AmazUtah_NLP at SemEval-2024 Task 9: A MultiChoice Question Answering System for Commonsense Defying Reasoning

Mina Ghashami*
Amazon Web Services
ghashami@amazon.com

Soumya Smruti Mishra*
Amazon Web Services
soumish@amazon.com

Abstract

The SemEval 2024 BRAINTEASER task represents a pioneering venture in Natural Language Processing (NLP) by focusing on lateral thinking, a dimension of cognitive reasoning that is often overlooked in traditional linguistic analyses. This challenge comprises of Sentence Puzzle and Word Puzzle subtasks and aims to test language models' capacity for divergent thinking.

In this paper, we present our approach to the BRAINTEASER task. We employ a holistic strategy by leveraging cutting-edge pre-trained models in multiple choice architecture, and diversify the training data with Sentence and Word Puzzle datasets. To gain further improvement, we fine-tuned the model with synthetic humor/jokes dataset and the RiddleSense dataset which helped augmenting the model's lateral thinking abilities. Empirical results show that our approach achieve 92.5% accuracy in Sentence Puzzle subtask and 80.2% accuracy in Word Puzzle subtask.

1 Introduction

The success of language models has inspired the Natural Language Processing community to attend to tasks that require implicit and complex reasoning. Human reasoning encompasses two types of reasoning: lateral and vertical thinking approaches. Lateral thinking demand out-of-the-box thinking. It is a form of creative reasoning that deviates from traditional, logical processes and has received little attention from NLP community. Vertical thinking on the other hand, relies on logical reasoning, and have been relatively popular in the past few years.

The BRAINTEASER dataset by (Jiang et al., 2023) stands as a crucial benchmark for evaluating question-answering systems. It particularly assesses these systems on their ability for lateral thinking — pushing them to transcend conventional commonsense reasoning towards more innovative and creative approaches to problem-solving. Additionally, as part of this effort to test language models' lateral thinking capabilities, the SemEval 2024 BRAINTEASER task (Jiang et al., 2024), offers a focused challenge derived from the broader dataset, further probing the creative reasoning abilities of these models. This task is crucial because it addresses a gap in Natural Language Processing (NLP) where most tasks focus on linear, logical (vertical) thinking, neglecting the complex, divergent aspects of human cognition represented by lateral thinking. By encompassing two subtasks — Sentence Puzzle and Word Puzzle — BRAINTEASER aims to test a model's ability to go beyond conventional commonsense associations, requiring an understanding of both standard meanings and the ability to reinterpret them in novel ways. This is vital for advancing the field of NLP, as it pushes the boundaries of what artificial intelligence can achieve in terms of mimicking the nuanced and creative aspects of human thought.

Our system adopts a multifaceted strategy for this challenge, centering on the use of advanced pre-trained models like BERT (Devlin et al., 2019) and DeBERTaV3 (He et al., 2023) through HuggingFace's (Wolf et al., 2020) AutoModelForMultipleChoice and AutoModelForSequenceClassification. This approach is enhanced by a diverse training regimen that mixes Sentence and Word Puzzle datasets, ensuring a broad exposure to different types of lateral thinking challenges. Additionally, the model is fine-tuned with a humor/jokes dataset generated by GPT-4 (OpenAI et al., 2024) and the RiddleSense (Lin et al., 2021) dataset, which introduces elements of creativity, unconventional thinking, and complex puzzle-solving. This comprehensive strategy aims to equip the model with enhanced lateral thinking abilities, crucial for tackling the creative and nuanced demands of the BRAINTEASER task in SemEval 2024.

*Both authors contributed equally to this work.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1436–1442
June 20-21, 2024

In our participation in the BRAINTEASER task, we discovered that our system, particularly when finetuned with AutoModelForMultipleChoice, outperformed the baseline instruction-tuned systems mentioned in the original paper. This approach demonstrated a significant advantage in handling multiple-choice tasks. However, we faced challenges with AutoModelForSequenceClassification, suggesting an area for improvement. The incorporation of additional synthetic data and open-source dataset like RiddleSense positively influenced our performance. Quantitatively, our system achieved a commendable 6th place in the Sentence Puzzle and 10th in the Word Puzzle, indicating stronger proficiency in sentence-based challenges and room for growth in word-based puzzles.

Our code and data will be available at https://github.com/soumyasmruti/semeval-2024-brainteaser after cleaning and de-anonimization.

2 Background

The task involves two types of brain teasers: Sentence Puzzle and Word Puzzle. In Sentence Puzzle, the input is a sentence-based question that defies commonsense, with multiple-choice answers. For instance, "A man shaves everyday, yet keeps his beard long." The choices include "He is a barber," "He wants to maintain his appearance," and so on. The Word Puzzle involves a word-based teaser, like "What part of London is in France?" with choices focusing on letters in the words (e.g., "The letter N"). The output in both cases is the selection of the correct choice that represents lateral thinking.

In order to counter the potential for Large Language Models (LLMs) memorizing solutions, BRAINTEASER (Jiang et al., 2023) incorporates two novel methods of puzzle generation: semantic and context reconstruction. These techniques generate variations of puzzles that preserve the core challenge of overturning conventional commonsense reasoning without altering the fundamental nature of the puzzles. This approach is aimed at enhancing the robustness of the puzzles against the memorization capabilities of LLMs, ensuring that the puzzles continue to effectively test the models' ability to engage in lateral thinking by challenging ingrained commonsense assumptions. This is to ensure the model is evaluating reasoning ability rather than memorization.

Systems are evaluated based on two accuracy metrics: Instance-based Accuracy, considering each question (original and adversarial) as a separate instance, and Group-based Accuracy, where a system must correctly solve all questions in a group (original and its adversarial versions) to score.

3 Related Work

We can broadly categorize the reasoning landscape of language models into two groups. The first, is ‘commonsense reasoning‘, also known as ‘vertical reasoning‘. This refers to the ability to make deductions based on everyday knowledge. The second category is ‘lateral reasoning‘; i.e. a creative problem-solving approach that involves looking at situations from unconventional perspectives.

Researchers have explored various approaches to endow LLMs with commonsense reasoning abilities (Rae et al., 2021). One prominent approach is the use of knowledge graphs, which represent structured knowledge in the form of entities and their relationships (Ilievski et al., 2021). Authors in (Wang et al., 2021) proposed a method for incorporating commonsense knowledge from ConceptNet (Speer et al., 2018) into language models, leading to improved performance on commonsense reasoning tasks.

Another approach involves fine-tuning pretrained LLMs on commonsense reasoning datasets. Authors in this paper (Huang et al., 2019) introduced the COSMOS QA dataset, which consists of multiple-choice questions that require commonsense reasoning. They showed that fine-tuning pretrained LLMs on this dataset can significantly improve their commonsense reasoning capabilities.

Researchers have also investigated the use of prompting techniques to elicit commonsense reasoning from LLMs without explicit fine-tuning. (Zhou et al., 2022) proposed a method called "Conditional Prompt-Tuning" that enables LLMs to perform commonsense reasoning by conditioning on carefully designed prompts. In another work (Wei et al., 2022), chain-of-thought prompting showed how to unlock LLM's reasoning ability via effective prompting techniques.

There hasn't been extensive research on ‘lateral thinking‘ of LLMs. Very recenlty, OlaGPT (Xie et al., 2023) proposed a cognitive architecture framework in which they summarize various methods of human reasoning into Chain-of-Thought (CoT) templates, to maximize the LLMs' reasoning effect.

Overall, while LLMs have shown flashes of non-linear, exploratory thinking on some benchmarks, lateral thinking as a holistic cognitive process remains an open challenge.

4 Methodology

In this section, we describe different methods and approaches we employed in solving the Brain-Teaser puzzle.

4.1 Sequence Classification with BERT

In this approach, we enhanced the performance of a sequence classification model through the instruction fine-tuning process. We leveraged the powerful contextual embeddings provided by BERT (Devlin et al., 2019). Our methodology involved initializing the model with pre-trained BERT weights and employing the streamlined 'AutoModelForSequenceClassification' class from the Hugging Face Transformers library, which linearly projects the embedding from the language model encoder to each document into the class logits for that document. We instructed the model with selecting the most appropriate answer from a set of four choices provided alongside a given question. Despite the meticulous fine-tuning process our experimental results revealed sub-optimal performance.

4.2 MultipleChoice QA with BERT and DeBERTa

We leveraged the versatile 'AutoModelForMultipleChoice' architecture from Hugging Face's library, which integrates a pre-trained transformer model with a specialized classification head. This architecture was pivotal in adapting the model for our multiple-choice task, which involved combining both Word Puzzle and Sentence Puzzle datasets to diversify our training data.

To ensure optimal performance, we split our training data into separate training and validation sets. Throughout the training process, we utilized the validation set to fine-tune hyperparameters, ensuring the model's efficacy.

The AutoModelForMultipleChoice architecture comprises a pre-trained base transformer augmented with a classification head. This head, typically consisting of neural network components such as linear layers and activation functions, enables the model to make informed multiple-choice predictions.

Our model initialization involved embedding pre-trained DeBERTa representations, followed by further training on the designated training dataset. This approach facilitated the model's adaptation to our specific task requirements, ultimately enhancing its performance.

4.2.1 Augmenting with RiddleSence and Humor Data

Next, we decided to use two additional data sources to augment our training data. This was with the aim of expanding the diversity of our dataset, enriching it with a wide range of humor styles, scenarios, and perspectives. This augmentation not only increases the robustness and variety of our model but also enhances its adaptability to different contexts. We utilized the public Riddlesense dataset as well as creating humor style data by prompting GPT 4.

The Ridlesense dataset consists of Riddles which are a form of puzzle where a question, often presented in a cryptic or metaphorical manner, challenges the reader to find a clever or unexpected answer.

To create the humor style QA, we prompted GPT 4. Crafting jokes content often requires a touch of ingenuity, an out-of-the-box approach, and a healthy dose of lateral thinking, and GPT-4 allowed us to explore unconventional and amusing angles to questions and answers. It's like having a comedy writer who never runs out of fresh and unexpected punchlines. The details about how the dataset was generate is provided in Appendix A.

We then used the same AutoModelForMultiple-Choice architecture and trained the model on augmented training data.

5 Experimental setup

5.1 Datasets Description

The task dataset and additional datasets used in our approaches are detailed in Table 1, with all datasets being in the English language. We did not perform any extra pre-processing on the original training or test data. To generate humor data, we used GPT-4 (OpenAI et al., 2024) using prompt engineering. Regarding the RiddleSense (Lin et al., 2021) dataset, which originally had five labels, we adapted it to a four-label format. This was achieved by reassigning questions with the fifth label as the correct answer to the fourth choice. Consequently, all fifth-choice answers across questions were remapped to their corresponding fourth choices, and all original fifth choices were discarded. Riddlesense and humor datasets, were

	Sentence Puzzle			Word Puzzle		
Dataset	**Train**	**Validation**	**Test**	**Train**	**Validation**	**Test**
Provided	405	102	120	316	80	96
Humor Data GPT4	211	-	-	211	-	-
Riddlesense	4531	-	-	-	-	-

Table 1: Dataset Statistics, '-' means the data was not used for the stage of the task.

selected for their similarity to the original training data, offering commonsense-defying puzzles. For details on the train-validation-test split, please refer to Table 1. We also experimented by adding SWAG (Zellers et al., 2018) and CODAH (Chen et al., 2019) datasets, but found that they reduced overall performance.

5.2 Implementation Details

The raw text was tokenized using a byte-level Byte-Pair Encoding (BPE) vocabulary with 50,257 merge rules, and inputs longer than 1024 tokens were truncated.

Our models were based on the BERT-base and DeBERTaV3 base architectures. The BERT model comprises 12 layers, 768-dimensional embeddings, and 12 attention heads, totaling 117M parameters. The DeBERTaV3 base model features 12 layers and a hidden size of 768, with 110M backbone parameters and a 128K token vocabulary introducing an additional 98M parameters in the embedding layer.

Both models were initialized with pre-trained weights in the AutoModelForMultipleChoice architecture. We conducted a random hyperparameter search, exploring batch sizes of [4, 16, 32] and learning rates of [5e-5, 1e-4, 2e-4]. The configurations yielding the highest validation accuracy were selected for each model size.

We utilized Amazon SageMaker for training, opting for the ml.p3.8xlarge instance for BERT-based approaches and the ml.p3.16xlarge instance for training our DeBERTaV3-based approaches. The training time for the BERT models with the original data was under 20 minutes, while the DeBERTa-based approaches were trained in under one hour. This efficient use of resources enabled us to achieve significant performance improvements with minimal cost and time.

6 Results

In Table 2, we demonstrate the performance of our model, where the provided numbers represent the accuracy for various groups. "Original," "Semantic," and "Context" denote the original question, its semantic reconstruction, and context reconstruction, respectively. These three categories are based on instance-based accuracy, where each question is treated as a separate instance. The score reports the accuracy for both the original question and its adversarial counterparts. "Orig. + Sem." represents group-based accuracy, where the original question and its semantic reconstruction are considered and calculated together. Similarly, "Orig. + Sem. + Con." includes the previous group along with the contextual reconstruction of the original question.

In the table, "AMSC" represents AutoModelForSequenceClassification, and "AMMC" represents AutoModelForMultipleChoice. The models used are bert-base-uncased and microsoft/deberta-v3-base. The notation "train-data-wp+sp" indicates that the training data for this approach includes both sentence puzzle and word puzzle training data provided by the organizers of the task. "Humor" represents the synthetic dataset generated by prompting GPT-4, and "RiddleSense" refers to the open-source RiddleSense dataset (Lin et al., 2021). The scores of human performance and the baseline system, as provided in the original paper (Jiang et al., 2023), are depicted in gray. Scores obtained by our system are shown in black, with the best performances for each task highlighted in bold.

6.1 Subtask A : Sentence Puzzle

Initially, we trained our models only on the provided sentence puzzle dataset but soon realized that combining both the sentence puzzle and word puzzle datasets yielded better validation scores. Consequently, we used the bert-base model with AutoModelForSequenceClassification, achieving an overall accuracy of 50.8%. Given that the dataset is in a multiple-choice format, we experimented

Approaches	Sentence Puzzle						Word Puzzle					
	Original	**Semantic**	**Context**	**Orig. + Sem.**	**Orig. + Sem. + Con.**	**Overall**	**Original**	**Semantic**	**Context**	**Orig. + Sem.**	**Orig. + Sem. + Con.**	**Overall**
Human	.907	.907	944	.907	.889	.920	.917	.917	.917	.917	.900	.917
ChatGPT	.608	.593	.679	.507	.397	.627	.561	.524	.518	.439	.292	.535
RoBERTa-L	.435	.402	.464	.330	.201	.434	.195	.195	.232	.146	.061	.207
BERT-base + AMSC + train-data-wp+sp	.475	.55	.5	.35	.25	.508	.281	.312	.375	.031	0	.323
BERT-base + AMMC + train-data-wp+sp	.650	.625	.625	.600	.500	.600	.438	.375	.406	.344	.375	.406
DeBERTaV3 + AMMC + train-data-wp+sp	.900	.900	.850	.900	.825	.883	.75	.75	.625	.719	.500	.708
DeBERTaV3 + AMMC + train-data-wp+sp + Humor + RiddleSense	**.925**	**.950**	**.900**	**.925**	**.875**	**.925**	-	-	-	-	-	-
DeBERTaV3 + AMMC + train-data-wp + Humor	-	-	-	-	-	-	**.844**	**.812**	**.750**	**.781**	**.594**	**.802**

Table 2: SemEval2024 Task 9: BRAINTEASER results table, which shows the performance of different approaches on the test set. Orig. = Original, Sem. = Semantic, Con. = Context, AMSC = AutoModelForSequenceClassification, AMMC = AutoModelForMultipleChoice

with AutoModelForMultipleChoice using the same bert model. This change significantly improved performance, increasing accuracy by 10 points to 60%. Encouraged by this, we opted for the larger DeBERTaV3 model under the AutoModelForMultipleChoice configuration. This model, combined with the original dataset, significantly boosted performance, raising overall accuracy to 83.3%. After incorporating additional datasets containing humor-style questions and the RiddleSense dataset, our best accuracy score reached 92.5%. Our approach ranked 6th among the 31 teams that participated in the task and outperformed the baseline zero shot ChatGPT by almost 50 percentage points.

6.2 Subtask B : Word Puzzle

The word puzzle setup followed almost the same approach as sentence puzzle but during validation process we found the best model was the one which was trained with only original training data from word puzzle dataset and adding humor dataset. Adding RiddleSense data and sentence pzzle data didn't improve the score of the word puzzle in validation process, therefore we didn't submit that output. Our approach for this subtask didn't perform that well when compared to other teams, we ranked 10th among the 23 teams that participated in this task, but outperformed the baseline zero shot ChatGPT by almost 40 percentage points.

7 Conclusion

In this work, we present our novel system designed for the SemEval 2024 BRAINTEASER task, which notably achieved 6th place in the Sentence Puzzle and 10th in the Word Puzzle categories. Our approach leverages advanced pre-trained models like BERT and DeBERTa, optimized through HuggingFace's AutoModelForMultipleChoice and AutoModelForSequenceClassification. This strategy was further enhanced by incorporating a diverse training regimen, blending Sentence and Word Puzzle datasets with a unique humor/jokes dataset and the RiddleSense dataset. This mix has been instrumental in equipping our model with the lateral thinking capabilities essential for this task. While our system excelled in the Sentence Puzzle, reflecting a stronger grasp in sentence-based lateral thinking, the performance in the Word Puzzle highlighted areas for improvement, particularly in word-based lateral reasoning. The additional challenge posed by adversarial versions of puzzles, involving both Semantic and Context Reconstruction, underscores the complexity of this task. Our system's performance underscores the efficacy of our training approach in enhancing lateral thinking in language models, a significant step forward in NLP. Future work will focus on refining our approach for word-based puzzles and further enhancing the model's ability to navigate complex, creative reasoning paths, thereby advancing the field's understanding of AI's potential in mimicking nuanced aspects of human cognition.

References

Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. 2019. Codah: An adversarially authored question-answer dataset for common sense. *arXiv preprint arXiv:1904.04365*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. *arXiv preprint arXiv:1909.00277*.

Filip Ilievski, Pedro Szekely, and Bin Zhang. 2021. Cskg: The commonsense knowledge graph. In *The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings 18*, pages 680–696. Springer.

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 *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP 2021): Findings*. To appear.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2018. Conceptnet 5.5: An open multilingual graph of general knowledge.

Bin Wang, Guangtao Wang, Jing Huang, Jiaxuan You, Jure Leskovec, and C-C Jay Kuo. 2021. Inductive learning on commonsense knowledge graph completion. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

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

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. Huggingface's transformers: State-of-the-art natural language processing.

Yuanzhen Xie, Tao Xie, Mingxiong Lin, WenTao Wei, Chenglin Li, Beibei Kong, Lei Chen, Chengxiang Zhuo, Bo Hu, and Zang Li. 2023. Olagpt: Empowering llms with human-like problem-solving abilities. *arXiv preprint arXiv:2305.16334*.

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*.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825.

A Humor Dataset Details

We used the following prompts to generate Jokes or Humor style dataset. We experimented with multiple prompts and gather all the output in a json file and analyzed them manually.

PROMPT 1 - Could you create a dataset for me that includes humor-styled questions, each with multiple choices and an answer? The dataset should be in JSON format.

PROMPT 2 - Could you create a dataset of 40 jokes for me in JSON format? Each joke should include four options and the correct answer.

PROMPT 3 - Could you generate an additional 20 jokes with multiple choices and an answer? Please ensure there are no duplicates and that none of them are the same as those previously generated.

We initially prompted GPT 4 to generate 200 questions at once but that didn't go well. The output contained duplicate questions after 15 / 16 unique ones. Basically, the model kept repeating itself. So we mostly used PROMPT 3 multiple times to generate high quality data. Before adding each we checked for duplicates again manually. Provided below are some of the jokes generated by the prompt.

"joke": "Why did the bicycle fall over?", "options": ["A. Because it was two-tired.", "B. It had a flat.", "C. It was unbalanced.", "D. It slipped."], "answer": "A"

"joke": "What's orange and sounds like a parrot?", "options": ["A. A carrot", "B. An orange bird", "C. A tangerine", "D. A flamingo"], "answer": "A" ,