KnowComp at SemEval-2024 Task 9: Conceptualization-Augmented Prompting with Large Language Models for Lateral Reasoning

Weiqi Wang, Baixuan Xu, Haochen Shi, Jiaxin Bai, Qi Hu, Yangqiu Song
Department of Computer Science and Engineering, HKUST, Hong Kong SAR, China
{wwangbw, bxuan, hshiah, jbai, qhuaf, yqsong}@cse.ust.hk

Abstract

Lateral thinking is essential in breaking away from conventional thought patterns and finding innovative solutions to problems. Despite this, language models often struggle with reasoning tasks that require lateral thinking. In this paper, we present our system for SemEval-2024 Task 9's BrainTeaser challenge, which requires language models to answer brain teaser questions that typically involve lateral reasoning scenarios. Our framework is based on large language models and incorporates a zero-shot prompting method that integrates conceptualizations of automatically detected instances in the question. We also transform the task of question answering into a declarative format to enhance the discriminatory ability of large language models. Our zero-shot evaluation results with ChatGPT indicate that our approach outperforms baselines, including zero-shot and few-shot prompting and chain-of-thought reasoning. Additionally, our system ranks ninth on the official leaderboard, demonstrating its strong performance.

1 Introduction

Recently, the Natural Language Processing (NLP) community has witnessed remarkable advancements driven by large language models, such as GPT-3.5 (OpenAI, 2022) and GPT4 (OpenAI, 2023), that demonstrated impressive capabilities in tasks like text generation (Chung et al., 2023; Maynez et al., 2023; Maiorino et al., 2023), translation (Mu et al., 2023; Bawden and Yvon, 2023; Zhang et al., 2023), reasoning (Huang and Chang, 2023; Chan et al., 2024; Gaur and Saunshi, 2023; Ho et al., 2023; Shi et al., 2023), complex reasoning (Bai et al., 2023; Fang et al., 2024), analogical understanding (Cheng et al., 2023; Ye et al., 2024)and sentiment analysis (Carneros-Prado et al., 2023; Deng et al., 2023). However, these models predominantly rely on conventional sequential thinking, often struggling to exhibit the creativity and innovative problem-solving abilities that humans possess. This limitation has spurred researchers to explore the realm of lateral thinking within the NLP domain (Veale and Li, 2013).

Lateral thinking, a concept popularized by De Bono (1970), refers to the ability to break free from established thought patterns and approach problems from unconventional angles. It encourages the exploration of unorthodox ideas, perspectives, and solutions, leading to breakthroughs and the discovery of new opportunities that may have otherwise remained hidden (Lawrence et al., 2016). Harnessing the power of lateral thinking can significantly enhance the capabilities of language models, enabling them to tackle complex, non-linear challenges by thinking "outside the box." However, engaging in this type of reasoning presents a significant challenge, as it demands the ability to contradict common knowledge—a skill highly valued by cutting-edge language models like ChatGPT (OpenAI, 2022) and GPT4 (OpenAI, 2023). Challenging traditional modes of commonsense reasoning poses a serious obstacle for these language models, as it requires them to set aside their inherent strengths and approach the problem from a different perspective.

In light of this direction, Jiang et al. (2023) have recently introduced BrainTeaser, a human-curated benchmark that evaluates the lateral thinking ability of language models. This benchmark encompasses sentence and word puzzles in a question-answering format that challenge common sense, demanding language models to demonstrate innovative thinking in order to provide accurate and insightful responses. The findings of this study expose a significant disparity in the lateral thinking capacities of even large-scale language models, including those augmented with commonsense knowledge (Wang et al., 2023a), when compared to human performance. This gap in accuracy exceeds 40%, emphasizing the necessity for novel

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1639–1645
June 20-21, 2024 ©2024 Association for Computational Linguistics

approaches to enhance the reasoning capabilities of language models.

We propose a new approach to enhance the lateral thinking capability of language models by applying conceptualization (He et al., 2022). Conceptualization is the process of abstracting instances into high-level concepts, which introduces abstract knowledge associated with the concept for the instance (Tenenbaum et al., 2011). Our method involves instructing ChatGPT to perform conceptualization over the premises in the question via a step-by-step process that identifies instances, conceptualizes them into concepts, generates relevant abstract knowledge, and merges them back into the prompt. To make the judgment less biased among choices, we transform the questions into declarative formats. We test our framework with ChatGPT (OpenAI, 2022) in a zero-shot manner, where no training data is used. Our experiment results show that our framework achieves an overall accuracy of 78.3% for sentence puzzles and 85.4% for word puzzles, ranking ninth and eighth in the official leaderboard, respectively.

2 Related Works

2.1 Lateral Reasoning

Lateral reasoning, also known as "thinking outside the box," has garnered significant attention in cognitive psychology and educational research (Evans and Alderson, 2000). Over the past decades, researchers have explored various aspects of lateral reasoning, aiming to understand its underlying processes and develop effective strategies to enhance individuals' lateral thinking abilities (Millar and Taylor, 1995). It is known to be challenging as such type of reasoning usually defies commonsense knowledge, which is knowledge about facts in the world that is typically shared among individuals (Mueller, 2014; Fang et al., 2021b,a). In the domain of NLP, Jiang et al. (2023) are the first to construct evaluation benchmarks that evaluate such cognitive ability. They formulate the task as a question-answering task and design a data collection protocol to crawl sentence puzzles and word puzzles from the web with quality filtering. Experiment results on various language models show the difficulty of their collected dataset.

2.2 Conceptualization

Conceptualization aims to abstract a set of entities or events into a general concept, thereby forming abstract commonsense knowledge within its original context (Murphy, 2004). Existing works primarily focused on entity-level conceptualization (Durme et al., 2009; Song et al., 2011, 2015; Liu et al., 2022), with He et al. (2022) pioneering the construction of an event conceptualization benchmark by extracting concepts for social events from WordNet (Miller, 1995) synsets and Probase (Wu et al., 2012). Wang et al. (2023b,a) further proposed a semi-supervised framework for conceptualizing CSKBs and demonstrated that abstract knowledge can enhance commonsense inference modeling and question answering. Wang et al. (2024) proposed distilling such type of knowledge from large language models to improve commonsense reasoning. Wang et al. (2023c) and Yu et al. (2023) also leveraged similar method to acquire abstract knowledge as high-level knowledge representation. In this paper, we share similar aspirations from previous works and leverage the power of conceptualization to assist large language models in performing lateral reasoning.

	Sentence Puzzle	Word Puzzle
#Data	120	96

Table 1: Number of data in the testing set of the BrainTeaser (Jiang et al., 2023) benchmark.

3 Task Definition and Dataset

We follow the identical task definition as proposed by Jiang et al. (2023), where each data entry can be viewed as a Question-Answering (QA) task. In each QA pair, the question describes a specific context or puzzle, and the answer serves as the lateral explanation or solution to the puzzle. The goal is to find an explanation that supports and does not contradict a given set of premises (P), which includes explicitly stated clauses and implicitly derived clauses through default commonsense inferences or associations. The set of premises (P) plays a crucial role in the puzzle. It encompasses the atomic premise set, which includes explicitly stated clauses (p_1, p_2, p_3) provided by the context, as well as implicit clauses (p_4, p_5) obtained through default commonsense inferences or associations. These implicit premises can sometimes lead to incorrect assumptions or constraints that hinder finding the correct solution (Bar-Hillel et al., 2018). The puzzle is presented in a multiple-choice format, where the answer choices represent potential explanations or solutions. This format is chosen to make the task more amenable to automated evaluation and facilitate human comprehension.

We use the dataset presented by Jiang et al. (2023, 2024) as our evaluation benchmark and follow the the original released split of data. Since we approach this task by following a zero-shot manner, no training and validation data is used. As shown in Table 1, there are 120 sentence puzzles and 96 word puzzles in the testing set. On average, the questions in this dataset consist of 34.88 tokens, while the corresponding answers have an average length of 9.11 tokens.

4 Method

In this section, we introduce our proposed method. Our method can be divided into three steps: (1) automatically identify instances in the premises in the question and conceptualize them; (2) transform the QA pair into declarative statements; and (3) Prompt ChatGPT in a zero-shot manner to obtain its prediction.

4.1 Conceptualization Augmentation

Our approach to conceptualization follows the method proposed by Wang et al. (2024). First, we provide ChatGPT with a question from the BrainTeaser QA pairs and instruct it to identify relevant keywords and instances in the question. Specifically, we ask it to focus on instances that are pertinent to the question at hand. Next, we utilize the prompt from Wang et al. (2024) to guide ChatGPT in generating conceptualizations for the identified instances. We also instruct ChatGPT to generate abstract knowledge that is relevant to the context of the question. Both the generated conceptualizations and abstract knowledge are integrated into the prompts to assist in the reasoning process. For example, consider the question "A man shaves everyday, yet keeps his beard long" in a sentence puzzle. ChatGPT identifies *shave* and *beard* as the two key instances. The instance "shave" is then conceptualized to "shaving," which further implies that *shaving causes a man's beard go short.*

4.2 Declarative Transformation

We then convert each puzzle into a declarative format and modifying the task to involve selecting the most plausible statement from the options, rather than the traditional question-and-answer format. To achieve this, we present ChatGPT with the question and one of the potential answers, and instruct it to generate a declarative statement that conveys the same meaning as the given question and answer with minimal alterations. For instance, consider the question "In a small village, two farmers are working in their fields - a diligent farmer and a lazy farmer. The hardworking farmer is the son of the lazy farmer, but the lazy farmer is not the father of the hardworking farmer. Can you explain this unusual relationship?" and one of the options, "The lazy farmer is his mother." In response, ChatGPT produces the statement "In a small village, there are two farmers working in their fields - a diligent farmer and a lazy farmer. The hardworking farmer is the son of the lazy farmer, but the lazy farmer is not the father of the hardworking farmer. This peculiar relationship can be clarified by asserting that the lazy farmer is, in fact, the mother of the hardworking farmer."

4.3 Zero-shot Prompting

Finally, we prompt ChatGPT again to ask it to select the most plausible one from the given three statements. For each statement, we also append the derived conceptualizations and associated abstract knowledge into the statement such that they can also be considered during the selection process. We also ask ChatGPT to focus on whether the abstract knowledge has any conflict to the statement presented, which aims at identifying conflicts between commonsense knowledge and the presented statement.

5 Experiments

In this section, we present details of experiments we conducted on the BrainTeaser benchmark.

5.1 Setup

We access ChatGPT through Microsoft Azure APIs[1]. The code of the accessed version for ChatGPT is `gpt-35-turbo-20230515`. The maximum generation length is set to 100 tokens and the temperature is set to 1.0. All other hyperparameters remain unchanged as default. We experiment with three random seeds and report the best performances achieved according to the leaderboard's ranking. For the evaluation metric, we keep using accuracy as the metric and also evaluate the puzzles in instance-based and group-based fashions.

[1] https://azure.microsoft.com/en-us/products/ai-services/

		Instance-based			Group-based		Overall
Category	**Model**	**Original**	**Semantic**	**Context**	**Ori & Sem**	**Ori & Sem & Con**	
				Sentence Puzzle			
Random	-	25.52	24.88	22.81	5.58	1.44	24.40
Instruction	FlanT5(11B; zero-shot)	33.49	31.58	36.84	22.01	11.00	33.97
	FlanT5(11B; two-shot)	37.80	33.49	38.76	26.79	13.40	36.68
	FlanT5(11B; four-shot)	38.28	34.45	41.15	26.79	13.40	37.96
	FlanT5(11B; six-shot)	38.28	34.45	41.63	27.27	13.88	38.12
	FlanT5(11B; eight-shot)	38.76	33.01	41.63	26.79	14.35	37.80
	T0(11B)	22.01	22.01	29.67	16.27	11.00	24.56
	T0P(11B)	23.92	22.49	34.93	17.70	11.96	27.11
	T0PP(11B)	26.32	27.27	37.80	19.14	11.96	30.46
	ChatGPT(zero-shot)	60.77	59.33	67.94	50.72	39.71	62.68
	ChatGPT(two-shot)	61.72	60.77	68.90	51.67	40.67	63.80
	ChatGPT(four-shot)	59.33	55.98	62.20	47.85	32.06	59.17
	ChatGPT(six-shot)	60.29	59.81	66.51	51.20	40.19	62.20
	ChatGPT(eight-shot)	63.16	62.68	67.46	54.55	44.02	64.43
Commonsense	RoBERTa-L(CSKG)	35.41	36.84	44.98	28.71	18.18	39.07
	CAR	10.53	10.53	11.48	5.74	2.39	10.85
Ours	ChatGPT w. Concept.	**82.50**	**77.50**	**75.00**	**72.50**	**62.50**	**78.30**
Human*	-	90.74	90.74	94.44	90.74	88.89	91.98
				Word Puzzle			
Random	-	26.02	27.85	22.51	7.32	1.83	25.34
Instruction	FlanT5(11B; zero-shot)	42.68	32.93	43.90	28.66	20.12	39.84
	FlanT5(11B; two-shot)	44.51	34.76	45.73	30.49	18.90	41.67
	FlanT5(11B; four-shot)	43.29	35.98	47.56	30.49	20.73	42.28
	FlanT5(11B; six-shot)	44.51	36.59	47.56	29.88	17.68	42.89
	FlanT5(11B; eight-shot)	45.73	33.54	46.95	27.44	16.46	42.07
	T0(11B)	17.07	14.02	23.17	9.76	6.10	18.09
	T0P(11B)	28.66	26.22	34.15	19.51	12.80	29.67
	T0PP(11B)	33.54	31.10	39.63	20.12	10.98	34.76
	ChatGPT(zero-shot)	56.10	52.44	51.83	43.90	29.27	53.46
	ChatGPT(two-shot)	55.49	53.66	51.22	44.51	30.49	53.46
	ChatGPT(four-shot)	54.27	53.66	51.83	43.90	28.05	53.25
	ChatGPT(six-shot)	56.71	51.83	**54.27**	45.12	28.66	54.27
	ChatGPT(eight-shot)	58.54	56.71	54.27	48.17	34.76	56.50
Commonsense	RoBERTa-L(CSKG)	18.90	16.46	30.49	12.80	6.10	21.95
	CAR	38.41	31.10	20.12	26.22	6.10	29.88
Ours	ChatGPT w. Concept.	**84.40**	**90.60**	**81.20**	**84.40**	**65.60**	**85.40**
Human*	-	91.67	91.67	91.67	91.67	89.58	91.67

Table 2: Main zero-shot results over two BrainTeaser subtasks across all models in all metrics: Ori = Original, Sem = Semantic, Con = Context, Concept = Conceptualization. The best performance among all models is in bold, and the second-best performance is underlined. Most of the results are reported by Jiang et al. (2023).

5.2 Baselines

For baselines, we largely follow Jiang et al. (2023) and use the officially reported results as baselines. These include instruction-based language models such as ChatGPT (OpenAI, 2022), T0 (Sanh et al., 2022), and FlanT5 (Chung et al., 2022), which were evaluated in a zero setting using specific instruction templates. In addition, commonsense models were also evaluated, including RoBERTa-L (CSKG; Ma et al., 2021) and CAR (Wang et al., 2023a), which were enhanced with commonsense knowledge and achieved impressive zero-shot performance on multiple tasks. The models were evaluated using a scoring method defined in previous studies and the choice with the highest score is selected. Meanwhile, we also report the performances of ChatGPT in a few-shot setting with up to eight shots.

5.3 Results and Analysis

Table 2 presents the results of our study. Our method significantly improves the performance of ChatGPT, outperforming all baselines. In fact, it surpasses all large language models in a zero-shot scenario and even outperforms ChatGPT itself with eight-shot prompting. For sentence puzzles, we observe an overall improvement of 13.87%, while for word puzzles, there is a 28.90% improvement. However, our method still falls short of human performance, indicating room for further improvement. Interestingly, we notice a larger improvement in word puzzles compared to sentence puzzles. This gain may be attributed more to our declarative transformation than to conceptualization, which theoretically offers little help in solving word puzzles.

6 Conclusion

In conclusion, this paper describes the solution by the KnowComp group to task 9 of SemEval-2024. Our method tackles the task of lateral thinking by leveraging the framework of conceptualization, which is a traditional reasoning method performed by humans, to assist large language models in answering brain teaser questions in a zero-shot manner. Experiment results show the superiority of our method, outperforming all previous zero-shot baselines with the same large language model as the backbone.

Acknowledgements

The authors of this paper were supported by the NSFC Fund (U20B2053) from the NSFC of China, the RIF (R6020-19 and R6021-20), and the GRF (16211520 and 16205322) from RGC of Hong Kong. We also thank the support from the UGC Research Matching Grants (RMGS20EG01-D, RMGS20CR11, RMGS20CR12, RMGS20EG19, RMGS20EG21, RMGS23CR05, RMGS23EG08).

References

Jiaxin Bai, Xin Liu, Weiqi Wang, Chen Luo, and Yangqiu Song. 2023. Complex query answering on eventuality knowledge graph with implicit logical constraints. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Maya Bar-Hillel, Tom Noah, and Shane Frederick. 2018. Learning psychology from riddles: The case of stumpers. *Judgment and Decision Making*, 13(1):112–122.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of BLOOM. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation, EAMT 2023, Tampere, Finland, 12-15 June 2023*, pages 157–170. European Association for Machine Translation.

David Carneros-Prado, Laura Villa, Esperanza Johnson, Cosmin C. Dobrescu, Alfonso Barragán, and Beatriz García-Martínez. 2023. Comparative study of large language models as emotion and sentiment analysis systems: A case-specific analysis of GPT vs. IBM watson. In *Proceedings of the 15th International Conference on Ubiquitous Computing & Ambient Intelligence (UCAmI 2023) - Volume 2, Riviera Maya, Mexico, 28-29 November, 2023*, volume 842 of *Lecture Notes in Networks and Systems*, pages 229–239. Springer.

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2024. Exploring the potential of chatgpt on sentence level relations: A focus on temporal, causal, and discourse relations. In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 684–721. Association for Computational Linguistics.

Jiayang Cheng, Lin Qiu, Tsz Ho Chan, Tianqing Fang, Weiqi Wang, Chunkit Chan, Dongyu Ru, Qipeng Guo, Hongming Zhang, Yangqiu Song, Yue Zhang, and Zheng Zhang. 2023. Storyanalogy: Deriving story-level analogies from large language models to unlock analogical understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 11518–11537. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

John Joon Young Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 575–593. Association for Computational Linguistics.

Edward De Bono. 1970. Lateral thinking. *New York*, page 70.

Xiang Deng, Vasilisa Bashlovkina, Feng Han, Simon Baumgartner, and Michael Bendersky. 2023. Llms to the moon? reddit market sentiment analysis with large language models. In *Companion Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 1014–1019. ACM.

Benjamin Van Durme, Phillip Michalak, and Lenhart K. Schubert. 2009. Deriving generalized knowledge from corpora using wordnet abstraction. In *EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, Athens, Greece, March 30 - April 3, 2009*, pages 808–816. The Association for Computer Linguistics.

Kenneth E Evans and Andrew Alderson. 2000. Auxetic materials: functional materials and structures from lateral thinking! *Advanced materials*, 12(9):617–628.

Tianqing Fang, Zeming Chen, Yangqiu Song, and Antoine Bosselut. 2024. Complex reasoning over logical queries on commonsense knowledge graphs. *arXiv preprint arXiv:2403.07398*.

Tianqing Fang, Weiqi Wang, Sehyun Choi, Shibo Hao, Hongming Zhang, Yangqiu Song, and Bin He. 2021a. Benchmarking commonsense knowledge base population with an effective evaluation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8949–8964. Association for Computational Linguistics.

Tianqing Fang, Hongming Zhang, Weiqi Wang, Yangqiu Song, and Bin He. 2021b. DISCOS: bridging the gap between discourse knowledge and commonsense knowledge. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 2648–2659. ACM / IW3C2.

Vedant Gaur and Nikunj Saunshi. 2023. Reasoning in large language models through symbolic math word problems. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5889–5903. Association for Computational Linguistics.

Mutian He, Tianqing Fang, Weiqi Wang, and Yangqiu Song. 2022. Acquiring and modelling abstract commonsense knowledge via conceptualization. *CoRR*, abs/2206.01532.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14852–14882. Association for Computational Linguistics.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1049–1065. Association for Computational Linguistics.

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, EMNLP 2023, Singapore, December 6-10, 2023*, pages 14317–14332. Association for Computational Linguistics.

Moyra Lawrence, Sylvain Daujat, and Robert Schneider. 2016. Lateral thinking: how histone modifications regulate gene expression. *Trends in Genetics*, 32(1):42–56.

Jingping Liu, Tao Chen, Chao Wang, Jiaqing Liang, Lihan Chen, Yanghua Xiao, Yunwen Chen, and Ke Jin. 2022. Vocsk: Verb-oriented commonsense knowledge mining with taxonomy-guided induction. *Artif. Intell.*, 310:103744.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. 2021. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13507–13515. AAAI Press.

Antonio Maiorino, Zoe Padgett, Chun Wang, Misha Yakubovskiy, and Peng Jiang. 2023. Application and evaluation of large language models for the generation of survey questions. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, pages 5244–5245. ACM.

Joshua Maynez, Priyanka Agrawal, and Sebastian Gehrmann. 2023. Benchmarking large language model capabilities for conditional generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9194–9213. Association for Computational Linguistics.

BJ Millar and NG Taylor. 1995. Lateral thinking: the management of missing upper lateral incisors. *British Dental Journal*, 179(3):99–106.

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

Yongyu Mu, Abudurexiti Reheman, Zhiquan Cao, Yuchun Fan, Bei Li, Yinqiao Li, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2023. Augmenting large language model translators via translation memories. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10287–10299. Association for Computational Linguistics.

Erik T Mueller. 2014. *Commonsense reasoning: an event calculus based approach*. Morgan Kaufmann.

Gregory Murphy. 2004. *The big book of concepts*. MIT press.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI*.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Haochen Shi, Weiqi Wang, Tianqing Fang, Baixuan Xu, Wenxuan Ding, Xin Liu, and Yangqiu Song. 2023. QADYNAMICS: training dynamics-driven synthetic QA diagnostic for zero-shot commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 15329–15341. Association for Computational Linguistics.

Yangqiu Song, Haixun Wang, Zhongyuan Wang, Hongsong Li, and Weizhu Chen. 2011. Short text conceptualization using a probabilistic knowledgebase. In *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 2330–2336. IJCAI/AAAI.

Yangqiu Song, Shusen Wang, and Haixun Wang. 2015. Open domain short text conceptualization: A generative + descriptive modeling approach. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 3820–3826. AAAI Press.

Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. 2011. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285.

Tony Veale and Guofu Li. 2013. Creating similarity: Lateral thinking for vertical similarity judgments. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 660–670. The Association for Computer Linguistics.

Weiqi Wang, Tianqing Fang, Wenxuan Ding, Baixuan Xu, Xin Liu, Yangqiu Song, and Antoine Bosselut. 2023a. CAR: conceptualization-augmented reasoner for zero-shot commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 13520–13545. Association for Computational Linguistics.

Weiqi Wang, Tianqing Fang, Chunyang Li, Haochen Shi, Wenxuan Ding, Baixuan Xu, Zhaowei Wang, Jiaxin Bai, Xin Liu, Jiayang Cheng, Chunkit Chan, and Yangqiu Song. 2024. CANDLE: iterative conceptualization and instantiation distillation from large language models for commonsense reasoning. *CoRR*, abs/2401.07286.

Weiqi Wang, Tianqing Fang, Baixuan Xu, Chun Yi Louis Bo, Yangqiu Song, and Lei Chen. 2023b. CAT: A contextualized conceptualization and instantiation framework for commonsense reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13111–13140. Association for Computational Linguistics.

Zhaowei Wang, Haochen Shi, Weiqi Wang, Tianqing Fang, Hongming Zhang, Sehyun Choi, Xin Liu, and Yangqiu Song. 2023c. Abspyramid: Benchmarking the abstraction ability of language models with a unified entailment graph. *CoRR*, abs/2311.09174.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Qili Zhu. 2012. Probase: a probabilistic taxonomy for text understanding. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2012, Scottsdale, AZ, USA, May 20-24, 2012*, pages 481–492. ACM.

Xiao Ye, Andrew Wang, Jacob Choi, Yining Lu, Shreya Sharma, Lingfeng Shen, Vijay Tiyyala, Nicholas Andrews, and Daniel Khashabi. 2024. Analobench: Benchmarking the identification of abstract and long-context analogies. *CoRR*, abs/2402.12370.

Changlong Yu, Weiqi Wang, Xin Liu, Jiaxin Bai, Yangqiu Song, Zheng Li, Yifan Gao, Tianyu Cao, and Bing Yin. 2023. Folkscope: Intention knowledge graph construction for e-commerce commonsense discovery. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1173–1191. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.