ANALYSIS OF "CHAIN-OF-THOUGHT PROMPTING ELICITS REASONING IN LARGE LANGUAGE MODELS"

Ali Khalaji
ali.khalaji@stud.tu-darmstadt.de
February 13, 2024

Abstract

In recent years, the reasoning capabilities of language models have been significantly improved. Notably, the use of exemplars in few-shot prompting has led to substantial progress. However, this method demands additional human annotations, resulting in extended and costly instruction-tuning for language models. Despite previous attempts at prompt design, several reasoning tasks show limited improvement, represented by a flat scaling curve. In this review, I provide an analysis of a novel approach called “chain-of-thought prompting”. The innovative technique aims to prompt language models with only a few exemplars, including intermediate steps of reasoning, following human problem-solving methods, producing superior outcomes in arithmetic, commonsense, and symbolic reasoning tasks compared to standard prompting, and achieving state-of-the-art reasoning accuracy. The findings of the paper demonstrate that chain-of-thought prompting significantly improves the performance of language models across various reasoning tasks, particularly when the model is scaled and has a minimum of 100 billion parameters.

1 Introduction

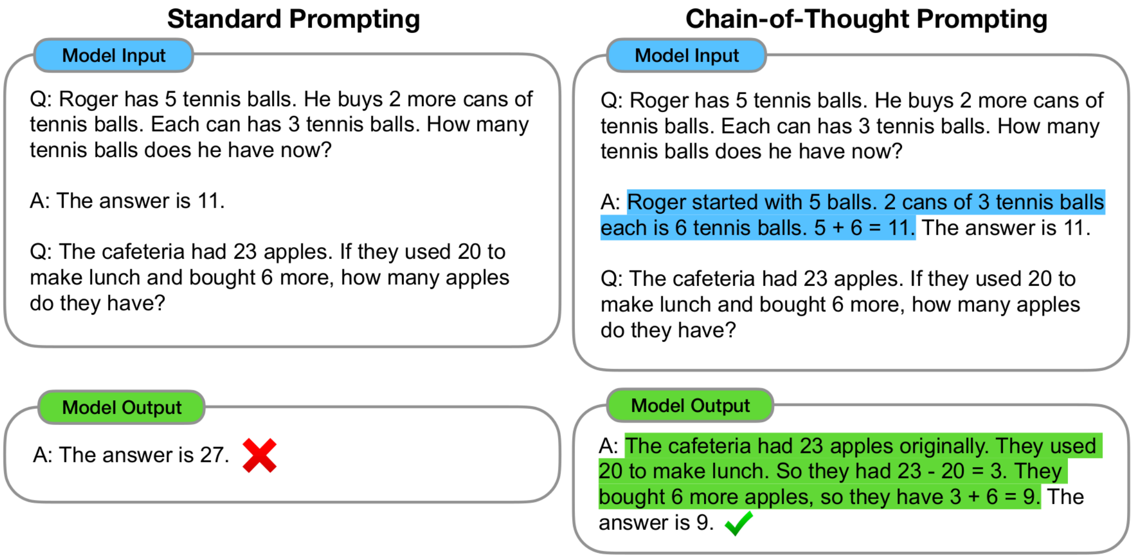
Large language models (LLMs) have gained considerable attention in recent years due to their central role in solving complex reasoning tasks. They have a number of advantages, including accuracy, speed, and flexibility. These models are trained by being exposed to large amounts of text data (Chang et al., 2023). In essence, language models learn from a wide variety of written information, enabling them to understand and process human language for a wide range of applications (Radford et al., 2019). Despite their strengths, LLMs may encounter challenges in solving math problems or answering factual questions which can result in incorrect or misleading answers.

Scaling models has been demonstrated to be beneficial (Rosenfeld et al., 2019), leading to improved performance and efficiency, as shown by Kaplan et al. (2020). However, simply increasing model size alone is insufficient to achieve desired results in challenging reasoning tasks. Addressing these challenges demands additional considerations beyond only size scaling to attain the desired level of performance (Wei et al., 2022).

Another property of LLMs is to prompt the model with examples, as stated by Brown et al. (2020), rather than relying on fine-tuning. Few-shot learning has demonstrated promising results for a variety of simple question-answering tasks. Nevertheless, it tends to underperform on tasks requiring complex reasoning and often shows limited improvement as the scale of the language model increases (McKenzie et al., 2023). To address this limitation, consider providing an explanation on how to solve the problem step-by-step within example prompts, encouraging the model to apply this method to similar problems. This approach, referred to as “chain-of-thought prompting”, is proposed by Wei et al. (2023). Before moving on to the chain-of-thought method and discussing it in detail, we will explore related works that share some similar ideas in enhancing the reasoning ability of language models.

2 Related Work

One prior direction of research that inspired chain-of-thought prompting is using intermediate steps. Reynolds and McDonell (2021) discussed methods of prompt programming and augmented few-shot exemplars with intermediate steps through natural language to break down a problem into components before making a decision. Nye et al. (2021) trained transformers to perform multi-step computations by encoding intermediate steps into a temporal buffer which can be used as long as needed. Another significant contribution in this field was



Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

A: The answer is 27. ✗

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

Figure 1: Chain-of-thought prompting vs. Standard prompting

presented by Zhou et al. (2023). This work involved breaking down complex problems into a series of simpler subproblems and utilizing the answer for each subproblem in a nested manner.

Leveraging the advantages of decomposing complex problems and combining them with the rationalization of language models has become increasingly popular in recent years. For solving question-answer pairs in math word problems, Ling et al. (2017) introduced a step-by-step explanation approach to how the calculations result in correct answers. Rajani et al. (2019) trained a language model to generate explanations, that are used by a classifier to make predictions. This approach was also adopted by Cobbe et al. (2021), who trained an additional verifier to rerank the generated solutions and select the highest-ranked ones.

In the next section, we will explore how the proposed chain-of-thought prompting method works. Despite some prior research, according to the results of Wei et al. (2023), it can unlock various natural language reasoning abilities in off-the-shelf language models of sufficient scale through simple prompting without extensive labeled annotations.

3 Chain-of-thought Method

Figure 1 demonstrates how chain-of-thought prompting works. In comparison with standard prompting (Brown et al., 2020), the model is fed with an example prompt as a triple including example question, chain-of-thought, and task statement. The provided chain-of-thought explains to the language model step-by-step how to deal with the task in the way that normally human solve such a problem. The methodology can be defined as follows:

- Guiding a language model to rationalize its reasoning by breaking down complex problems into smaller, more manageable steps.
- Solving each step before proceeding to the next.
- Using the result of each step for the subsequent steps before arriving at the final answer.

The objective of prompt design is to identify the optimal prompt, allowing the language model to effectively solve the given task in the most effective manner (Liu et al., 2021). Emphasizing a prompting-only approach is crucial because it eliminates the need for a vast training dataset, enabling a model to handle various tasks without sacrificing generality. Furthermore, as noted by Zhao et al. (2023) we can perform instruction tuning on LLMs with task descriptions expressed in natural language. The language-oriented nature of chain-of-thought prompting makes it versatile, and applicable to any task that can be solved through human language, especially those that can be broken down into multiple steps. We will further explore this aspect in the next section.

4 Experiment and Results

To assess the effectiveness of chain-of-thought prompting, the researchers conducted evaluations using various benchmarks, including arithmetic reasoning, commonsense reasoning, and symbolic reasoning. The assessment involved five language models, each with distinct architectures and parameter sizes:

- **GPT-3** (Brown et al., 2020) with parameter configurations of 350M, 1.3B, 6.7B, and 175B.
- **LamDA** (Thoppilan et al., 2022) with parameter sizes of 422M, 2B, 8B, 68B, and 137B.
- **PaLM** (Chowdhery et al., 2022) with parameters ranging from 8B to 540B.
- **UL2** (Tay et al., 2023) with 20B parameters.
- **Codex** (Ye et al., 2023) with 12M parameters.

In my opinion, assessing the method across different language models with varying scaling factors is valuable. Each language model comes with its distinct strengths and weaknesses, and such evaluations can provide insights into how scaling factors influence its overall performance.

4.1 Arithmetic Reasoning

Arithmetic reasoning tasks present a challenge for language models in problem-solving. Due to the complexity of numerical relationships and logical patterns, LMs may struggle to solve such tasks accurately. Chain-of-thought prompting has been explored for the named language models on the following arithmetic reasoning benchmarks:

GSM8K (Cobbe et al., 2021), **SVAMP** (Patel et al., 2021), **ASDiv** (Miao et al., 2020), **AQuA** (Ling et al., 2017), **MAWPS** (Koncel-Kedziorski et al., 2016).

The baseline employed in the experiment relies on standard prompting from Brown et al. (2020), augmented by the use of a chain-of-thought approach that explains how to think about the question asked, presented in the form of a rationale. To apply the chain of thought, the authors provide the model with a set of eight manually generated chains of thought. Figure 2 shows the results obtained using chain-of-thought prompting on arithmetic reasoning benchmarks.

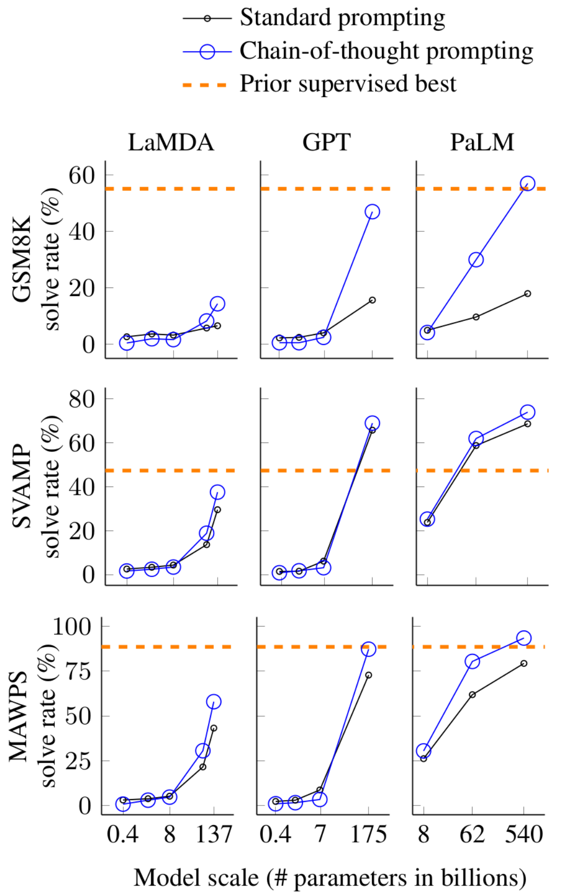
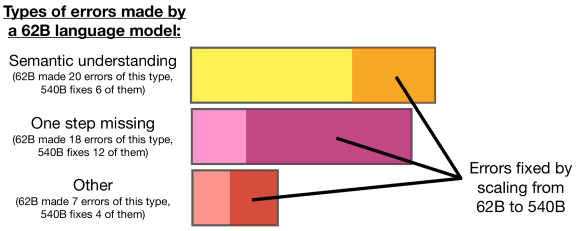


Figure 2: Chain-of-thought prompting vs. Standard prompting on math problems

According to Wei et al. (2023), as can be seen in every language model tested, the greater the number of parameters a model has, the more substantial the gains achieved by using the method. Specifically, the method performs better on benchmarks where the standard query did not give satisfactory results due to the complexity of the problem.

One notable advantage conferred by the utilization of chain-of-thought prompting is the ability to cor-



Types of errors made by a 62B language model:
Semantic understanding
(62B made 20 errors of this type, 540B fixes 6 of them)
One step missing
(62B made 18 errors of this type, 540B fixes 12 of them)
Other
(62B made 7 errors of this type, 540B fixes 4 of them)
Errors fixed by scaling from 62B to 540B

Figure 3: Errors caused by using PaLM 6bB can be corrected by scaling up to 540B.

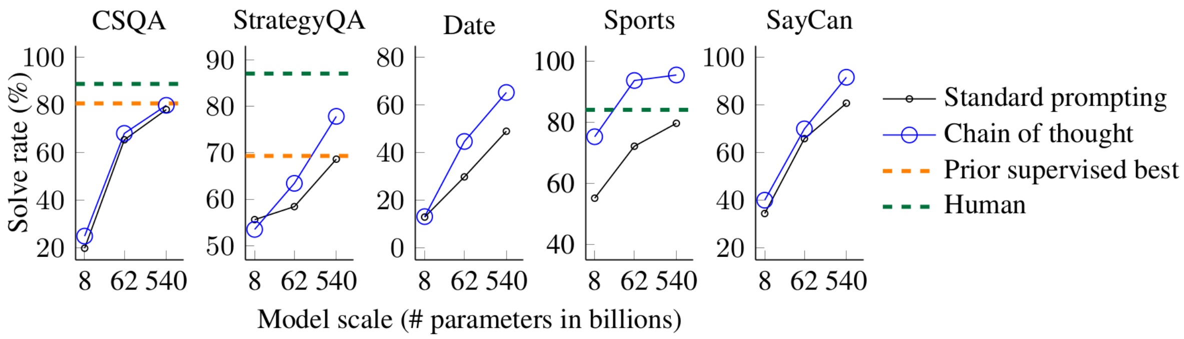


Figure 4: Chain-of-thought prompting for commonsense reasoning tasks.

rect errors when using larger language models.

Errors encountered in an experiment were classified as semantic understanding errors, one-step missing errors, and others. Figure 3 shows that scaling PaLM to 540B parameters successfully addressed a significant proportion of errors in all three categories.

The results suggest that as language models increase in scale, they acquire a wider range of semantic and logical skills (Li et al., 2020). The success of chain-of-thought reasoning is influenced by factors such as model size and training computation and involves abilities such as semantic understanding, symbol mapping, topic coherence, arithmetic skills, and faithfulness (Wei et al., 2023).

4.2 Commonsense Reasoning

In this section, I will delve into the results of chain-of-thought prompting on benchmarks related to commonsense reasoning. As defined by Kuo and Chen (2023), commonsense reasoning involves making assumptions about everyday situations and their fundamental nature. To explore various types of commonsense reasoning tasks, the authors applied the method to the following benchmarks:

CSQA (Talmor et al., 2019), **StrategyQA** (Geva et al., 2021), **Date** and **Sport** (Srivastava et al., 2023), and **SayCan** (Ahn et al., 2022).

The results in Figure 4 show that chain-of-thought prompting is a scale-dependent emergent ability, showing improved performance with sufficiently large models. It also tends to suppress the performance obtained with standard prompting, particularly for general background knowledge questions. I will discuss further investigation of the effect of the scaling factor on chain-of-thought prompting in more detail in the Discussion section.

4.3 Symbolic Reasoning

The next experiment conducted in the paper involved chain-of-thought prompting across two symbolic reasoning tasks: *Last Letter Concatenation* and *Coin Flip*. Since the standard prompting has nearly achieved a 100% success rate on the in-domain evaluation set, I will now go on to the out-of-domain (OOD) test set.

As highlighted by Zhu et al. (2022), the OOD test set comprises data distinct from the training data and represents a different distribution. This set evaluates the ability of a machine learning model to generalize beyond its trained domains. The impact of chain-of-thought is substantial when applied to the OOD test set. As shown in Table 1, the model can effectively generalize tasks beyond the training data if it is appropriately scaled.

4.4 Ablation Study

To investigate the reasons for the improved performance of language models in reasoning, the authors conducted an ablation study. They presented alternative variations of chain-of-thought, systematically removing components to comprehensively analyze their impact on the overall improvement observed.

Equation only, The model outputs only a mathematical equation before the answer has been provided, which did not significantly help for GSM8K. This suggests that GSM8K's complex questions require natural language reasoning steps, emphasizing the importance of contextual understanding in certain scenarios. The performance under the Equation Only condition falls between that of stan-

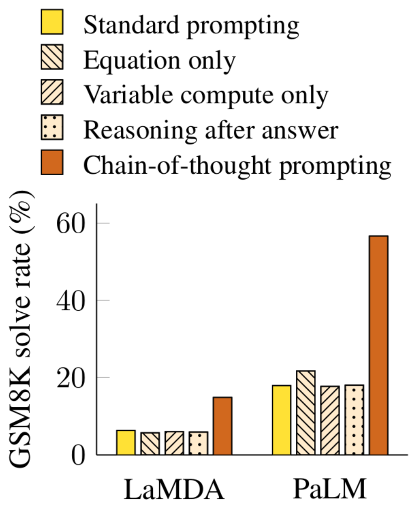


Figure 5: Examination of diverse prompting variations through an ablation study.

dard prompting and chain-of-thought prompting.

Chan-of-thought after answer, Authors conducted a test to examine if the chain-of-thought enables the model to access prior knowledge. However, as depicted in Figure 5, this is not supported by the results. Nonetheless, the sequential reasoning facilitated by chain-of-thought prompting proves to be more valuable than simply triggering prior knowledge.

Robustness Check To demonstrate the robustness of the method with various prompts, the authors utilized three sets of eight prompts, each generated by different annotators. They also included three sets of eight example prompts randomly drawn from GSM8K. The results, as shown in Figure 6, indicate that, despite expected variations, consistent with Le Scao and Rush (2021), the chain of thought consistently outperforms standard prompts across all sets of examples in arithmetic, commonsense and symbolic reasoning

It should be noted that the prompts in the GSM8K dataset (Cobbe et al., 2021) have been written by crowd workers without a machine learning background. Furthermore, in almost all cases, the order of exemplar prompts does not negatively affect the performance of chain-of-thought prompting. However, an examination of the results from the robustness analysis in Table 2 reveals an exception in symbolic reasoning, attributed to the nature of this classification task. It can be argued, according to Zhao et al. (2021) few-shot learning is unstable when using different prompts due to biases in language models. These models tend to rely heavily on the most recent tokens in the input sequence.

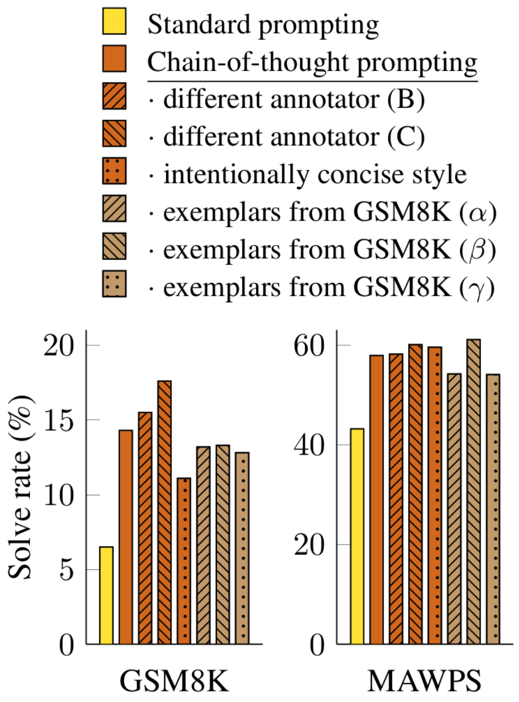


Figure 6: Robustness analysis: chain-of-thought prompting with diverse arithmetic prompts.

5 Analysis

In the previous section, we examined the effects of chain-of-thought prompting on few-shot prompting. This resulted in a significant improvement in the reasoning ability of language models with at least 100B parameters. Additionally, the ability of the method to correct reasoning errors in large-scale models is a crucial aspect. As highlighted by Wei et al. (2023), this approach holds promise for explainability, as the generated explanation can guide the model in demonstrating how to arrive at a specific answer. However it is not faithful as newly investigated by Turpin et al. (2023), chain-of-thought prompting can also systematically introduce biases to the model during the inference process. Regulating the process by which explanations are generated is necessary to improve their faithfulness and ensure avoidance of motivated reasoning (Turpin et al., 2023).

In my personal assessment, while the method's robustness was assessed with diverse sets of prompts, whether human-generated or randomly derived from benchmark training sets, exploring the metrics for selecting appropriate prompts and guiding the language model during reasoning remains an ongoing area for investigation. Notably, for smaller-scale models, the use of an illogical Chain of Thought (CoT) can lead to lower performance in comparison to standard prompting.

5.1 Limitations

Chain-of-thought prompting can be less effective for models with fewer than 100B parameters, making it more suitable for larger models in practical applications. Smaller models often struggle, showing problems such as repetition of information, incomplete logic, or lack of semantic understanding (Brown et al., 2020), making it harder to provide clear and understandable answers. The size of a model is influenced by factors such as the amount of training, the pre-training data, and the model structure chosen (Wei et al., 2023). Furthermore, increasing the size of language models is computationally demanding and presents substantial hardware challenges. Therefore, alternative approaches are expected to play a central role in shaping the future of emergent abilities in large language models (Wei et al., 2022). The additional overheads of chain-of-thought prompting are the increased computational costs associated with decomposing the problems and manually composed prompts.

5.2 Future Work

While this paper has provided insights into improving the reasoning ability of large language models, according to my observations, the field is rapidly evolving. Future research should delve deeper into how a language model can automatically generate reasoning chains and potentially optimize this over a validation set, as proposed by Zhang et al. (2022). Additionally, there is a need to explore how reasoning can be induced in smaller models. Addressing potential mistakes raised by chain-of-thoughts is crucial, as newly (Yao et al., 2023) augment chain-of-thought with integrated memory to record the history of reasoning paths, allowing for backtracking in the event of mistakes. However, I think the computational overheads of the backtracking process must be considered. Furthermore, investigating iterative prompting frameworks to elicit reasoning in a step-by-step manner would be a valuable avenue for future inquiry.

6 Conclusion

In this paper analysis, I discussed the results, benefits, and limitations of chain-of-thought prompting method. We observed that chain-of-thought prompting is an emergent capability of model scale, and performance gains are achieved when used with models of 100B parameters. It shows promise for improving the reasoning capabilities of large language models and, by eliciting intermediate steps of reasoning, making them better suited to complex, multi-step problems. Using chain-of-thought prompting can be beneficial when a task is challenging and requires multi-step reasoning, especially when the scaling curve of the model shows a relatively flat trend.

Use of AI writing Assistances

To improve the overall quality of my paper analysis, I used AI writing tools for language checking. These tools fixed any grammar mistakes and spelling helped me find appropriate words that precisely expressed my viewpoints and ensured that the paper was well-organized. They also helped me to follow the rules of academic writing and made my work look more professional. With the help of AI tools, my analysis became clearer and more suitable for academic readers.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. 2022. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. ArXiv:2204.01691 [cs].

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. ArXiv:2005.14165 [cs].

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A Survey on Evaluation of Large Language Models. ArXiv:2307.03109 [cs].

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. ArXiv:2204.02311 [cs].

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. ArXiv:2110.14168 [cs].

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. ArXiv:2101.02235 [cs].

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. ArXiv:2001.08361 [cs, stat].

Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. 2016. MAWPS: A Math Word Problem Repository. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1152–1157, San Diego, California. Association for Computational Linguistics.

Hui-Chi Kuo and Yun-Nung Chen. 2023. Zero-Shot Prompting for Implicit Intent Prediction and Recommendation with Commonsense Reasoning. ArXiv:2210.05901 [cs].

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online. Association for Computational Linguistics.

Zhuohan Li, Eric Wallace, Sheng Shen, Kevin Lin, Kurt Keutzer, Dan Klein, and Joseph E. Gonzalez. 2020. Train Large, Then Compress: Rethinking Model Size for Efficient Training and Inference of Transformers. ArXiv:2002.11794 [cs].

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program Induction by Rationale Generation : Learning to Solve and Explain Algebraic Word Problems. ArXiv:1705.04146 [cs].

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pretrain, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. ArXiv:2107.13586 [cs].

Ian R. McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, Andrew Gritsevskiy, Daniel Wurgaft, Derik Kauffman, Gabriel Recchia, Jiacheng Liu, Joe Cavanagh, Max Weiss, Sicong Huang, The Floating Droid, Tom Tseng, Tomasz Korbak, Xudong Shen, Yuhui Zhang, Zhengping Zhou, Najoung Kim, Samuel R. Bowman, and Ethan Perez. 2023. Inverse Scaling: When Bigger Isn't Better. ArXiv:2306.09479 [cs].

Shen-yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2020. A Diverse Corpus for Evaluating and Developing English Math Word Problem Solvers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 975–984, Online. Association for Computational Linguistics.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2021. Show Your Work: Scratchpads for Intermediate Computation with Language Models. ArXiv:2112.00114 [cs].

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP Models really able to Solve Simple Math Word Problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. ArXiv:1906.02361 [cs].

Laria Reynolds and Kyle McDonell. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. ArXiv:2102.07350 [cs].

Jonathan S. Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. 2019. A Constructive Prediction of the Generalization Error Across Scales. ArXiv:1909.12673 [cs, stat].

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, et al. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. ArXiv:2206.04615 [cs, stat].

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*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. UL2: Unifying Language Learning Paradigms. ArXiv:2205.05131 [cs].

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. LaMDA: Language Models for Dialog Applications. ArXiv:2201.08239 [cs].

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. ArXiv:2305.04388 [cs].

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. ArXiv:2206.07682 [cs].

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. ArXiv:2201.11903 [cs].

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. ArXiv:2305.10601 [cs].

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models. ArXiv:2303.10420 [cs].

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic Chain of Thought Prompting in Large Language Models. ArXiv:2210.03493 [cs].

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-Shot Performance of Language Models. ArXiv:2102.09690 [cs].

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. ArXiv:2303.18223 [cs].

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. ArXiv:2205.10625 [cs].

Zining Zhu, Soroosh Shahtalebi, and Frank Rudzicz. 2022. OOD-Probe: A Neural Interpretation of Out-of-Domain Generalization. ArXiv:2208.12352 [cs].

Appendix

		Last Letter Concatenation						Coin Flip (state tracking)					
		2		OOD: 3		OOD: 4		2		OOD: 3		OOD: 4	
Model		standard	CoT	standard	CoT	standard	CoT	standard	CoT	standard	CoT	standard	CoT
UL2	20B	0.6	**18.8**	0.0	0.2	0.0	0.0	70.4	67.1	51.6	52.2	48.7	50.4
LaMDA	420M	0.3	**1.6**	0.0	0.0	0.0	0.0	52.9	49.6	50.0	50.5	49.5	49.1
	2B	2.3	**6.0**	0.0	0.0	0.0	0.0	54.9	**55.3**	47.4	48.7	49.8	50.2
	8B	1.5	**11.5**	0.0	0.0	0.0	0.0	52.9	**55.5**	48.2	49.6	51.2	50.6
	68B	4.4	**52.0**	0.0	**0.8**	0.0	**2.5**	56.2	**83.2**	50.4	**69.1**	50.9	**59.6**
	137B	5.8	**77.5**	0.0	**34.4**	0.0	**13.5**	49.0	**99.6**	50.7	**91.0**	49.1	**74.5**
PaLM	8B	2.6	**18.8**	0.0	0.0	0.0	**0.2**	60.0	**74.4**	47.3	**57.1**	50.9	**51.8**
	62B	6.8	**85.0**	0.0	**59.6**	0.0	**13.4**	91.4	**96.8**	43.9	**91.0**	38.3	**72.4**
	540B	7.6	**99.4**	0.2	**94.8**	0.0	**63.0**	98.1	**100.0**	49.3	**98.6**	54.8	**90.2**

Table 1: Standard vs. chain-of-thought prompts aid longer inference in two symbolic tasks.

	Commonsense			Symbolic	
	Date	Sports	SayCan	Concat	Coin
Standard prompting	21.5 ±0.6	59.5 ±3.0	80.8 ±1.8	5.8 ±0.6	49.0 ±2.1
Chain of thought prompting	26.8 ±2.1	85.8 ±1.8	91.7 ±1.4	77.5 ±3.8	99.6 ±0.3
Ablations					
· variable compute only	21.3 ±0.7	61.6 ±2.2	74.2 ±2.3	7.2 ±1.6	50.7 ±0.7
· reasoning after answer	20.9 ±1.0	63.0 ±2.0	83.3 ±0.6	0.0 ±0.0	50.2 ±0.5
Robustness					
· different annotator (B)	27.4 ±1.7	75.4 ±2.7	88.3 ±1.4	76.0 ±1.9	77.5 ±7.9
· different annotator (C)	25.5 ±2.5	81.1 ±3.6	85.0 ±1.8	68.1 ±2.2	71.4 ±11.1

Table 2: Results of ablation and robustness study for commonsense and symbolic reasoning datasets.