TOWARDS UNDERSTANDING CHAIN-OF-THOUGHT PROMPTING: AN EMPIRICAL STUDY OF WHAT MATTERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Chain-of-Thought (CoT) prompting, which encourages language models (LMs) to generate intermediate rationales for the final answer through in-context demonstrations, dramatically improves large LMs' ability to solve reasoning tasks. Despite its success, there is little understanding on what makes CoT prompting effective and which aspects of the demonstrated reasoning steps contribute to its performance. In this paper, we show that prompting with invalid demonstrations affects little in CoT reasoning, achieving over 80–90% of the performance obtained using the original CoT under various metrics, while still generating coherent lines of reasoning during inference. Further experiments show that other aspects of the rationales, such as being relevant to the query and correctly ordering the reasoning steps, are the actual key to the effectiveness of CoT. Overall, these findings deepen our understanding of CoT prompting, while leading to new questions regarding large LMs' capability to learn to reason in context and reflections on benchmarking few-shot reasoning.[1]

1 INTRODUCTION

Large language models (LLMs) can perform new tasks during inference when prompted with a few demonstrations (Brown et al., 2020). Chain-of-Thought (CoT) prompting (Wei et al., 2022) can (Figure 1) dramatically improve the ability of LLMs to do complex and multi-step reasoning, by including a *rationale* (colored part in Figure 1) for each example, i.e., a series of reasoning steps, which encourages the LLM to explicitly generate its reasoning process before the answer. Despite its successes, there is little understanding of what makes CoT prompting effective and which aspects of the demonstrated reasoning steps contribute to its performance.

	Query	Standard	Chain-of-Thought	Invalid Reasoning
In-context Demonstration	Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?	The answer is 39.	Originally, Leah had 32 chocolates and her sister had 42. So in total they had 32 + 42 = 74. After eating 35, they had 74 - 35 = 39 pieces left in total. The answer is 39.	Originally, Leah had 32 chocolates and her sister had 42. So her sister had 42 - 32 = 10 chocolates more than Leah has. After eating 35, since 10 + 35 = 45, they had 45 - 6 = 39 pieces left in total. The answer is 39.
Inference by LLM	Julie is reading a 120-page book. Yesterday, she read 12 pages and today, she read twice as many pages as yesterday. If she wants to read half of the remaining pages tomorrow, how many pages should she read?	The answer is 18. ✗	Julie is reading a 120-page book. Yesterday, she read 12 pages and today, she read 24 pages. So she read a total of 12 + 24 = 36 pages. Now she has 120 - 36 = 84 pages left. Since she wants to read half of the remaining pages, she should read 84 / 2 = 42 pages. The answer is 42. ✓	Yesterday, Julie read 12 pages. Today, she read 12 * 2 = 24 pages. So she read a total of 12 + 24 = 36 pages. Now she needs to read 120 - 36 = 84 more pages. She wants to read half of the remaining pages tomorrow, so she needs to read 84 / 2 = 42 pages tomorrow. The answer is 42. ✓

Figure 1: Example demonstrations and model inference results for standard prompting, Chain-of-Thought (CoT) prompting, and CoT with invalid reasoning (§3), where the rationale is in color (green: valid, yellow: invalid). Valid reasoning for the demonstrations matters only a small portion to the performance of CoT—by providing rationales with invalid reasoning, LLMs achieve over 80–90% of the performance of CoT while performing logically sound and pertinent lines of reasoning.

[1]We will release our code and model input/output upon acceptance.

In this paper, we design a series of ablation experiments where we deliberately change different aspects of the demonstrated rationales and measure how the model performance varies accordingly (§3, §4). On two representative multi-step reasoning tasks—arithmetic reasoning and multi-hop factual question answering (QA), we find that **the validity of reasoning matters only a small portion to the performance**—by providing rationales with completely invalid reasoning steps, the LLM can still achieve over 80–90% of the performance of CoT under various metrics while generating coherent lines of reasoning towards the answer (§3). Through further examinations, we identify and formulate other aspects of a CoT rationale (§4), and find that **being relevant to the query and correctly ordering the reasoning steps are the key** for the effectiveness of CoT prompting.

Overall, our findings suggest that what LLMs *learn* about how to reason under CoT prompting could be limited. Rather, the rationales may mainly serve as the role of an output format/space that regularizes the LM generation to be in correct ordering and relevant to the query. Our work suggests a new way of interpreting the evaluation scores in view of the prior knowledge LLMs possess, while leading to new questions regarding LLMs' capabilities of learning to reason in context and reflections on benchmarking few-shot reasoning (§5). We discuss related work in §A.6.

2 Study Formulation & Experimental Setup

Components of a CoT Rationale. We identify two distinct components of a CoT rationale as follows (Figure 1; examples in Table 1).

(1) Bridging objects: the key objects the LLM model needs to traverse in order to make a successful final prediction. For arithmetic reasoning, they are numeric parts (numbers & equations), and for factual QA, they are the subject & object entities.

(2) Language templates: the complementary parts of bridging objects, which serve as textual hints and relations/predicates that derive the correct bridging objects along the reasoning process.

CoT provides correct bridging objects and language templates as demonstrations to show the LLM how to reason. While CoT achieves impressive performance, we are interested in the following questions: *Are ground truth bridging objects/language templates important? If not, what would be the key aspects that are needed for the LLM to reason properly?* These questions are the main focus of our study, which will be discussed in detail in §3 and §4.

Arithmetic Reasoning	Multi-hop QA
Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?	Q: Who is the grandchild of Dambar Shah?
A: Originally, Leah had 32 chocolates and her sister had 42. So in total they had 32 + 42 = 74. After eating 35, they had 74 - 35 = 39 pieces left in total. The answer is 39.	A: Dambar Shah (? - 1645) was the father of Krishna Shah. Rudra Shah was the child of Krishna Shah (? - 1661). So the final answer (the name of the grandchild) is: Rudra Shah.

Table 1: Bridging objects and language templates of a Chain-of-Thought rationale. Here we illustrate with one in-context exemplar for each task we experiment with.

Datasets & In-context Exemplars. We experiment on two representative tasks involving multi-step reasoning: arithmetic reasoning & multi-hop factual question answering (QA). We select benchmarks on which CoT prompting brings significant improvements over standard prompting. For arithmetic reasoning, we experiment on GSM8K (Cobbe et al., 2021), one of the most challenging mathematical reasoning benchmarks. For multi-hop QA, we experiment on Bamboogle, a dataset of compositional questions constructed by Press et al. (2022). Due to budget considerations, we uniformly sample 800 out of the 1319 test examples for evaluation for GSM8K. We base our experiments on the original prompt exemplars released by Wei et al. (2022) and Press et al. (2022) with slight editing to make their structures more consistent and reduce redundancy. More details are included in Appendix A.1.

Backbone Language Model. We use InstructGPT-175B (Ouyang et al., 2022; Brown et al., 2020) `text-davinci-002` for results in the main paper. We also report results and additional discussions with three other LLMs with 175B or more parameters in Appendix A.5.

Evaluation. For extrinsic evaluation, we use **Answer Accuracy** for GSM8K and **Answer F1** for Bamboogle. For intrinsic evaluation (§A.3), we measure the Recall/F1 (**Inter. Recall/F1**) of the bridging objects which need to be derived by the LLM (i.e., those that do not appear in the query).

	GSM8K			Bamboogle	
	Inter. Recall	**Inter. F1**	**Answer Acc.**	**Inter. Recall**	**Answer F1**
STD (Standard prompting)	N/A	N/A	15.4	N/A	20.6
CoT (Chain-of-Thought prompting)	43.9	48.3	48.5	45.2	45.2
1 Invalid Reasoning	39.8	43.9	39.5	44.4	39.4
2 No *coherence* for bridging objects	35.3	39.2	35.8	40.8	37.4
3 No relevance for bridging objects	21.4	26.2	27.5	39.6	34.0
4 No *coherence* for language templates	24.1	28.3	25.8	35.2	32.1
5 No relevance for language templates	29.5	34.0	32.8	40.4	29.4
6 No *coherence*	25.2	29.4	23.1	39.6	33.8
7 No relevance	9.6	11.9	11.0	36.8	23.9

Table 2: Intrinsic and extrinsic evaluation results under InstructGPT (`text-davinci-002`) for all settings in our experiments. Results for three other LLMs can be found in Appendix A.5.

	1 (#3)	2 (#200)	3 (#268)	4 (#179)	5 (#85)	6 (#46)	7 (#15)	8 (#4)
CoT	0.67	0.65	0.52	0.38	0.33	0.29	0.21	0.28
Invalid Reasoning	0.78	0.58	0.48	0.37	0.24	0.28	0.20	0.25

Table 3: Model performance (**Inter. F1**) under CoT and invalid reasoning setting across different reasoning depths on GSM8K. The number of samples for each reasoning depth is shown in the parenthesis. Overall, small performance drops are observed across different difficulty levels.

3 How Much Does Valid Reasoning Matter?

Intuitively, one of the most important aspects of a Chain-of-Thought rationale would be its logically valid and sound reasoning. If we provide rationales with invalid reasoning steps in the demonstrated examples instead, we should expect the LLM to fail to reason properly and gain little or even negative improvements compared with standard prompting (where no rationale is given). To test this intuition, we design an ablation study where we construct invalid reasoning steps for the demonstrated rationales, and measure its influence on model behavior.

Constructing Invalid Chain of Reasoning. We manually write rationales with invalid reasoning for *all* the in-context demonstration examples. We keep the premise steps which are copies/paraphrases of facts from the query, and change the subsequent steps such that they do not logically derive the final answer. Importantly, we are *not* adopting an adversarial/counterfactual perturbation setting where minimal alterations are applied to make the reasoning invalid; instead, we apply rather drastic changes where we change both the bridging objects and language templates and hence little valid reasoning exists to help solve the query. We include some examples in Table 5 and the full prompts in Appendix A.7.

Results. Table 2 shows that LLMs can achieve surprisingly high performance when provided with invalid reasoning steps for the demonstrations (1). In particular, under **Inter. Recall/Inter.F1**, i.e., intrinsic evaluation, which is arguably a more faithful measurement of the rationale quality (§A.3), all LLMs we tested can retain over 90% of the performance achieved under CoT prompting.

Since GSM8K has large variations in the difficulty levels of the problem instances,[2] we additionally report the performance breakdown based on the difficulty level of each instance in Table 3. Overall, small performance drops are observed across different difficulty levels.

Qualitative analysis. Qualitative examination confirms that rationales generated from the invalid demonstrations look indistinguishable from the rationales generated under the original CoT demonstrations. In almost all cases where the predicted final answer is correct, the rationales do reach the answer with valid and sound reasoning steps (as in CoT), drastically different from those in the given demonstrations. For cases where the final answer is wrong, the errors the LLM makes are also in the same types with the errors made under CoT prompting.

[2]We use the number of reasoning steps required to solve a problem as its difficulty level.

4 WHAT ARE THE KEY ASPECTS OF CHAIN-OF-THOUGHTS?

Our findings open the question: *If valid reasoning is not required, what are the key aspects that determine the effectiveness of CoT prompting?* Re-examining the CoT rationales and our ablation setting in §3, we formulate two notions that capture two other aspects of a rationale in what follows.

Relevance. Whether the rationale is based on the corresponding component from the query. For bridging objects, this means whether the rationale uses the same objects mentioned in the query (numbers for arithmetic reasoning and entities for factual QA); for language templates, this means whether they are about the same set of entities/relations as the query, and allude to the question being asked. For example, a template about *"Patricia"* and *"hair"* would not have relevance to a query about *"Leah"* and *"Chocolates"*, and a template seeking the *"brother-in-law"* of the topic entity does not have relevance to a query asking for the *"grandchild"* (Table 5).

Coherence. Whether the rationale is in the correct order, i.e., later steps could not be pre-conditions for earlier steps and reversely, earlier steps could not be based on later steps. For example, a rationale where *"32 + 42 = 74"* appears before the introduction of *"32"* or *"42"* would not have coherence on bridging objects, and similarly for language templates.

We design a set of ablation settings to examine the impact of these two aspects for different components of a CoT-like rationale. Details on how we designed prompts to ablate each aspect are included in §A.4. Results are reported in Table 2; we summarize the main findings in what follows.

Relevance and coherence are key for the performance of CoT prompting. Most settings for this section (2-7) have rather large performance drops from CoT, where the low-performing ones approach or even underperform standard prompting. This suggests that overall, relevance and coherence are key for the performance of CoT.

Keeping relevance is crucial. The no relevance setting 7 where both components of the rationale have no relevance achieves significantly poorer performance than other ablation settings, and even underperforms standard prompting where no rationale is given on GSM8K. Manual examination of generated rationale under this setting suggests that the LLM generates irrelevant rationales (both bridging objects and language templates) for 15 out of 20 examples. Many of the irrelevant rationales have recurring topics (e.g., *"cats and dogs"*, *"passengers and buses"*) which we hypothesize are frequent patterns in the portion relevant to mathematics in the pretraining corpora. Overall, this suggests that a certain level of relevance is crucial for the LLM to stick to the query being asked.

Relevance matters more than coherence for bridging objects. Providing incoherent bridging objects (2) achieves better performance than providing irrelevant bridging objects (3), especially on the more challenging GSM8K dataset (39.2 *v.s.* 26.2 **Inter. F1**). which indicates that it is important for the bridging objects to be relevant, but not as important to have them in the right order to guide the LLM along the reasoning process. We quantitatively measure the coverage of bridging objects from the query for the generated rationales, and find that the settings with no relevance for bridging objects (3, 7) do have significantly lower coverage (<60%) than other settings (≈80%).

Coherence of language templates is important. Coherence of language templates 4 matters a lot to the performance of CoT prompting. By examining the predicted rationales, we find that the LLM is indeed generating rationales with incoherent language templates (14 out of 20 examples).

5 DISCUSSION & CONCLUSION

In this paper, we aim to better understand CoT prompting through a series of ablation experiments that unveil the impact of different aspects of a CoT rationale. Given the surprisingly high performance obtained by ablating the validity of reasoning (§3), it can be concluded that what the LLM *learns* from the demonstrations about how to reason properly is limited—rather, the LLM has already gained a lot of such reasoning ability from pretraining (at least for tasks we experiment on), and the provided reasoning steps may mainly serve as the role of output format/space that regularizes the LLM to generate rationales that look step-by-step while being coherent and relevant to the query (§4). Our findings indicate that evaluations on alternative benchmarks where LLMs have less prior knowledge are needed to more faithfully assess LLMs' abilities on learning to reason from few-shot demonstrations. We include additional results and discussion in §A.5.

References

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.

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. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*, 2022.

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *arXiv preprint arXiv:2208.01066*, 2022.

Joel Jang, Seonghyeon Ye, and Minjoon Seo. Can large language models truly understand prompts? a case study with negated prompts. *arXiv preprint arXiv:2209.12711*, 2022.

Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, Kang Min Yoo, and Taeuk Kim. Ground-truth labels matter: A deeper look into input-label demonstrations. In *EMNLP*, 2022.

Aman Madaan and Amir Yazdanbakhsh. Text and patterns: For effective chain of thought, it takes two to tango. *arXiv preprint arXiv:2209.07686*, 2022.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *EMNLP*, 2022.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*, 2022.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. Impact of pretraining term frequencies on few-shot reasoning. *arXiv preprint arXiv:2202.07206*, 2022.

Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *arXiv preprint arXiv:2210.01240*, 2022.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2300–2344, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.167. URL `https://aclanthology.org/2022.naacl-main.167`.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=_VjQlMeSB_J`.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=RdJVFCHjUMI`.

Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Ves Stoyanov, Greg Durrett, and Ramakanth Pasunuru. Complementary explanations for effective in-context learning. *arXiv preprint arXiv:2211.13892*, 2022.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.

A Appendix

A.1 Chain of Thought Exemplars

We base our experiments on the original prompt exemplars released by Wei et al. (2022); Press et al. (2022) with slight editing to make the structure more consistent and reduce redundancy, which makes our ablations more convenient to conduct. The edited CoT prompts for arithmetic reasoning and multi-hop QA could be found in Table 9 and Table 10 respectively. We mainly perform the following edits: 1) shift premise steps (copy/paraphrase of facts from the query) to the beginning steps of the rationale; 2) add/expand the language templates for steps with no/over-concise language templates; 3) remove unnecessary steps/information that are unhelpful for answering the query.

Overall, these edits only slightly affect the performance of CoT. A comparison of the performance is shown in Table 4.

	GSM8K			Bamboogle	
	Inter. Recall	**Inter. F1**	**Answer Acc.**	**Inter. Recall**	**Answer F1**
Chain-of-Thought (Original)	44.5	48.7	48.1	44.8	43.1
Chain-of-Thought (After Editing)	43.9	48.3	48.5	45.2	45.2

Table 4: Performance comparison (under `text-davinci-002`) of the Chain-of-Thought exemplars before/after our editing.

A.2 Example Prompts for Each Setting

Table 5 includes example demonstrations for each setting we experimented with.

A.3 More Details on Intrinsic Evaluation

We use Recall/F1 of the bridging objects as the metrics for intrinsic evaluation of the generated rationales. For GSM8K, since annotations for ground truth reasoning steps are available, we use the derived numbers in the annotated steps as a proxy for bridging objects.[3] For Bamboogle, we manually annotate the bridging objects (intermediate entities) and measure their recall. While the metrics don't take into account the quality of the language templates, we examine the predicted rationales for 20 random examples under each setting we tested except standard prompting (which does not generate any rationale), and find that for all the examples, whenever the LLM reaches a correct bridging object, the corresponding language template within the step is also correct. This suggests that overall, the correctness of bridging objects is a very good indicator of the quality of the reasoning steps.

[3]We do not use whole equations since we observe that the LLM may express the mathematical equation in different ways, e.g., "5 plus 3 is 8", "5 + 3 = 8".

Prompt Setting	**Example Query (Arithmetic Reasoning)** *Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?*	**Example Query (Factual QA)** *Who is the grandchild of Dambar Shah?*
STD (Standard prompting)	39	So the final answer is: Rudra Shah.
CoT (Chain-of-Thought)	Originally, Leah had 32 chocolates and her sister had 42. So in total they had 32 + 42 = 74. After eating 35, they had 74 - 35 = 39 pieces left in total. The answer is 39.	Dambar Shah (? - 1645) was the father of Krishna Shah. Rudra Shah was the child of Krishna Shah (? - 1661). So the final answer (the name of the grandchild) is: Rudra Shah.
(1) Invalid Reasoning	Originally, Leah had 32 chocolates and her sister had 42. So her sister had 42 - 32 = 10 chocolates more than Leah has. After eating 35, since 10 + 35 = 45, they had 45 - 6 = 39 pieces left in total. The answer is 39.	Dambar Shah (? - 1645) was the king of the Gorkha Kingdom. The Gorkha Kingdom was established by Prince Dravya Shah. Dravya Shah has a child named Rudra Shah. So the final answer (the name of the grandchild) is: Rudra Shah.
(2) No *coherence* for bridging objects	Originally, Leah had 32 + 42 = 74 chocolates and her sister had 32. So in total they had 74 - 35 = 39. After eating 35, they had 42 pieces left in total. The answer is 39.	Krishna Shah was the father of Rudra Shah. Dambar Shah (? - 1645) was the child of Krishna Shah (? - 1661). So the final answer (the name of the grandchild) is: Rudra Shah.
(3) No relevance for bridging objects	Originally, Leah had 19 chocolates and her sister had 31. So in total they had 19 + 31 = 50. After eating 29, they had 50 - 29 = 21 pieces left in total. The answer is 21.	Metis Amando was the father of David Amando. Randall Amando was the child of David Amando. So the final answer (the name of the grandchild) is: Randall Amando.
(4) No *coherence* for language templates	After eating 32, they had 42 pieces left in total. Originally, Leah had 32 + 42 = 74 chocolates and her sister had 35. So in total they had 74 - 35 = 39. The answer is 39.	Dambar Shah (? - 1645) was the child of Krishna Shah. Krishna Shah (? - 1661) was the father of Rudra Shah. So the final answer (the name of the grandchild) is: Rudra Shah.
(5) No relevance for language templates	Patricia needs to donate 32 inches, and wants her hair to be 42 inches long after the donation. Her hair is 35 inches long currently. Her hair needs to be 32 + 42 = 74 inches long when she cuts it. So she needs to grow 74 - 35 = 39 more inches. The answer is 39.	The husband of Dambar Shah (? - 1645) is Krishna Shah. Krishna Shah (? - 1661) has a brother called Rudra Shah. So the final answer (the name of the brother-in-law) is: Rudra Shah.
(6) No *coherence*	After eating 32 + 42 = 74, they had 32 pieces left in total. Originally, Leah had 74 - 35 = 39 chocolates and her sister had 35. So in total they had 42. The answer is 39.	Krishna Shah was the child of Rudra Shah. Dambar Shah (? - 1645) was the father of Krishna Shah (? - 1661). So the final answer (the name of the grandchild) is: Rudra Shah.
(7) No relevance	Patricia needs to donate 19 inches, and wants her hair to be 31 inches long after the donation. Her hair is 29 inches long currently. Her hair needs to be 19 + 31 = 50 inc long when she cuts it. So she needs to grow 50 - 29 = 21 more inches. The answer is 21.	The husband of Metis Amando is David Amando. David Amando has a brother called Randall Amando. So the final answer (the name of the brother-in-law) is: Randall Amando.

Table 5: Examples for all settings in our experiments.

A.4 Details on Ablations in §4

In order not to introduce mixed effects which could make the results not well-controlled, we base the ablation settings on top of the CoT prompts instead of the setting in §3.

Given the two components (bridging objects and language templates) and the two aspects (relevance and coherence) of the rationale, there are naturally four ablation settings where each could examine one aspect of a certain component. We also experiment with two other settings: no relevance where neither bridging objects nor language templates have relevance, and *no coherence* which is defined analogously ((6), (7) in Table 5).

Destroying relevance. We perform random substitutions to ablate the relevance of a certain component. For ablating the relevance of bridging objects, we randomly sample alternatives (numbers for GSM8K, entities for Bamboogle) for those from the query, and change the bridging objects in the subsequent steps correspondingly to maintain the coherence of the rationale. Using our running example, we randomly replace the bridging objects from the query: *"32"* → *"19"*, *"42"* → *"31"* and *"35"* → *"29"*, then change the bridging object from the first entailment step from *"32 + 42 = 74"* to *"19 + 31 = 50"*, and so on so forth. To ablate the relevance of language templates, for GSM8K, we randomly sample an annotated rationale from the training set, and use its template in place of the original template. For Bamboogle, we manually replace the template with an alternative which is irrelevant to the query.

Destroying coherence. Ablating the coherence is rather straightforward, where we randomly shuffle the components and permute their orderings.

A.5 Additional Results & Discussion

Table 6 includes results for `text-davinci-003`, `text-davinci-002`'s very recent improved version.

Comparing with the results from `text-davinci-002` (Table 2), it could be seen that `text-davinci-003` brings large performance improvements, especially under the ablation settings. In particular, providing invalid reasoning for the rationales (①) overall only marginally harms the performance, and even outperforms CoT on GSM8K under intrinsic evaluation. This suggests that `text-davinci-003` is equipped with even stronger multi-step "reasoning" abilities on the evaluated tasks through pre-training, and learns little about how to reason from the demonstrations.

For the remaining settings where we ablate the relevance/coherence (②-⑦), the same trend can be observed on the challenging GSM8K dataset, e.g., the model still suffers a lot when providing rationales that are irrelevant or have incoherent language templates. For the relatively easier Bamboogle dataset, the high model capacity indicated by its impressive performance has basically erased significant impacts from the ablations, with the only standing observation that the model still needs the rationales to be relevant to maintain its performance.

Overall, from the performance achieved by `text-davinci-002` and `text-davinci-003`, we can observe a general trend where LLMs suffer less from the ablations when they have more prior knowledge about the task. To further explore this, we test on `InstructX`, the instruction-tuned version of a recent proprietary LLM (`X`) that is directly trained on both arithmetic reasoning and factual QA in CoT fashion during instruction tuning, and hence has immense knowledge on these tasks. The results are shown in Table 7. It could be seen that none of the ablations has significant impacts on the model performance, which further strengthens this pattern. On the positive side, this indicates that LLMs can effectively utilize their prior knowledge to solve new problems; however, this also leads to the concern that LLMs may over-rely on their prior knowledge and ignore important information in the context, including those that are crucial for specifying the task semantics (Jang et al., 2022).

We also test on the original proprietary model `X`, which is a non-instruction-finetuned LLM that exhibits strong CoT reasoning ability. The results are included in Table 8. Overall, similar observations could be found, which suggests that our findings are not exclusive to instruction-tuned models. There are some inconsistencies between the performance from `X` and InstructGPT on Bamboogle, where the importance of coherence and relevance for bridging objects is flipped. This could be the consequence of instruction tuning, and differences in pretraining corpora and model scales.

We integrate all our discussions as follows via a series of questions that are opened up from our study.

***Do* LLMs learn to reason from CoT demonstrations?** Given the surprisingly high performance obtained by ablating the validity of reasoning for the in-context rationales (§3), it can be concluded that what the LLM learns from the demonstrations about how to reason properly is limited—rather, the LLM has already gained a lot of such complex reasoning ability from pretraining (at least for tasks we experiment on), and the provided reasoning steps serve more as the role of an output format/space, that regularizes the LLM to generate rationales that look step-by-step while being coherent and relevant to the query (§4). Moreover, results obtained from recent stronger models including `text-davinci-003` and `InstructX` (see Appendix A.5) suggest that LLMs suffer further less from the ablations when they have more prior knowledge about the task. In particular, for `InstructX` which is directly trained on both arithmetic reasoning and factual QA in CoT fashion and hence has immense knowledge on these tasks, it could be seen that none of the ablations has significant impacts on its performance. On the positive side, this indicates that LLMs can effectively utilize their prior knowledge to solve new problems. However, from another perspective, if we view the invalid reasoning setting as a *task* where the goal is to generate invalid reasoning steps for the query, then the LLM has basically failed to capture the task as it still tries to predict valid reasoning steps. This leads to the concern that LLMs may over-rely on their prior knowledge and ignore important information in the context that are presumably rare in the pretraining distribution, including those that are crucial for specifying the task semantics (Jang et al., 2022).

***Can* LLMs learn to reason in-context?** We note that what we find does not in any way diminish the *potential* of learning to reason in context for LLMs; recent work has also shown evidence that learning in context is possible and could be powerful (Garg et al., 2022; Akyürek et al., 2022). Rather, our findings show that the existing successes of CoT are not sufficient for establishing that LLMs are good *few-shot learners* of reasoning; instead, the pretraining corpora have already forged them to be good reasoners on the tasks being evaluated, and the main role that the demonstrations play is to elicit such reasoning skills.

Reflections on benchmarking few-shot reasoning. An important topic on benchmarking in the era of large pre-trained language models is to quantify the level of prior knowledge the LLM has gained about the end task being evaluated, which is crucial for assessing how well can the model truly extrapolate from pretraining and acquire new skills (Chollet, 2019). One direct way is to look into the pretraining corpora when it is accessible, e.g., Razeghi et al. (2022) investigates the correlation between the model performance and the frequency of terms from the test instances in the pretraining data. However, the pretraining corpora are not always accessible, and low-level statistics are usually not adequate when the topics of interest are abstract and high-level skills such as reasoning. Along this direction, our work could be regarded as a way to approximately quantify the prior knowledge that the LLM possesses on multi-step reasoning. Our findings indicate that evaluations on alternative benchmarks where LLMs have less prior knowledge are needed to more faithfully assess the LLMs' abilities on learning to reason from few-shot demonstrations.

	GSM8K			**Bamboogle**	
	Inter. Recall	**Inter. F1**	**Answer Acc.**	**Inter. Recall**	**Answer F1**
STD (Standard prompting)	N/A	N/A	15.2	N/A	25.1
CoT (Chain-of-Thought prompting)	48.4	53.1	54.5	61.6	59.5
① Invalid Reasoning	50.2	53.5	51.5	60.8	56.4
② No *coherence* for bridging objects	46.5	51.5	50.4	59.2	55.2
③ No relevance for bridging objects	32.5	38.3	47.2	60.4	56.9
④ No *coherence* for language templates	37.8	43.3	41.9	57.2	51.4
⑤ No relevance for language templates	44.6	49.9	51.8	62.4	59.3
⑥ No *coherence*	34.5	39.4	31.0	57.6	55.2
⑦ No relevance	15.5	17.8	16.2	50.0	49.0

Table 6: Intrinsic and extrinsic evaluation results under `text-davinci-003` for all settings. Discussions are included in Appendix A.5.

	GSM8K			**Bamboogle**	
	Inter. Recall	**Inter. F1**	**Answer Acc.**	**Inter. Recall**	**Answer F1**
STD (Standard prompting)	N/A	N/A	21.8	N/A	36.5
CoT (Chain-of-Thought prompting)	72.2	73.0	63.8	57.6	56.9
① Invalid Reasoning	71.8	72.6	64.4	55.6	52.8
② No *coherence* for bridging objects	72.1	72.9	65.8	51.6	49.3
③ No relevance for bridging objects	71.1	71.9	64.6	54.0	52.8
④ No *coherence* for language templates	71.6	72.2	63.9	54.0	52.0
⑤ No relevance for language templates	71.9	72.7	64.9	55.2	53.5
⑥ No *coherence*	71.7	72.5	64.2	54.4	54.0
⑦ No relevance	70.7	71.6	64.5	50.0	51.9

Table 7: Intrinsic and extrinsic evaluation results under `InstructX`, the instruction-tuned version of `X` (a proprietary large language model with 500+ billion parameters) for all settings. Discussions are included in Appendix A.5.

A.6 Related Work

A few recent work focuses on understanding/analyzing CoT prompting. Madaan & Yazdanbakhsh (2022) investigates the importance of different components of the demonstrated CoT rationales by changing them to be *counterfactual*. They only experiment with limited ways of changing the rationales to be *wrong* including using incorrect calculations (e.g., *"5 + 4 = 7"*) or entities. For most of their settings, even though the rationales are made counterfactual, they are still *correct* since the query is changed accordingly (see, e.g., Table 48 of their paper). Concurrent to our work, Ye et al. (2022) also explores how corrupting the CoT rationales affects the model performance. They experiment with using incorrect calculations and *dropping* (parts of) the bridging objects/language templates,

	GSM8K			Bamboogle	
	Inter. Recall	**Inter. F1**	**Answer Acc.**	**Inter. Recall**	**Answer F1**
STD (Standard prompting)	N/A	N/A	15.0	N/A	31.0
CoT (Chain-of-Thought prompting)	36.6	40.6	37.0	54.0	54.8
① Invalid Reasoning	33.9	36.9	31.8	50.4	46.1
② No *coherence* for bridging objects	30.3	35.0	33.5	33.6	25.7
③ No relevance for bridging objects	15.5	20.1	21.2	47.2	47.7
④ No *coherence* for language templates	23.1	27.3	21.9	40.4	35.5
⑤ No relevance for language templates	19.5	22.9	20.4	38.4	30.6
⑥ No *coherence*	23.9	28.3	24.1	39.6	33.6
⑦ No relevance	12.1	16.4	16.4	28.4	14.3

Table 8: Intrinsic and extrinsic evaluation results under X, a proprietary large language model with 500+ billion parameters. Discussions are included in Appendix A.5.

which are different from our ablation designs. Saparov & He (2022) investigates systematically evaluating CoT by creating a synthetic QA dataset based on first-order logic, which allows for parsing the generated rationales into symbolic proofs for formal analysis. Overall, to our knowledge, we are the first to show that it is possible to have CoT rationales that are wrong and drastically deviate from the gold ones while still maintaining high model performance.

There have also been several subsequent work of Chain-of-Thought prompting since its introduction, including better decoding strategies (Wang et al., 2022), automatic exemplar constructions (Zhang et al., 2022) and program-based CoT (Chen et al., 2022). In this paper, we are primarily focused on understanding the effectiveness of the original CoT prompting method where we use the same experimental settings (e.g., greedy decoding) and base our experiments on the same few-shot exemplars used. We believe our findings could also apply to some of the subsequent variants of CoT prompting.

In general in-context learning (ICL), Min et al. (2022) shows that for a wide range of tasks in natural language understanding with categorical label space (classification and multi-choice), ground truth input-label mappings matter very little for end-task performance, and other aspects such as the label space, overall format and the distribution of text are the key. Building on this work, Kim et al. (2022) finds that the correct input-label correspondence could have varying impacts based on the task and experimental configurations. Webson & Pavlick (2022) finds that for instruction models, the performance on natural language inference tasks has small degradations under irrelevant or misleading instructions. Xie et al. (2022) provides theoretical analysis of ICL by formulating it as Bayesian inference. Our work could be viewed as an attempt to empirically understand ICL in sequence generation tasks requiring multi-step reasoning.

A.7 Full List of Prompts

Full prompts for the settings in our experiments are included in Table 9-24.

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?
A: There are 15 trees originally. Then there were 21 trees after the Grove workers planted some more. So there must have been 21 - 15 = 6 trees that were planted. The answer is 6.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
A: There are originally 3 cars. Then 2 more cars arrive. Now 3 + 2 = 5 cars are in the parking lot. The answer is 5.

Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
A: Originally, Leah had 32 chocolates and her sister had 42. So in total they had 32 + 42 = 74. After eating 35, they had 74 - 35 = 39 pieces left in total. The answer is 39.

Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
A: Jason had 20 lollipops originally. Then he had 12 after giving some to Denny. So he gave Denny 20 - 12 = 8 lollipops. The answer is 8.

Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?
A: Shawn started with 5 toys. He then got 2 toys each from his mom and dad. So he got 2 * 2 = 4 more toys. Now he has 5 + 4 = 9 toys. The answer is 9.

Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?
A: There were originally 9 computers. For each day from monday to thursday, 5 more computers were installed. So 4 * 5 = 20 computers were added. Now 9 + 20 = 29 computers are now in the server room. The answer is 29.

Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?
A: Michael started with 58 golf balls. He lost 23 on Tuesday, and lost 2 more on wednesday. So he had 58 - 23 = 35 at the end of Tuesday, and 35 - 2 = 33 at the end of wednesday. The answer is 33.

Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
A: Olivia had 23 dollars. She bought 5 bagels for 3 dollars each. So she spent 5 * 3 = 15 dollars. Now she has 23 - 15 = 8 dollars left. The answer is 8.

Table 9: Full prompt for Chain-of-Thought prompting in our experiments (arithmetic reasoning).

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?
Answer: Theodor Haecker was 65 years old when he died. Harry Vaughan Watkins was 69 years old when he died. So the final answer (the name of the person) is: Harry Vaughan Watkins.

Question: Why did the founder of Versus die?
Answer: Versus was founded by Gianni Versace. Gianni Versace was shot and killed on July 15, 1997. So the final answer (reason of death) is: Shot.

Question: Who is the grandchild of Dambar Shah?
Answer: Dambar Shah (? - 1645) was the father of Krishna Shah. Rudra Shah was the child of Krishna Shah (? - 1661). So the final answer (the name of the grandchild) is: Rudra Shah.

Question: Are both director of film FAQ: Frequently Asked Questions and director of film The Big Money from the same country?
Answer: The director of the film FAQ: Frequently Asked Questions is Carlos Atanes. The director of the film The Big Money is John Paddy Carstairs. The nationality of Carlos Atanes is Spanish. The nationality of John Paddy Carstairs is British. Spanish is not equal to British. So the final answer (whether they have the same nationality) is: No.

Table 10: Full prompt for Chain-of-Thought prompting in our experiments (factual QA).

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?
A: There are 15 trees originally. Then there were 21 trees after the Grove workers planted some more. Now 15 + 21 = 36. Since there were 6 workers in the grove, so the grove workers planted 36 / 6 = 6 trees today. The answer is 6.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
A: There are originally 3 cars. Then 2 more cars arrive. Now 3 * 2 = 6 cars come. So 6 - 1 = 5 cars are in the parking lot. The answer is 5.

Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
A: Originally, Leah had 32 chocolates and her sister had 42. So her sister had 42 - 32 = 10 chocolates more than Leah has. After eating 35, since 10 + 35 = 45, they had 45 - 6 = 39 pieces left in total. The answer is 39.

Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
A: Jason had 20 lollipops originally. Then he had 12 after giving some to Denny. Now 20 + 12 = 32. Jason has 4 times what Denny has, so he gave Denny 32 / 4 = 8 lollipops. The answer is 8.

Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?
A: Shawn started with 5 toys. He then got 2 toys each from his mom and dad. Now 5 - 2 = 3. So he has 3 * 3 = 9 toys now for Christmas. The answer is 9.

Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?
A: There were originally 9 computers. For each day from monday to thursday, 5 more computers were installed. Now 9 * 5 = 45 computers. Since 4 * 4 = 16, now 45 - 16 = 29 computers are now in the server room. The answer is 29.

Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?
A: Michael started with 58 golf balls. He lost 23 on Tuesday, and lost 2 more on wednesday. So compared with wednesday, he lost 23 - 2 = 21 more balls on Tuesday. So he had 58 - 21 = 37 golf balls at the end of wednesday. The answer is 37.

Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
A: Olivia had 23 dollars. She bought 5 bagels for 3 dollars each. So she earned 23 - 5 = 18 dollars. Now 18 / 3 = 6. So she has 6 + 2 = 8 dollars left. The answer is 8.

Table 11: Full prompt for "invalid reasoning" setting (arithmetic reasoning).

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?
Answer: Theodor Haecker wrote an essay, Kierkegaard and the Philosophy of Inwardness in 1913. Harry Vaughan Watkins played his final Wales international against England in January 1906. So the final answer (the name of the person) is: Theodor Haecker.

Question: Why did the founder of Versus die?
Answer: Versus was a diffusion line of the Italian luxury fashion house Versace, which began in 2009. 2009 is the year American singer Michael Jackson died of acute propofol and benzodiazepine intoxication. So the final answer (reason of death) is: Intoxication.

Question: Who is the grandchild of Dambar Shah?
Answer: Dambar Shah (? - 1645) was the king of the Gorkha Kingdom. The Gorkha Kingdom was established by Prince Dravya Shah. Dravya Shah has a child named Rudra Shah. So the final answer (the name of the grandchild) is: Rudra Shah.

Question: Are both director of film FAQ: Frequently Asked Questions and director of film The Big Money from the same country?
Answer: FAQ: Frequently Asked Questions is a feature-length dystopian movie. The Big Money is a 1958 comedy film. Dystopian stories mostly take place in British. Comedy stories mostly happen in Australia. British is not equal to Australia. So the final answer (whether they have the same nationality) is: No.

Table 12: Full prompt for "invalid reasoning" setting (factual QA).

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?
A: There are 21 - 15 = 6 trees originally. Then there were 15 trees after the Grove workers planted some more. So there must have been 21 trees that were planted. The answer is 6.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
A: There are originally 3 + 2 = 5 cars. Then 3 more cars arrive. Now 2 cars are in the parking lot. The answer is 5.

Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
A: Originally, Leah had 32 + 42 = 74 chocolates and her sister had 32. So in total they had 74 - 35 = 39. After eating 35, they had 42 pieces left in total. The answer is 39.

Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
A: Jason had 20 - 12 = 8 lollipops originally. Then he had 20 after giving some to Denny. So he gave Denny 12 lollipops. The answer is 8.

Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?
A: Shawn started with 4 toys. He then got 5 + 4 = 9 toys each from his mom and dad. So he got 5 more toys. Now he has 2 * 2 = 4 toys. The answer is 9.

Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?
A: There were originally 5 computers. For each day from monday to thursday, 4 * 5 = 20 more computers were installed. So 9 + 20 = 29 computers were added. Now 9 computers are now in the server room. The answer is 29.

Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?
A: Michael started with 2 golf balls. He lost 23 on Tuesday, and lost 35 - 2 = 33 more on wednesday. So he had 58 at the end of Tuesday, and 58 - 23 = 35 at the end of wednesday. The answer is 33.

Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
A: Olivia had 5 * 3 = 15 dollars. She bought 5 bagels for 23 - 15 = 8 dollars each. So she spent 3 dollars. Now she has 23 dollars left. The answer is 8.

Table 13: Full prompt for “no coherence for bridging objects” setting (arithmetic reasoning).

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?
Answer: 65 was Harry Vaughan Watkins years old when he died. 65 was 69 years old when he died. Theodor Haecker is bigger than 69. So the final answer (the name of the person) is: Harry Vaughan Watkins.

Question: Why did the founder of Versus die?
Answer: Versus was shot and founded. Gianni Versace was killed on July 15, 1997 by Gianni Versace. So the final answer (reason of death) is: Shot.

Question: Who is the grandchild of Dambar Shah?
Answer: Krishna Shah was the father of Rudra Shah. Dambar Shah (? - 1645) was the child of Krishna Shah (? - 1661). So the final answer (the name of the grandchild) is: Rudra Shah.

Question: Are both director of film FAQ: Frequently Asked Questions and director of film The Big Money from the same country?
Answer: The director of John Paddy Carstairs is John Paddy Carstairs. The director of British is Spanish. The nationality of Carlos Atanes is British. The nationality of John Paddy Carstairs is film FAQ: Frequently Asked Questions. Carlos Atanes is not equal to film The Big Money. So the final answer (whether they have the same nationality) is: No.

Table 14: Full prompt for “no coherence for bridging objects” setting (factual QA).

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?
A: There are 4 trees originally. Then there were 8 trees after the Grove workers planted some more. So there must have been 8 - 4 = 4 trees that were planted. The answer is 4.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
A: There are originally 18 cars. Then 9 more cars arrive. Now 18 + 9 = 27 cars are in the parking lot. The answer is 27.

Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
A: Originally, Leah had 19 chocolates and her sister had 31. So in total they had 19 + 31 = 50. After eating 29, they had 50 - 29 = 21 pieces left in total. The answer is 21.

Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
A: Jason had 37 lollipops originally. Then he had 14 after giving some to Denny. So he gave Denny 37 - 14 = 23 lollipops. The answer is 23.

Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?
A: Shawn started with 8 toys. He then got 6 toys each from his mom and dad. So he got 6 * 2 = 12 more toys. Now he has 8 + 12 = 20 toys. The answer is 20.

Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?
A: There were originally 23 computers. For each day from monday to thursday, 10 more computers were installed. So 4 * 10 = 40 computers were added. Now 23 + 40 = 63 computers are now in the server room. The answer is 63.

Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?
A: Michael started with 46 golf balls. He lost 27 on Tuesday, and lost 6 more on wednesday. So he had 46 - 27 = 19 at the end of Tuesday, and 19 - 6 = 13 at the end of wednesday. The answer is 13.

Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
A: Olivia had 48 dollars. She bought 7 bagels for 6 dollars each. So she spent 7 * 6 = 42 dollars. Now she has 48 - 42 = 6 dollars left. The answer is 6.

Table 15: Full prompt for "no relevance for bridging objects" setting (arithmetic reasoning).

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?
Answer: Albin Barack was 49 years old when he died. Carl Clemens was 55 years old when he died. 55 is bigger than 49. So the final answer (the name of the person) is: Carl Clemens.

Question: Why did the founder of Versus die?
Answer: The gang was founded by John Vitti. John Vitti drowned and got killed on February 2009. So the final answer (reason of death) is: drowning.

Question: Who is the grandchild of Dambar Shah?
Answer: Metis Amando was the father of David Amando. Randall Amando was the child of David Amando. So the final answer (the name of the grandchild) is: Randall Amando.

Question: Are both director of film FAQ: Frequently Asked Questions and director of film The Big Money from the same country?
Answer: The director of "The Forgortten Bride" is Paul Cuevas. The director of "Grace and the Rose" is Ronnie Dixon. The nationality of Paul Cuevas is Australia. The nationality of Ronnie Dixon is France. Australia is not equal to France. So the final answer (whether they have the same nationality) is: No.

Table 16: Full prompt for "no relevance for bridging objects" setting (factual QA).

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?
A: Then there were 15 trees after the Grove workers planted some more. So there must have been 21 trees that were planted. There are 21 - 15 = 6 trees originally. The answer is 6.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
A: Then 3 more cars arrive. Now 2 cars are in the parking lot. There are originally 3 + 2 = 5 cars. The answer is 5.

Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
A: After eating 32, they had 42 pieces left in total. Originally, Leah had 32 + 42 = 74 chocolates and her sister had 35. So in total they had 74 - 35 = 39. The answer is 39.

Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
A: Then he had 20 after giving some to Denny. So he gave Denny 12 lollipops. Jason had 20 - 12 = 8 lollipops originally. The answer is 8.

Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?
A: Now he has 5 toys. So he got 2 more toys. Shawn started with 2 * 2 = 4 toys. He then got 5 + 4 = 9 toys each from his mom and dad. The answer is 9.

Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?
A: So 9 computers were added. Now 5 computers are now in the server room. There were originally 4 * 5 = 20 computers. For each day from monday to thursday, 9 + 20 = 29 more computers were installed. The answer is 29.

Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?
A: So he had 58 at the end of Tuesday, and 23 at the end of wednesday. He lost 2 on Tuesday, and lost 58 - 23 = 35 more on wednesday. Michael started with 35 - 2 = 33 golf balls. The answer is 33.

Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
A: Now she has 23 dollars left. So she spent 5 dollars. Olivia had 3 dollars. She bought 5 * 3 = 15 bagels for 23 - 15 = 8 dollars each. The answer is 8.

Table 17: Full prompt for "no coherence for language template" setting (arithmetic reasoning).

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?
Answer: Theodor Haecker is bigger than 65. Harry Vaughan Watkins was 69 years old when he died. 69 was 65 years old when he died. So the final answer (the name of the person) is: Harry Vaughan Watkins.

Question: Why did the founder of Versus die?
Answer: Versus was killed on July 15, 1997. Gianni Versace was founded by Gianni Versace and shot. So the final answer (reason of death) is: Shot.

Question: Who is the grandchild of Dambar Shah?
Answer: Dambar Shah (? - 1645) was the child of Krishna Shah. Krishna Shah (? - 1661) was the father of Rudra Shah. So the final answer (the name of the grandchild) is: Rudra Shah.

Question: Are both director of film FAQ: Frequently Asked Questions and director of film The Big Money from the same country?
Answer: The nationality of film FAQ: Frequently Asked Questions is not equal to Carlos Atanes. The nationality of film The Big Money is John Paddy Carstairs. The director of Carlos Atanes is Spanish. The director of John Paddy Carstairs is British. Spanish is British. So the final answer (whether they have the same nationality) is: No.

Table 18: Full prompt for "no coherence for language template" setting (factual QA).

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?
A: Then there were 21 - 15 = 6 trees after the Grove workers planted some more. So there must have been 15 trees that were planted. There are 21 trees originally. The answer is 6.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
A: Then 3 + 2 = 5 more cars arrive. Now 3 cars are in the parking lot. There are originally 2 cars. The answer is 5.

Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
A: After eating 32 + 42 = 74, they had 32 pieces left in total. Originally, Leah had 74 - 35 = 39 chocolates and her sister had 35. So in total they had 42. The answer is 39.

Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
A: Then he had 20 - 12 = 8 after giving some to Denny. So he gave Denny 20 lollipops. Jason had 12 lollipops originally. The answer is 8.

Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?
A: Now he has 4 toys. So he got 5 + 4 = 9 more toys. Shawn started with 5 toys. He then got 2 * 2 = 4 toys each from his mom and dad. The answer is 9.

Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?
A: So 5 computers were added. Now 4 * 5 = 20 computers are now in the server room. There were originally 9 + 20 = 29 computers. For each day from monday to thursday, 9 more computers were installed. The answer is 29.

Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?
A: So he had 2 at the end of Tuesday, and 23 at the end of wednesday. He lost 35 - 2 = 33 on Tuesday, and lost 58 more on wednesday. Michael started with 58 - 23 = 35 golf balls. The answer is 33.

Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
A: Now she has 5 * 3 = 15 dollars left. So she spent 5 dollars. Olivia had 23 - 15 = 8 dollars. She bought 3 bagels for 23 dollars each. The answer is 8.

Table 19: Full prompt for "no relevance for language template" setting (arithmetic reasoning).

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?
Answer: Theodor Haecker has 65 golf balls. Harry Vaughan Watkins has 69 golf balls. 69 balls are more than 65 balls. So the final answer (the person who has more golf balls) is: Harry Vaughan Watkins.

Question: Why did the founder of Versus die?
Answer: The leader of Versus was Gianni Versace. Gianni Versace shot three people and got into jail. So the final answer (reason for imprisonment) is: Shot.

Question: Who is the grandchild of Dambar Shah?
Answer: The husband of Dambar Shah (? - 1645) is Krishna Shah. Krishna Shah (? - 1661) has a brother called Rudra Shah. So the final answer (the name of the brother-in-law) is: Rudra Shah.

Question: Are both director of film FAQ: Frequently Asked Questions and director of film The Big Money from the same country?
Answer: The author of the film FAQ: Frequently Asked Questions is Carlos Atanes. The author of film The Big Money is John Paddy Carstairs. The wife of Carlos Atanes is from Spanish. The wife of John Paddy Carstairs is from British. Spanish is warmer than British. So the final answer (the country which is warmer) is: Spanish.

Table 20: Full prompt for "no relevance for language template" setting (factual QA).

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?
A: Then there were 21 - 15 = 6 trees after the Grove workers planted some more. So there must have been 15 trees that were planted. There are 21 trees originally. The answer is 6.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
A: Then 3 + 2 = 5 more cars arrive. Now 3 cars are in the parking lot. There are originally 2 cars. The answer is 5.

Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
A: After eating 32 + 42 = 74, they had 32 pieces left in total. Originally, Leah had 74 - 35 = 39 chocolates and her sister had 35. So in total they had 42. The answer is 39.

Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
A: Then he had 20 - 12 = 8 after giving some to Denny. So he gave Denny 20 lollipops. Jason had 12 lollipops originally. The answer is 8.

Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?
A: Now he has 4 toys. So he got 5 + 4 = 9 more toys. Shawn started with 5 toys. He then got 2 * 2 = 4 toys each from his mom and dad. The answer is 9.

Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?
A: So 5 computers were added. Now 4 * 5 = 20 computers are now in the server room. There were originally 9 + 20 = 29 computers. For each day from monday to thursday, 9 more computers were installed. The answer is 29.

Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?
A: So he had 2 at the end of Tuesday, and 23 at the end of wednesday. He lost 35 - 2 = 33 on Tuesday, and lost 58 more on wednesday. Michael started with 58 - 23 = 35 golf balls. The answer is 33.

Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
A: Now she has 5 * 3 = 15 dollars left. So she spent 5 dollars. Olivia had 23 - 15 = 8 dollars. She bought 3 bagels for 23 dollars each. The answer is 8.

Table 21: Full prompt for “no coherence” setting (arithmetic reasoning).

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?
Answer: 65 is bigger than Harry Vaughan Watkins. 65 was 69 years old when he died. Theodor Haecker was 69 years old when he died. So the final answer (the name of the person) is: Harry Vaughan Watkins.

Question: Why did the founder of Versus die?
Answer: Versus was shot and killed on July 15, 1997. Gianni Versace was founded by Gianni Versace. So the final answer (reason of death) is: Shot.

Question: Who is the grandchild of Dambar Shah?
Answer: Krishna Shah was the child of Rudra Shah. Dambar Shah (? - 1645) was the father of Krishna Shah (? - 1661). So the final answer (the name of the grandchild) is: Rudra Shah.

Question: Are both director of film FAQ: Frequently Asked Questions and director of film The Big Money from the same country?
Answer: The nationality of John Paddy Carstairs is not equal to John Paddy Carstairs. The nationality of British is Spanish. The director of Carlos Atanes is British. The director of John Paddy Carstairs is film FAQ: Frequently Asked Questions. Carlos Atanes is film The Big Money. So the final answer (whether they have the same nationality) is: No.

Table 22: Full prompt for “no coherence” setting (factual QA).

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?
A: Tom started with 4 apples. Then he had 8 after borrowing some from Amy. So he borrowed Amy 8 - 4 = 4. The answer is 4.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
A: Benjamin has 18 gloves originally. Then he got 9 more gloves. So he has 18 + 9 = 27 gloves now. The answer is 27.

Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
A: Patricia needs to donate 19 inches, and wants her hair to be 31 inches long after the donation. Her hair is 29 inches long currently. Her hair needs to be 19 + 31 = 50 inches long when she cuts it. So she needs to grow 50 - 29 = 21 more inches. The answer is 21.

Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
A: There were 37 trains originally. Then there were 14 after some were driven away. So there should be 37 - 14 = 23 that were driven away. The answer is 23.

Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?
A: The taxi has a ride fee of 8 dollars. Michelle rode the taxi for 6 miles with 2 dollars per mile. So the taxi charge is 6 * 2 = 12. So the total amount that Michelle paid for the ride was 8 + 12 = 20. The answer is 20.

Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?
A: Haley is currently 23 inches tall. She grows at the rate of 10 inches every year for 4 years. So she will have grown by 10 * 4 = 40 inches. Her height after 4 years will be 23 + 40 = 63 inches. The answer is 63.

Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?
A: Abigail had 46 dollars in her purse originally. She spent 27$inthestore, andhas$6 left now. After going shopping, she had 46 - 27 = 19 dollars left. So she lost 19 - 6 = 13 dollars. The answer is 13.

Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
A: George earned 48 in total. He sold 7 cars for 6 dollars each. So he earned 7 * 6 = 42 dollars from them. The lego set cost was then 48 - 42 = 6. The answer is 6.

Table 23: Full prompt for “no relevance” setting (arithmetic reasoning).

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?
Answer: Albin Barack has 49 golf balls. Carl Clemens has 55 golf balls. 55 balls are more than 49 balls. So the final answer (the person who has more golf balls) is: Carl Clemens.

Question: Why did the founder of Versus die?
Answer: The leader of the gang was John Vitti. John Vitti drowned three people and got into jail. So the final answer (reason for imprisonment) is: drowning.

Question: Who is the grandchild of Dambar Shah?
Answer: The husband of Metis Amando is David Amando. David Amando has a brother called Randall Amando. So the final answer (the name of the brother-in-law) is: Randall Amando.

Question: Are both director of film FAQ: Frequently Asked Questions and director of film The Big Money from the same country?
Answer: The author of “The Forgortten Bride” is Paul Cuevas. The author of “Grace and the Rose” is Ronnie Dixon. The wife of Paul Cuevas is from Spanish. The wife of Ronnie Dixon is from British. Spanish is warmer than British. So the final answer (the country which is warmer) is: Spanish.

Table 24: Full prompt for “no relevance” setting (factual QA).