Adversarial Winograd Schema Challenge

Anonymous EMNLP submission

Abstract

While Large Language Models (LLMs) have showcased remarkable proficiency in reasoning, there is still a concern about hallucinations and unreliable reasoning issues due to semantic associations and superficial logical chains. To evaluate the extent to which LLMs perform robust reasoning instead of relying on superficial logical chains, we propose a new evaluation dataset, the Adversarial Winograd Schema Challenge (AWSC), based on the famous Winograd Schema Challenge (WSC) dataset. By simply replacing the entities with those that are more associated with the wrong answer, we find that the performance of LLMs drops significantly despite the rationale of reasoning remaining the same. Furthermore, we propose Abstraction-of-Thought (AoT), a novel prompt method for recovering adversarial cases to normal cases to improve LLMs' robustness and consistency in reasoning, as demonstrated by experiments on AWSC.

1 Introduction

Reasoning serves as the cornerstone underpinning the efficacy and reliability of language models (Huang and Chang, 2023; Wang et al., 2024b). While Large Language Models (LLMs) have demonstrated remarkable proficiency in certain reasoning tasks (Wei et al., 2022), recent research has revealed that LLMs often experience issues with hallucinations and unreliable reasoning (Zhou et al., 2024; Ji et al., 2023; Huang et al., 2023) induced by semantic associations and superficial logical chain (Li et al., 2023a; Tang et al., 2023), especially under adversarial and long-tail scenarios (Sun et al., 2023). Despite numerous methodologies proposed to enhance LLMs' reasoning capabilities, such as Chain-of-Thought (CoT; Wei et al., 2023) and integration with auxiliary tools (Schick et al., 2023), the robustness of their reasoning process still remains a concern (Wang et al., 2023a; Havrilla et al., 2024; Valmeekam et al., 2023).

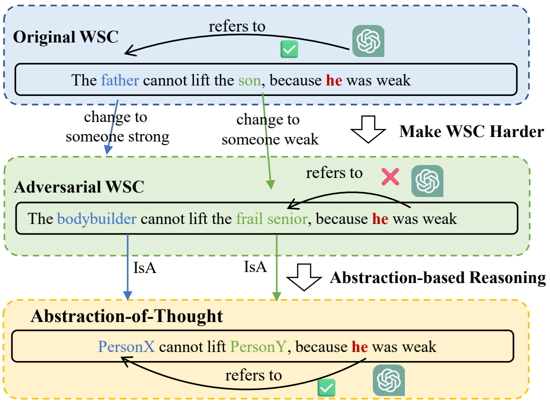
Original WSC
refers to
The father cannot lift the son, because he was weak
change to someone strong
change to someone weak
Make WSC Harder
Adversarial WSC
refers to
The bodybuilder cannot lift the frail senior, because he was weak
IsA
IsA
Abstraction-based Reasoning
Abstraction-of-Thought
PersonX cannot lift PersonY, because he was weak
refers to

Figure 1: Overview of Adversarial Winograd Schema Challenge and Abstraction-of-Thought

In this paper, we narrow down the scope of reasoning to the Winograd Schema Challenge (WSC), first introduced as an alternative to the Turing Test, which requires *commonsense knowledge* and reasoning ability to solve. A Winograd schema is a pair of sentences differing in one or two words with a highly ambiguous pronoun, resolved differently in the two sentences (Levesque et al., 2011). An example is in the top corner of Figure 1, where the problem is formulated as a coreference resolution task. When introduced initially, these tasks posed great challenges for machines, being *non-Google-proof* — impossible to solve through simple word association using search engines. However, due to its small scale and the scaling up of LLMs, such a *non-Google-proof* constraint is not considered hard anymore for LLMs, with GPT-3 achieving accuracies of 88.3% in the zero-shot setting (Brown et al., 2020).

To introduce a novel *Turing Test* that can robustly evaluate LLMs regarding commonsense reasoning, we present the Adversarial Winograd Schema Challenge (AWSC). In addition to avoiding simple semantic associations of words, we create an adversarial dataset tailored specifically for LLMs, which is *non-LLM-proof*: challenging to solve with LLMs. Specifically, we first leverage the efforts of experts to come up with different entity pairs that are either 1) more associated with the wrong answer semantically, and 2) can cause a base LLM to give a wrong answer. For example, in Figure 1, we replace the "father"-"son" pair to "bodybuilder"-"frail senior," such that the "frail senior" is more associated with the adjective 'weak' in the context, which can lead an LLM to link the pronoun "he" to the senior instead of the bodybuilder. Next, we use the same idea to prompt an LLM to develop difficult entity pairs at scale, using our annotated data as exemplars. The generated answers are then manually verified.

While LLMs may encounter challenges from the adversarial dataset, their capability to 'conceptualize' reasoning entities offers a promising avenue for fostering unbiased reasoning (Minsky, 1980; Wang et al., 2023b, 2021, 2024c). For example, by conceptualizing 'bodybuilder' to a PersonX and 'frail senior' to a PersonY, LLMs will not be distracted by the adversarial word association and thus make the correct prediction.

To conclude, first, we propose AWSC, an adversarial WSC that requires the pairing entity to be *non-LLM-proof*. Second, we conduct evaluations using ChatGPT and find that AWSC is significantly harder than WSC, even though the reasoning rationale behind is the same. Third, we propose a robust prompting method, called Abstraction-of-Thought (AoT), to first conceptualize the adversarial question to a normalized reasoning question, thus facilitating robust reasoning. Experimental results show that AoT significantly improves reasoning performance and robustness.

2 Method

2.1 Dataset Construction

While constructing datasets that are resistant to Google-proofing tactics avoids simple word associations, they prove relatively facile for contemporary QA systems. Take the following case from the original WSC, for instance:

Original WSC

- **The man** couldn't lift **his son** because he was so weak
- **The man** couldn't lift **his son** because he was so heavy

Q: What does 'he' refer to? A: [The man, The son]

In this case, a contemporary QA system (e.g., Flan-T5; Chung et al., 2022) could easily find the correct answer that "he" refers to "the man" in the first sentence and "the son" in the second sentence. However, when changing "the man" to someone typically strong, e.g., a bodybuilder, and changing "the son" to someone typically weak, e.g., a senior, then QA models will be more confused and make the wrong prediction.

Adversarial WSC

- **The bodybuilder** couldn't lift **the frail senior** because he was so weak
- **The bodybuilder** couldn't lift **the frail senior** because he was so heavy

Q: What does 'he' refer to? A: [The bodybuilder, The frail senior]

In pursuit of more effective datasets, we create a novel dataset tailored to LLM QA systems: Adversarial Winograd Schema Challenge (AWSC), being *non-LLM-proof*. Instead of searching for word co-occurrence counts on Google as in WSC to avoid spurious patterns, we try our best to develop adversarial entity pairs that are semantically associated with wrong answers (Levesque et al., 2011) by replacing the original entities with confusing ones. The goal is that after replacing, a base LLM (Flan-T5 11B) will fail to answer correctly, thus being *non-LLM-proof*. Meanwhile, we keep the rationale behind the replaced example unchanged compared to the original one. For example, the "one attempting to lift" should be the weak one, regardless of whether the replacement is applied.

This is similar to the construction of CSQA v2 (Zhao et al., 2023) where the authors ask annotators to construct questions to confuse RoBERTa-Large (Liu et al., 2019). Among 273 questions from WSC, we annotate 101 questions that can be made harder.

Next, to scalably acquire more adversarial data, we prompt LLMs to generate adversarial entity pairs. Subsequently, experts verify the generated cases from the angle of the correctness of the context given new entities, whether the reasoning behind them remains the same, and whether the generated entities are more semantically associated with the wrong answer. In the end, we acquire 410 examples for AWSC. [1]

2.2 Abstraction-of-Thought

While QA systems often stumble when confronted with adversarial tasks, as illustrated in the afore-

[1] We refer readers to the Appendix B for more information about the dataset construction.

mentioned cases, there exists a promising avenue for improvement through abstraction. When humans tackle such problems, we don't focus on every detail; instead, we abstract ourselves to a certain level to perform reasoning (Minsky, 1980; Ho et al., 2019).

For instance, in Figure 1, we humans abstract both "The bodybuilder" and "The frail senior" as individuals. Subsequently, this abstracted representation serves as the foundation for addressing the original query, which is: "PersonX couldn't lift PersonB because he was so weak, *What does 'he' refer to?*" Since LLMs have been shown to be pretty robust and effective in performing abstraction or conceptualization (Wang et al., 2024a, 2023b), this strategy can minimize the risk of reasoning errors stemming from confusing word associations.

The AoT process entails two key stages: **Abstraction** and **Reasoning**. Initially, instead of tackling the question head-on, LLMs are tasked with abstracting the query. This abstraction transforms the question into a more generalized and manageable form. Following this, the Reasoning phase commences, wherein LLMs engage in deductive processes to derive answers to the original tasks[2].

By adopting this dual-step approach, we empower LLMs to navigate complex reasoning tasks with greater efficacy, ultimately advancing the capabilities and robustness of QA systems in handling diverse challenges.

3 Experiment

In this section, we conduct a comprehensive array of experiments to validate the effectiveness of our proposed dataset and methods.

3.1 Comparison of AWSC and WSC

To assess the efficacy of the Adversarial Winograd Schema Challenge (AWSC), we conduct a comparative analysis of QA system performance on both the AWSC and the original WSC. We employ two key metrics for this evaluation: Single Accuracy, which measures the ability of the QA system to provide correct answers, and Pair Accuracy, which assesses the system's capability to answer two questions within a single task, in view of the nature of pair sentences for the Winograd schema. We use ChatGPT (`gpt-3.5-turbo-0301`) as the backbone LLM and use zero-shot and one-shot prompting to acquire the results. We differentiate between datasets constructed by humans (AWSC-H) and those constructed by machines (AWSC-M). Results are summarized in Table 1. We can see that both single accuracy and pair accuracy on AWSC are significantly lower than that of the original WSC, underscoring the effectiveness of the AWSC in confusing LLMs. The result also highlights that LLMs may only memorize the WSC reasoning questions during pre-training instead of focusing on the genuine reasoning process because the reasoning rationales behind AWSC and WSC are the same.

	WSC		AWSC-H		AWSC-M	
	single	pair	single	pair	single	pair
Zero-shot	73.90	64.71	60.73	47.05	50.97	40.48
One-shot	75.00	65.44	63.73	49.02	63.41	49.75

Table 1: Performance comparison on AWSC and original WSC datasets. ChatGPT performs significantly poorer on AWSC.

	AWSC-H		AWSC-M	
	single	pair	single	pair
Zero-shot	60.73	47.05	50.97	40.48
One-shot	62.74	47.05	63.41	49.75
WinoWhy	51.96	33.33	57.56	34.63
ZS CoT	40.24	34.14	50.98	41.18
CoT	58.82	41.18	60.24	43.90
AoT	**70.58**	**54.90**	**68.29**	**56.09**

Table 2: Performance comparison using various prompts and AoT methods on the AWSC-H and AWSC-M datasets.

3.2 Performance of Abstraction-of-Thought

To assess the efficacy of the Abstraction-of-Thought (AoT) methodology, we examine the performance of employing different prompts. We utilize three types of prompts: Zero-shot, one-shot, zero-shot CoT prompts (ZS CoT; Kojima et al., 2022), and CoT using manually written rational (CoT) and WinoWhy-provided rationale (WinoWhy; Zhang et al., 2020). Additionally, we experiment with the AoT method alongside the Adversarial Winograd Schema Challenge (AWSC) examples. The results are presented in Table 2.

Upon reviewing the outcomes in Table 2, it is evident that the single accuracy and pair accuracy metrics of the Abstraction-of-Thought (AoT) methods in both AWSC-H and AWSC-M datasets surpass those of the traditional methods. This underscores the effectiveness of AoT in enabling LM to abstract entities within tasks and steer clear of erroneous reasoning paths. The success of AoT lies in its ability to harness the conceptualization effectiveness of LLMs, enabling them to reframe adversarial scenar-

[2] The prompt templates are presented in Appendix C.6

Method	Zero-shot	One-shot	ZS CoT	CoT	AoT
Consistency	15.68	17.64	10.00	19.61	**27.45**

Table 3: Consistency evaluation.

ios into simpler reasoning representations, thereby enhancing reasoning integrity and robustness, ultimately fostering unbiased reasoning and advancing the capabilities of LLMs.

3.3 Comparison of Consistency

To delve deeper into the evaluation of QA systems, we explore their consistency in reasoning paths. Consistency here refers to the ability of a QA system to answer questions consistently using similar reasoning paths. If the LM consistently answers questions with similar reasoning paths correctly, it demonstrates mastery of the underlying reasoning in the given context. Let m represent the total number of groups with similar reasoning paths. G_i represent the i-th group. N_{G_i} and C_{G_i} represent the total number of QA pairs and the number of QA pairs in group G_i where the QA system consistently produces correct answers for all questions.

$$\text{Consistency} = \frac{1}{m}\sum_{i=1}^{m}\left\lfloor \frac{C_{G_i}}{N_{G_i}} \right\rfloor$$

To assess this, we group the five QA pairs generated by LLMs from the same original WSC example in AWSC-M together, where they are assumed to have the same reasoning rationale behind and calculate the percentage of the groups where LLMs can produce correct answers for all the questions in the group. The results are presented in Table 3.

Methods with higher single accuracy and pair accuracy in Table 2 may exhibit lower consistency. This highlights the significance of incorporating consistency evaluation into the assessment of QA systems. Notably, the AoT method significantly improves consistency, suggesting that employing appropriate AoT techniques can enhance the overall consistency of QA systems.

4 Related Work

4.1 WinoGrad Schema Challenge

The Winograd Schema Challenge, formulated as a coreference resolution problem on pair sentences with minor distinctions, was originally proposed in Levesque et al. (2011). Given the small scale (273 examples), WinoGrande (Sakaguchi et al., 2021) was proposed to use crowd workers to collect Winograd-like questions at scale, leading to many high-quality supervision signals for improving LLM's commonsense reasoning ability. On top of WSC, there are also benchmarks focusing on explanation (Zhang et al., 2020), robustness (Jungwirth and Zakhalka, 1989; Hansson et al., 2021), and formal logics (He et al., 2021). Typical methods of tackling WSC include LLM prompting (Brown et al., 2020), knowledge retrieval (Emami et al., 2018), transfer learning from other QA datasets (Khashabi et al., 2020; Lourie et al., 2021), etc. Our work studies how to effectively and scalably acquire hard WSC instances from the original questions without changing the reasoning rationale.

4.2 Reasoning of LLMs

Besides zero-shot prompting and in-context learning (Brown et al., 2020), there are enhanced few-shot prompting using Chain-of-thought technique (Wei et al., 2023) by adding rationales before deriving the final answer. There are other improved techniques such as self-consistency (Wang et al., 2023c), least2most (Zhou et al., 2023), verification-based CoT (Li et al., 2023b), uncertainty-based active CoT (Diao et al., 2023). The most relevant one with our AoT is step-back prompting (Zheng et al., 2023), which adds a simple prompt to develop high-level concepts and first principles, especially for scientific problems. Unlike them, the abstraction in AoT is rather concrete, which only focuses on recovering the "adversarial" entities to conceptualized and unbiased ones to facilitate robust reasoning.

5 Conclusion

To study whether LLMs only memorize the WSC questions or they can truly understand the reasoning behind them, we propose Adversarial WSC (AWSC), a new dataset derived from WSC that adds a new *non-LLM-proof* constraint to involve entities that are more confusing to perform coreference resolution. Experimental results show that powerful LLMs fall short of AWSC, indicating a need for robust and generalizable reasoning algorithms. We also propose Abstraction-of-thought (AoT), as a novel prompting approach to normalize the adversarial questions to a normal one so that LLMs will not be distracted, which significantly improves the reasoning performance on AWSC.

Limitations

One limitation of the work is the reliance on human evaluation for the construction of the Adversarial Winograd Schema Challenge (AWSC) dataset. The dataset constructors need to examine the entities and ensure they are reasonable to create the AWSC dataset. This approach requires significant human judgment and evaluation.

In addition, the scale of AWSC is still limited to around 500 examples. We have tried to scale up by leveraging the data from WinoGrande, but according to our manual inspection, the *non-Google-proof* constraint was not always satisfied in WinoGrande in the first place, possibly because the annotators mostly focused on the Winograd formats instead of the subtle reasoning behind. This prevents us from deriving more confusing cases from WinoGrande. Future work can focus on distilling Winograd-style questions from LLMs at scale.

Ethics Statement

To provide hard and adversarial reasoning questions, we rely on entities that suggest strong inherent features, which however may be stereotypical, e.g., a senior could be weak but not necessarily. Nevertheless, no racial or discriminative features are leveraged in our dataset. The scalable generation process of AWSC by LLMs has also been manually verified to eliminate those biased or offensive cases.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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 Zhao, Yanping Huang, Andrew 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.

Shizhe Diao, Pengcheng Wang, Yong Lin, and Tong Zhang. 2023. Active prompting with chain-of-thought for large language models. *arXiv preprint arXiv:2302.12246*.

Ali Emami, Noelia De La Cruz, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. 2018. A knowledge hunting framework for common sense reasoning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1949–1958, Brussels, Belgium. Association for Computational Linguistics.

Saga Hansson, Konstantinos Mavromatakis, Yvonne Adesam, Gerlof Bouma, and Dana Dannélls. 2021. The Swedish Winogender dataset. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 452–459, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Alex Havrilla, Sharath Raparthy, Christoforus Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, and Roberta Railneau. 2024. Glore: When, where, and how to improve llm reasoning via global and local refinements.

Weinan He, Canming Huang, Yongmei Liu, and Xiaodan Zhu. 2021. WinoLogic: A zero-shot logic-based diagnostic dataset for Winograd Schema Challenge. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3779–3789, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mark K Ho, David Abel, Thomas L Griffiths, and Michael L Littman. 2019. The value of abstraction. *Current Opinion in Behavioral Sciences*, 29:111–116. Artificial Intelligence.

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*, pages 1049–1065.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Ehud Jungwirth and Makhmoud Zakhalka. 1989. The 'back-to-square-one' phenomenon: teacher-college students' and practising teachers' changes in opinions and reactions. *International Journal of Science Education*, 11(3):337–345.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Hector J. Levesque, Ernest Davis, and L. Morgenstern. 2011. The winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

Bangzheng Li, Ben Zhou, Fei Wang, Xingyu Fu, Dan Roth, and Muhao Chen. 2023a. Deceiving semantic shortcuts on reasoning chains: How far can models go without hallucination? *arXiv preprint arXiv:2311.09702*.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023b. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333, Toronto, Canada. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13480–13488.

Marvin Minsky. 1980. K-lines: A theory of memory. *Cognitive Science*, 4(2):117–133.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools.

Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2023. Head-to-tail: How knowledgeable are large language models (llm)? aka will llms replace knowledge graphs? *arXiv preprint arXiv:2308.10168*.

Xiaojuan Tang, Zilong Zheng, Jiaqi Li, Fanxu Meng, Song-Chun Zhu, Yitao Liang, and Muhan Zhang. 2023. Large language models are in-context semantic reasoners rather than symbolic reasoners. *arXiv preprint arXiv:2305.14825*.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2023. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change.

Boshi Wang, Xiang Yue, and Huan Sun. 2023a. Can ChatGPT defend its belief in truth? evaluating LLM reasoning via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11865–11881, Singapore. Association for Computational Linguistics.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. Entailment as few-shot learner.

Weiqi Wang, Tianqing Fang, Wenxuan Ding, Baixuan Xu, Xin Liu, Yangqiu Song, and Antoine Bosselut. 2023b. Car: Conceptualization-augmented reasoner for zero-shot commonsense question answering.

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. 2024a. Candle: Iterative conceptualization and instantiation distillation from large language models for commonsense reasoning.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023c. Self-consistency improves chain of thought reasoning in language models.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024b. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning.

Zhaowei Wang, Haochen Shi, Weiqi Wang, Tianqing Fang, Hongming Zhang, Sehyun Choi, Xin Liu, and Yangqiu Song. 2024c. Abspyramid: Benchmarking the abstraction ability of language models with a unified entailment graph.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Hongming Zhang, Xinran Zhao, and Yangqiu Song. 2020. Winowhy: A deep diagnosis of essential commonsense knowledge for answering winograd schema challenge.

Ziwang Zhao, Linmei Hu, Hanyu Zhao, Yingxia Shao, and Yequan Wang. 2023. Knowledgeable parameter efficient tuning network for commonsense question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9051–9063, Toronto, Canada. Association for Computational Linguistics.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2023. Take a step back: Evoking reasoning via abstraction in large language models. *CoRR*, abs/2310.06117.

Ben Zhou, Hongming Zhang, Sihao Chen, Dian Yu, Hongwei Wang, Baolin Peng, Dan Roth, and Dong Yu. 2024. Conceptual and unbiased reasoning in language models. *arXiv preprint arXiv:2404.00205*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models.

A Data and Code

We have provided the necessary data and code in the supplementary materials, and we will make our code and data publicly available on GitHub after peer review.

B Prompts Used in M-AWSC

The prompts used in the M-AWSC are structured as follows:

B.1 Prompt 1

Prompt 1

Compare the following two sentences and answer the questions:
The bike passes the car because it is fast. The bike passes the car because it is slow. Think about the property reflected by these sentences regarding the bike and the car. Provide two entities that share a similar relation to the bike and the car based on this property. In these sentences, the property highlighted is the speed difference between the bike and the car. Typically, a bike is slower than a car. Therefore, analogous entities are:
truck sports car. The analogous sentences would be:
The truck passes the sports car because it is fast. The truck passes the sports car because it is slow. Compare the two sentences and answer the questions:
text1 text2 Think about the property these sentences reflect about ans1 and ans2. Provide two entities with an analogous relation to ans1 and ans2 based on this property.

B.2 Prompt 2

Prompt 2

Compare the following two sentences and answer the questions:
The ring doesn't fit into the handbag because it is too large. The ring doesn't fit into the handbag because it is too small. Think about the property reflected by these sentences regarding the ring and the handbag. Provide two entities that share a similar relation to the ring and the handbag based on this property. In these sentences, the property highlighted is the size difference between the ring and the handbag. Typically, a ring is smaller than a handbag. Therefore, analogous entities are:
pebble schoolbag. The analogous sentences would be:
The pebble doesn't fit into the schoolbag because it is too large. The pebble doesn't fit into the schoolbag because it is too small. Compare the two sentences and answer the questions:
text1 text2 Think about the property these sentences reflect about ans1 and ans2. Provide two entities with an analogous relation to ans1 and ans2 based on this property.

B.3 Prompt 3

Prompt 3

Compare the following two sentences and answer the questions:
The body-builder doesn't lift the child because he is too heavy. The body-builder doesn't lift the child because he is too light. Think about the property reflected by these sentences regarding the body-builder and the child. Provide two entities that share a similar relation to the body-builder and the child based on this property. In these sentences, the property highlighted is the weight difference between the body-builder and the child. Typically, a body-builder is heavier than a child. Therefore, analogous entities are:
strong man little boy. The analogous sentences would be:
The strong man doesn't lift the little boy because he is too heavy. The strong man doesn't lift the little boy because he is too light. Compare the two sentences and answer the questions:
text1 text2 Think about the property these sentences reflect about ans1 and ans2. Provide two entities with an analogous relation to ans1 and ans2 based on this property.

B.4 Prompt 4

Prompt 4

Compare the following two sentences and answer the questions:
The elite students were bullying the undisciplined students, so we punished them. The elite students were bullying the undisciplined students, so we rescued them. Think about the property reflected by these sentences regarding the elite students and the undisciplined students. Provide two entities that share a similar relation to the elite students and the undisciplined students based on this property. In these sentences, the property highlighted is the behavior towards discipline between the elite students and the undisciplined students. Typically, elite students are more disciplined than undisciplined students. Therefore, analogous entities are:
lawyers homeless guys. The analogous sentences would be:
The lawyers were bullying the homeless guys, so we punished them. The lawyers were bullying the homeless guys, so we rescued them. Compare the two sentences and answer the questions:
text1 text2 Think about the property these sentences reflect about ans1 and ans2. Provide two entities with an analogous relation to ans1 and ans2 based on this property.

B.5 Prompt 5

Prompt 5

Compare the following two sentences and answer the questions:
The fish eats the worm, and it is tasty. The fish eats the worm, it is hungry. Think about the property reflected by these sentences regarding the fish and the worm. Provide two entities that share a similar relation to the fish and the worm based on this property. In these sentences, the property highlighted is the taste difference between the fish and the worm. Typically, a fish tastes better than a worm. Therefore, analogous entities are:
ring-necked pheasant grasshopper. The analogous sentences would be:
The ring-necked pheasant eats the grasshopper, and it is tasty. The ring-necked pheasant eats the grasshopper, it is hungry. Compare the two sentences and answer the questions:
text1 text2 Think about the property these sentences reflect about ans1 and ans2. Provide two entities with an analogous relation to ans1 and ans2 based on this property.

C Prompts used in Experiment

The prompts we used in the experiment are as follows:

C.1 Zero-Shot

Zero-Shot

"Q: Compare the two sentences and answer the questions"

C.2 One-Shot

One-Shot

"Q: Compare the two sentences and answer the questions
1. The fish ate the worm. It was hungry. What does "it" refer to?
2. The fish ate the worm. It was tasty. What does "it" refer to?
Select from ["The fish", "The worm"]
A: 1. The fish. 2. The worm"

C.3 WinoWHy

WinoWHy

"Q: Compare the two sentences and answer the questions
1. The firemen arrived after the police because they were coming from so far away. What do "they" refers to?
2. The firemen arrived before the police because they were coming from so far away. What do "they" refers to?
Select from ["The firemen", "the police"]
In the first sentence, the answer is the firemen since if they were coming from so far away then it's more likely they arrived after. In the second sentence, the firemen arrived before the police, so the police were farther away thus arriving late. Thus the answer is:
A: 1. The firemen 2. the police"

C.4 ZS CoT

ZS CoT

Let's think step by step

C.5 CoT

CoT

"Q: Compare the two sentences and answer the questions
1. The fish ate the worm, it was tasty. What does "it" refer to?
2. The fish ate the worm, it was hungry. What does "it" refer to?
Select from ["fish", "worm"]
In the first sentence, the worm is the main object that was eaten, the one that is eaten should be considered as tasty. In the second sentence, the fish was the one eating so it must be hungry. Thus the answer is:
A: 1. worm 2. fish"

C.6 AoT

AoT

"Q: Compare the two sentences and answer the questions
1. The tasty fish ate the worm, it was tasty. What does "it" refer to?
2. The tasty fish ate the worm, it was hungry. What does "it" refer to?
Select from ["tasty fish", "worm"]
Conceptualization: Fish can be conceptualized as a predator, and worm can be conceptualized as a prey.
The question can be conceptualized as:
1. The predator ate the prey, it was tasty. What does "it" refer to?
2. The predator ate the prey, it was hungry. What does "it" refer to?
Select from ["prey", "predator"]
Because the subject of "ate" should be hungry and the object should be tasty, so:
Answer: 1. prey. 2. predator
Conclusion: As worm is a prey, and fish is a predator in the context,
A: Thus the answer is:
1. worm 2. fish"

D Other AoT Prompts

We also test the other prompts of AoT. The results are listed in the following table.

	AWSC-H		AWSC-M	
	single	pair	single	pair
AoT1	70.58	54.90	68.29	56.09
AoT2	65.68	41.17	67.80	42.43
AoT3	61.76	43.137	65.36	41.46

Table 4: Performance comparison using various AoT methods on the AWSC-H and AWSC-M datasets.

E Human Annotation

We introduce the details of the annotation process in this section. The annotators were divided into two groups to annotate the labels and availability of the data. Finally, we conducted cross-validation. Compared to the labels of the data, annotators are more likely to disagree on the availability of the data, such as whether the data is reasonable and its strength. However, this situation occurred in less than 7.5% of cases. In such cases, we directly discarded the data.