Chain-of-Thought Prompting May Be Harmful for Medium-Sized Instruction-Tuned LMs on Commonsense QA Tasks

Alberto Mario Ceballos Arroyo, Hamza Tahboub, Byron Wallace, Huaizu Jiang
Northeastern University

Abstract

Chain-of-thought (CoT) prompting has been proposed as a means of improving the performance of language models on a variety of tasks. Eliciting such "reasoning" has been shown to sometimes yield dramatic empirical gains on downstream tasks for massive language models (60B+ parameters). Recent work has shown that CoT reasoning can be induced in smaller language models (~1-30B parameter, which we will refer to as "medium"-sized) via instruction fine-tuning. However, while one can indeed elicit reasoning sequences from such models via the "let's think step-by-step" incantation, it is not clear that doing so consistently improves downstream task performance. In this work, we evaluate the relative performance of five medium-sized language models on five commonsense question-answering datasets (which span multiple domains). Perhaps surprisingly, we find that CoT prompting degrades performance by an average of 5.7% across these models and tasks. This degradation is consistent; for example, CoT is harmful for Flan-T5 (3B and 11B) across all commonsense-based datasets considered. Indeed, for 22 out of 25 model/dataset pairs evaluated, CoT prompting yields worse results than "direct" answering. Our results suggest that while CoT prompting of medium-sized LMs has shown promising results in some scenarios, one should adopt this approach with caution, especially for commonsense reasoning tasks. We hope that this work fosters future research into *why* CoT appears harmful in some cases, and in turn leads to efforts to improve the zero-shot capabilities of small-to-medium LMs.

1 Introduction

Chain-of-thought (CoT) prompting has been found to be an effective technique for improving the performance of large language models (LMs) by inducing intermediate "reasoning"-like behavior in outputs. Across multiple large LMs, CoT has been shown to improve performance on multi-step tasks such as commonsense question answering (QA) and arithmetic reasoning (Kojima et al., 2022; Wei et al., 2022c).

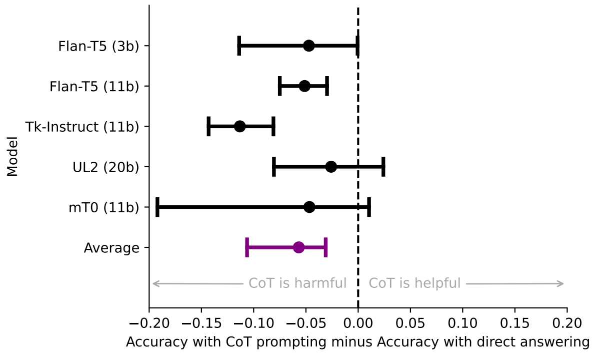


Figure 1: **Differences in accuracies between CoT prompting and direct answering for commonsense reasoning** (0 means no difference; negative values indicate that CoT prompting performs worse). We report averages (circles) and ranges (bars) across five datasets for each of five "modestly-sized" models. We also report an overall average across models. CoT prompting yields consistently worse performance across all models and datasets considered.

Most work on CoT has involved very *large* LMs (>60B parameters). The ability to "reason"—or at least to output strings indicative of doing so—appears to emerge naturally in large LMs when they are pre-trained over sufficiently large corpora with self-supervised objectives (Wei et al., 2022b,c). CoT outputs can then be elicited from such massive models via prompts crafted to do so, for example, by containing "let's think step-by-step", as shown in Fig. 2.

Unfortunately, CoT reasoning capabilities (and the associated potential of improved task performance) do not seem to emerge naturally in smaller LMs, which are far more computationally efficient (and so, more accessible). Indeed, past research has suggested that attempting to induce CoT reasoning from models containing fewer than 10B parameters can harm their performance (Magister

Model	CoT Acc.	Direct Acc.	CoT Same	CoT Better	CoT Worse	CoT Δ
Flan-T5 (3B)	81.05%	85.76%	87.20%	4.03%	8.75%	-4.71%
Flan-T5 (11B)	81.09%	86.21%	86.48%	4.19%	9.31%	-5.12%
T*k*-INSTRUCT (11B)	52.86%	64.18%	63.90%	12.38%	23.70%	-11.32%
UL2 (20B)	38.86%	41.44%	64.86%	16.27%	18.86%	-2.59%
mT0 (11B)	63.67%	68.33%	82.64%	6.34%	11.00%	-4.66%

Table 1: **How often CoT prompting affects the accuracy of different models compared with direct answering**: no effect (CoT Same), improves the accuracy (CoT Better), or harms the accuracy (CoT Worse). We also report the difference of accuracies when using CoT instead of answering directly. We present an average over 5 datasets: CommonsenseQA, SocialIQA, PIQA, HellaSwag, and CosmosQA. The full table is provided in the Appendix.

et al., 2022; Wei et al., 2022c). Concurrently, recent work has demonstrated the potential of *instruction fine-tuning* (Wei et al., 2022a), which entails fine-tuning models on a collection of supervised corpora with tasks phrased as instructions, enables small-to-moderate LMs to achieve performance competitive with much larger models (Chung et al., 2022). Moreover, this supervision paradigm appears to “unlock” zero-shot CoT capabilities in smaller models to some extent.

However, it remains unclear whether CoT prompting actually improves the performance of small-to-medium LMs on downstream tasks. For example, close inspection of results reported in existing work reveals that smaller ($\leq$11B parameters) variants of Flan-T5, an instruction-tuned variant of T5 (Raffel et al., 2020), perform *consistently worse* under CoT prompting rather than eliciting answers directly (Table 5 in Chung et al. 2022).

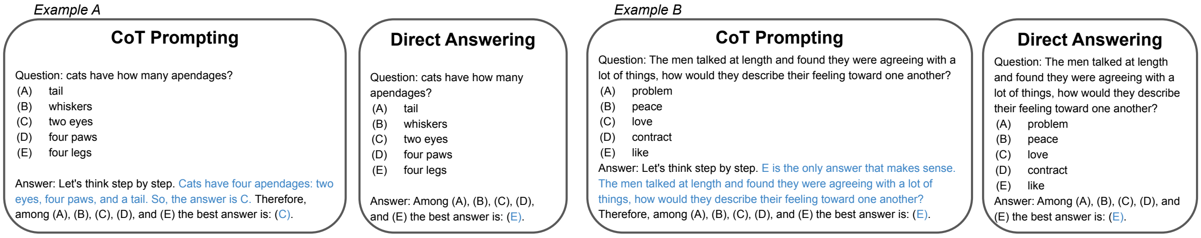
In this work we seek to evaluate whether CoT prompting improves the performance of instruction-tuned small-to-medium LMs. We focus on this model class because such models appear to provide *accessible* SOTA or near SOTA performance across many NLP tasks (at least in zero- and few-shot settings) (Chung et al., 2022). We consider two task types where CoT has been shown to benefit large LMs: Commonsense QA and algebraic reasoning problems. Both naturally lend themselves to step-by-step reasoning and have been used in prior CoT work. We run examples from datasets representing these tasks through five modestly sized (instruction-tuned) LMs in a zero-shot setting, and evaluate outputs produced with and without the “let’s think step-by-step” prompt. Statistics related to the datasets are given in Table 2.

We find that for *all* commonsense reasoning tasks—i.e., all datasets other than the one related to algebraic reasoning—-CoT prompting performs worse than direct answering (by an average of 5.7 points in absolute performance). Moreover, manual error analysis demonstrates that one of the main bottleneck of using CoT prompting is producing correct reasoning steps (rationales). Our results suggest that while CoT prompting of modestly-sized LMs has shown promising results in some scenarios, one should adopt this approach with caution, especially for commonsense reasoning tasks. We hope that this work motivates research into *why* CoT appears harmful in some cases, and in turn leads to efforts to improve the zero-shot capabilities of small-to-medium LMs.

2 Related work

Instruction and multi-task fine-tuned models. Providing language instructions to models has proven helpful for NLP tasks. Specifically, recent work has proposed training LMs on datasets comprising a multitude of supervised tasks with associated instructions to enable models to generalize better when provided instructions for unseen tasks.

Muennighoff et al. (2022) fine-tuned mT5 on xP3, a collection of 46 datasets encompassing various tasks and languages. Wang et al. (2022b) trained T*k*-INSTRUCT, a model based on T5 (Raffel et al., 2020) and fine-tuned on a meta dataset encompassing more than 1,600 tasks with a handful of positive and negative samples each. Chung et al. (2022) released versions of T5 and PaLM fine-tuned on instruction versions of more than 1,800 tasks, including 9 tasks with rationales, resulting in significantly improved generalization capabilities over the original models, while Iyer et al. (2022) fine-tuned the 30 and 175 billion parameter versions of OPT on 2,000 tasks by providing specific instructions and, in some cases, chains of reasoning. In the aforementioned work, scaling up to thousands of tasks proved critical in improving the generalization capabilities of such models and contributed to unlocking the capability to do CoT prompting.


Example A
CoT Prompting
Direct Answering
Example B
CoT Prompting
Direct Answering

Question: cats have how many apendages?
(A) tail
(B) whiskers
(C) two eyes
(D) four paws
(E) four legs

Answer: Let's think step by step. Cats have four apendages: two eyes, four paws, and a tail. So, the answer is C. Therefore, among (A), (B), (C), (D), and (E) the best answer is: (C).

Question: cats have how many apendages?
(A) tail
(B) whiskers
(C) two eyes
(D) four paws
(E) four legs

Answer: Among (A), (B), (C), (D), and (E) the best answer is: (E).

Question: The men talked at length and found they were agreeing with a lot of things, how would they describe their feeling toward one another?
(A) problem
(B) peace
(C) love
(D) contract
(E) like
Answer: Let's think step by step. E is the only answer that makes sense. The men talked at length and found they were agreeing with a lot of things, how would they describe their feeling toward one another? Therefore, among (A), (B), (C), (D), and (E) the best answer is: (E).

Question: The men talked at length and found they were agreeing with a lot of things, how would they describe their feeling toward one another?
(A) problem
(B) peace
(C) love
(D) contract
(E) like
Answer: Among (A), (B), (C), (D), and (E) the best answer is: (E).

Figure 2: **Examples of CoT prompting and direct answering on the CommonsenseQA dataset.** The text in light blue was generated by the Flan-T5 11B model. We include an open parenthesis in our prompt to make it more likely for the model to output one of the valid options, thus simplifying the answer parsing.

Chain-of-thought prompting in small and medium-sized language models. As self-supervised, generative pre-training is not enough to unlock CoT prompting capabilities in smaller models, several strategies have been proposed to address this need. Chung et al. (2022) showed that even models with less than a billion parameters were capable of zero-shot CoT prompting by first carrying out instruction fine-tuning on a huge collection of datasets with instructions and then on a smaller set of datasets which had rationales about the correct answers. Tay et al. (2022) achieved a similar goal in a 20 billion parameter model by pre-training from scratch on data combining multiple tasks under three denoising objectives with varying corruption rates, whereas work by Magister et al. (2022); Ho et al. (2022); Li et al. (2022) showed that curated reasoning chains output by large LMs can be used to "teach" smaller models (with as few as 300 million parameters) to do CoT reasoning through teacher forcing pipelines.

3 Experimental Setup

3.1 Types of Prompts

Direct answering. Given a model and a test instance, we prompt the model to provide an answer directly. In this prompt style, we set up a template for each dataset with the question or context followed by two to five lettered options (depending on the dataset). We then prompt the model to output only the letter corresponding to the best option.

Chain-of-Thought (CoT) prompting. For this setup we prompt the model with the question or context followed by the options. Then, we add a new line with the standard “let’s think step-by-step” instruction. We then allow the model to generate $\leq$128 tokens, and append the output of this to the initial input and feed the result back to the model along with a prompt asking it to output the letter corresponding to the best answer.

3.2 Evaluations

To measure the impact of CoT prompting, we first compute the overall accuracy with and without CoT prompting, respectively, for all models and datasets evaluated. For each setup, we gather the samples for which the given model produces a correct answer and those for which it produces an incorrect answer. Given these, we calculate the proportion of samples on which the CoT improves the accuracy, harms the accuracy, or has no effect. Importantly, this enables us not to calculate the difference in performance when switching from direct answering to CoT prompting.

3.3 Models

We run experiments using the following models.

Flan-T5. We use the 3B and 11B versions of Flan-T5 (Chung et al., 2022), which is a family of instruction-tuned T5 models with the encoder-decoder architecture (Raffel et al., 2020) on a mixture of 146 distinct categories of tasks, including 9 datasets with rationales supporting answers.

T*k*-INSTRUCT. Similar to Flan-T5, T*k*-INSTRUCT is a family of models (Wang et al., 2022a) fine-tuned on top of T5, on a mixture of over 1,600 tasks. We use only the 11B version of these models.

UL2. We use the 20B-parameter variant of UL2, which is pre-trained under a denoising task. It is trained to recover the missing words in a given incomplete input under three settings, with varying degrees of noise (Tay et al., 2022).

mT0. This is a family of models based on T0 (Sanh et al., 2022). Its generalization capability is greatly increased by fine-tuning over a wider

Dataset	Domain	# Options	Split	# Examples	CoT Δ
CommonsenseQA	general commonsense	5	test	2194	-5.634%
SocialIQA	social commonsense	3	val	1954	-3.612%
PIQA	physical commonsense	2	val	1838	-10.632%
HellaSwag	sentence completion	4	val	10042	-5.424%
Cosmos QA	contextual commonsense	4	val	2985	-3.108%
AQuA	algebraic word problems	5	test	254	0.078%

Table 2: **Statistics of all the datasets used in this work and the average differences of accuracies with and without using CoT prompting**, averaged across all 5 models.

variety of tasks. We use only the 11B-parameter version in this work (Muennighoff et al., 2022).

4 Results

The implementation details for all our experiments are available in Appendix A. The datasets on which we carried out the experiments are described in Table 2 and Appendix B.

Commonsense QA. We provide a summary of results for all models considered on five commonsense reasoning tasks in Table 1 and Figure 1 (full results in the Appendix). Using CoT prompting consistently harms model performance on the commonsense QA datasets considered. More precisely, CoT prompting degrades model performance by anywhere from 2.6 to 11.3 points (average 5.68). This result is somewhat counterintuitive given that such tasks seem to require multiple steps of reasoning, and chains of thought are supposed to help the model produce intermediate rationales. Indeed, our findings suggest that even when CoT capabilities are “unlocked” in smaller LMs, it may be inadvisable to elicit such “reasoning”.

Algebraic reasoning. However, on algebraic word problems, we observe that CoT prompting provides a significant boost to the Flan-T5 models, as can be seen in Table 3. This is a sharp contrast to the commonsense QA case, and suggests that CoT prompting of modestly sized models may be beneficial for *certain* types of tasks, but perhaps not others. This result is consistent with findings of earlier work that has shown models to achieve significant performance gains in algebraic tasks from zero-shot CoT prompting (Kojima et al., 2022).

Manual error analysis. To attempt to characterize *why* CoT prompting harms performance on commonsense reasoning datasets, we sampled 98 random examples from the CommonsenseQA dataset and fed them to Flan-T5 (11B) on them using CoT prompting. We manually labelled each example regarding whether the correct answer was given and whether a correct reasoning chain was given. We find that 20.4% of the actual questions were badly written or contained errors, such as having two or more correct (and sometimes even repeated) choices. Excluding these, 50% of generations included accurate answers and logical accompanying rationales. 10.2% of outputs contained incorrect answers and incorrect rationales; 11.2% contained a correct answer but an incorrect rationale; and 3.1% produced an incorrect answer despite producing a correct rationale. Finally, in 5.1% of the examples the model produced a correct answer after generating a rationale that only repeated the given question.

Interestingly, Flan-T5 managed to override its own rationales in a significant number of cases where it was not helpful at all to produce one (16.3%), and there were relatively few instances where an incorrect answer was generated after producing a sound rationale. Nevertheless, there was a significant number of instances where the model generated incorrect rationales leading to incorrect answers. Figure 2 includes one example of Flan-T5 generating reasoning that merely repeats information from the prompt and another where incorrect reasoning causes the model to select the wrong option.

5 Conclusions

In this work, we empirically evaluated the performance of instruction-tuned small-to-medium sized LMs for commonsense QA tasks with CoT prompting in comparison with direct answering. We found that, perhaps surprisingly, CoT prompting yields substantially *worse* performance on commonsense QA tasks across five model variants. Our analysis also provides insight on the failure cases when using CoT prompting for such models. Ultimately, we hope this motivates research into *why* CoT often harms performance for such models, and ultimately into methods to address this issue.

Limitations

In this work, we considered five medium-sized instruction-tuned LMs, and evaluated their performance across six datasets in all. We only investigated the zero-shot setting following the intriguing success of models like Flan-T5 and UL2 in unlocking the zero-shot CoT capabilities; it is possible that CoT could yield benefits for smaller models when they are fine-tuned on the commonsense datasets, which we leave as the future work.

We also note that all results here are on English corpora, and our findings may not generalize to other languages.

Ethics Statement

Our work is centered on studying the impact of using CoT prompting with medium-sized instruction-tuned LMs for commonsense reasoning tasks. By studying this, we highlight that NLP practicioners should be careful when using such approaches, which might mitigate the risks associated with the behaviors we describe in our findings.

References

Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M. Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan Fries, Maged S. Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-Jian Jiang, and Alexander M. Rush. 2022. PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts.

Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about Physical Commonsense in Natural Language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439.

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.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large Language Models Are Reasoning Teachers.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. 2022. OPT-IML: Scaling Language Model Instruction Meta Learning through the Lens of Generalization.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, Wenhu Chen, and Xifeng Yan. 2022. Explanations from Large Language Models Make Small Reasoners Better.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching Small Language Models to Reason.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. Crosslingual Generalization through Multitask Finetuning.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):140:5485–140:5551.

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

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense Reasoning about Social Interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

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.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2022. UL2: Unifying Language Learning Paradigms.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, A. Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, M. Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddharth Deepak Mishra, Sujan C. Reddy, Sumanta Patro, Tanay Dixit, Xu dong Shen, Chitta Baral, Yejin Choi, Hannaneh Hajishirzi, Noah A. Smith, and Daniel Khashabi. 2022a. Benchmarking generalization via in-context instructions on 1,600+ language tasks.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. 2022b. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.

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. 2022b. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022c. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Appendix

A Implementation details of LM prompting

We use the Hugging Face Transformers (v. 4.20.0) implementations of all models considered (Wolf et al., 2020). We set up the prompt templates for all datasets using the PromptSource library (Bach et al., 2022). For simplicity and computational efficiency—and in keeping with prior work (Li et al., 2022; Wei et al., 2022c)—we use only greedy decoding. For direct answering, we generate a maximum of 5 tokens, whereas for CoT we allow the model to generate up to 128 tokens (to accommodate "reasoning" strings).

All experiments were carried out on a computer with two NVIDIA A100 40GB GPUs. Running the experiments reported in Table 3 took around 10 hours.

B Datasets

We evaluate the relative performance of direct answering vs CoT prompting on five datasets encompassing several domains of commonsense reasoning, as well as a single dataset requiring algebraic thinking. Table 2 summarizes the sizes and domains of each dataset as well as the partitions we used for our experiments.

CommonsenseQA CommonsenseQA is a general question-answering dataset in which 12,247 questions are paired with five candidate options. The dataset aims to ask questions that require background knowledge that is trivial to humans but is not easy to gather directly on the web. The questions and answers are not authored in relation to a specific text; rather, they are generated freely by workers (Talmor et al., 2019).

SocialIQA A question-answering dataset of 38,000 questions focused on testing the capability of language models to reason about the social implications of people's actions (Sap et al., 2019). For each question, a context describing a social situation is provided, and the model is required to select the most likely answer from three candidate options.

PIQA Inspired by `instructables.com`, PIQA is a dataset that measures the physical commonsense knowledge of language models (Bisk et al., 2020). Given a natural language goal or objective and two possible solutions, the model must choose the most appropriate solution, of which exactly one is correct.

HellaSwag Contains 70k examples of questions in which the model must choose the most logical sentence from four given options to continue a given scene or situation. The task relies on general commonsense knowledge and is generally trivial for humans (Zellers et al., 2019).

Cosmos QA Comprises 35,600 reading-comprehension problems that require commonsense reasoning. Four candidate options are given for each problem, and the correct answer is never mentioned explicitly in the accompanying text. Therefore, the model must rather read between the lines and employ "contextual commonsense reasoning" to arrive at the most appropriate conclusion (Huang et al., 2019).

AQuA A dataset consiting of 100,000 algebraic word problems. Each problem is accompanied by five candidate options, only one of which is correct (Ling et al., 2017). This is not a commonsense-based dataset. Rather, we use it to compare our findings in regards to commonsense QA datasets to other domains in which CoT prompting has been shown more consistently to provide a significant and substantial advantage, even in smaller models.

C Full results

Model	Dataset	CoT Acc.	Direct Acc.	CoT Same	CoT Better	CoT Worse	CoT Δ
Flan-T5 (3B)	ECQA	83.95%	95.34%	85.77%	1.41%	12.80%	-11.39%
Flan-T5 (11B)	ECQA	86.13%	91.97%	89.14%	2.50%	8.34%	-5.84%
T*k*-INSTRUCT (11B)	ECQA	55.01%	63.12%	68.16%	11.85%	19.96%	-8.11%
UL2 (20B)	ECQA	37.41%	39.15%	67.72%	15.26%	17.00%	-1.74%
mT0 (11B)	ECQA	71.31%	72.40%	90.86%	4.01%	5.10%	-1.09%
Flan-T5 (3B)	SocialIQA	73.94%	78.65%	85.15%	5.06%	9.77%	-4.71%
Flan-T5 (11B)	SocialIQA	78.14%	82.39%	87.35%	4.19%	8.44%	-4.25%
T*k*-INSTRUCT (11B)	SocialIQA	49.02%	60.79%	60.99%	13.61%	25.38%	-11.77%
UL2 (20B)	SocialIQA	41.65%	39.24%	66.37%	18.01%	15.60%	2.41%
mT0 (11B)	SocialIQA	73.37%	73.11%	90.83%	4.70%	4.44%	0.26%
Flan-T5 (3B)	PIQA	79.97%	84.05%	83.72%	6.09%	10.17%	-4.08%
Flan-T5 (11B)	PIQA	77.20%	84.70%	82.58%	4.95%	12.45%	-7.50%
T*k*-INSTRUCT (11B)	PIQA	61.09%	75.40%	57.93%	13.87%	28.18%	-14.31%
UL2 (20B)	PIQA	56.03%	64.09%	58.65%	16.64%	24.70%	-8.06%
mT0 (11B)	PIQA	55.54%	74.75%	65.12%	7.83%	27.04%	-19.21%
Flan-T5 (3B)	HellaSwag	83.69%	87.02%	87.61%	4.52%	7.85%	-3.33%
Flan-T5 (11B)	HellaSwag	80.33%	85.38%	82.09%	6.42%	11.47%	-5.05%
T*k*-INSTRUCT (11B)	HellaSwag	45.57%	54.58%	64.99%	12.99%	22.00%	-9.01%
UL2 (20B)	HellaSwag	23.96%	29.37%	64.93%	14.82%	20.23%	-5.41%
mT0 (11B)	HellaSwag	33.16%	37.48%	71.34%	12.16%	16.48%	-4.32%
Flan-T5 (3B)	CosmosQA	83.68%	83.74%	93.76%	3.08%	3.14%	-0.06%
Flan-T5 (11B)	CosmosQA	83.65%	86.63%	91.25%	2.88%	5.86%	-2.98%
T*k*-INSTRUCT (11B)	CosmosQA	53.60%	67.00%	67.43%	9.58%	22.98%	-13.40%
UL2 (20B)	CosmosQA	35.23%	35.37%	66.62%	16.61%	16.75%	-0.14%
mT0 (11B)	CosmosQA	84.95%	83.91%	95.06%	2.98%	1.94%	1.04%
Flan-T5 (3B)	AQuA	25.97%	23.21%	69.67%	16.53%	13.77%	2.76%
Flan-T5 (11B)	AQuA	25.97%	24.01%	68.10%	16.92%	14.96%	1.96%
T*k*-INSTRUCT (11B)	AQuA	25.18%	25.18%	66.13%	16.92%	16.92%	0.00%
UL2 (20B)	AQuA	21.64%	24.79%	74.79%	11.02%	14.17%	-3.15%
mT0 (11B)	AQuA	16.52%	17.70%	93.29%	2.75%	3.93%	-1.18%

Table 3: Full results for all models and datasets.