Concept-Reversed Winograd Schema Challenge: Evaluating and Improving Robust Reasoning in Large Language Models via Abstraction

Kaiqiao Han[1]*, **Tianqing Fang**[2,3]*, **Zhaowei Wang**[2], **Yangqiu Song**[2], **Mark Steedman**[4]
[1]Zhejiang University [2]HKUST [3]Tencent AI Lab [4] University of Edinburgh
kaiqiaohan@zju.edu.cn, {tfangaa, zwanggy, yqsong}@cse.ust.hk

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 Concept-Reversed Winograd Schema Challenge (CR-WSC), based on the famous Winograd Schema Challenge (WSC) dataset. By simply reversing the concepts to 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 using conceptual abstraction to improve LLMs' robustness and consistency in reasoning, as demonstrated by experiments on CR-WSC.[1].

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., 2023; 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

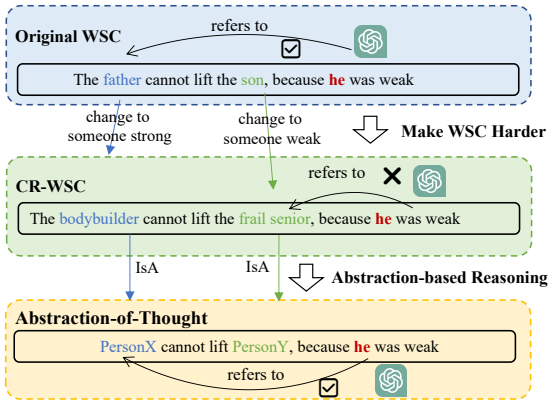

Figure 1: Overview of Concept-Reversed Winograd Schema Challenge and Abstraction-of-Thought

Chain-of-Thought (CoT; Wei et al., 2023) and integration with tools and model (Schick et al., 2023; Chai et al., 2023; Huang et al., 2024), the robustness of their reasoning process still remains a concern (Wang et al., 2023a; Havrilla et al., 2024; Valmeekam et al., 2023).

In this paper, we narrow down the scope of reasoning to the Winograd Schema Challenge (WSC), a classic reasoning challenge 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, 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 (Levesque et al., 2011). 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

* Equal Contribution
[1]Code and data are available at https://github.com/HKUST-KnowComp/Adv-WSC

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 229–243
April 29 - May 4, 2025

et al., 2020).

To introduce a novel *Turing Test* that can robustly evaluate LLMs regarding commonsense reasoning, we present the Concept-Reversed Winograd Schema Challenge (CR-WSC). 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 ask NLP experts to come up with different concept pairs that 1) has reversed attributes associated with the true answer (more semantically associated with the wrong answer), and 2) can cause a base LLM to give a wrong answer. For example, in Figure 1, we replace the "father"-"son" pair with "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., 2021, 2024d). 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 CR-WSC, an adversarial Winograd Schema Challenge that requires the pairing entity to be *non-LLM-proof*. Second, we conduct evaluations using LLMs and find that CR-WSC is significantly harder than WSC, even though the reasoning rationale and logic behind it are the same. Third, we propose a robust prompting method, called Abstraction-of-Thought (AoT), to first abstract 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]

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 because in the training data, statements of the form "X couldn't lift Y because he was weak/heavy" often co-occur with statements about X being weak or Y being heavy, but not vice versa. 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 because the inherent assumptions about the strength of bodybuilders and the weakness and frailty of seniors work against the commonsense knowledge the model relies on for predicting who can lift whom.

CR-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: Concept-Reversed Winograd Schema Challenge (CR-WSC), being *non-LLM-proof*. Instead of searching for word co-occurrence counts on Google as in WSC to avoid spurious patterns, we ask annotators to try their best to develop adversarial entity pairs that are semantically associated with wrong answers by replacing the original entities with confusing ones. The goal is that after replacing, an 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 anno-

tators 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 in this *non-LLM-proof* way. Next, to scalably acquire more adversarial data, we prompt LLMs to generate adversarial entity pairs. Subsequently, expert annotators 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. We recruit two annotators, both graduate students specializing in NLP, to carry out the annotations. They work independently on the annotations and attempt to resolve any discrepancies afterward, ultimately agreeing to disagree when necessary. In the end, we acquire 410 examples for CR-WSC[2].

2.2 Abstraction-of-Thought

While QA systems often stumble when confronted with adversarial tasks, as illustrated in the aforementioned 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 their types. 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[3]. By adopting this dual-step approach, we empower LLMs to navigate reasoning tasks with greater efficacy, advancing the capabilities and robustness of QA systems in handling diverse challenges.

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

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

	WSC		CR-WSC-H		CR-WSC-M	
	single	pair	single	pair	single	pair
GPT3.5 (0-s)	73.90	64.71	60.73	47.05	50.97	40.48
GPT3.5 (1-s)	75.00	65.44	63.73	49.02	63.41	49.75
GPT4 (0-s)	85.92	80.88	53.92	37.25	54.63	28.29
GPT4 (1-s)	91.91	86.03	76.47	68.62	74.63	60.94

Table 1: Performance comparison on CR-WSC and original WSC datasets. ChatGPT and GPT4 both perform significantly poorer on CR-WSC. 0-s indicates zero-shot and 1-s indicates one-shot.

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 CR-WSC and WSC

To assess the efficacy of the Concept-Reversed Winograd Schema Challenge (CR-WSC), we conduct a comparative analysis of QA system performance on both the CR-WSC 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, given the nature of pair sentences for the Winograd schema. We use ChatGPT (`gpt-3.5-turbo-0301`) and GPT4 (`gpt-4-turbo-2024-04-09`) as the backbone LLM and use zero-shot and one-shot prompting to acquire the results. We differentiate between datasets constructed by humans (CR-WSC-H) and those constructed by machines (CR-WSC-M). Results are summarized in Table 1. We can see that both single accuracy and pair accuracy on CR-WSC are significantly lower than that of the original WSC, underscoring the effectiveness of the CR-WSC in confusing LLMs. The result also highlights that LLMs may only memorize the WSC questions during pre-training instead of focusing on genuine reasoning because the reasoning rationales behind CR-WSC and WSC are the same.

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

	GPT3.5				Llama3.1				Mistral-7B			
	CR-WSC-H		CR-WSC-M		CR-WSC-H		CR-WSC-M		CR-WSC-H		CR-WSC-M	
	single	pair	single	pair	single	pair	single	pair	single	pair	single	pair
Zero-shot	60.73	47.05	50.97	40.48	31.37	11.76	32.43	6.83	30.39	7.84	24.39	6.83
One-shot	62.74	47.05	63.41	49.75	64.71	52.94	59.27	47.32	50.00	13.73	44.63	16.10
WinoWhy	51.96	33.33	57.56	34.63	77.45	**68.62**	**72.20**	57.07	25.49	5.88	**47.80**	13.17
ZS CoT	40.24	34.14	50.98	41.18	45.10	45.10	36.10	31.22	23.53	3.92	24.63	6.83
CoT	58.82	41.18	60.24	43.90	76.47	64.71	71.95	56.09	48.04	13.73	43.17	14.63
AoT	**70.58**	**54.90**	**68.29**	**56.09**	**78.43**	**68.62**	71.95	**57.56**	**52.94**	**19.61**	42.20	**20.49**

Table 2: Performance comparison using various prompts and AoT methods on the CR-WSC-H and CR-WSC-M datasets across GPT3.5, Llama3.1, and Mistral-7B-Instruct-v0.2 models.

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 Concept-Reversed Winograd Schema Challenge (CR-WSC) examples. The results are presented in Table 2. We use the closed-sourced ChatGPT (`gpt-3.5-turbo-0301`), open-sourced Llama-3.1 (`Meta-Llama-3.1-70B-Instruct-Turbo`), and Mistral 7B (`Mistral-7B-Instruct-v0.2`)[4] as representatives.

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 CR-WSC-H and CR-WSC-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 scenarios 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 further evaluate QA systems, we examine their consistency in reasoning paths, meaning the system can answer similar questions using similar reasoning paths. Consistency indicates mastery of reasoning in a given context. Let m represent the number of groups with similar reasoning paths, G_i the i-th group, and N_{G_i} and C_{G_i} the total and correct QA pairs in group G_i, respectively. Consistency is calculated as: Consistency $= \frac{1}{m}\sum_{i=1}^{m}\left\lfloor\frac{C_{G_i}}{N_{G_i}}\right\rfloor$.

We group the five QA pairs from the same WSC example in CR-WSC-M, assuming they share the same reasoning rationale. Results in Table 5 show

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

Table 3: Consistency Evaluation.

that methods with higher single and pair accuracy (Table 2) can still have lower consistency. This underscores the importance of evaluating consistency in QA systems. Notably, the AoT method greatly improves consistency, showing its effectiveness in enhancing overall system performance.

3.4 Error Analysis

To deepen our understanding of LLM's reasoning errors and the AoT method, we provide examples of the CoT and AoT methods to compare how LLM applies these methods differently and examples where AoT fails, as examples shown in Table 4.

Firstly, we present examples of the CoT and AoT methods to compare how LLM applies these two methods in Example 1. In CoT, LLM simply judges based on logical relationships or correlations, which can easily mislead it into providing incorrect answers. Conversely, by using AoT, LLM eliminates adversarial inputs, thereby enabling it to offer correct reasoning paths.

In the following example, LLM does not grasp the appropriate logical relationships and is misled by multiple "reject" in the input, resulting in a CoT reasoning error. However, after conceptualizing through AoT, LLM eliminates adversarial influences and provides accurate answers.

Furthermore, we provide an example of AoT failures where, in some cases, it does not provide the appropriate level of abstraction, failing to eliminate adversarial issues and thereby failing to answer questions correctly in example 2. How to enable LLM to achieve the appropriate level of abstraction is a future research direction worth exploring.

In the second example, AoT conceptualizes "The bodybuilder" as a strong individual and "the frail

[4]https://aimlapi.com/

Example 1:
James asked the rejected person for a favor but he was refused. James asked the rejected person for a favor but he refused. Q: What does 'he' refer to?
AoT Conceptualization: In these sentences, James is the one making a request, and the rejected person is the one being asked for a favor. The pronoun "he" can refer to either James or the rejected person depending on the context provided by the sentence. 1. James asked the rejected person for a favor but he was refused. Here, "he was refused" implies that the action of refusing was directed towards the subject making the request. Thus, "he" refers to James, who made the request and was refused. 2. James asked the rejected person for a favor but he refused. In this sentence, "he refused" indicates that the subject performing the action of refusing is the one who was asked for a favor. Therefore, "he" refers to the rejected person, who refused to grant the favor. **Conclusion:** 1. James 2. the rejected person
CoT Explanation: In the first sentence, "he was refused" implies that James was the one who experienced refusal, suggesting that the rejected person refused James's request. Therefore, "he" refers to James. In the second sentence, "he refused" indicates that the subject of the sentence, James, was the one who did the refusing. Therefore, "he" refers to James. **Answer:** 1. James 2. James
Example 2:
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?
AoT Conceptualization: The bodybuilder can be conceptualized as a strong individual, and the frail senior can be conceptualized as a physically weaker individual. The question can be conceptualized as: The strong individual couldn't lift the weak individual because he was so weak. What does "he" refer to? The strong individual couldn't lift the weak individual because he was so heavy. What does "he" refer to? Select from ["weak individual", "strong individual"] Because the subject of "couldn't lift" should be strong and the object should be weak, so: Answer: 1. frail senior. 2. frail senior
CoT: Conclusion: As the frail senior is the weaker individual in the context, Answer: Thus the answer is: 1. frail senior 2. frail senior

Table 4: Case studies of differences between AoT and CoT.

senior" as a physically weaker individual. This does not eliminate adversarial issues, leading LLM to still perform inaccurately when answering the question.

4 Related Work

4.1 WinoGrad Schema Challenge

The Winograd Schema Challenge (WSC) was first proposed in Levesque et al. (2011). Due to its small scale, WinoGrande (Sakaguchi et al., 2021) was introduced to expand it. Additional benchmarks focus on explanation (Zhang et al., 2020), robustness (Jungwirth and Zakhalka, 1989; Hansson et al., 2021), and formal logic (He et al., 2021). Common approaches include LLM prompting, knowledge retrieval, and transfer learning from other datasets. Our work explores scalable ways to generate difficult examples without altering reasoning logic.

4.2 Reasoning of LLMs

In addition to zero-shot prompting and in-context learning (Brown et al., 2020), methods like Chain-of-Thought (CoT) reasoning (Wei et al., 2023), self-consistency (Wang et al., 2023c), and active CoT (Diao et al., 2023) have improved few-shot prompting. The most related technique to our AoT is step-back prompting (Zheng et al., 2024), which encourages high-level thinking. AoT focuses on transforming adversarial entities into unbiased ones to strengthen reasoning robustness.

5 Conclusion

To determine if LLMs truly understand reasoning or simply memorize questions, we introduce CR-WSC, a new dataset with confusing entities for coreference resolution. Experiments show that even powerful LLMs struggle with CR-WSC, highlighting the need for more robust reasoning methods. We propose AoT, a prompting technique that normalizes adversarial questions to improve LLM reasoning ability in complex reasoning questions.

Limitations

One limitation of the work is the reliance on human evaluation for the construction of the Concept-Reversed Winograd Schema Challenge (CR-WSC) dataset. The dataset constructors need to examine the entities and ensure they are reasonable to create the CR-WSC dataset. This approach requires

significant human judgment and evaluation. However, All evaluation sets should be manually verified to ensure the accuracy of evaluation and maintain the high quality of datasets—many well-used datasets with manual annotation, such as MMLU, Big-Bench, and MMMU (Hendrycks et al., 2021; Srivastava et al., 2023; Yue et al., 2024).

In addition, the scale of CR-WSC 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

In our efforts to generate challenging and adversarial reasoning questions, we leverage entities with strong inherent characteristics. However, we recognize that such traits can sometimes be perceived as stereotypical; for instance, a senior individual might be depicted as weak, even though this is not necessarily accurate. Importantly, our dataset does not incorporate any racial or discriminatory features. Furthermore, the scalable generation process for our Concept-Reversed Winograd Schema Challenge Dataset (CR-WSC), executed by LLMs, has undergone meticulous manual verification to ensure the exclusion of biased or offensive content.

We employ a multi-layered approach to dataset creation to maintain ethical standards and avoid perpetuating stereotypes. Our team actively engages in reviewing and refining the dataset, ensuring that the content produced aligns with our commitment to fairness and inclusivity. This thorough oversight helps to identify and address any potential issues before they impact the final dataset. Addressing stereotypes and biases begins with their identification. Recognizing these issues is a crucial initial step, enabling individuals and organizations to devise strategies to mitigate them and foster more inclusive and equitable environments (Mehrabi et al., 2021b,a; Zhao et al., 2017).

Furthermore, our research introduces the Abstraction-of-Thought (AoT) framework as a method for transforming adversarial questions within the CR-WSC dataset into more neutral and conceptually focused reasoning problems. By emphasizing conceptual reasoning over surface-level biases, AoT aids in preventing the reinforcement of stereotypes and biases in both the dataset and the resulting models.

This multi-pronged approach, combining manual verification and AoT techniques, demonstrates our commitment to creating high-quality, ethical, and unbiased datasets and AI systems.

Acknowledgement

We thank the anonymous reviewers and chairs for their constructive suggestions. Yangqiu Song was supported by the ITSP Platform Research Project (ITS/189/23FP) from ITC of Hong Kong, SAR, China, and the AoE (AoE/E-601/24-N), the RIF (R6021-20) and the GRF (16211520 and 16205322) from RGC of Hong Kong, SAR, China. We thank Prof. Ernest Davis for his insightful feedback on this work.

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.

Ziwei Chai, Tianjie Zhang, Liang Wu, Kaiqiao Han, Xiaohai Hu, Xuanwen Huang, and Yang Yang. 2023. Graphllm: Boosting graph reasoning ability of large language model.

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*.

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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

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*.

Xuanwen Huang, Kaiqiao Han, Yang Yang, Dezheng Bao, Quanjin Tao, Ziwei Chai, and Qi Zhu. 2024. Can gnn be good adapter for llms?

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.

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. 2023. Deceiving semantic shortcuts on reasoning chains: How far can models go without hallucination? *arXiv preprint arXiv:2311.09702*.

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.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021a. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Ninareh Mehrabi, Pei Zhou, Fred Morstatter, Jay Pujara, Xiang Ren, and Aram Galstyan. 2021b. Lawyers are dishonest? quantifying representational harms in commonsense knowledge resources.

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.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng,

Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

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, Wei Fan, Qing Zong, Hongming Zhang, Sehyun Choi, Tianqing Fang, Xin Liu, Yangqiu Song, Ginny Y. Wong, and Simon See. 2024c. Absinstruct: Eliciting abstraction ability from llms through explanation tuning with plausibility estimation.

Zhaowei Wang, Haochen Shi, Weiqi Wang, Tianqing Fang, Hongming Zhang, Sehyun Choi, Xin Liu, and Yangqiu Song. 2024d. 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.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi.

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

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints.

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. 2024. Take a step back: Evoking reasoning via abstraction in large language models.

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*.

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-CR-WSC

The prompts used in the M-CR-WSC 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.
Answer:
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.

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.
Answer:
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.

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.
Answer:
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.

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.
Answer:
In these sentences, the property highlighted is the difference in behavior or discipline between the **elite students** and the **undisciplined students**. Typically, **elite students** are more disciplined compared to **undisciplined students**. Therefore, analogous entities are:
lawyers
homeless individuals.
The analogous sentences would be:
The **lawyers** were bullying the **homeless individuals**, so we punished them.
The **lawyers** were bullying the **homeless individuals**, so we rescued them.

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.
Answer:
In these sentences, the property highlighted is the taste or satisfaction derived from eating the **worm** by the **fish**. Typically, the fish finds the worm either tasty or satisfies its hunger. 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.

C Prompts used in Experiment 3.2

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.

	CR-WSC-H		CR-WSC-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 5: Performance comparison using various AoT methods on the CR-WSC-H and CR-WSC-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.

F Case Study

To deepen our understanding of LLM's reasoning errors and the AoT method, we provide examples of the CoT and AoT methods to compare how LLM applies these methods differently and examples where AoT fails.

We categorized failure cases into two types:

Inability to achieve the appropriate level of abstraction: Example: In the sentence, "The body-builder couldn't lift the frail senior because he was so heavy," AoT might incorrectly focus on physical strength instead of the contextual weight factor, leading to an incorrect reference assignment.

Ineffective elimination of adversarial influences: Example: In cases with multiple conflicting "reject" terms designed to confuse reasoning paths, AoT might fail to disentangle these distractions, resulting in incorrect conclusions.

These insights underscore the need for further refinement to enhance AoT's robustness, which we aim to address in future work.

Firstly, we present examples of the CoT and AoT methods to compare how LLM applies these two methods. In CoT, LLM simply judges based on logical relationships or correlations, which can easily mislead it into providing incorrect answers. Conversely, by using AoT, LLM eliminates adversarial inputs, thereby enabling it to offer correct reasoning paths.

In the following example, LLM did not grasp the

appropriate logical relationships and was misled by multiple "reject" inputs, resulting in a CoT reasoning error. However, after conceptualizing through AoT, LLM eliminated adversarial influences and provided accurate answers.

Furthermore, we provide examples of AoT failures where, in some cases, it did not provide the appropriate level of abstraction, failing to eliminate adversarial issues and thereby failing to answer questions correctly. How to enable LLM to achieve the appropriate level of abstraction is a future research direction worth exploring.

In the following example, AoT conceptualized "The bodybuilder" as a strong individual and "the frail senior" as a physically weaker individual. This did not eliminate adversarial issues, leading LLM to still perform inaccurately when answering the question.

AoT Case

James asked the rejected person for a favor but he was refused.
James asked the rejected person for a favor but he refused.
Q: What does 'he' refer to
Answer:
Conceptualization:
In these sentences, James is the one making a request, and the rejected person is the one being asked for a favor. The pronoun ""he"" can refer to either James or the rejected person depending on the context provided by the sentence.
1. James asked the rejected person for a favor but he was refused. - Here, ""he was refused"" implies that the action of refusing was directed towards the subject making the request. Thus, ""he"" refers to James, who made the request and was refused.
2. James asked the rejected person for a favor but he refused. - In this sentence, ""he refused"" indicates that the subject performing the action of refusing is the one who was asked for a favor. Therefore, ""he"" refers to the rejected person, who refused to grant the favor.
Conclusion: 1. James 2. the rejected person"

CoT Case

James asked the rejected person for a favor but he was refused.
James asked the rejected person for a favor but he refused.
Q: What does 'he' refer to
Answer:
In the first sentence, ""he was refused"" implies that James was the one who experienced refusal, suggesting that the rejected person refused James's request. Therefore, ""he"" refers to James.
In the second sentence, ""he refused"" indicates that the subject of the sentence, James, was the one who did the refusing. Therefore, ""he"" refers to James.
A: 1. James 2. James

AoT Failure Case

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
Answer:
Conceptualization: The bodybuilder can be conceptualized as a strong individual, and the frail senior can be conceptualized as a physically weaker individual. The question can be conceptualized as:
1. The strong individual couldn't lift the weak individual because he was so weak. What does ""he"" refer to?
2. The strong individual couldn't lift the weak individual because he was so heavy. What does ""he"" refer to?
Select from [""weak individual"", ""strong individual""]
Because the subject of ""couldn't lift"" should be strong and the object should be weak, so:
Answer: 1. frail senior. 2. frail senior
Conclusion: As the frail senior is the weaker individual in the context, A: Thus the answer is: 1. frail senior 2. frail senior"

G Abstraction Ability of LLMs

The ability to perform abstraction in reasoning has been an active area of research for LLMs. Abstraction can assist LLMs in solving scientific problems

by allowing them to think about the underlying theorems and principles behind the questions, thereby improving their reasoning capabilities in real-world problem-solving (Zheng et al., 2024). Existing research has demonstrated that incorporating abstraction can indeed enhance the reasoning ability of LLMs, and this has been validated in fine-tuning paradigms (Wang et al., 2024c). Improvements have been observed across various tasks, including question-answering (Wang et al., 2023b).