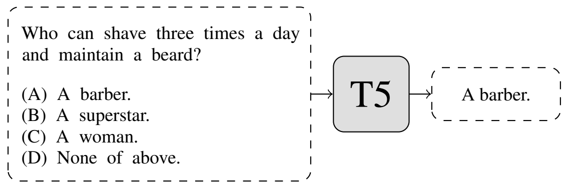

Figure 2: Fine-tuning T5-based models for multiple choice question answering.

3 Task Overview

3.1 Adversarial Examples

The BrainTeaser dataset includes two types of adversarial examples for each original data: *Semantic Reconstruction* and *Context Reconstruction*. In semantic reconstruction, the original question is rephrased so that it conveys the same meaning with the same answer. Extraneous options (i.e., other choices) are kept unchanged in this construction method. In context reconstruction, on the other hand, both the original question and choices are changed so that they describe a new situational context with the same reasoning path as the original question.

3.2 Dataset

The BrainTeaser dataset (Sentence Puzzle) consists of train and test splits, containing 169 and 40 original data along with their adversarial examples, totaling up to 507 and 120, respectively. The test set was released after the evaluation phase was over. Furthermore, a subset of the training data consisting of 102 examples was selected as the validation set during the evaluation phase. However,

Model	BS	LR
ALBERT v2 xlarge	48	1e-5
ALBERT v2 xxlarge	48	1e-5
DeBERTa v3 base	48	25e-6
DeBERTa v3 large	48	11e-6
RoBERTa base	64	1e-5
RoBERTa large	64	1e-5
Flan T5 base	24	5e-4
Flan T5 large	8	4e-4
Unified QA v2 base	24	5e-4
Unified QA v2 large	8	4e-4

Table 1: The hyper-parameters used for fine-tuning our models. LR indicates the Learning Rate and BS shows the Batch Size.

as described in Section 4.2, we chose to employ k-fold cross-validation instead of relying solely on the validation set for model development.

3.3 Evaluation Metrics

The task organizers have defined two types of accuracy metrics to evaluate the performance of models: *Instance-based accuracy*, where each question is considered a separate instance, and *Group-based accuracy*, where each question and its adversarial instances form a group and systems are given an accuracy of one only when they correctly predict all questions in the group.

We refer to the instance-based accuracy on all examples as `overall` accuracy and the instance-based accuracy on `original/semantic/context` examples as `ori/sem/con` accuracy. Correspondingly, `ori-sem` and `ori-sem-con` denote the group-based accuracy of their corresponding questions.

4 Experimental Setup and Results

4.1 Implementation Details

All models were implemented in Python using the Transformers (Wolf et al., 2020) library. AdamW (Loshchilov and Hutter, 2017) was used for optimization, and all models were fine-tuned for 4 epochs. Due to resource constraints, we only tuned the effective batch size and Learning Rate (LR) of models using grid search. See Table 1 for the list of hyper-parameters used for fine-tuning models.

Dataset(s)	# Samples	CV Accuracy
RS	3,510	**81.43**
CSQA	9,741	79.66
PIQA	16,113	79.48
SIQA	33,410	79.95
HellaSWAG	39,905	78.48
SWAG	73456	76.51
BrainTeaser		75.53

Table 2: The 5-fold cross-validation accuracies of models fine-tuned on a union of different commonsense datasets and BrainTeaser (BT), compared with the accuracy of a model fine-tuned on BrainTeaser only.

4.2 Reliability of Experiments

During the development of our models, we noticed that the limited number of training and validation examples led to noisy results when evaluating the original validation set. Consequently, relying solely on this set for model development was deemed unreliable. Therefore, we used 5-fold cross-validation to perform our experiments in the evaluation phase of the competition. Data folds were created by splitting the 169 groups into five sections, ensuring that questions from the same group would not appear in both the training and validation sets. Moreover, we observed that the random initialization of linear projection layers in BERT-based models causes significant variations in the performance of models. Therefore, we repeated the experiments related to BERT-based models three times and averaged the results to increase the reliability.

4.3 Auxiliary Datasets

In contrast to prior vertical thinking datasets, such as PIQA (Bisk et al., 2020) and RiddleSense (Lin et al., 2021), solving BrainTeaser's lateral thinking puzzles requires more creativity and defying preconceptions (Jiang et al., 2023). Our hypothesis is, however, that although combining vertical thinking datasets with BrainTeaser may not directly improve our model's performance, it can provide our model with some knowledge that might be helpful during the reasoning process. For instance, solving the example puzzle in Figure 2 requires the model to have some common sense about what barbers do and what they do not. Another reason why using auxiliary datasets during fine-tuning might be helpful is that fine-tuning large models on small datasets, such as BrainTeaser's training set, can increase the risk of overfitting, which may be prevented by using more training data.

Some datasets that cover various aspects of commonsense reasoning are RiddleSense (RS) (Lin et al., 2021) for computational creativity, CommonSenseQA (CSQA) (Talmor et al., 2018), SWAG (Zellers et al., 2018), and HellaSWAG (Zellers et al., 2019) for general commonsense knowledge, Social IQA (SIQA) (Sap et al., 2019) for social psychology knowledge, and Physical IQA (PIQA) (Bisk et al., 2020) for physical knowledge. To determine which ones can be effective for our task, we fine-tuned a `Flan-T5-base` model on the union of BrainTeaser's training set and each of the mentioned dataset's training data, and compared their accuracies with a similar model fine-tuned on BrainTeaser only (Table 2). As expected, fine-tuning on a combination of BrainTeaser and commonsense datasets enhances the model's performance in all cases. It is also notable that, despite being the smallest dataset, RiddleSense improves the model's accuracy more than any other dataset, possibly because of its distribution overlap with BrainTeaser, as they both have been collected from public websites and deal with computational creativity.

Following (Khashabi et al., 2020), we generate training batches so that each one contains almost the same number of examples from each dataset.

The datasets mentioned in our study serve as valuable resources for enhancing the performance of our multiple-choice QA (MCQA) models. Among these datasets, RS, CSQA, and PIQA are inherently structured as MCQA datasets, making them suitable for direct use in our experiments. However, to incorporate SWAG, HellaSWAG, and SIQA into our study, we need to transform their formats into MCQA. For SWAG, we consider `sent1` as the question and concatenate `sent2` with all potential endings to create the options. Similarly, in HellaSWAG, `ctx-a` is treated as the question, while `ctx-b` is prepended to each possible ending to form the options. Finally, in SIQA, the combination of the `context` and `question` fields in each sample constructs the final question.

4.4 Model Selection

As discussed in Section 2, we fine-tuned two groups of models, BERT-based and T5-based models. Following the results of the previous section (Section 4.3), all models were fine-tuned on a combination of BrainTeaser and RiddleSense. Despite the potential performance improvement from including other datasets, we limited our training set to RiddleSense and BrainTeaser for computational feasibility.

Metric	Accuracy	Ranking
ori	92.5	4
sem	95.0	3
con	82.5	6
ori-sem	92.5	4
ori-sem-con	82.5	5
overall	90.0	7

Table 3: The accuracies and rankings of our submission based on different official metrics. Refer to Section 3.3 for more details about the evaluation metrics.

The reported results in Table 4 indicate that Unified-QA's performance is approximately on par with or outperforms Flan T5. This is expected because Unified-QA-v2 was specifically trained for question answering on many QA datasets, including CSQA, PIQA, and SIQA (Khashabi et al., 2022), which can enhance the performance on BrainTeaser as shown in the previous section. In the case of BERT-based models, not only does DeBERTa-v3 surpass all other BERT-based models, but it also achieves the highest test accuracy among all models and slightly outperforms the human performance, suggesting the effectiveness of its architecture for this task.

5 Results and Error Analysis

5.1 Competition Results

We submitted our DeBERTa-v3-large [2] model (Table 4) during the competition, ranking 7 in the official leaderboard. See Table 3 for more details.

5.2 Error Analysis

There is a 12.5% gap between the accuracies of our best DeBERTa-v3 model on `ori-sem` and `con` (see Table 5), signifying that even though our model learns the semantics of puzzles very well, it sometimes fails to generalize the underlying reasoning paths to other similar situations. This gap is much narrower (5%) for our Unified-QA-v2 model, which outperforms the DeBERTa-v3 on context-

[2] Please note that the DeBERTa-v3-large checkpoint used in our submission was selected before the release of the official test set. For analysis of our best checkpoint, refer to Section 5.2.

Model	# Params	CV Accuracy	Test Accuracy [1]
ALBERT v2 xlarge	59M	79.38	75.83
ALBERT v2 xxlarge	223M	76.06	83.33
RoBERTa base	125M	81.42	80.83
RoBERTa large	355M	83.47	86.67
DeBERTa v3 base	184M	85.90	87.50
DeBERTa v3 large [2]	434M	**89.47**	**93.33**
Flan T5 base	223M	81.43	82.50
Flan T5 large	750M	82.22	84.17
Unified QA v2 base	223M	80.49	84.17
Unified QA v2 large	734M	80.64	90.08
Human (Jiang et al., 2023)	-	-	91.98

Table 4: The overall 5-fold cross-validation and test accuracies of BERT-based and T5-based models
[1] Best accuracies on the official test set released after the evaluation phase
[2] Our submission during the evaluation phase

reconstruction adversarial examples by 2.5% despite underperforming it on original and semantic-reconstruction examples, suggesting that T5-based models may learn to generalize the reasoning paths in the BrainTeaser task better than BERT-based models.

The Unified-QA-v2 model also outperforms DeBERTa-v3 on questions to which *"None of above."* is the answer (see Table 5), which is expected because T5-based models have access to all possible choices while BERT-based models can only see one choice at a time (see Figure 1 and Figure 2).

Five of the six groups that included incorrect predictions from DeBERTa-v3 and Unified-QA-v2 (see Table 5) are identical, and among the errors made in these five groups, six out of seven wrong predictions belong to the same questions, which indicates that the two models almost made the same mistakes. Analyzing those six questions shows us that half of them are related to the models' understanding of math.

Metric	DeBERTa-v3	Unified-QA-v2
ori	97.5	92.5
sem	97.5	92.5
con	85.0	87.5
ori-sem	97.5	92.5
ori-sem-con	85.0	85.0
overall	93.3	90.8
choice d[1]	87.0	93.0
false answers	8	11
false groups	6	6

Table 5: A comparison between the performance of our best models - [1]Overall accuracy of questions to which *"None of above."* is the answer.

6 Conclusion

In this study, we investigated the effectiveness of fine-tuning various language models (including BERT-based and T5-based models) on the BrainTeaser benchmark. We demonstrated the efficacy of multi-task fine-tuning on additional commonsense datasets and its impact on performance in BrainTeaser.

Although our best models achieved performance surpassing human levels, it's important to note that our study was limited to language models with fewer than one billion parameters and training sets comprising at most two datasets combined. Future research could explore extending this study in these directions, as well as investigating other aspects of computational creativity and question-answering.

We hope that our work inspires future research in these areas and contributes to the ongoing advancement of natural language understanding and reasoning.

References

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the*

AAAI conference on artificial intelligence, volume 34, pages 7432–7439.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering. *arXiv preprint arXiv:2005.00646*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

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, Yeganeh Kordi, and Hannaneh Hajishirzi. 2022. UnifiedQA-v2: Stronger generalization via broader cross-format training. *arXiv preprint arXiv:2202.12359*.

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. *arXiv preprint arXiv:2005.00700*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. *arXiv preprint arXiv:1909.02151*.

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. *arXiv preprint arXiv:2101.00376*.

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*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. SocialIQa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. CommonSenseQA: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.