NIMZ at SemEval-2024 Task 9: Evaluating Methods in Solving Brainteasers Defying Commonsense

Zahra Rahimi, Mohammad Moein Shirzady, Zeinab Sadat Taghavi and Hossein Sameti
Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
{zarahimi, mohammad.shirzady99, sameti}@sharif.edu,
zeinabtaghavi1377@gmail.com

Abstract

The goal and dream of the artificial intelligence field have long been the development of intelligent systems or agents that mimic human behavior and thinking. Creativity is an essential trait in humans that is closely related to lateral thinking. The remarkable advancements in Language Models have led to extensive research on question-answering and explicit and implicit reasoning involving vertical thinking. However, there is an increasing need to shift focus towards research and development of models that can think laterally. One must step outside the traditional frame of commonsense concepts in lateral thinking to conclude. Task 9 of SemEval-2024 is Brainteaser (Jiang et al., 2024), which requires lateral thinking to answer riddle-like multiple-choice questions. In our study, we assessed the performance of various models for the Brainteaser task. We achieved an overall accuracy of 75% for the Sentence Puzzle subtask and 66.7% for the Word Puzzle subtask. All the codes, along with the links to our saved models, are available on our GitHub[1].

1 Introduction

With recent advancements in deep learning and especially language models, extensive research has been conducted about reasoning in various natural language processing tasks, including question answering. These reasoning methods adopt vertical thinking. However, lateral thinking is another type often associated with creativity. In the 9th task of SemEval, Brainteaser (Jiang et al., 2024), a task of answering multiple-choice riddle-like questions is defined. To answer these questions, the model needs to employ lateral thinking. This method of thinking differs from vertical thinking in that the reasoning process is not linear. To arrive at a conclusion, one must examine the subject from a perspective beyond the usual conventional thinking paradigms (Waks, 1997). An example of a comparison between the two types of thinking is provided in Figure 2 in the Appendix. Lateral thinking demands a mind that is open, flexible, and creative. Equipping AI models with cognitive abilities such as lateral thinking can enhance problem-solving, adaptability, and coping with new situations and challenges.

In this work, we have evaluated the performance of three categories of models on answering brainteaser questions. We trained and evaluated two language models, BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019), the model presented in Yasunaga et al. (2021) (QA-GNN), and a T5 (Raffel et al., 2019) model for sentence puzzle and word puzzle subtasks. In the QA-GNN method, the ConceptNet knowledge graph (Speer et al., 2017) is used as the source of commonsense knowledge. Through the brainteaser task, we gained insights into two types of thinking - vertical and lateral. We also learned the significance of implementing lateral thinking in AI systems to bridge the gap between human and AI performance. Furthermore, this task piqued our interest in the captivating subject of creativity in artificial intelligence models. We achieved an overall accuracy of 75% and ranked 20th for the Sentence Puzzle subtask. For the Word Puzzle subtask, we ranked 19th and achieved an overall accuracy of 66.7%. All the codes, along with the links to our saved models, are available on our GitHub.

2 Background

The goal and dream of the artificial intelligence field has long been the development of intelligent systems or entities with human-like behavior and thinking. According to existing research, there are two types of thinking in humans: vertical and lateral. Most of the existing research focuses on vertical thinking. Vertical thinking involves a logi-

[1] https://github.com/z-rahimi-r/NIMZ-at-SemEval-Task-9-BRAINTEASER

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 148–154

cal and sequential approach, while lateral thinking requires creativity and flexibility to explore problems from unique and unconventional perspectives (Waks, 1997). The Brainteaser dataset (Jiang et al., 2023) contains 1100 riddle-like English questions requiring lateral thinking. The nature of questions often defies commonsense when approached with vertical thinking. The Brainteaser task includes two subtasks: sentence puzzle and word puzzle. The details of the dataset are presented in the Table 1.

Most research focuses on vertical thinking, using commonsense for implicit and explicit reasoning tasks such as commonsense question answering. Commonsense intelligence is intuitively reasoning about everyday situations and events, which requires knowledge of how the world works (Choi, 2022). In the task of commonsense question answering, two popular methods are fine-tuning language models and using graph neural network (GNN) models. In recent years, the use of knowledge graphs, the primary sources of commonsense knowledge, has increased. Commonsense knowledge stored in language model parameters is mainly descriptive and taxonomic knowledge, often explicitly stated in the language content that these models have been trained on (Hwang et al., 2021). The method presented in COMET (Bosselut et al., 2019) can be a means to teach language models other types of knowledge. The success of COMET can be attributed to the combination of neural and symbolic representations of knowledge, as well as the use of language to represent symbolic knowledge (Choi, 2022). The COMET model is fine-tuned on the ATOMIC knowledge graph (Hwang et al., 2021). This knowledge graph serves as a customized textbook for language models to learn commonsense knowledge and how the world works (Choi, 2022).

In the second popular category of methods, a knowledge graph is used as the complementary source of knowledge with the help of graph neural networks as the medium to harvest this knowledge (Feng et al., 2020; Wang et al., 2022; Zhang et al., 2022). One advantage of using graph neural networks is their interpretability. In QA-GNN (Yasunaga et al., 2021), the RoBERTa LM is used with graph neural networks. Each answer option is checked independently in their method to determine if it is the answer. For each answer option, a subgraph is extracted from the ConceptNet. This subgraph consists of the entities in question and the answer option, all the entities within two hops

	Sentence Puzzle	Word Puzzle
Train	507	395
Test	120	96

Table 1: Dataset Statistics

from question and answer entities on the ConceptNet graph, and the relations between them. In the presented method, the question and the answer option are concatenated and encoded using RoBERTa LM (Liu et al., 2019), then placed as a context node in the subgraph. Since some nodes in the subgraph are more related to the question and its answer, the RoBERTa LM is used to calculate a score for each node in the subgraph. This score is used as an additional feature to the node embeddings to increase the influence of more related entities. Training is done through the message-passing method. Finally, the score of each option being the answer is calculated and the answer option with the highest score will be the final answer to the question. The approach described in Zhang et al. (2023) is similar to QA-GNN but with one key difference. While QA-GNN evaluates each answer option independently using a local graph, this method also includes a global graph that allows for simultaneous evaluation and comparison of all answer options, leading to refined probabilities. Refining the probabilities of each answer option in this way can produce a more accurate result. They consider this method similar to how humans eliminate less likely options. The most similar available study to the Brainteaser task is Riddlesense (Lin et al., 2021), where a riddle dataset is presented. To solve the riddles, one needs advanced natural language understanding, commonsense, and counterfactual reasoning skills, which are complex cognitive processes. They have trained and evaluated several language models, GNN-based models, and text-to-text models on the Riddlesense dataset.

2.1 MCQA in LLMs

Inference from LLMs for multiple choice question answering is done using two methods: Multiple Choice Prompting (MCP) and Cloze Promting (CP) (Robinson et al., 2022). MCP involves presenting a question with several answer options to an LLM and asking it to select the most appropriate answer from the given choices. The other method, CP, involves creating a sentence or passage with a blank that the model needs to fill in with an appropriate

word or phrase. Robinson et al. (2022) criticizes using cloze-style prompts for evaluating LLMs, suggesting that this approach may not fully leverage these models' capabilities for MCQA tasks. However, the evaluation of LLMs with the MCP method has the problem that the order of presenting the options can change the final answer of the LLM. They have evaluated different LLMs, and based on the results, the model's size and providing examples (few-shot inference) to the language model can improve its performance and reduce the dependence of the final answer on the order of options.

2.2 How creative are LLMs?

Margaret Boden's criteria for creativity _novelty, value, and surprise_ are utilized to evaluate the creative capabilities of LLMs. Franceschelli and Musolesi (2023) discusses how much SOTA LLMs satisfy these criteria. LLMs can indeed produce valuable content, as evidenced by their impact and the quality of their outputs. The novelty of an idea or product is being dissimilar to existing examples, the reference of which can either be the person who comes up with it (psychological creativity) or the entire human history (historical creativity). Novelty in LLMs can occur accidentally or as a result of out-of-distribution production or careful prompts, and the degree of novelty is inherently limited by the models' design, focusing on probabilistic outputs based on historical data. The definition of surprise is how unexpected an idea is. Three types of surprise are defined: Combinatorial creativity, which is producing an unfamiliar combination of familiar ideas; Exploratory creativity, which is finding new and undiscovered solutions within the current style of thinking; and Transformational creativity, which is related to changing the current style of thinking. The autoregressive nature of LLMs makes the production of surprising content by them unlikely and only limited to combinatorial creativity, making truly surprising or transformational creativity challenging to achieve. True creativity requires self-awareness and self-evaluation capabilities, which current LLMs lack (Franceschelli and Musolesi, 2023).

3 System Overview

In this section, we will present the systems used to tackle the brainteaser task. The three main approaches in question-answering tasks are fine-tuning language models, graph neural networks, and text-to-text transformers. So, we decided to evaluate the performance of these models on the brainteaser task. Although the role of commonsense in this task is as a distractor (Jiang et al., 2023), we decided to evaluate the impact of using commonsense knowledge through ConceptNet knowledge graph and graph neural networks. While the answer may challenge commonsense in the Brainteaser questions, it does not violate it. All the models are trained for sentence puzzles and word puzzles separately. The general sketch for each type of system is presented in Figure 1.

3.1 Language models: BERT and RoBERTa

We trained and evaluated two language models, BERT-Base (Devlin et al., 2019) and RoBERTa-Large (Liu et al., 2019) on the Brainteaser dataset. The training was done on two different in-house splits of the training data, and the model with the best performance on the validation data was saved for final evaluation on the test set. During the training and inference phase for the two language models of BERT and RoBERTa, the probability of each option being the answer is checked separately. To do that, the question and the answer option are concatenated with the token [SEP] placed between them and given to the language model as input. The score of that option being the answer, is calculated using the output representation of the [CLS] token through a linear layer. Finally, the option that has the highest probability will be the answer to the question.

3.2 LM + GNN: QA-GNN

The QA-GNN model (Yasunaga et al., 2021) uses RoBERTa LM and graph neural networks for reasoning. The knowledge source used in this method is the ConceptNet knowledge graph (Speer et al., 2017). In this method, a separate subgraph is extracted for each answer option. The question and answer option are concatenated, and the resulting embedding from RoBERTa is used as a context node in the graph. This node is only connected to the entities belonging to the answer option and the question (it is not connected to other entities extracted from the knowledge graph). To train the QA-GNN model, pre-processing must be done on the dataset first. For each question and answer option pair, their entities and all of their neighbor entities up to two hops in the ConceptNet knowledge graph are extracted, along with the relations between them. Training is done through the message-

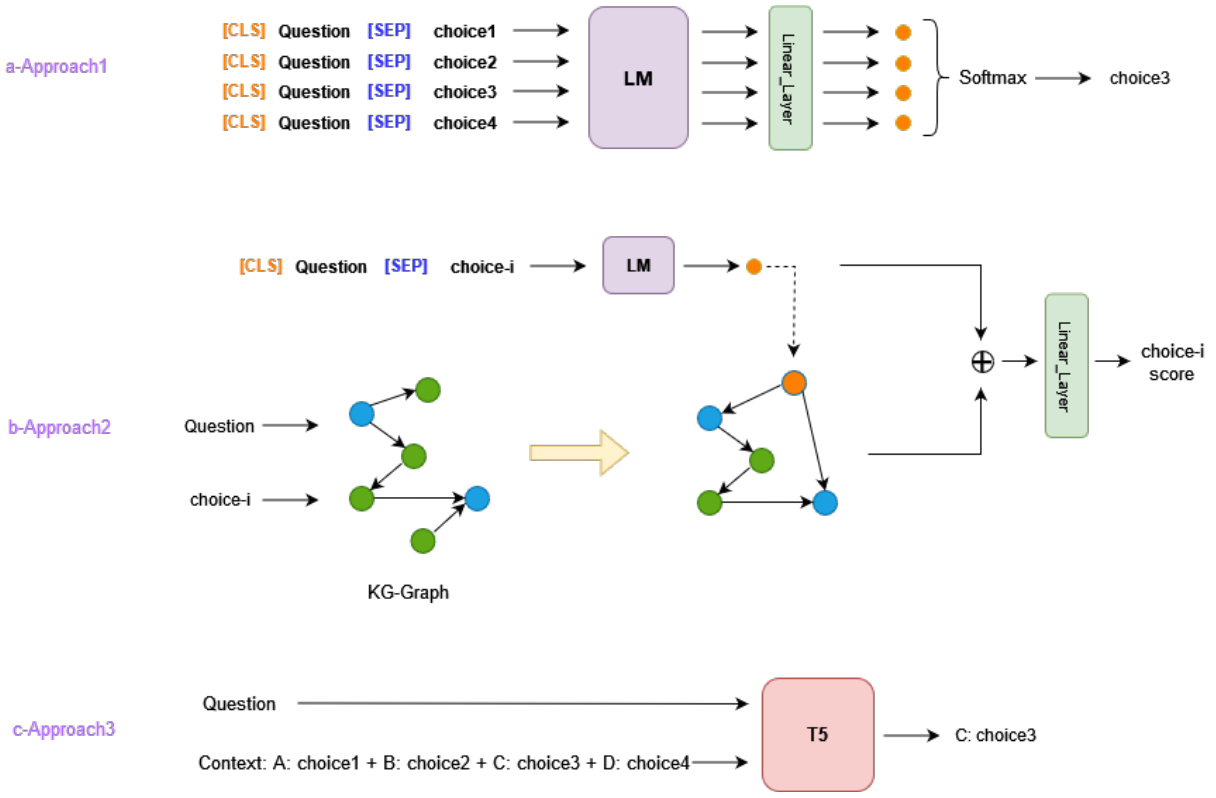

Figure 1: The three categories of methods evaluated for brainteaser task. **a-Approach1:** Fine-tuning LMs like BERT and RoBERTa; **b-Approach:** LM+GNN method, the blue circles are question and choice entities, and the green circles are extracted knowledge-graph entities; **c-Approach:** Fine-tuning a T5 model.

passing method. The score of each answer option, being the final answer, is calculated using the concatenation of the RoBERTa LM representation, context node representation learned through message-passing, and the pooled graph representation, through a linear layer. Finally, the option that has the highest score will be the answer to the question. The interested reader can refer to the original paper for more in-depth details.

3.3 Text-to-Text model

The third method we evaluated was the T5 text-to-text model (Raffel et al., 2019). In this method, the input question and the context, which includes all the options concatenated together, are passed to the T5 model as input. The answer will be in the form of a span extracted from the context, meaning the options. This model considers all options and makes a final decision, setting it apart from previous models.

4 Experimental Setup

We have trained and evaluated base and large sizes of BERT, RoBERTa, and T5 models using the Hugging-Face transformers library, with different hyperparameters to find the best setting. To train the QA-GNN model, we followed the procedure provided by the code available on the GitHub of Yasunaga et al. (2021). After preprocessing the Brainteaser dataset, the QA-GNN models were trained for 100 epochs with early-stopping. In the inference phase of the T5 model, in some cases, the extracted span was incomplete and did not include the letter of the answer option, in these cases the "none of above" option was selected. The code for the in-house train-dev split and the hyperparameters used for training the best-performing models are available in the notebooks on our GitHub[1].

4.1 Evaluation metrics

For each original question in the dataset, two additional adversarial variants are created: semantic reconstruction and contextual reconstruction. Semantic reconstruction rephrases the original question and does not change anything else. In contextual reconstruction, the context of the question does not change, but the surface form of the question and its answer options are changed. An example from

[1] https://github.com/z-rahimi-r/NIMZ-at-SemEval-Task-9-BRAINTEASER

		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
Baselines	Chat-GPT	0.608	0.593	0.679	0.507	0.397	0.627	0.561	0.524	0.518	0.439	0.292	0.535
	RoBERTa-Large	0.435	0.402	0.464	0.33	0.201	0.434	0.195	0.195	0.232	0.146	0.061	0.207
Eval.	BERT-Base	0.7	0.775	0.725	0.7	0.6	0.733	**0.7187**	**0.75**	0.531	**0.7187**	**0.4375**	**0.6666**
	QA-GNN	0.75	0.725	0.775	0.7	0.675	0.75	0.4375	0.4687	0.4375	0.4062	0.2187	0.4479
Post Eval.	RoBERTa-Large	**0.85**	**0.8**	**0.85**	**0.8**	**0.75**	**0.8333**	**0.7187**	0.6875	**0.5625**	0.625	0.375	0.6562
	T5-Large	0.55	0.625	0.525	0.5	0.275	0.5666	0.5937	0.5625	0.5312	0.4375	0.25	0.5625

Table 2: Results on Test set. The baselines are zero-shot results

the dataset is available in Table 3 in the Appendix. The purpose of designing these two variants is to test the robustness of the model. If the model has not memorized the content and is capable of lateral thinking, it will correctly answer these two adversarial variants of each question (Jiang et al., 2023). Models are evaluated using two accuracy metrics: instance-based accuracy metric and group-based accuracy metric. In instance-based accuracy, each question is evaluated separately. In group-based accuracy, a question is evaluated with its adversarial variants, and only if all three are answered correctly, it is scored One. Otherwise, it is scored Zero.

5 Results

We evaluated models from different categories on this task. Due to the riddle-like and unique nature of the questions, it was difficult for the models to generalize to new questions of the test set. We achieved an overall accuracy of 75% and ranked 20th for the Sentence Puzzle subtask. For the Word Puzzle subtask, we ranked 19th and achieved an overall accuracy of 66.7%. The QA-GNN model performed best for the sentence puzzle in the evaluation phase. Still, for the word puzzle, the BERT-base model had the best performance, and QA-GNN performed poorly, which could be due to the absence of reasoning paths on the knowledge graph between the concepts of the answer option and the question. The results of the two phases, evaluation and post-evaluation, are presented in the Table 2. Some wrongly predicted examples for the Word Puzzle subtask are presented in Table 4 in the Appendix.

6 Conclusion

In this study, we evaluated the performance of three main categories of popular methods in the question-answering task on the two subtasks of Sentence Puzzle and Word Puzzle of the SemEval- task 9 Brainteaser. We have achieved an overall accuracy of 75% for the Sentence Puzzle subtask and 66.7% for the Word Puzzle subtask. The nature of the Brainteaser questions is such that they challenge commonsense and require lateral thinking and intellectual creativity to be solved. Models other than LLMs tend to perform poorly in generalizing to new and different examples, especially when it comes to tasks that require creativity, such as puzzles and brainteasers. While LLMs tend to perform better, they still have limited capability when it comes to being creative. Regarding the suggestions for future work, we believe utilizing the chain-of-thought (Wu et al., 2023) method and teaching LLMs to reason step by step with the in-context-learning method can be effective. Another idea is to develop two modules for LLMs or AI agents. The first module will aid in the creative production of knowledge, while the second module will check the rationality of the produced knowledge and its consistency concerning the context of the desired problem. As mentioned earlier, the autoregressive nature of current LLMs and reliance on probabilistic solutions have limited their ability to produce creative content. So, there is a need to design new architectures and different training methods to overcome this limitation. This can be a helpful step towards enhancing creativity and lateral thinking in AI systems.

References

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. *ArXiv*, abs/1911.11641.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Yejin Choi. 2022. The Curious Case of Commonsense Intelligence. *Daedalus*, 151(2):139–155.

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.

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. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1295–1309, Online. Association for Computational Linguistics.

Giorgio Franceschelli and Mirco Musolesi. 2023. On the creativity of large language models.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*.

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.

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, Online. Association for Computational Linguistics.

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.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Joshua Robinson, Christopher Rytting, and David Wingate. 2022. Leveraging large language models for multiple choice question answering. *ArXiv*, abs/2210.12353.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: an open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Shlomo Waks. 1997. Lateral thinking and technology education. *Journal of Science Education and Technology*, 6:245–255.

Ruijie Wang, Luca Rossetto, Michael Cochez, and Abraham Bernstein. 2022. Qagcn: A graph convolutional network-based multi-relation question answering system. Technical Report arxiv.2206, University of Zurich.

Dingjun Wu, Jing Zhang, and Xinmei Huang. 2023. Chain of thought prompting elicits knowledge augmentation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6519–6534, Toronto, Canada. Association for Computational Linguistics.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online. Association for Computational Linguistics.

Qin Zhang, Shangsi Chen, Meng Fang, and Xiaojun Chen. 2023. Joint reasoning with knowledge subgraphs for multiple choice question answering. *Information Processing and Management*, 60(3):103297.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D. Manning, and Jure Leskovec. 2022. Greaselm: Graph reasoning enhanced language models for question answering. *ArXiv*, abs/2201.08860.

Adv. Strategy	Question	Answers
Original	How could a cowboy ride into town on Friday, stay two days, and ride out on Wednesday?	**His horse is named Wednesday.** While in town, he stays in bed for two days. Friday and Saturday are holidays. None of the above.
Semantic Reconstruction	How could a cowboy come into town on Friday, stay two days, and then ride away on Wednesday?	**His horse is named Wednesday.** While in town, he stays in bed for two days. Friday and Saturday are holidays. None of the above.
Context Reconstruction	How can a pilot take off in Los Angeles on Tuesday, fly for 48 hours, and land in Tokyo on Tuesday?	**The pilot's airplane is named Tuesday.** He flies straight for 24h and flies quickly for hours left. There was a one-week long holiday. None of the above.

Table 3: A sentence-based lateral thinking puzzle and its adversarial variations from Brainteaser (Jiang et al., 2023)

Question	Choice List
What do you call a toothless bear?	A brown bear. **A polar bear.** **A gummy bear.** None of above.
What kind of birds always make noise?	**Humming bird.** **Hawk.** Owl. None of above.
What is the best key for a satisfying meal?	**A joykey.** **A turkey.** A hockey. None of above.
What lacks legs and feet but has toes?	Cabbages. **Tomatoes.** **Onions.** None of above.

Table 4: Examples of wrong predictions of Word Puzzle

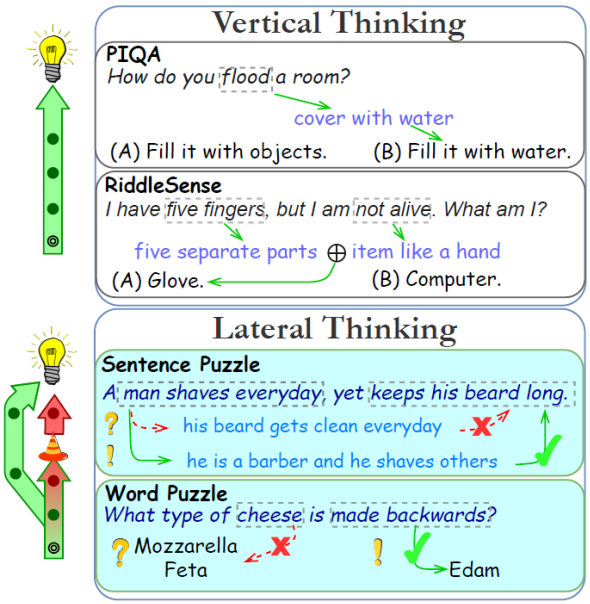

Figure 2: Comparing Vertical Thinking tasks (PIQA (Bisk et al., 2019) and RiddleSense (Lin et al., 2021)) to the BRAINTEASER lateral thinking task. (Jiang et al., 2023)

A Appendix

An example from the dataset is available in Table 3. Also, a few wrongly predicted examples for the Word Puzzle subtask are presented in Table 4. Figure 2 depicts a comparison of Vertical and Lateral thinking.