AuRoRA: A One-for-all Platform for Augmented Reasoning and Refining with Task-Adaptive Chain-of-Thought Prompting

Anni Zou[1,3]*, Zhuosheng Zhang[2]†, Hai Zhao[1,3]†, Xiaoshan Li[5]*, Minjie Bian[4,5]

[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2] School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University
[3] Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University
[4] Shanghai Technology Innovation Department, Shanghai Data Group Co., Ltd
[5] Shanghai Data Group Co., Ltd
{annie0103, zhangzs}@sjtu.edu.cn,zhaohai@cs.sjtu.edu.cn,{lixs, bianmj}@sdata.net.cn

Abstract

Large language models (LLMs) empowered by chain-of-thought (CoT) prompting have yielded remarkable prowess in reasoning tasks. Nevertheless, current methods predominantly lean on handcrafted or task-specific demonstrations, lack reliable knowledge basis and thus struggle for trustworthy responses in an automated pattern. While recent works endeavor to improve upon one certain aspect, they ignore the importance and necessity of establishing an integrated and interpretable reasoning system. To address these drawbacks and provide a universal solution, we propose AuRoRA: a one-for-all platform for augmented reasoning and refining based on CoT prompting that excels in adaptability, reliability, integrity, and interpretability. The system exhibits superior performances across six reasoning tasks and offers real-time visual analysis, which has pivotal academic and application value in the era of LLMs. The AuRoRA platform is available at https://huggingface.co/spaces/Anni123/AuRoRA.

Keywords: Chain of Thought Prompting, Large Language Models, Task Adaptation, Reasoning

1. Introduction

Large language models (LLMs) have showcased impressive competence across a wide range of reasoning tasks (Brown et al., 2020; Scao et al., 2022; Chowdhery et al., 2022; Touvron et al., 2023; Qin et al., 2023; OpenAI, 2023). Notably, these capabilities are bolstered by chain-of-thought (CoT) prompting (Kojima et al., 2023; Wei et al., 2023) that guides LLMs to conduct reasoning step by step based on a simple trigger or several in-context learning (ICL) demonstrations.

Despite the unparalleled achievements gained by current CoT prompting methods, obstacles persist in divergent aspects toward a more adaptive, reliable and interpretable resolution. On one hand, existing approaches primarily rely on manually crafted or task-specific demonstrations, thus failing to adapt themselves to diverse question types (Wei et al., 2023; Wang et al., 2023; Zhou et al., 2023; Zhang et al., 2023c). On the other hand, research has unveiled a discrepancy between the intrinsic knowledge within LLMs and the knowledge required for a given task (Zhang et al., 2023a; Zhu et al., 2023; McKenna et al., 2023), which results in the *hallucination* phenomenon (Ji et al., 2023; Zhang et al., 2023b). Hence, establishing an effective alignment between the acquired and required knowledge for the input question is of crucial importance. Although recent works have embarked on either pursuing generality (Zou et al., 2024) or enhancing reasoning reliability (Gou et al., 2023; Zhao et al., 2023; Yoran et al., 2023), they tend to prioritize one specific aspect and yet overlook the necessity of an integrated and interpretable system, which abounds with potential application values.

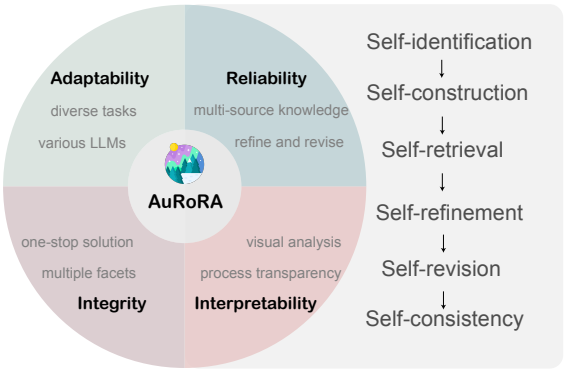

Figure 1: The major features of AuRoRA.

To address the above drawbacks, we propose AuRoRA: a one-for-all platform for augmented reasoning and refining based on CoT prompting.

* Equal contribution.
† Corresponding authors. This paper was partially supported by Joint Research Project of Yangtze River Delta Science and Technology Innovation Community (No. 2022CSJGG1400) and Joint Funds of the National Natural Science Foundation of China (No. U21B2020).

LREC-COLING 2024, pages 1801–1807
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

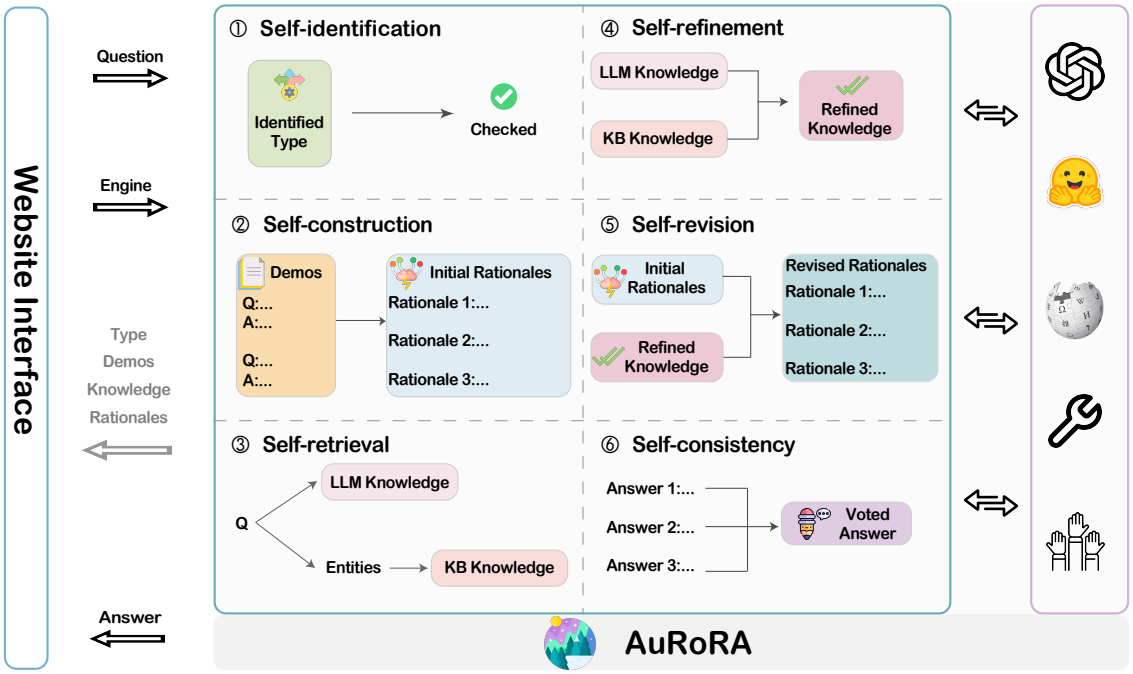

Figure 2: The workflow of AuRoRA platform.

AuRoRA works in six steps: (i) first, it identifies the type of input question; (ii) second, it automatically constructs the most representative ICL demonstrations according to the question type and derives the initial rationales; (iii) third, it retrieves relevant knowledge from LLMs and external knowledge bases; (iv) fourth, it refines the acquired multi-source knowledge to achieve verification and filtering; (v) fifth, it revises the initial rationales based on the high-quality knowledge; (vi) lastly, it outputs the final answer with highest self-consistency. We evaluate our system on six tasks encompassing arithmetic, commonsense, and symbolic reasoning. Experimental results show that AuRoRA consistently outperforms the zero-shot baseline by a large margin.

To sum up, AuRoRA has the prominent features as follows (as shown in Figure 1):

(i) **Adaptability**: liberates the need for manual labor and handily adapts to diverse tasks and LLMs;

(ii) **Reliability**: fuses multi-source knowledge and adopts a self-revision mechanism;

(iii) **Integrity**: serves as a universal one-stop solution that augments the reasoning capability from various facets;

(iv) **Interpretability**: provides user-friendly visual analysis and achieves process transparency.

2. The AuRoRA platform

This section will present the details of our proposed framework and describe the website interface.

2.1. System Workflow

Figure 2 depicts the workflow of AuRoRA platform. Its workflow consists of six steps:
(i) Self-identification: determines the type of input question which is confirmed by the user;
(ii) Self-construction: builds the most representative ICL demonstrations according to the identified type and derives the initial rationales;
(iii) Self-retrieval: extracts multi-source knowledge to further alleviate hallucination;
(iv) Self-refinement: refines multi-source knowledge for verification and filtering;
(v) Self-revision: revises the initial rationales based on the high-quality knowledge;
(vi) Self-consistency: returns the final answer with the high hest level of self-consistency.

AuRoRA provides a one-for-all solution for performing task-adaptive CoT reasoning. The workflow runs free from manual annotations, improves the credibility of the responses, and offers visual analysis with adequate transparency.

Self-identification Given an input question q, we leverage the demonstrations formulated as [`Question:`q_i`; Type:`t_i], where (q_i, t_i) are from distinct types including *arithmetic*, *commonsense-mc*, *commonsense-verify*, *symbolic-coin* and *symbolic-letter*. Then the input query is appended to the demonstrations, which is subsequently delivered to LLMs to infer the data type t.

Self-construction Given the question type t, we first locate the corresponding data pool DP_t.

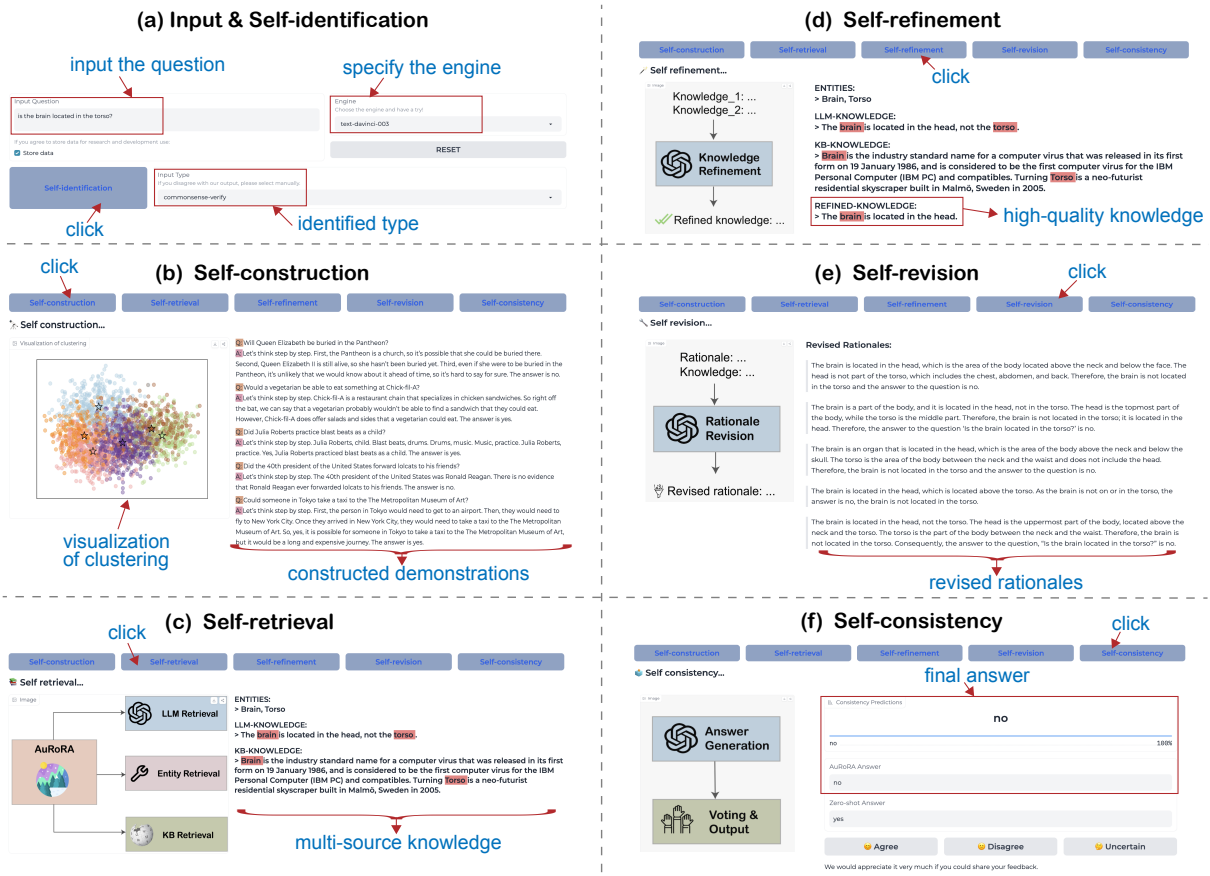

Figure 3: The website interface of AuRoRA platform.

Following Zhang et al. (2023c), we select k representative questions in DP_t. Afterward, we concatenate these k typical questions along with their rationales and answers acquired by Zero-shot-CoT (Kojima et al., 2023) and employ them as the ICL demonstrations for inference. To foster the divergent thinking pattern of LLMs, we borrow the idea from Wang et al. (2023) and sample a diverse set of candidate rationales: $\mathcal{R}_{ini} = \{r^1_{ini}, r^2_{ini}, \ldots, r^n_{ini}\}$, where n denotes the number of initial rationales.

Self-retrieval The goal of *self-retrieval* is to extract knowledge from varied sources related to q. The sources are generated by LLMs or retrieved from Wikipedia in our implementation. On one hand, a knowledge-guided trigger for question q is devised as: [$\texttt{Question}$: q; $\texttt{Prompt}_{\texttt{kn}}$], where $\texttt{Prompt}_{\texttt{kn}}$ stands for *"Necessary knowledge about the question by not answering the question:"*. The trigger is then fed into LLMs to acquire latent intrinsic knowledge $Know_{llm}$. On the other hand, we employ Wikepedia-API to search for the query entities and select top sentences from their Wikipedia pages. In this way, we access knowledge from external sources with higher factuality $Know_{kb}$.

Self-refinement With previously attained multi-source knowledge $\{Know_{llm}, Know_{kb}\}$, the objective of *self-refinement* is to generate brief and refined knowledge that balances factuality and relevance to the question. Concretely, we concatenate the diverse knowledge and ask LLMs to perform a refinement and verification: [$\texttt{Knowledge}_1$: $Know_{llm}$; $\texttt{Knowledge}_2$: $Know_{kb}$; $\texttt{Prompt}_{\texttt{ref}}$], where $\texttt{Prompt}_{\texttt{ref}}$ refers to *"By using Knowledge$_2$ to check Knowledge$_1$, output the brief and correct knowledge:"*. By this means, we attain the refined knowledge $Know_{ref}$.

Self-revision Now that we have a set of initial rationales $\mathcal{R}_{ini} = \{r^1_{ini}, r^2_{ini}, \ldots, r^n_{ini}\}$ and refined knowledge $Know_{ref}$, we revise the original rationales according to the high-quality knowledge. In detail, given $r^i_{ini} \in \mathcal{R}_{ini}$, we navigate LLMs to modify the antecedent rationales by simultaneously reviewing the question and knowledge: [$\texttt{Question}$: q; $\texttt{Knowledge}$: $Know_{ref}$; $\texttt{Original rationale}$: r^i_{ini}; $\texttt{Prompt}_{\texttt{rev}}$], where $\texttt{Prompt}_{\texttt{rev}}$ denotes *"With Knowledge given, output the revised rationale for Question in a precise and certain style by thinking step by step:"*. Hence, we manage to obtain a set of revised rationales $\mathcal{R}_{rev} = \{r^1_{rev}, r^2_{rev}, \ldots, r^n_{rev}\}$.

Table 1: Experimental results on different reasoning tasks.

Method	Arithmetic		Commonsense		Symbolic	
	MultiArith	SingleEq	CSQA	Strategy	Letter	Coin
Zero-shot	24.7	84.1	72.7	47.4	0.0	43.2
Zero-shot CoT	78.7	78.7	64.6	54.8	57.6	91.4
AuRoRA	91.7	92.9	75.4	66.0	60.7	99.1

Table 2: Typical methods of enhancing CoT prompting.

Method	Mixed Types	Automatic Prompt	Enhanced Knowledge	Revised Rationale	Consistent Answer
Auto-CoT (Zhang et al., 2023c)	✗	✓	✗	✗	✗
Self-Prompting (Li et al., 2023)	✗	✓	✗	✗	✗
CRITIC (Gou et al., 2023)	✗	✗	✓	✓	✗
Verify-and-Edit (Zhao et al., 2023)	✗	✗	✓	✓	✓
MCR (Yoran et al., 2023)	✗	✗	✓	✗	✓
Self-Consistency (Wang et al., 2023)	✗	✗	✗	✗	✓
GeM-CoT (Zou et al., 2024)	✓	✓	✗	✗	✗
AuRoRA (Ours)	✓	✓	✓	✓	✓

Self-consistency In the end, we continually adopt the self-consistency decoding strategy (Wang et al., 2023) and get a corresponding set of answers $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$. Then we take a majority vote over $\mathcal{A}$ and derive the final answer: $a_f = \arg\max_a \sum_{i=1}^{n} \mathbb{1}(a_i = a)$.

2.2. Website Interface

The online website of AuRoRA is illustrated in Figure 3. We provide an example of interacting with the system to get the corresponding responses.

Input Setting In the first place, the users enter the question in the input box on the top-left of the website and specify the engine of LLMs on the top-right. The users can check the `Store data` box if they agree to store their data for research and development use.

Interpretable Process After determining the input setting, the users sequentially click the buttons of *self-identification*, *self-construction*, *self-retrieval*, *self-refinement*, *self-revision*, and *self-consistency*. As is shown in Figure 3, all of the buttons are in **blue** with the corresponding label on them. The content of each step is thus displayed on the panel of the website, contributing to the interpretability of our system.

Output Feedback Eventually, clicking the *self-consistency* button presents the predicted answers by AuRoRA and the zero-shot baseline. Moreover, the users are free to share their feedback by clicking the buttons of *Agree*, *Disagree* and *Uncertain* at the bottom of the website.

3. Experiments

3.1. Setup

For LLM engines, we provide a series of currently available models of GPT-3 and GPT-3.5 from OpenAI API [1]. To validate the effectiveness of our system, we conduct experiments on six datasets including two arithmetic reasoning tasks (MultiArith (Roy and Roth, 2015), SingleEq (Koncel-Kedziorski et al., 2015)), two commonsense reasoning tasks (CSQA (Talmor et al., 2019), StrategyQA (Geva et al., 2021)),and two symbolic reasoning tasks (Last Letter Concatenation (Wei et al., 2023), Coin Flip (Wei et al., 2023)) Following the setting of Wei et al. (2023), the number of demonstrations k is 8 except for *symbolic-letter* (4), *commonsense-mc* (7) and *commonsense-verify* (6). The number of rationales is 5 considering the trade-off between performance and costs.

3.2. Main Results

Table 1 presents the results on six reasoning tasks, manifesting the consistent advancements of AuRoRA over the zero-shot baselines. Besides, since the goal of our work is to build a one-for-all system, we compare our proposed AuRoRA with the prevailing typical CoT methods, as demonstrated in Table 2. Notably, our system distinguishes itself for its remarkable versatility. It not only excels in scenarios involving mixed question types and obviates the need for manual demonstration crafting, but it also undergoes a multi-stage process, thereby fertilizing our integrated system from manifold angles.

[1] https://openai.com/blog/openai-api

Table 3: An example from StrategyQA.

Type	Content
Q	Is pickled cucumber ever red?
Original CoT	Pickled cucumber is usually green. Therefore, pickled cucumber is rarely red. The answer is **no**.
Self knowledge	Pickled cucumbers can be made in a variety of ways, and the color of the cucumbers can vary depending on the ingredients used in the pickling process.
KB knowledge	[...] The flesh of Kool-Aid pickles typically take on a pink or red color.
Refined knowledge	The ingredients used to pickled cucumbers will make them different colors, for example Kool-Aid pickles are pink or red.
Corrected CoT	Depending on the ingredients used in the vinegar solution, such as herbs and spices, the pickled cucumbers can take on a red color. So the answer is **yes**.
Answer	**Yes**(Gold) || **Yes**(AuRoRA) || **No**(Zero-shot Baseline)

Table 4: Accuracy of different system settings on ARC-c datasets.

System	Accuracy(%)
zero-shot	51.4
w/ maths setting	70.0
w/ self-identification(**ours**)	77.6

Table 5: Contributions of each process.

Feature	Constributed Process
Adaptability (input)	- Self-identification - Self-construction
Reliability (process)	- Self-retrieval - Self-refinement - Self-revision
Reliability (output)	- Self-consistency
Integrity	- All six steps
Interpretability	- All six steps

4. Analysis

4.1. Interpretability: Case Study

In order to further demonstrate the superiority and interpretability of our system, we provide a case study from StrategyQA, which is shown in Table 3. From this example, our system successfully corrects the factual errors entailed in baseline methods by retrieving knowledge from multiple sources and conducting refinement upon it. Hence, our system enjoys superior robustness to various input questions than baselines.

4.2. Generalization of our system

We conduct experiments to test generalization of our system. We test a mathematical reasoning system on ARC-c dataset. We implement a mathematical variant of AuRoRA by replacing the *self-identification* phase with direct mathematical demonstrations. As shown in Table 4, even with *missing* configurations, our system runs far better than the baseline (51.4 $\rightarrow$ 70.0), which demonstrates the favorable generalization capability and robustness of our system.

4.3. Contributions of each process

Table 5 elucidates the contributions of each process in detail, demonstrating that each processes plays a vital role in building the proposed one-for-all AuRoRA platform.

5. Related Works

Recent works have promoted the CoT method in three main aspects, namely sample selection, reasoning enhancement and decoding strategy. For sample selection, Auto-CoT (Zhang et al., 2023c) and Self-Prompting (Li et al., 2023) guided LLMs to perform reasoning by automatically constructing diverse and task-adapted samples. For reasoning enhancement, methods such as CRITIC (Gou et al., 2023), Verify-and-Edit (Zhao et al., 2023) and MCR (Yoran et al., 2023) corrected the rationales by interacting with external knowledge. For decoding strategy, self-consistency (Wang et al., 2023) generated diverse reasoning paths for aggregation.

6. Conclusion

In this work, we propose AuRoRA, a one-for-all platform for augmented reasoning and refining based on task-adaptive CoT prompting. The core objective is to provide a universal reasoning system featuring adaptability, reliability, integrity and interpretability. Moreover, AuRoRA simultaneously offers user-friendly operation and real-time visual analysis from the website interface. Our proposed AuRoRA channels the CoT prompting method towards an integrated and applicable pattern from a bigger picture.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *ArXiv preprint*, abs/2204.02311.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *ArXiv preprint*, abs/2305.11738.

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.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213.

Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. 2015. Parsing algebraic word problems into equations. *Transactions of the Association for Computational Linguistics*, 3:585–597.

Junlong Li, Zhuosheng Zhang, and Hai Zhao. 2023. Self-prompting large language models for zero-shot open-domain qa. *ArXiv preprint*, abs/2212.08635.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. *ArXiv preprint*, abs/2305.14552.

OpenAI. 2023. Gpt-4 technical report. *ArXiv preprint*, abs/2303.08774.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *ArXiv preprint*, abs/2302.06476.

Subhro Roy and Dan Roth. 2015. Solving general arithmetic word problems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1743–1752, Lisbon, Portugal. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv preprint*, abs/2211.05100.

Chenglei Si, Weijia Shi, Chen Zhao, Luke Zettlemoyer, and Jordan Boyd-Graber. 2023. Getting more out of mixture of language model reasoning experts. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8234–8249.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *ArXiv preprint*, abs/2201.08239.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric

Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Ori Yoran, Tomer Wolfson, Ben Bogin, Uri Katz, Daniel Deutch, and Jonathan Berant. 2023. Answering questions by meta-reasoning over multiple chains of thought. *ArXiv preprint*, abs/2304.13007.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023a. How language model hallucinations can snowball. *ArXiv preprint*, abs/2305.13534.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023b. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023c. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. *ArXiv preprint*, abs/2305.03268.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, et al. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

Zhaocheng Zhu, Yuan Xue, Xinyun Chen, Denny Zhou, Jian Tang, Dale Schuurmans, and Hanjun Dai. 2023. Large language models can learn rules. *arXiv preprint arXiv:2310.07064*.

Anni Zou, Zhuosheng Zhang, Hai Zhao, and Xiangru Tang. 2024. Generalizable chain-of-thought prompting in mixed-task scenarios with large language models.