

Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters

Boshi Wang¹ Sewon Min² Xiang Deng¹ Jiaming Shen³ You Wu³
Luke Zettlemoyer² Huan Sun¹

¹The Ohio State University ²University of Washington ³Google Research
{wang.13930, deng.595, sun.397}@osu.edu
{sewon, lsz}@cs.washington.edu, {jmshen, wuyou}@google.com

Abstract

Chain-of-Thought (CoT) prompting can dramatically improve the multi-step reasoning abilities of large language models (LLMs). CoT explicitly encourages the LLM to generate intermediate rationales for solving a problem, by providing a series of reasoning steps in the demonstrations. Despite its success, there is still little understanding of what makes CoT prompting effective and which aspects of the demonstrated reasoning steps contribute to its performance. In this paper, we show that CoT reasoning is possible even with invalid demonstrations—prompting with invalid reasoning steps can achieve over 80-90% of the performance obtained using 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 much more important for effective CoT reasoning. Overall, these findings both deepen our understanding of CoT prompting, and open up new questions regarding LLMs’ capability to learn to reason in context.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) improve the ability of sufficiently large LLMs to do complex and multi-step reasoning. In addition to (query, answer) example-pair demonstrations, CoT prompting includes a *rationale* (colored part in Figure 1) for each example, i.e., a series of reasoning steps towards the answer, which encourages the LLM to explicitly generate its intermediate reasoning process before predicting the final answer. Despite its successes, there is little understanding of what makes CoT prompting effective

¹Our code and model input/output are available [here](#).

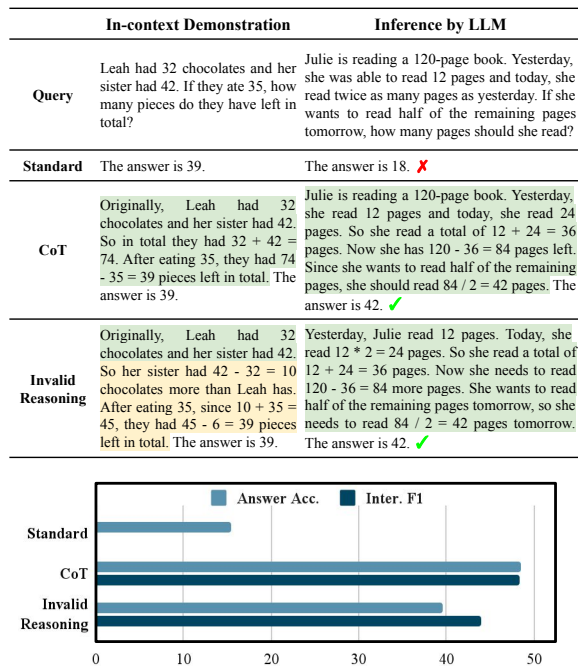


Figure 1: Results of standard prompting, Chain-of-Thought (CoT) prompting, and our ablation setting with invalid reasoning (§4). We show one demonstration example and one inference example for arithmetic reasoning, where the rationale is in color (green: valid, yellow: invalid). We find that valid reasoning for the demonstrations matters only a small portion to the performance of CoT—by providing rationales with invalid reasoning, LLMs can achieve over 80-90% of the performance of CoT under various metrics while performing logically sound and pertinent reasoning.

and which aspects of the demonstrated reasoning steps contribute to its performance. Recent findings also reveal that in-context learning could be very different from fine-tuning/training; for example, Min et al. (2022) and Webson and Pavlick (2022) show that providing random labels or misleading instructions in context only marginally harms model performance for certain tasks. Inspired by this work, we take a closer look at CoT prompting to study how and why it works.

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 (§4, §5). 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 (§4). Through further examinations, we identify and formulate other aspects of a CoT rationale (§5), 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, they have already gained a lot of such “reasoning abilities” from pretraining, and the demonstrations may mainly specify an output space/format that regularizes the model generation to look step-by-step while being in order and relevant to the query. Our work suggests a new way of interpreting the evaluation scores in view of the prior knowledge LLMs possess, **and leads to reflections on benchmarking few-shot reasoning which we discuss in §6.**

2 Background & Study Formulation

Chain-of-Thought (CoT) prompting. Different from the standard way of prompting language models where a set of (query, answer) pairs are given as demonstrations (Brown et al., 2020), CoT prompting (Wei et al., 2022) additionally includes a rationale (Figure 1, colored) for each example, encouraging the model to verbalize the intermediate reasoning steps for solving the task. Such a technique has been shown to improve the performance of LLMs with sufficient scale on complex reasoning, sometimes to a large degree especially on arithmetic reasoning, multi-hop question answering, and symbolic reasoning.

Components of a CoT rationale. We identify two distinct components of a CoT rationale (examples in Table 1):

- **Bridging objects:** the key and necessary objects that the model needs to traverse in order to make a successful final prediction. For arithmetic reasoning, the bridging objects are defined to be the

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.

numeric part (numbers & equations) of the rationale, and for factual QA, the bridging objects are defined to be the subject & object entities.

- **Language templates:** the complementary parts of bridging objects, which serve as textual hints and relations/predicates that guide the model to derive the correct bridging objects along the reasoning process.

Research questions. In Chain-of-Thought prompting, correct bridging objects and language templates are provided 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 §4 and §5.

3 Experimental Setup

3.1 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, as shown in previous work (Wei et al., 2022; Press et al., 2022); they are more suitable for our study, since our goal is to understand how different aspects of the Chain-of-Thought rationales contribute to the performance of CoT prompting. For arithmetic reasoning, we experiment on GSM8K (Cobbe et al., 2021), one of the most challeng-

ing mathematical reasoning benchmarks available which is also repeatedly adopted by prior work as a key benchmark for arithmetic reasoning; for multi-hop factual 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 GSM8K for evaluation. We evaluate on all 125 test samples for Bamboogle.

We base our experiments on the original prompt exemplars, i.e., the set of (query, rationale, answer) pairs released by Wei et al. (2022) and Press et al. (2022), with slight editing to make the structure more consistent and reduce redundancy, which makes our ablations more convenient to conduct. These edits only slightly affect the performance of CoT; we show our edited demonstration examples and include more details in Appendix A.1.

3.2 Backbone Language Model

We use InstructGPT-175B² (Ouyang et al., 2022; Brown et al., 2020) text-davinci-002 as our backbone LLM, which is one of the most performant and widely-used LLMs with public APIs and has demonstrated strong performance under CoT prompting (Wei et al., 2022). We report its results and analyze them in the main content. In addition, we also test on text-davinci-003 (a very recent improved version of text-davinci-002), PaLM (Chowdhery et al., 2022) and Flan-PaLM (Chung et al., 2022), where the results and discussion could be found in Appendix A.3. All generations are done by greedy decoding (i.e., sampling with zero temperature) as in the original CoT work (Wei et al., 2022).

3.3 Evaluation

Prior work mainly performs evaluation using the correctness of the final answer, which could be viewed as an *extrinsic* way of assessing the predicted rationales. However, this may not align well with the actual quality of the rationale in many cases, as mentioned in Huang and Chang (2022). For example, a rationale that is correct for all but the last step (and hence derives the wrong final answer) would still be assigned a zero score, while a rationale that is wrong/incomplete but reaches the correct final answer would be assigned a full

score. Therefore, in addition to extrinsic evaluation (**Answer Accuracy** for GSM8K, **Answer F1** for Bamboogle), we perform *intrinsic* evaluation where we measure the Recall/F1 (**Inter.³ Recall/F1**) of the bridging objects which need to be derived by the LLM (i.e., those that do not appear in the query). 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.⁴ For Bamboogle, we manually annotate the bridging objects (intermediate entities) and measure their recall. While it is still possible for the model to reach correct bridging objects with the wrong language templates, we manually verify that this rarely happens; details are included in Appendix A.2.

4 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), since we are teaching the LLM to reason in the wrong way which could be even worse than not doing so at all. 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.

4.1 Constructing Invalid Chain of Reasoning

We manually write rationales with invalid reasoning for *all* the in-context demonstration examples. Since our focus here is to investigate the importance of the validity of reasoning, we only ablate the parts in a CoT rationale which are involved with derivations that are logically sound and helpful for answering the query. More specifically, 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

²We also tried the original GPT-3 175B without instruction-finetuning in our preliminary experiments, but find that CoT prompting does not yield much performance gain than standard prompting, echoing Fu et al. (2022).

³Abbreviation for “Intermediate”.

⁴We 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”.

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. The full prompts in our setting are included in Appendix A.4.

For example, consider an in-context demonstration (see ① in Table 4) for arithmetic reasoning. Here the query is “*Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?*”. For the 1st entailment step which should sum “32” and “42” to get the total amount “ $32 + 42 = 74$ ” as in CoT, we instead write “*So her sister had $42 - 32 = 10$ chocolates more than Leah has.*” which has both the wrong bridging object and language template, and is completely unhelpful for solving the problem. The subsequent steps are written based on the previous steps, and in the end, answer the question whereas the rationale does not in any way lead to the answer logically. While the step itself still describes something that could be entailed in the example we just gave, this is not the case generally and most of the steps we write are neither helpful nor entailments from earlier steps. For example, the next step “*After eating 35, since $10 + 35 = 45$, they had $45 - 6 = 39$ pieces left in total*” makes use of unwarranted information (“6”) and has no valid entailment anywhere. We illustrate our construction using another example for factual QA, where the question is “*Who is the grandchild of Dambar Shah?*”. Here, we write a rationale that finds the kingdom of “*Dambar Shah*” and then a child of the person who established the kingdom, which does not lead to “*the grandchild of Dambar Shah*”.

4.2 Results & Analysis

Quantitative results. Table 2 summarizes the quantitative results for text-davinci-002. We include additional results and discussion for text-davinci-003, PaLM and Flan-PaLM in Appendix A.3. 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 (§3.3), all LLMs we tested can retain over 90% of the performance achieved under CoT prompting.

For GSM8K where there are large variations in the difficulty levels (here, we use the number of

reasoning steps required to solve a problem as its difficulty level) of the problem instances, we additionally examine the model performance separately for each difficulty level. The results are shown in Figure 2. The performance drop is also uniform across samples with different levels of difficulty. At the instance level, after omitting samples where both settings get the correct/wrong answer, there is a significant portion for the remaining ones (62/196 for GSM8K, 6/20 for Bamboogle) where CoT gets the wrong answer and the invalid reasoning setting gets the correct answer. This further strengthens the finding that there is no strong connection between the reasoning validity of the demonstrations and the quality of the model predictions.

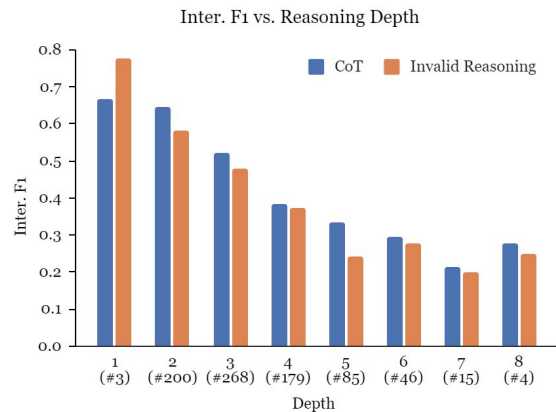


Figure 2: Model performance using CoT and demonstrations with invalid reasoning for examples with different reasoning depths on GSM8K. The number of samples for each reasoning depth is shown below (led by “#”). The performance drop is consistent across different levels of difficulty.

Qualitative analysis. By checking the generated rationales for the invalid reasoning setting, we find that overall they look indistinguishable from the rationales generated by CoT prompting. 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. To compare the distribution of errors between CoT and the invalid reasoning setting, we examine 20 samples from GSM8K where CoT gets the correct final answer and the invalid reasoning setting gets the wrong answer, and another 20 examples for the opposite case. We use the same error categorizations as in

	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 text-davinci-003, PaLM and Flan-PaLM could be found in Appendix A.3.

Error Types	CoT correct & IR wrong	CoT wrong & IR correct
Calculation	20%	20%
One step missing	35%	25%
Semantic understanding	45%	55%

Table 3: Distribution of error types of 20 examples from GSM8K where Chain-of-Thought (CoT) prompting reaches the correct answer and the Invalid Reasoning setting (IR) reaches a wrong answer, and 20 examples for the opposite case.

Wei et al. (2022) for the qualitative analysis, and summarize the results in Table 3. The distributions of errors in both cases are highly similar.

Summary. Combining the quantitative and qualitative results, we can see that there is a low chance for any systematic difference between CoT and the invalid reasoning setting to exist. The LLM still tries and manages to generate logically sound and pertinent reasoning decently, and ablating the validity of reasoning for the demonstrations only brings a small performance degradation. This opens the question: *If valid reasoning is not required, what are the key aspects that determine the effectiveness of CoT prompting?*

5 What are the Key Aspects of Chain-of-Thoughts?

Re-examining the rationales in our ablation setting in §4, we can find that even though the reasoning is invalid, they have the following properties:

- The rationales still use information from the query; more specifically, they still start from bridging objects mentioned in the query, and the

language templates are related to the query. Recall our running example for arithmetic reasoning (Table 4), even though the reasoning here is wrong, the numbers “32” and “42” are kept from the query, and the language templates are still about “Leah”, “Leah’s sister” and “Chocolates”, and try to seek the answer to the query. Therefore, the rationale is still relevant to the query being asked.

- Each step of a rationale still follows the previous steps. Using again the same example, the bridging object (equation in this case) “ $42 - 32 = 10$ ” in the first entailment step uses numbers from previous steps; likewise, the language template “*So her sister had _ chocolates more than Leah has*” is something that follows after the earlier steps. Hence, overall, the rationale still appears to be coherent.

We formulate two notions that capture these two aspects of a rationale in what follows.

Relevance. A component of the rationale has relevance if it is based on the corresponding component from the query. For bridging objects, this could be formally defined as using the exact same objects mentioned in the query (numbers for arithmetic reasoning and entities for factual QA); for language templates, they have relevance if they are still 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 similarly, a template that attempts to find the “brother-in-law” of the topic entity does not have relevance to a query which seeks the “grandchild” (Table 4).

Coherence. A component of the rationale has coherence if it 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.

In what follows, we design a set of ablation settings to examine the impact of these two aspects for different components of a CoT-like rationale.

5.1 Ablation Settings

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 §4.

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 4).

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.

5.2 Results & Analysis

The results could be found in Table 2, and we include additional results for text-davinci-003,

PaLM and Flan-PaLM in Appendix A.3. We summarize the main findings in what follows.

Relevance and coherence are key for the performance of CoT prompting. It can be seen that most of the 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 (STD) where no rationale is given on GSM8K. To see why such low performance happens, we manually examine the generated rationales under this setting for 20 examples on GSM8K. We find that the LLM is indeed generating irrelevant rationales (both bridging objects and language templates) for 15 out of 20 examples. Many of the 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 (below 60%) than other settings (around 80%).

Coherence of language templates is important. Different from the coherence of bridging objects ②, the 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), which negatively affects reasoning.

Prompt Setting	Example Query (Arithmetic Reasoning) <i>Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?</i>	Example Query (Factual QA) <i>Who is the grandchild of Dambar Shah?</i>
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 <i>coherence</i> for <i>bridging objects</i>	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 <i>relevance</i> for <i>bridging objects</i>	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 <i>coherence</i> for <i>language templates</i>	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 <i>relevance</i> for <i>language templates</i>	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 <i>coherence</i>	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 <i>relevance</i>	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 4: Examples for all settings in our experiments.

6 Discussion

The results from §4 and §5 open up new questions regarding learning to reason in context for LLMs, which we discuss next.

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, 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. Moreover, results obtained from recent stronger models including text-davinci-003 and Flan-PaLM (see Appendix A.3) suggest that LLMs

suffer further less from the ablations when they have more prior knowledge about the task. In particular, for Flan-PaLM which is directly trained on both arithmetic reasoning and factual QA in CoT fashion and hence has immense knowledge on these tasks (Chung et al., 2022), 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., 2023).

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., 2023). 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.

7 Related Work

There have been several subsequent work of Chain-of-Thought prompting since its introduction. Wang et al. (2023) proposes to sample a diverse set of reasoning paths instead of performing greedy decoding, and marginalize over the sampled paths to select the most consistent answer. Zhang et al. (2023) proposes a method for automatically constructing the in-context exemplars for CoT. Chen et al. (2022) explores program-based CoT which can better disentangle computation from reasoning. 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.

A few recent work focuses on understanding/analyzing CoT prompting. Madaan and 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 the model performance could be affected by corrupting the CoT rationales. They experiment with using incorrect calculations and *dropping* (parts of) the bridging objects/language templates, which are different from our ablation designs. Saparov and He (2023) 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.

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, Yoo et al. (2022) finds that the correct input-label correspondence could have varying impacts based on the task and experimental configurations, and Wei et al. (2023) finds that models with larger scale can override semantic priors and learn input-label mapping in context. Webson and 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 under-

stand ICL in sequence generation tasks requiring multi-step reasoning.

8 Conclusion

In this paper, we aim to better understand Chain-of-Thought prompting through a series of ablation experiments that unveil the impact of different aspects of a CoT rationale. We find that 1) the validity of reasoning in the prompting examples matters only a small portion to the performance; 2) relevance to the input query and following the order along the reasoning steps are the key to the effectiveness of CoT prompting. Overall, our findings deepen the understanding of CoT prompting, and open up new questions/reflections regarding LLMs’ capability of learning to reason in context.

Limitations

Experiments on other types of reasoning tasks.

In addition to the two representative reasoning tasks (arithmetic reasoning and multi-hop question answering) that we experiment on, there are also other tasks where CoT prompting brings significant improvements over standard prompting shown by previous work, many of which are symbolic reasoning tasks such as Last letter concatenation, Coin flip from Wei et al. (2022) and Temporal Sequences, Tracking Shuffled Objects from BIG-Bench (Srivastava et al., 2022; Suzgun et al., 2022). However, most (if not all) tasks there are highly *template-based* and hence the reasoning steps have little variations, both within each example and across different examples. This makes it difficult for us to conduct our ablation studies on these tasks. Take the example of Last letter concatenation, a task about concatenating the last letters of a given sequence of words (e.g., “Amy Brown” → “yn”). Here, every step in the rationale except the last is in the form “*The last letter of X is Y*” where *X* is some word in the given sequence and *Y* is the last letter of *X*. Hence, the language templates are the same and there is no sense of order among the steps (the order is completely characterized by the given sequence instead), and our ablation settings will not apply well. Extending our ablation designs to these “reduced” cases is one of the items we want to explore in the future.

A more systematic treatment of “invalid reasoning”. We manually write rationales with invalid reasoning for the experiments in §4 since automatically synthesizing such rationales turns out to be

challenging, mostly due to the informal nature of the tasks we experiment on (relatedly, the original CoT rationales are also human-written). We intend to give a more systematic treatment of the invalid reasoning setting in the future, e.g., following the categorizations of informal logical fallacies (Copi et al., 2016).

Improvements on intrinsic evaluation. Our intrinsic evaluation of the generated rationales is based on the correctness of bridging objects, which, even though is a good indicator of the quality of language templates (Appendix A.2) in our experiments, may not be a good metric in general cases. It also relies on ground truth bridging objects, which are usually not available and costly to annotate. Toward this end, one direction we want to explore further is to develop ways to conduct more comprehensive and reference-free intrinsic evaluations. Recent papers such as Golovneva et al. (2023) have also done promising work along this line.

Acknowledgements

The authors would like to thank the anonymous reviewers and colleagues from the OSU NLP group for their thoughtful comments. This research was supported in part by Google Faculty Award, Google Research Scholar Award, NSF IIS 1815674, NSF CAREER 1942980, NSF OAC-2112606, and Ohio Supercomputer Center (Center, 1987). The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notice herein.

References

- Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2023. [What learning algorithm is in-context learning? investigations with linear models](#). In *The Eleventh International Conference on Learning Representations*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. [Language models are few-shot learners](#). *Advances in neural information processing systems*, 33:1877–1901.
- Ohio Supercomputer Center. 1987. [Ohio supercomputer center](#).

- Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. [Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks](#). *arXiv preprint arXiv:2211.12588*.
- François Chollet. 2019. [On the measure of intelligence](#). *arXiv preprint arXiv:1911.01547*.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. [Palm: Scaling language modeling with pathways](#). *arXiv preprint arXiv:2204.02311*.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. [Scaling instruction-finetuned language models](#). *arXiv preprint arXiv:2210.11416*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *arXiv preprint arXiv:2110.14168*.
- Irving Copi, Carl Cohen, and Victor Rodych. 2016. *Introduction to logic*. Routledge.
- Yao Fu, Hao Peng, and Tushar Khot. 2022. [How does gpt obtain its ability? tracing emergent abilities of language models to their sources](#). *Yao Fu’s Notion*.
- Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. 2022. [What can transformers learn in-context? a case study of simple function classes](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 30583–30598. Curran Associates, Inc.
- Olga Golovneva, Moya Peng Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. [ROSCOE: A suite of metrics for scoring step-by-step reasoning](#). In *The Eleventh International Conference on Learning Representations*.
- Jie Huang and Kevin Chen-Chuan Chang. 2022. [Towards reasoning in large language models: A survey](#). *arXiv preprint arXiv:2212.10403*.
- Joel Jang, Seonghyeon Ye, and Minjoon Seo. 2023. [Can large language models truly understand prompts? a case study with negated prompts](#). In *Transfer Learning for Natural Language Processing Workshop*, pages 52–62. PMLR.
- Aman Madaan and Amir Yazdanbakhsh. 2022. [Text and patterns: For effective chain of thought, it takes two to tango](#). *arXiv preprint arXiv:2209.07686*.
- Sewon Min, Xixi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. [Rethinking the role of demonstrations: What makes in-context learning work?](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. [Training language models to follow instructions with human feedback](#). *arXiv preprint arXiv:2203.02155*.
- Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. [Measuring and narrowing the compositionality gap in language models](#). *arXiv preprint arXiv:2210.03350*.
- Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. [Impact of pretraining term frequencies on few-shot numerical reasoning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Abulhair Saparov and He He. 2023. [Language models are greedy reasoners: A systematic formal analysis of chain-of-thought](#). In *The Eleventh International Conference on Learning Representations*.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. [Beyond the imitation game: Quantifying and extrapolating the capabilities of language models](#). *arXiv preprint arXiv:2206.04615*.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. [Challenging big-bench tasks and whether chain-of-thought can solve them](#). *arXiv preprint arXiv:2210.09261*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). In *The Eleventh International Conference on Learning Representations*.
- Albert Webson and Ellie Pavlick. 2022. [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*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In

- Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023. [Larger language models do in-context learning differently](#). *arXiv preprint arXiv:2303.03846*.
- Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. [An explanation of in-context learning as implicit bayesian inference](#). In *International Conference on Learning Representations*.
- Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Ves Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2022. [Complementary explanations for effective in-context learning](#). *arXiv preprint arXiv:2211.13892*.
- Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. 2022. [Ground-truth labels matter: A deeper look into input-label demonstrations](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2422–2437, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. [Automatic chain of thought prompting in large language models](#). In *The Eleventh International Conference on Learning Representations*.

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 5.

A.2 More Details on Intrinsic Evaluation

We use Recall/F1 of the bridging objects as the metrics for intrinsic evaluation of the generated rationales. 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.

A.3 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 Flan-PaLM (Chung et al., 2022), the instruction-tuned version of PaLM (Chowdhery et al., 2022) 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., 2023).

We also test PaLM, 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 PaLM 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.

A.4 Full List of Prompts

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

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

	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 <i>coherence</i> for bridging objects	46.5	51.5	50.4	59.2	55.2
③ No <i>relevance</i> for bridging objects	32.5	38.3	47.2	60.4	56.9
④ No <i>coherence</i> for language templates	37.8	43.3	41.9	57.2	51.4
⑤ No <i>relevance</i> for language templates	44.6	49.9	51.8	62.4	59.3
⑥ No <i>coherence</i>	34.5	39.4	31.0	57.6	55.2
⑦ No <i>relevance</i>	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.3.

	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 <i>coherence</i> for bridging objects	72.1	72.9	65.8	51.6	49.3
③ No <i>relevance</i> for bridging objects	71.1	71.9	64.6	54.0	52.8
④ No <i>coherence</i> for language templates	71.6	72.2	63.9	54.0	52.0
⑤ No <i>relevance</i> for language templates	71.9	72.7	64.9	55.2	53.5
⑥ No <i>coherence</i>	71.7	72.5	64.2	54.4	54.0
⑦ No <i>relevance</i>	70.7	71.6	64.5	50.0	51.9

Table 7: Intrinsic and extrinsic evaluation results under Flan-PaLM (Chung et al., 2022), the instruction-tuned version of PaLM for all settings. Discussions are included in Appendix A.3.

	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 <i>coherence</i> for bridging objects	30.3	35.0	33.5	33.6	25.7
③ No <i>relevance</i> for bridging objects	15.5	20.1	21.2	47.2	47.7
④ No <i>coherence</i> for language templates	23.1	27.3	21.9	40.4	35.5
⑤ No <i>relevance</i> for language templates	19.5	22.9	20.4	38.4	30.6
⑥ No <i>coherence</i>	23.9	28.3	24.1	39.6	33.6
⑦ No <i>relevance</i>	12.1	16.4	16.4	28.4	14.3

Table 8: Intrinsic and extrinsic evaluation results under PaLM. Discussions are included in Appendix A.3.

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 $3 + 2 = 5$ cars come. So $5 - 1 = 4$ cars are in the parking lot. The answer is 4.
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 $5 + 2 = 7$. So he has $7 + 2 = 9$ 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 $9 + 5 = 14$ computers. Since $4 + 4 = 8$, now $14 + 8 = 22$ computers are now in the server room. The answer is 22.
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 $8 - 4 = 4$. 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 <i>in the store</i> , and has 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).

ACL 2023 Responsible NLP Checklist

A For every submission:

- ☒ A1. Did you describe the limitations of your work?
The Limitation section follows right after the Conclusion section.
- ☒ A2. Did you discuss any potential risks of your work?
We used standard techniques for obtaining model generations for our experiments on publicly accessible datasets (also adopted by previous work), which do not involve any kind of harmful/biased content. Our methods also don't in any way induce such contents from models.
- ☒ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1 is the Introduction section.
- ☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Section 3.1. We used standard datasets for our experiments.

- ☒ B1. Did you cite the creators of artifacts you used?
Section 3.1.
- ☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
The datasets we used are all publicly available for research purposes and we don't modify any of their content for our experiments. Due to space constraints, we omit this information in our paper.
- ☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We use the datasets under the same intended usage that they were created with, and we don't modify any content of these datasets in our experiments. Due to space constraints, we omit this information in our paper.
- ☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The datasets used by us don't contain any such personal information. Due to space constraints, we omit this information in our paper.
- ☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We refer the readers to the original papers which released these datasets for such documentation. Due to space constraints, we omit this information in our paper.
- ☒ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3.1.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☒ Did you run computational experiments?

Section 4,5 and Appendix A.1, A.3.

- ☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?

Section 3.2.

- ☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

Section 3.2.

- ☒ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

Due to limited budgets, we report all results with a single run. This is transparent from the paper.

- ☒ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

We didn't use any packages involved with setting custom configurations for our experiments. We also attached the code in the supplementary materials.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

- ☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

No response.

- ☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

No response.

- ☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

No response.

- ☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

No response.

- ☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

No response.