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

Yifan Jiang[1], **Filip Ilievski**[1,2], **Kaixin Ma**[3]
[1] Information Sciences Institute, Viterbi School of Engineering, University of Southern California
[2]Department of Computer Science, Faculty of Science, Vrije Universiteit Amsterdam
[3]Tencent AI Lab, Bellevue, WA
yifjia@isi.edu, f.ilievski@vu.nl, kaixinma@global.tencent.com

Abstract

While vertical thinking relies on logical and commonsense reasoning, lateral thinking requires systems to defy commonsense associations and overwrite them through unconventional thinking. Lateral thinking has been shown to be challenging for current models but has received little attention. A recent benchmark, BRAINTEASER, aims to evaluate current models' lateral thinking ability in a zero-shot setting. In this paper, we split the original benchmark to also support fine-tuning setting and present SemEval Task 9: **BRAINTEASER(S)**,[1] the first task at this competition designed to test the system's reasoning and lateral thinking ability. As a popular task, BRAINTEASER(S)'s two subtasks receive 483 team submissions from 182 participants during the competition. This paper provides a fine-grained system analysis of the competition results, together with a reflection on what this means for the ability of the systems to reason laterally. We hope that the BRAINTEASER(S) subtasks and findings in this paper can stimulate future work on lateral thinking and robust reasoning by computational models.

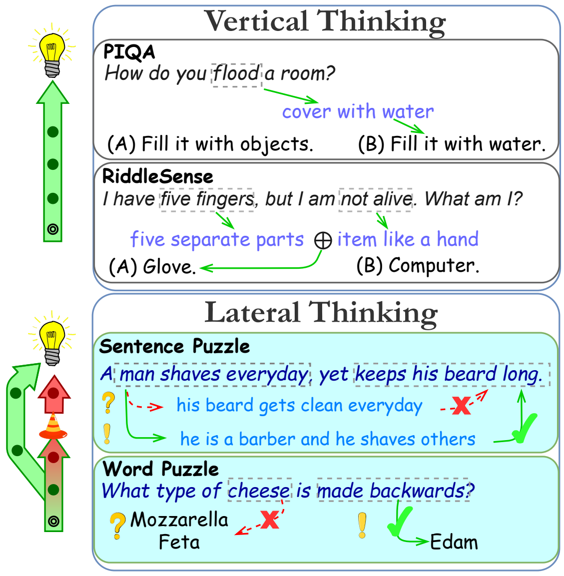

Figure 1: Figure from the first lateral thinking benchmark BRAINTEASER (Jiang et al., 2023c), contrasting existing Vertical Thinking tasks (PIQA (Bisk et al., 2020) and RiddleSense (Lin et al., 2021)) to lateral thinking. Solving BRAINTEASER's lateral puzzles requires default commonsense thinking to be deprecated.

1 Introduction

Vertical thinking requires logical and commonsense reasoning, i.e., making plausible sequential associations of different pieces of commonsense knowledge. As presented in Figure 1 (top), we can easily infer that flooding a room requires filling it with water, based on common sense, and inanimate objects with five fingers are gloves in the riddle. In contrast, lateral thinking is a creative and divergent process that requires thinking out of the box and defying common sense. For example, as shown in Figure 1 (bottom), one needs to overwrite the commonsense associations of *man shaves* to *he shaves himself*, and regard the man as somebody who shaves others all day (e.g., a barber) to answer the lateral puzzle.

While there are many datasets focusing on commonsense reasoning (Talmor et al., 2019; Bisk et al., 2020; Sap et al., 2019b) and numerous studies on improving commonsense reasoning ability of artificial systems (Ma et al., 2021a,b; Zhang et al., 2022), lateral thinking challenges have received little attention and are often filtered out as noise during preprocessing (Vajjala and Meurers, 2012; Speer et al., 2017; Sap et al., 2019a). Consequently, artificial systems' ability to solve lateral thinking problems remains understudied.

To bridge this gap, in (Jiang et al., 2023c), we introduce a novel BRAINTEASER benchmark with two tasks of different granularity: Sentence Puz-

[1]We use BRAINTEASER to represent the original benchmark and BRAINTEASER(S) to represent the data in SemEval task for clarity.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1994–2008
June 20-21, 2024 ©2024 Association for Computational Linguistics

zles and Word Puzzles (cf. Figure 1). The task is formulated in a multiple-choice QA setting for a straightforward human and automatic evaluation. The dataset is constructed via a three-stage pipeline to ensure that the questions are valid and challenging.

We organize our SemEval Task with **BRAINTEASER(S)**, which contains the same data as the BRAINTEASER benchmark to *study model's lateral thinking ability*. Differing from the original benchmark that only focuses on the zero-shot setting, BRAINTEASER(S) divides this data into train/trial/test sets and has no limitation on the method adaptation. The goal of this paper is to describe the SemEval task and provide an analysis of the participant results. We provide details of the data construction pipeline in Section 2 and the SemEval Task description in Section 3. We present the overall leaderboard result and fine-grained method analysis in Section 4. Finally, we discuss the summarized result and conclude with high-level insight to stimulate future works on lateral thinking. For further information, we refer the reader to our source code,[2] task website,[3] and competition website.[4]

2 Source Dataset

We use our recently introduced BRAINTEASER dataset (Jiang et al., 2023c) as the basis for our evaluation. In this section, we briefly describe the data construction pipeline and we refer interested readers to (Jiang et al., 2023c) for full details.

The data construction pipeline has three stages. In the first stage, we collect lateral thinking puzzles from public websites such as `riddles.com` and `rd.com` and conduct filtering and deduplication. Then, the remaining questions are manually verified to ensure that they fit in the sentence or word puzzle categories.

Since the collected puzzles are open-ended questions, which poses great challenges for evaluation. These open-ended puzzles are then converted to multiple-choice questions in the second stage. Specifically, we leverage tools such as COMET (Hwang et al., 2021), WordNet and Wikipedia to construct distractors for every question. For sentence puzzles, we collect distractors that overwrite non-central premises of the question, and for word puzzles, we collect distractors that are semantically similar to the correct answer to ensure they are challenging for systems.

Finally, in stage three, we construct additional data to mitigate the risk of memorization by large pretrained language models. In particular, for each question, we rephrase the original question using an open-source rephrasing tool without changing its answers or distractors.[5] This set is referred to as *Semantic Reconstruction*. Additionally, we leverage GPT-4 to reconstruct each question into a new context such that the misleading question premise is kept. In this case, both the question and the correct answer become different, but the reasoning path remains the same. After reconstruction, the distractors are collected in the same way as described earlier. This set is referred to as *Context Reconstruction*. A strong reasoning model is expected to solve all variants of the question consistently, as their reasoning patterns are identical despite being phrased differently. In total, we construct 1,119 data samples, including reconstruction variants. We report the key statistics in Table 1.

Table 1: Key statistics of the BRAINTEASER dataset. Choices combine the correct answer with all the distractors.

	Sentence	Word
# Puzzles	627	492
Average Question Tokens	34.88	10.65
% Long Question (>30 tokens)	48.32%	2.23%
Average Answer Tokens	9.11	3.0
Std of Choice Tokens	2.36	0.52

3 Task Description

3.1 Task Definition and Organization

In BRAINTEASER(S), we utilize both subtasks in the BRAINTEASER benchmark for evaluation: Sentenze Puzzle (*SP*) and Word Puzzle (*WP*). Both subtasks are multiple-choice QA tasks. We run our SemEval task on CodaLab. Our task is divided into two primary phases: (i) The Practice Phase runs from September 2023 to January 2024, and (ii) The Evaluation Phase runs from 10th Jan 2024 to 31st Jan 2024. We open the Post-Evaluation Phase after 31st Jan 2024 to encourage further research.

3.2 Evaluation Metrics and Data Splits

Evaluation Metrics We evaluate all systems using the same accuracy metrics as Jiang et al. (2023c):

[2]https://github.com/1171-jpg/BrainTeaser

[3]https://brainteasersem.github.io/

[4]https://codalab.lisn.upsaclay.fr/competitions/15566

[5]`https://quillbot.com/`

Table 2: Data statistics of each data split and baseline of BRAINTEASER(S).

	SP	WP
BRAINTEASER	627	492
Data Split of BRAINTEASER(S)		
Train	507	396
↪ Trial (*subset of train*)	120	96
Test	120	120
Baseline overall accuracy		
Human	0.920	0.917
ChatGPT (BRAINTEASER)	0.627	0.535
RoBERTa-L (BRAINTEASER)	0.434	0.207

Instance-based Accuracy considers each (original or reconstruction) question separately. We report instance-based accuracy on the original puzzles and their semantic and context reconstructions. *Group-based Accuracy* considers each original puzzle and its variants as a group. The model will score 1 only when it successfully solves all three puzzles in the group, otherwise, its score is 0. *Overall Accuracy* computes accuracy over all instances.

Data Split To enable BRAINTEASER(S) to support both fine-tuning and zero/few-shot setting, we further divided the original BRAINTEASER dataset into 3 data splits: train, trial, and test set, as shown in Table 2. The train set consists of 507 sentence puzzles and 396 word puzzles. We reuse a portion of the train set as a trial set, which contains 120 sentence puzzles and 96 word puzzles. The test set has 120 data for both subtasks. We release questions and answers from the train and trial set during the Practice Phase. We only release the questions of the test set during the Evaluation Phase and release the whole dataset after the Evaluation Phase ends.

Baseline We provide three baselines (Table 2, see Appendix A for details) to show the gap between humans and SOTA models. To get a comprehensive and robust evaluation performance for each subtask, the human evaluation is computed over 102 data randomly sampled from the original **BRAINTEASER** benchmark, ChatGPT and RoBERTa-L (Liu et al., 2019) performance are also computed over the **BRAINTEASER** in zero-shot setting, i.e. the original unpartitioned data of (Jiang et al., 2023c).

4 Participant System and Results

4.1 Participant Overview

We have 182 participants in total. In the Practice Phase, we have no limitation on the number of submissions to support exploration and enable participants to understand the submission format. We receive 243 submissions for *SP* and 155 for *WP*. In the Evaluation Phase, we allow up to three submissions per team and keep the submission with the best overall accuracy. Our final leaderboard has 48 team submissions for *SP* and 37 for *WP*.

4.2 Leaderboard Results

Table 3 (see Appendix A for full table) displays the top ten models for each subtask, ranked by overall accuracy. The best-performing model in *SP* excels in all six metrics, whereas the leading models in *WP* excel in all but context reconstruction. In the **instance-based accuracy metrics**, most top-performing models (75%) in two subtasks show better performance on original and semantic reconstruction compared to context reconstruction. Most models (80% in *SP*; 70% in *WP*) show the same trend across the entire leaderboard. In the **group-based accuracy metric**, half of the top models in both tasks align with their original instance-based accuracy for the grouped original and semantic reconstruction (Ori&Sem). Only one model in *WP* maintains its performance on all reconstructions (Ori&Sem&Con). Across the leaderboard, more than 80 percent of models in both subtasks show a decrease in Ori&Sem accuracy, ranging from 0.025 to 0.175 in *SP* and 0.031 to 0.281 in *WP*. Nearly all models show a significant drop in Ori&Sem&Con accuracy, with declines varying from 0.025 to 0.275 in *SP* and 0.031 to 0.344 in *WP*.

4.3 Fine-grained System Analysis

In this section, we provide system analysis for the models from the 28 system description papers from participants.*

Method Adaptation and Architecture Selection For both subtasks, the chosen adaptation methods among participants are either fine-tuning models (60%) or prompting models (65%) in a zero-shot (Sanh et al., 2021) or few-shot manner (Brown et al., 2020). Half of the participants try multiple adaptations and submit the best one. For the fine-tuning architecture, participants select either small-size models (<1B) including BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2020) or large-size models (>=1B) such as FLAN-T5 (Chung et al., 2022) and Mistral

* The rank discussed later in this section is based on systems with description papers.

Table 3: Top ten leaderboard results for both subtasks, including user submissions without system description papers. Ori = Original, Sem = Semantic, Con = Context. Team name with (*) submit the system description paper. The first, second and third submissions per category are represented by highlight, **bold** and underline, respectively.

Team Name	Overall	Instance-based			Group-based	
		Original	**Semantic**	**Context**	**Ori & Sem**	**Ori & Sem & Con**
Sentense Puzzle						
abdelhak*	0.983	1.000	1.000	0.950	1.000	0.950
HW-TSC*	**0.967**	1.000	**0.975**	**0.925**	**0.975**	0.900
Maxine	0.958	**0.975**	**0.975**	**0.925**	0.950	0.900
YingluLi	0.950	**0.975**	0.950	**0.925**	0.950	0.900
Theo	0.950	0.950	0.950	0.950	0.950	**0.925**
somethingx95	0.942	0.950	0.950	**0.925**	0.950	0.900
gerald	0.942	0.950	0.950	**0.925**	0.950	0.900
AmazUtah_NLP*	0.925	0.925	0.950	0.900	0.925	0.875
BITS Pilani*	0.900	**0.975**	0.925	0.800	0.925	0.775
ALF*	0.900	0.925	0.950	0.825	0.925	0.825
Word Puzzle						
Theo	0.990	1.000	1.000	**0.969**	1.000	0.969
gerald	0.990	1.000	1.000	**0.969**	1.000	0.969
somethingx95	**0.979**	1.000	1.000	0.938	1.000	**0.938**
zero_shot_is_all_you_need*	**0.979**	1.000	1.000	0.938	1.000	**0.938**
MasonTigers*	**0.979**	**0.969**	**0.969**	1.000	**0.969**	0.969
HW-TSC*	0.969	**0.969**	0.938	1.000	0.938	**0.938**
Maxine	0.969	**0.969**	0.938	1.000	0.938	**0.938**
YingluLi	0.969	**0.969**	0.938	1.000	0.938	**0.938**
kubapok	0.948	0.906	1.000	0.938	0.906	0.844
BITS Pilani*	0.917	0.938	0.938	0.875	0.938	0.812

7B (Jiang et al., 2023a). For the prompting architecture, the majority (90%) use closed-source LLMs such as GPT-4 (OpenAI et al., 2023), GPT-3.5, GeminiPro (Team et al., 2023), Claude (Anthropic, 2024), and Copilot.[6] Techniques like Chain-of-Thought (Wei et al., 2022a), Ensemble (Wang et al., 2022), and RECONCILE (Chen et al., 2023) are widely adopted for prompt engineering. Figure 2 provides a visualization of the overall accuracy distribution for each architecture. For fine-tuning architecture, fine-tuning on large models shows better performance with a tight accuracy range compared to small ones. Fine-tuning on small models shows competitive performance (three in the top five*) in *SP* but a significant drop in *WP*. Among the prompting designs, both zero-shot and few-shot show promising results (seven in the top nine systems*) on two subtasks, with the latter one having a wider accuracy range.

External Dataset Half of the participants (54%) implement their systems only on the original target task, but some further introduce external datasets (35%) to enhance their models' performance. Participants generate humor-style synthetic data using LLMs, crawl riddle websites, or use RiddleSense (Lin et al., 2021) to invoke models' lateral thinking abilities. Other commonsense datasets such as BIRD-QA (Chen and Zulkernine, 2021) or knowledge graphs including ConceptNet (Speer et al., 2017) and WordNet (Miller, 1995) are used to provide general concepts of key instances in questions. Using humor-style datasets tends to be useful on both subtasks, especially for fine-tuning models. Meanwhile, synthetic explanations derived from LLMs are used in prompting to evoke chain-of-thought (Wei et al., 2022b) reasoning abilities.

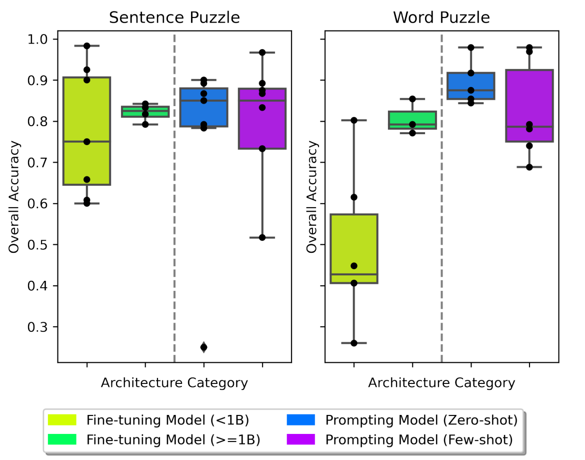

Figure 2: The overall accuracy distribution of each architecture selection.

Data Reconstruction Some participants (18%) reconstruct the original data or change the four-

[6] https://copilot.microsoft.com/

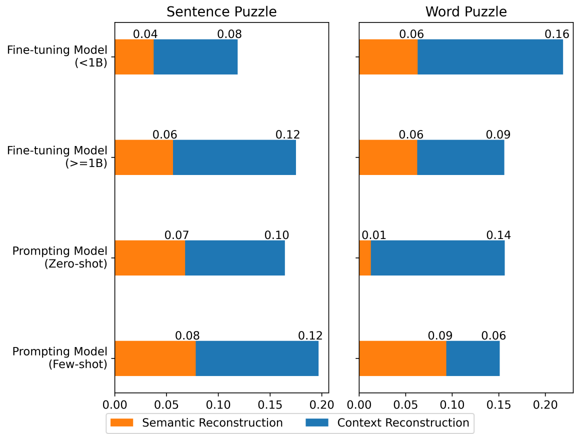

Figure 3: The drop in performance after introducing each reconstruction in group metric.

choice question format. Wang et al. (2024a) use back translation to enlarge the dataset size. Chakraborty et al. (2024) simplify each question into the binary choice problem and Reyes et al. (2024) solve the question under a classification approach with three class labels. Removing the unsure choice is also widely adopted for prompting, where the systems only choose unsure when they fail on the other three choices. Due to a limited number of data reconstruction samples, we cannot conclude which approach can improve performance.

Consistency of Model Predictions In Figure 3, we compare the drop in performance when considering reconstruction variants with group metrics to understand whether the models can solve lateral thinking puzzles by following a consistent reasoning path. On semantic reconstructions, the fine-tuning model has a smaller drop than zero/few-shot prompting in general. Fine-tuning on small models and zero-shot prompting work best on each subtask. On context reconstruction, all architectures show a more significant decline in performance. Fine-tuning on small models and few-shot prompting yield minimal drops in *SP* and *WP*, yet exhibit the largest declines in other subtasks.

5 Discussion

We start the discussion with the question: "*Is lateral thinking solved*?" The best-performing systems reach 100% on both tasks, making it seem that the task is solved. However, there remain many questions to explore. Our discussion targets 5 questions to provide overall insights: 1) What's the difference between the **BRAINTEASER(S)** SemEval Task and the original **BRAINTEASER** benchmark? 2) What's the difference between the best systems for sentence puzzles and word puzzles? 3) Are model predictions consistent with individual and group partitions? 4) What does fine-tuning mean for lateral thinking tasks? 5) What challenges still exist in the realm of lateral thinking?

5.1 Difference with the Original BRAINTEASER (Jiang et al., 2023c)

The **BRAINTEASER** benchmark (Jiang et al., 2023c) is proposed to evaluate LLMs' lateral thinking ability in **zero- and few-shot** settings while in **BRAINTEASER(S)** we release 80 percent of the data for training and we put no limitation on method adaptation. Although releasing data encourages more possibilities for participants, it also narrows down our hidden test set, making the comparison between system performance on BRAINTEASER(S) and the LLMs evaluation results on the BRAINTEASER benchmark unfair. With only 120 samples in the BRAINTEASER(S) test set, the probability of achieving high performance by some of the large number of systems becomes relatively large. Moreover, we expect that most of the lateral patterns will be recurring between the training and the test data, which especially benefits fine-tuning methods. With these caveats in mind, we hope the result and analysis on BRAINTEASER(S) can provide meaningful ideas and insight on lateral thinking and be verified systematically on the whole BRAINTEASER benchmark.

5.2 Effective System Choices and Differences

From subsection 4.3, we know architecture selection yields different distributions of performances on each subtask. On sentence puzzles, fine-tuning small models (Kelious and Okirim, 2024; Mishra and Ghashami, 2024; Farokh and Zeinali, 2024) with additional dataset providing competitive results. On word puzzles, either zero-shot (Moosavi Monazzah and Feghhi, 2024; Venkatesh and Sharma, 2024) or few-shot (Li et al., 2024; Raihan et al., 2024) prompting leads to top-performing results. In general, even small models obtaining language understanding during pretraining can adapt to sentence puzzles via fine-tuning, and additional humor-style datasets can evoke more lateral thinking abilities. On word puzzles, fine-tuned models have difficulties focusing on letter composition which hugely deviates from their pertaining dataset. Even the top-scoring fine-tuning model (Kelious and Okirim, 2024) on *SP* fails to perform well on *WP*. On the other hand, the prompting method leverages the information stored in LLMs' parameters and their access to large pre-training data to mitigate the difficulty of word puzzles. However, the nature of the frozen model not only reduces the effectiveness of the external datasets but also limits further improvement and requires meticulous prompting engineering to ensure stable performance.

5.3 Prediction Consistency

Reconstruction of the original brainteaser puzzles allows us to distinguish between memorizing the training corpus and the ability of models to generalize to unseen samples. As indicated in subsection 4.2, most models struggle with consistent lateral thinking. Context reconstruction poses greater challenges than semantic reconstruction due to the need for lateral reasoning adaptation to novel settings. Context reconstruction of word puzzles is the most challenging, highlighting the risks of overfitting and memorization. Figure 3 shows architectures have different consistency issues. Fine-tuned models have a significant drop in context reconstruction in *WP* because the novelty of puzzles limits models to training corpus. Few-shot prompting can be beneficial for consistency in word puzzles but useless in sentence puzzles. LLMs' ability to follow pattern (Mirchandani et al., 2023) leads them to focus on the surface form in word puzzles, which brings improvement in consistency. Few-shot prompting can hardly provide general patterns of sentence puzzles due to its uniqueness, and the example in the demonstration can mislead the model.

5.4 Impact of Fine-Tuning

Even though recently in-context learning (ICL) (Brown et al., 2020) has achieved great progress on reasoning tasks (Talmor et al., 2019; Bisk et al., 2020), we are happy to see half of the participants implement their system in fine-tuning approaches and showing promising performance. Fine-tuning on small models can lead to a wide accuracy distribution, which requires careful design on hyperparameters and the training process. Exposure to external datasets can stabilize and enhance performance. Fine-tuning on large models shows tight accuracy distribution but lacks top-performing models, which suggests the need for more fine-tuning data to "distort" the default commonsense (Kumar et al., 2022) and evoke lateral thinking out-of-distribution (Jiang et al., 2023b). Also, the large gap between instance- and group-based metric (Figure 3) points out that short-cut learning still exists among these methods.

5.5 Challenges in Lateral Thinking

We summarize the discussion with the challenges that remain unsolved and require further effort to evoke the models' lateral thinking abilities. 1) The system performances and our analysis are based on a small set of original BRAINTEASER benchmark (subsection 5.1). A more general and systematic analysis should be performed with the entire original BRAINTEASER data or even an enlarged version of it, starting from prompting models. 2) There is still a lack of a general approach demonstrating a stable and competitive performance on both subtasks. No existing method can merge the advantages of each architecture on each subtask (subsection 5.2). 3) Each model fails to generate consistent predictions similar to humans, even under simple semantic reconstructions (subsection 5.3). 4) Fine-tuning methods suffer from learning shortcuts while prompting methods have problems finding general lateral thinking patterns akin to humans (see also (Lewis and Mitchell, 2024)) (subsection 5.4).

6 Conclusions and Future Perspectives

This paper summarizes SemEval 2024 Task 9, BRAINTEASER(S), a novel task defying common sense. We present the motivation, data design, data construction, evaluation process, competition systems, participant results, result analysis, and discussion. BRAINTEASER(S) was popular among participants and received 483 submissions from 182 teams during the competition, with various method adaptations and architecture selections demonstrating different advantages on each subtask and evaluation metric. The best-performing systems have impressive performance on both subtasks, which reach 100% accuracy on lateral thinking puzzles from the web. However, our fine-grained analysis highlights the remaining questions and challenges for further research. Importantly, BRAINTESER(S) SemEval result is evaluated over a subset (20%) of original BRAINTEASER benchmark. Even on this subset and despite the access to 80% of the data for training, models still struggle to reason consistently on semantic and context reconstruction. Future work should investigate flexible ways to combine lateral and vertical thinking, construct better evaluation metrics for creative and open-ended generations, build connections within reconstruction based on analogical reasoning (Sourati et al., 2023) and explore a dynamic, multi-stage process where the model (or human) can request clarifications or obtain contextual hints. The BRAINTEASER(S) SemEval Task, together with its source BRAINTEASER task, is the first step toward injecting AI systems with lateral thinking ability. We hope that the competition results and analysis can inspire future research on developing and evaluating lateral thinking models.

Ethical Considerations

As our brain teasers are "folk knowledge" and are published on a range set of websites, it is hard to check their original licenses comprehensively. Yet, the website owners declare permission to print and download material for **non-commercial use** without modification on the material's copyright. Therefore, we provide the corresponding copyright statements and website URLs for each original brain teaser and its adversarial version. In addition, we ask the task participants to sign a document claiming that the only aim of the data usage is research. We note that, despite our best efforts, the task data may still contain bias in terms of gender or politics. We will indicate that future research should use the task data with caution.

7 Acknowledgements

We appreciate Baktash Ansari, Dilip Venkatesh, Soumya Smruti Mishra, Harshit Gupta, and Pouya Sadeghi for their support as emergency reviewers for the competition. This research was sponsored by the Defense Advanced Research Projects Agency via Contract HR00112390061, Defense Advanced Research Projects Agency with award N660011924033 and Strengthening Teamwork for Robust Operations in Novel Groups via number W911NF-19-S-0001.

References

Mohammad Hossein Abbaspour, Erfan Moosavi Monazzah, and Sauleh Eetemadi. 2024. Iust-nlplab at semeval-2024 task 9: Brainteaser by mpnet (sentence puzzle). In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1095–1098, Mexico City, Mexico. Association for Computational Linguistics.

Baktash Ansari, Mohammadmostafa Rostamkhani, and Sauleh Eetemadi. 2024. Bamo at semeval-2024 task 9: Brainteaser: A novel task defying common sense. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 224–232, Mexico City, Mexico. Association for Computational Linguistics.

Anthropic. 2024. Introducing claude 2.1.

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*, 05, pages 7432–7439.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Trina Chakraborty, Marufur Rahman, and Md Omar Faruqe. 2024. Deja vu at semeval 2024 task 9: A comparative study of advanced language models for commonsense reasoning. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1229–1234, Mexico City, Mexico. Association for Computational Linguistics.

Alvin Chen, Ray Groshan, and Sean Von Bayern. 2024. Mothman at semeval-2024 task 9: An iterative system for chain-of-thought prompt optimization. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1888–1900, Mexico City, Mexico. Association for Computational Linguistics.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *arXiv preprint arXiv:2309.13007*.

Yuhao Chen and Farhana Zulkernine. 2021. Bird-qa: a bert-based information retrieval approach to domain specific question answering. In *2021 IEEE International Conference On Big Data (Big Data)*, pages 3503–3510. IEEE.

Kyu Hyun Choi and Seung-Hoon Na. 2024. Geminipro at semeval-2024 task 9: Brainteaser on gemini. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1626–1630, Mexico City, Mexico. Association for Computational Linguistics.

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*.

Seyed Ali Farokh and Hossein Zeinali. 2024. Alf at semeval-2024 task 9: Exploring lateral thinking capabilities of lms through multi-task fine-tuning. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1534–1539, Mexico City, Mexico. Association for Computational Linguistics.

Harshit Gupta, Manav Chaudhary, Shivansh Subramanian, Tathagata Raha, and Vasudeva Varma. 2024. irel at semeval-2024 task 9: Improving conventional prompting methods for brain teasers. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1769–1777, Mexico City, Mexico. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Ethan Heavey, James Hughes, and Milton King. 2024. Stfx-nlp at semeval-2024 task 9: Brainteaser: Three unsupervised riddle-solvers. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 28–33, Mexico City, Mexico. Association for Computational Linguistics.

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*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023a. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Yifan Jiang, Filip Ilievski, and Kaixin Ma. 2023b. Transferring procedural knowledge across commonsense tasks. *arXiv preprint arXiv:2304.13867*.

Yifan Jiang, Filip Ilievski, Kaixin Ma, and Zhivar Sourati. 2023c. 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.

Abdelhak Kelious and Mounir Okirim. 2024. Abdelhak at semeval-2024 task 9 : Decoding brainteasers, the efficacy of dedicated models versus chatgpt. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 200–205, Mexico City, Mexico. Association for Computational Linguistics.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*.

Martha Lewis and Melanie Mitchell. 2024. Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models. *arXiv preprint arXiv:2402.08955*.

Yinglu Li, Zhao Yanqing, Min Zhang, Yadong Deng, Aiju Geng, Xiaoqin Liu, Mengxin Ren, Yuang Li, Su Chang, Xiaofeng Zhao, Xiaosong Qiao, Ming Zhu, Yilun Liu, Mengyao Piao, Feiyu Yao, shimin tao, Hao Yang, and Yanfei Jiang. 2024. Hw-tsc at semeval-2024 task 9: Exploring prompt engineering strategies for brain teaser puzzles through llms. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1657–1662, Mexico City, Mexico. 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. *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*.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. 2021a. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 15, pages 13507–13515.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Satoru Ozaki, Eric Nyberg, and Alessandro Oltramari. 2021b. Exploring strategies for generalizable commonsense reasoning with pre-trained models. *EMNLP 2021*.

Suyash Vardhan Mathur, Akshett Jindal, and Manish Shrivastava. 2024. Davinci at semeval-2024 task 9: Few-shot prompting gpt-3.5 for unconventional reasoning. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1202–1206, Mexico City, Mexico. Association for Computational Linguistics.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Suvir Mirchandani, Fei Xia, Pete Florence, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, Andy Zeng, et al. 2023. Large language models as general pattern machines. In *7th Annual Conference on Robot Learning*.

Soumya Mishra and Mina Ghashami. 2024. Amazutah_nlp at semeval-2024 task 9: A multichoice question answering system for commonsense defying reasoning. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1447–1453, Mexico City, Mexico. Association for Computational Linguistics.

Erfan Moosavi Monazzah and Mahdi Feghhi. 2024. Zero shot is all you need at semeval-2024 task 9: A study of state of the art llms on lateral thinking puzzles. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1901–1905, Mexico City, Mexico. Association for Computational Linguistics.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, and etc Lama Ahmad. 2023. Gpt-4 technical report.

Ioannis Panagiotopoulos, George Filandrianos, Maria Lymperaiou, and Giorgos Stamou. 2024. Ails-ntua at semeval-2024 task 9: Cracking brain teasers: Transformer models for lateral thinking puzzles. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1744–1757, Mexico City, Mexico. Association for Computational Linguistics.

Zahra Rahimi, Mohammad Moein Shirzady, Zeinab Taghavi, and Hossein Sameti. 2024. Nimz at semeval-2024 task 9: Evaluating methods in solving brainteasers defying commonsense. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 148–154, Mexico City, Mexico. Association for Computational Linguistics.

Md Nishat Raihan, Dhiman Goswami, Al Nahian Bin Emran, Sadiya Sayara Chowdhury Puspo, Amrita Ganguly, and Marcos Zampieri. 2024. Masontigers at semeval-2024 task 9: Solving puzzles with an ensemble of chain-of-thought prompts. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1360–1365, Mexico City, Mexico. Association for Computational Linguistics.

Cecilia Reyes, Orlando Ramos-Flores, and Diego Martínez-Maqueda. 2024. Iimas at semeval-2024 task 9: A comparative approach for brainteaser solutions. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1110–1115, Mexico City, Mexico. Association for Computational Linguistics.

Mohammadmostafa Rostamkhani, Shayan Mousavinia, and Sauleh Eetemadi. 2024. Rosha at semeval-2024 task 9: Brainteaser: A novel task defying common sense. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1027–1031, Mexico City, Mexico. Association for Computational Linguistics.

Pouya Sadeghi, Amirhossein Abaskohi, and Yadollah Yaghoobzadeh. 2024. utebc-nlp at semeval-2024 task 9: Can llms be lateral thinkers? In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1778–1789, Mexico City, Mexico. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *AAAI Conference on Artificial Intelligence*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473.

Vineet Saravanan and Steven Wilson. 2024. Ounlp at semeval-2024 task 9: Retrieval-augmented generation for solving brain teasers with llms. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 206–212, Mexico City, Mexico. Association for Computational Linguistics.

Marco Siino. 2024. Deberta at semeval-2024 task 9: Using deberta for defying common sense. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 290–296, Mexico City, Mexico. Association for Computational Linguistics.

Zhivar Sourati, Filip Ilievski, Pia Sommerauer, and Yifan Jiang. 2023. Arn: A comprehensive framework and benchmark for analogical reasoning on narratives.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, 1.

Kejsi Take and Chau Tran. 2024. Riddlemasters at semeval-2024 task 9: Comparing instruction finetuning with zero-shot approaches. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1393–1398, Mexico City, Mexico. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of NAACL-HLT*, pages 4149–4158.

Gemini Team, Rohan Anil, Sebastian Borgeaud, and etc Yonghui Wu. 2023. Gemini: A family of highly capable multimodal models.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the seventh workshop on building educational applications using NLP*, pages 163–173.

Dilip Venkatesh and Yashvardhan Sharma. 2024. Bits pilani at semeval-2024 task 9: Prompt engineering with gpt-4 for solving brainteasers. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 803–807, Mexico City, Mexico. Association for Computational Linguistics.

Jie Wang, Jin Wang, and Xuejie Zhang. 2024a. Ynu-hpcc at semeval-2024 task 9: Using pre-trained language models with lora for multiple-choice answering tasks. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 458–463, Mexico City, Mexico. Association for Computational Linguistics.

Weiqi Wang, Baixuan Xu, Haochen Shi, Jiaxin Bai, Qi Hu, and Yangqiu Song. 2024b. Knowcomp at semeval-2024 task 9: Conceptualization-augmented prompting with large language models for lateral reasoning. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1650–1656, Mexico City, Mexico. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022a. Chain-of-thought prompting elicits reasoning in large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Qi Yang, Jingjie Zeng, Liang Yang, and Hongfei Lin. 2024. yangqi at semeval-2024 task 9: Simulate human thinking by large language model for lateral thinking challenges. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 233–238, Mexico City, Mexico. Association for Computational Linguistics.

Jiarui Zhang, Filip Ilievski, Kaixin Ma, Jonathan Francis, and Alessandro Oltramari. 2022. A study of zero-shot adaptation with commonsense knowledge. *Automated Knowledge Base Construction(AKBC)*.

Micah Zhang, Shafiuddin Rehan Ahmed, and James H. Martin. 2024. Ftg-cot at semeval-2024 task 9: Solving sentence puzzles using fine-tuned language models and zero-shot cot prompting. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1235–1241, Mexico City, Mexico. Association for Computational Linguistics.

A CodaLab Leaderboard

In the main part of the paper, we only analyse the results for part of the participants' submission due to page limitation. Table 4 and 5 show a complete set of user names and results of the participants in the CodaLab competition for two subtasks, including users who did not submit a system description. The human evaluation is computed over 102 data randomly sampled from the **whole dataset**. The random base is average over three different seeds. The ChatGPT and RoBERTa-L baseline is computed over the whole dataset using OPENAI API[7] from 2023/5/01 to 2023/5/15.

We visualize each team's overall accuracy in each subtask according to the model adaptation category in Figure 4. In Sentence Puzzle, 12 teams employed fine-tuning, and 15 adopted zero/few-shot approaches. Fine-tuning achieved 1st, 3rd, and 5th positions on the leaderboard, whereas zero/few-shot have 7 places in the top ten. For Word Puzzle, 9 teams used fine-tuning, and 11 opted for zero/few-shot, with the latter dominating the top five ranks, outperforming fine-tuning.

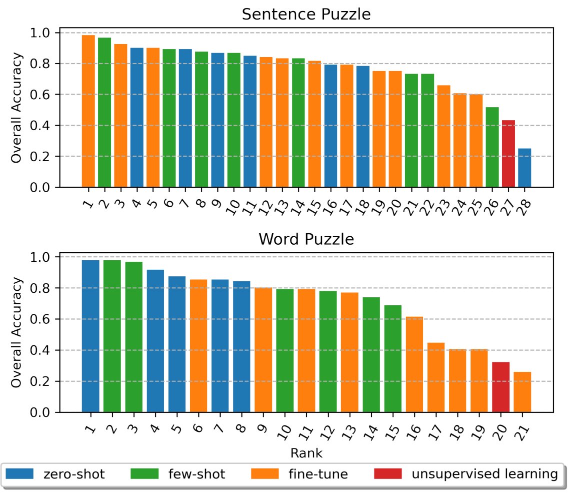

Figure 4: The overall accuracy performance of each team based on method adaptations.

B Participant Systems

In this section, we list the systems of all participants who submitted a system description paper. The **team name** represents each system, appended with the corresponding rank in [bracket], keywords in (parentheses), and a short description for further reference. *SP X* and*WP X* represent the ranks in sentence and word puzzles based on overall performance, respectively.

[7] https://platform.openai.com/docs/api-reference

Abdelhak [SP 1;WP 16] (Kelious and Okirim, 2024) (*Fine-tuned*;*DeBERTa*;*Zero-shot*;*ChatGPT*;*Temperature Anlysis*) They fine-tuned the pre-trained language model DeBERTa-v3-base in the multiple-choice setting. They further experimented with the relationship between temperature and lateral thinking with ChatGPT in a zero-shot setting.

HW-TSC [SP 2;WP 3] (Li et al., 2024)(*Fine-tuned*;*Mixstral*;*Zero-shot*;*Few-shot*;*GPT-3.5*;*GPT-4*;*Prompting Engineering*;*Ensemble*) They first experimented with fine-tuning Mixtral overall whole training set. They turned to GPT-3.5 and GPT-4 due to poor fine-tuning results. They identified and categorized over 20 challenging training instances to include in an extended prompt. Finally, they submitted their result with GPT-4 in the few-shot setting with a well-designed prompting demonstration as well as the ensemble method.

AmazUtah_NLP [SP 6;WP 10] (Mishra and Ghashami, 2024) (*Fine-tuned*;*DeBERTa*;*BERT*;*External Data*;*Synthetic Data*;*RiddleSense*) They fine-tuned DeBERTa and BERT in the multiple-choice setting. They utilized the public puzzle dataset RiddleSense as well as creating humor-style data by prompting GPT 4 as the external dataset. They also experimented by adding commonsense datasets SWAG and CODAH but found the introduction reduced overall performance.

BITS Pilani [SP 7;WP 5] (Venkatesh and Sharma, 2024) (*Zero-shot*;*GPT-4*;*Prompting Engineering*) They used OpenAI's GPT-4 model along with prompt engineering in the zero-shot setting to solve these brainteasers.

ALF [SP 7] (Farokh and Zeinali, 2024) (*Fine-tuned*;*ALBERT*;*RoBERTa*;*DeBERTa*;*Flan T5*;*Unified QA*;*External Data*;*RiddleSense*) Their experiments focused on two prominent families of pre-trained models, BERT and T5, and fine-tuned ALBERT, RoBERTa, DeBERTa, Flan T5 and Unified QA in the multiple-choice setting. They explored the potential benefits of multi-task finetuning on commonsense reasoning datasets, including RiddleSense, CSQA, PIQA, SIQA, Hellaswag, and SWAG, to enhance performance.

uTeBC-NLP [SP 8] (Sadeghi et al., 2024) (*Fine-tuned*;*Zephyr-7B-β*;*Zero-shot*;*Few-shot*;*GPT-3.5*;*GPT-*

4;*RAG*;*External Data*;*Synthetic Data*;*Prompting Engineering*;*COT*;*Lateral thinking enhancement analysis*) They explored Chain of Thought (CoT) strategies, enhancing prompts with detailed task descriptions, and retrieval augmented generation for generating in-context samples. Their experiments involve GPT-3.5 and GPT-4. They also showcased that fine-tuning Zephyr-7B-β with a lateral thinking approach significantly enhances the model's performance on other commonsense datasets.

yangqqi [SP 8;WP 6] (Yang et al., 2024) (*Zero-shot*;*ChatGPT*;*RAG*;*Self-Adaptive ICL*;*Prompting Engineering*;*External Data*;*ConceptNet*) They proposed the SHTL system to mimic human lateral thinking ability for solving brain teaser questions. They first retrieved related knowledge concepts from ConceptNet and used SAICL to find the optimal organization for each single test sample. At last, they provide ChatGPT with the related knowledge concepts and find the options to solve the conflicts contained in the related knowledge concepts effectively.

Mothman [SP 9] (Chen et al., 2024) (*Zero-shot*;*Few-shot*;*GPT-4*;*Prompting Engineering*;*COT*;) They proposed a system for iterative chain-of-thought prompt engineering which optimizes prompts using a flexible evaluation strategy on both model outputs and input data. They obtain feedback from human evaluation to modify the prompting demonstration interactively to guide GPT-4 to focus on challenging problems. They also proposed a new COT strategy requiring GPT-4 to produce rationals for both correct and incorrect options.

Zero_Shot_is_All_You_Need [SP 10;WP 2] (Moosavi Monazzah and Feghhi, 2024) (*Zero-shot*;*Bing*;*Gemini*;*Mixtral*;*Mixtral*;*ChatGPT*;*Phi-2*;*Prompting Engineering*;*Ensemble*;*Debate*) They examined the zero-shot ability of current state-of-the-art LLMs, Bing, Gemini, Mixtral, ChatGPT and Phi-2 to solve this task. They also tried ensemble and debate prompting engineering methods.

OUNLP [SP 10;WP 11] (Saravanan and Wilson, 2024) (*Zero-shot*;*Few-shot*;*GPT-3.5*;*GPT-4*;*Gemini*;*languagemodels*;*Prompting Engineering*;*COT*;*RECONCILE*;*External Data*;*crawled riddles*) They experimented with a series of structured prompts ranging from basic to those integrating task descriptions and explanations(COT). They use the most similar or the most different training example as the demonstration in the one-shot prompting. They downloaded a collection of riddles from the web as an external data source. In the end, they simulated a council scenario to evoke discussion between different models but didn't observe significant improvement.

BAMO [SP 11] (Ansari et al., 2024) (*Fine-tuned*;*RoBERTa*;*BERT*;*Zero-shot*;*Open Chat*;*Llama-2-70b*;*Mixtral*;*GPT3.5*;*Claud*;*Microsoft Copilot*;*Prompting Engineering*;*ReConcile*) They fine-tuned 2 models, BERT and RoBERTa Large, and employed a Chain of Thought (CoT) zero-shot prompting approach with 6 large language models, such as GPT-3.5, Mixtral, and Llama2. Finally, they utilized ReConcile prompting amount three models.

YNU-HPCC [SP 12;WP 13] (Wang et al., 2024a) (*Fine-tuned*;*DeBERTa*;*External Data*;*Back translation*) They fine-tuned DeBERTa in different training strategies and enhanced the training set with back translation.

FtG-CoT [SP 13] (Zhang et al., 2024) (*Fine-tuned*;*BERT*;*Zero-shot*;*Few-shot*;*GPT-3.5*;*Prompting Engineering*;*COT*) They first fine-tuned BERT in a multi-class classification setting and fine-tuned GPT-3.5 with chain-of-thought generated by zero-shot prompting. Then they picked the set of training demonstrations provided in the few-shot prompt based on the BERT encoding cosine similarity to the test question.

MasonTigers [SP 13;WP 2] (Raihan et al., 2024) (*Zero-shot*;*Few-shot*;*GPT-4.5*;*Claude*;*Mixtral*;*Prompting Engineering*;*COT*) They explored various prompting strategies to guide the models, including zero-shot, few-shot, and chain-of-thought prompting. The Ensemble method was adopted to enhance COT performance.

AILS-NTUA [SP 14;WP 7] (Panagiotopoulos et al., 2024) (*Fine-tuned*;*DeBERTa*;*RoBERTa*;*BERT*;*Mixtral*;*Llama 2*;*Phi-2*) They evaluated a plethora of pre-trained transformer-based language models of different sizes and pre-train dataset through fine-tuning. They also delved into models' frequent failures to obtain a deeper understanding of reasoning cues that make models struggle the most.

RiddleMaster [SP 15;WP 8] (Take and Tran, 2024) (*Fine-tuned*;*Mistral*;*Zero-shot*;*GPT-4*;*Prompting Engineering*;*COT*;*Ensemble*) They compared multiple zero-shot approaches using

GPT-4 as well as fine-tuned Mistral output.

UMBCLU[8] [SP 15;WP 11] (*Fine-tuned*;*Flan-T5*;*Data Augmentation*) They fine-tuned and evaluated various T5 family models on both the word and sentence puzzle tasks and showed that training on the alternative contexts improves a model's lateral reasoning capability.

KnowComp [SP 16;WP 7] (Wang et al., 2024b) (*Zero-shot*;*ChatGPT*;*Prompting Engineering*) They first prompted ChatGPT to identify relevant instances in the question and generate conceptualizations for the identified instances. They then converted each puzzle into a declarative format and modified the task to involve selecting the most plausible statement from the options.

NIMZ [SP 20;WP 19] (Rahimi et al., 2024) (*Fine-tuned*;*BERT*;*RoBERTa*;*T5*;*QA-GNN*;*External Data*;*ConceptNet*) They fine-tuned BERT, RoBERTa and T5 and evaluated their performance. They used ConceptNet as an external knowledge source and fine-tuned graph neural network QA-GNN and suggested its superiority on sentence puzzle.

Deja-Vu [SP 20;WP 20] (Chakraborty et al., 2024) (*Fine-tuned*;*BERT*;*RoBERTa*;*XLNet*;*BART*;*T5*;*Data Augmentation*) They fine-tuned five transformer-based language models and found the integration of sentence and word puzzles into a single dataset led to a noticeable decrease in accuracy.

GeminiPro [SP 21;WP 12] (Choi and Na, 2024) (*Zero-shot*;*Few-shot*;*Gemini*;*Prompting Engineering*) They tested Gemini's performance in zero-shot and few-shot settings. They experimented with whether tailor-made demonstrations to specific tasks can alleviate confusion and aid in 049 problem-solving.

iREL [SP 21;WP 14] (Gupta et al., 2024) (*Zero-shot*;*Few-shot*;*Gemini*;*Prompting Engineering*;*COT*) They tested Gemini's performance in zero-shot and few-shot settings. Especially in the few-shot setting, reasoning from Gemini and GPT-4 are integrated into the demonstration, selected by static or dynamic strategy.

IIMAS [SP 23;WP 22] (Reyes et al., 2024) (*Fine-tuned*;*BERT*;*RoBERTa*;*ChatGPT*;*Gemini*;*Data Augmentation*) They 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. During the training, they transformed the data format for specific templates.

IUST-NLPLAB [SP 24] (Abbaspour et al., 2024) (*Fine-tuned*;*MPNET*;*Zero-shot*;*GPT-3.5*) They first introduced a zero-shot approach leveraging the capabilities of the GPT3.5 model. Additionally, they presented three finetuning methodologies utilizing MPNET as the underlying architecture, each employing a different loss function.

ROSHA [SP 25;WP 20] (Rostamkhani et al., 2024) (*Fine-tuned*;*RoBERTa*;*Zero-shot*;*GPT-3.5*;*Gemini*;*Mixtral*;*GPT-4*;*External Data*;*BiRdQA*;*RiddleSense*;*Prompting Engineering*;*Reconcile*) They applied the XLM-RoBERTa model both to the original training dataset and concurrently to the original dataset alongside the BiRdQA dataset and the RiddleSense for comprehensive model training. They also tested the Reconcile prompting strategy with GPT-3.5, Gemini as well as Mixtral and zero-shot on GPT-4.

DaVinci [SP 26;WP 15] (Mathur et al., 2024) (*Few-shot*;*GPT-3.5*;*Prompting Engineering*) They used few-shot prompting on GPT-3.5 with rationale and gained insights regarding the difference in the nature of the two types of questions.

StFX-NLP [SP 27;WP 21] (Heavey et al., 2024) (*unsupervised*;*External Data*;*WordNet*) They explored three unsupervised learning models. Two of these models incorporate word sense disambiguation and part-of-speech tagging, specifically leveraging SensEmBERT and the Stanford log-linear part-of-speech tagger. The third model relies on a more traditional language modelling approach.

DeBERTa [SP 28] (Siino, 2024) (*Zero-shot*;*DeBERTa*) They used DeBERTa in zero-shot setting.

[8]The paper was withdrawn.

Table 4: Oveview of results of Sentence-puzzle subtask, including user submissions without system description papers. Ori = Original, Sem = Semantic, Con = Context. Team name with (*) submitted the system description paper. The first, second and third submissions per category are represented by highlight, **bold** and <u>underline</u>, respectively.

Team Name	**Overall**	Instance-based			Group-based	
		Original	**Semantic**	**Context**	**Ori & Sem**	**Ori & Sem & Con**
Abdelhak*	0.983	1.000	1.000	0.950	1.000	0.950
HW-TSC*	**0.967**	1.000	**0.975**	**0.925**	**0.975**	<u>0.900</u>
Maxine	<u>0.958</u>	**0.975**	**0.975**	**0.925**	<u>0.950</u>	<u>0.900</u>
YingluLi	0.950	**0.975**	<u>0.950</u>	**0.925**	<u>0.950</u>	<u>0.900</u>
Theo	0.950	<u>0.950</u>	<u>0.950</u>	0.950	<u>0.950</u>	**0.925**
somethingx95	0.942	<u>0.950</u>	<u>0.950</u>	**0.925**	<u>0.950</u>	<u>0.900</u>
gerald	0.942	<u>0.950</u>	<u>0.950</u>	**0.925**	<u>0.950</u>	<u>0.900</u>
AmazUtah_NLP*	0.925	0.925	<u>0.950</u>	<u>0.900</u>	0.925	0.875
BITS Pilani*	0.900	**0.975**	0.925	0.800	0.925	0.775
ALF*	0.900	0.925	<u>0.950</u>	0.825	0.925	0.825
uTeBC-NLP*	0.892	**0.975**	0.875	0.825	0.850	0.750
jkarolczak	0.892	**0.975**	0.875	0.825	0.875	0.775
kubapok	0.892	0.925	0.900	0.850	0.900	0.825
yangqi*	0.892	0.900	0.900	0.875	0.900	0.875
Mothman*	0.875	**0.975**	0.850	0.800	0.850	0.700
zero_shot_is_all_you_need*	0.867	<u>0.950</u>	0.825	0.825	0.800	0.725
OUNLP*	0.867	<u>0.950</u>	0.875	0.775	0.850	0.725
justingu	0.850	<u>0.950</u>	0.825	0.775	0.825	0.700
BAMO*	0.850	0.900	0.825	0.825	0.825	0.700
YNU-HPCC*	0.842	0.900	0.825	0.800	0.825	0.725
FtG-CoT*	0.833	0.900	0.825	0.775	0.800	0.675
MasonTigers*	0.833	0.850	0.825	0.825	0.800	0.700
AILS-NTUA*	0.817	0.850	0.825	0.775	0.825	0.700
RiddleMaster*	0.792	0.800	0.775	0.800	0.725	0.650
UMBCLU*	0.792	0.750	0.850	0.775	0.725	0.600
johnp	0.783	0.850	0.775	0.725	0.750	0.675
MABUSETTEH	0.783	0.800	0.775	0.775	0.775	0.700
KnowComp*	0.783	0.825	0.775	0.750	0.725	0.625
ehsan.tavan	0.775	0.800	0.800	0.725	0.775	0.675
amr8ta	0.775	0.775	0.775	0.775	0.750	0.650
yiannispn	0.767	0.800	0.800	0.700	0.750	0.625
haha123	0.758	0.825	0.775	0.675	0.750	0.625
adriti	0.758	0.750	0.725	0.800	0.725	0.675
TienDat23	0.758	0.725	0.800	0.750	0.675	0.525
Deja_Vu*	0.750	0.775	0.700	0.775	0.700	0.625
NIMZ*	0.750	0.750	0.725	0.775	0.700	0.675
iREL*	0.733	0.775	0.725	0.700	0.700	0.575
GeminiPro*	0.733	0.750	0.750	0.700	0.700	0.600
caoyongwang	0.725	0.800	0.700	0.675	0.700	0.550
IIMAS*	0.658	0.650	0.675	0.650	0.600	0.500
IUST-NLPLAB*	0.608	0.625	0.625	0.575	0.625	0.500
ROSHA*	0.600	0.625	0.575	0.600	0.500	0.375
Team DaVinci*	0.517	0.575	0.550	0.425	0.500	0.300
StFX-NLP*	0.433	0.425	0.400	0.475	0.350	0.200
Team 9	0.250	0.275	0.275	0.200	0.100	0.000
DeBERTa*	0.250	0.225	0.250	0.275	0.200	0.075
amirhallaji	0.242	0.225	0.200	0.300	0.050	0.025
maryam.najafi	0.233	0.225	0.275	0.200	0.100	0.025
Human (Jiang et al., 2023c)	0.920	0.907	0.907	0.944	0.907	0.889
GPT-4 (BREAINTEASER)	0.898	0.942	0.900	0.852	0.880	0.775
GPT-4 (BREAINTEASER(S))	0.858	0.925	0.825	0.825	0.8	0.775
ChatGPT (BREAINTEASER)	0.627	0.608	0.593	0.679	0.507	0.397
RoBERTa-L (BREAINTEASER)	0.434	0.435	0.402	0.464	0.330	0.201
Random	0.244	0.255	0.249	0.228	0.056	0.014

Table 5: Oveview of results of Word-puzzle subtask, including user submissions without system description papers. Ori = Original, Sem = Semantic, Con = Context. Team name with (*) submitted the system description paper. The first, second and third submissions per category are represented by highlight, **bold** and underline, respectively.

Team Name	**Overall**	Instance-based			Group-based	
		Original	**Semantic**	**Context**	**Ori & Sem**	**Ori & Sem & Con**
Theo	0.990	1.000	1.000	**0.969**	1.000	0.969
gerald	0.990	1.000	1.000	**0.969**	1.000	0.969
somethingx95	**0.979**	1.000	1.000	0.938	1.000	**0.938**
zero_shot_is_all_you_need*	**0.979**	1.000	1.000	0.938	1.000	**0.938**
MasonTigers*	**0.979**	**0.969**	**0.969**	1.000	**0.969**	0.969
HW-TSC*	0.969	**0.969**	0.938	1.000	0.938	**0.938**
Maxine	0.969	**0.969**	0.938	1.000	0.938	**0.938**
YingluLi	0.969	**0.969**	0.938	1.000	0.938	**0.938**
kubapok	0.948	0.906	1.000	0.938	0.906	0.844
BITS Pilani*	0.917	0.938	0.938	0.875	0.938	0.812
justingu	0.917	0.938	0.938	0.875	0.906	0.781
jkarolczak	0.875	0.906	0.938	0.781	0.875	0.688
yangqi*	0.875	0.906	0.938	0.781	0.906	0.688
ehsan.tavan	0.875	0.906	0.875	0.844	0.812	0.750
AILS-NTUA*	0.854	0.875	0.906	0.781	0.812	0.719
johnp	0.854	0.875	0.906	0.781	0.812	0.719
caoyongwang	0.854	0.844	0.844	0.875	0.781	0.719
KnowComp*	0.854	0.844	0.906	0.812	0.844	0.656
RiddleMaster*	0.844	0.844	0.844	0.844	0.781	0.656
yiannispn	0.833	0.844	0.844	0.812	0.719	0.625
AmazUtah_NLP*	0.802	0.844	0.812	0.750	0.781	0.594
OUNLP*	0.792	0.781	0.812	0.781	0.719	0.531
UMBCLU*	0.792	0.781	0.750	0.844	0.719	0.625
TienDat23	0.792	0.844	0.750	0.781	0.750	0.625
GeminiPro*	0.781	0.781	0.719	0.844	0.594	0.594
YNU-HPCC*	0.771	0.781	0.719	0.812	0.719	0.625
iREL*	0.740	0.719	0.719	0.781	0.562	0.531
Team DaVinci*	0.688	0.719	0.719	0.625	0.594	0.469
Abdelhak*	0.615	0.625	0.625	0.594	0.562	0.406
amr8ta	0.604	0.625	0.625	0.562	0.594	0.438
adriti	0.604	0.656	0.625	0.531	0.625	0.375
MABUSETTEH	0.583	0.594	0.625	0.531	0.562	0.281
NIMZ*	0.448	0.438	0.469	0.438	0.406	0.219
Deja_Vu*	0.406	0.375	0.469	0.375	0.344	0.125
ROSHA*	0.406	0.438	0.375	0.406	0.375	0.250
StFX-NLP*	0.323	0.406	0.219	0.344	0.125	0.062
IIMAS*	0.260	0.250	0.250	0.281	0.125	0.062
Human (Jiang et al., 2023c)	0.917	0.917	0.917	0.917	0.917	0.896
GPT-4 (BREAINTEASER)	0.736	0.811	0.756	0.640	0.689	0.494
GPT-4 (BREAINTEASER(S))	0.854	0.875	0.875	0.813	0.781	0.625
ChatGPT (BREAINTEASER)	0.535	0.561	0.524	0.518	0.439	0.293
RoBERTa-L (BREAINTEASER)	0.207	0.195	0.195	0.232	0.146	0.061
Random	0.260	0.279	0.225	0.073	0.018	0.253