

A Frequently Asked Questions

A.1 Why does increasing model scale improve chain-of-thought prompting?

The finding that successful chain-of-thought reasoning predictably emerges only at certain model scales is intriguing. Scaling up language models has been shown to confer benefits such as improved performance and sample efficiency (Kaplan et al., 2020), but chain-of-thought reasoning is emergent in the sense that its success cannot be predicted only by extrapolating the performance of small scale models, as chain of thought actually hurts performance for most models smaller than 10B parameters.

The question of why model scale improves chain-of-thought prompting is certainly multi-faceted, and we made a preliminary attempt to shed insight into it via error analysis. This small analysis involved manually reading 45 errors made by PaLM 62B and categorizing them into semantic understanding (20 errors), one step missing (18 errors), and other errors (7 errors). The “other category” included hallucinations, repetitive outputs, and symbol mapping errors. This categorization is a coarse one borrowed from the initial error analysis done on LaMDA in Appendix D.2, for which categories were conceived based on what improvements were needed to make the chain of thought correct.

As shown in Figure 9, scaling PaLM to 540B parameters fixed a substantial portion of errors in all three categories. Examples of semantic understanding and one-step missing errors that were fixed by scaling PaLM to 540B are given in Figure 10. This result appears consistent with a hypothesis that language models acquire a range of semantic understanding and logical reasoning skills as a function of model scale (though note that model scale is often conflated with other factors, such as amount of training compute).

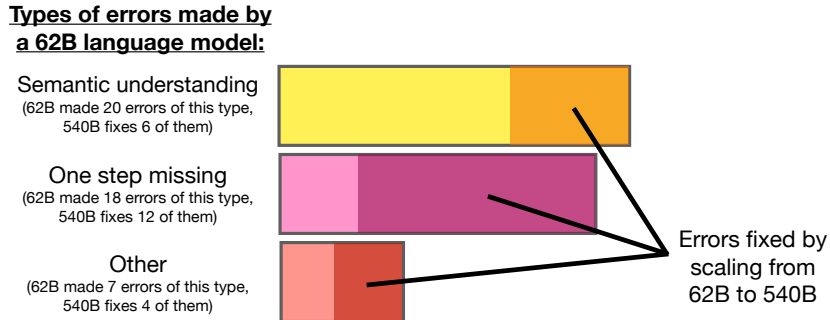


Figure 9: Error analysis of 45 problems that PaLM 62B got incorrect. These errors were categorized that semantic understanding, one step missing, and other. The other category includes hallucinations, repetitive outputs, and symbol mapping errors. Scaling PaLM to 540B fixed a substantial portion of errors in all categories.

There are also three notable points regarding why small language models fail. The first observation is that small language models fail at even relatively easy symbol mapping tasks. As demonstrated in Section 5, for even symbolic reasoning tasks that only require generalization to new examples using the same chain of thought logical structure that was given in the few-shot exemplars, small language models still failed. The second observation is that small language models seem to have inherently weaker arithmetic abilities, as shown by Brown et al. (2020), the ability to do simple arithmetic operations (without semantic understanding) requires sufficient model scale. Finally, we noticed qualitatively that small language models often did not generate a final answer that could be parsed, due to either repetitions or logic that never arrived at a final answer.

In summary, the success of chain-of-thought reasoning as a result of model scale is a complicated phenomena that likely involves a variety of emergent abilities (semantic understanding, symbol mapping, staying on topic, arithmetic ability, faithfulness, etc). Future work could more thoroughly investigate what properties of pretraining data, model architecture, and optimization objective causally enable such reasoning capabilities.

<p>Question</p> <p>Tracy used a piece of wire 4 feet long to support tomato plants in the garden. The wire was cut into pieces 6 inches long. How many pieces did she obtain?</p>	<p>62B Model Output</p> <p>She cut the wire into 6 inch pieces. This means she got $4 * 6 = 24$ pieces. The answer is 24. (Incorrect: semantic understanding error)</p>	<p>540B Model Output</p> <p>The wire was 4 feet long. This means it was $4 * 12 = 48$ inches long. It was cut into pieces 6 inches long. This means she obtained $48 / 6 = 8$ pieces. The answer is 8. (Correct)</p>
<p>Question</p> <p>Tom's ship can travel at 10 miles per hour. He is sailing from 1 to 4 PM. He then travels back at a rate of 6 mph. How long does it take him to get back?</p>	<p>62B Model Output</p> <p>He travels from 1 to 4 PM. That means he travels for 3 hours. He travels back at 6 mph. That means he travels for $3 * 6 = 18$ miles. The answer is 18. (Incorrect: semantic understanding error)</p>	<p>540B Model Output</p> <p>He travels at 10 mph from 1 to 4 PM. This means he travels 3 hours. 3 hours at 10 mph means he travels $3 * 10 = 30$ miles. He then travels back at 6 mph. This means he travels 6 miles per hour. He has to travel 30 miles, so it takes him $30 / 6 = 5$ hours. The answer is 5. (Correct)</p>
<p>Question</p> <p>Stephen placed an online order for groceries. His final bill came to \$40.00. Because this was through a delivery vendor, they tacked on a 25% fee to his final total and charged him \$3.00 in delivery fees. Stephen also added a \$4.00 tip. After the extra fees, what was the final price of Stephen's groceries?</p>	<p>62B Model Output</p> <p>The final price of Stephen's groceries was \$40.00. The delivery vendor added 25% to the final total, so the final total was 125% of \$40.00. This means the final total was $125% * \\$40.00 = \\50.00. The delivery vendor also charged \$3.00 in delivery fees. So the final price of Stephen's groceries was $\\$50.00 + \\$3.00 = \\$53.00$. The answer is \$53.00. (Incorrect: one step missing error)</p>	<p>540B Model Output</p> <p>Stephen's bill was \$40.00. The delivery vendor added 25% to the bill. This means the bill was increased by 25% of \$40.00. This is $0.25 * \\$40.00 = \\10.00. The delivery fee was \$3.00. The tip was \$4.00. So the total bill was $\\$40.00 + \\$10.00 + \\$3.00 + 540b \text{ model } \\$4.00 = \\$57.00$. The answer is 57. (Correct)</p>
<p>Question</p> <p>There are four schools competing at a basketball tournament. Each school has sent a girls' basketball team and a boys' basketball team and each team has 5 players each. Each school has also sent a coach for each team. In total, how many people have all of the schools sent?</p>	<p>62B Model Output</p> <p>There are 4 schools. Each school has sent a girls' basketball team and a boys' basketball team. Each team has 5 players each. That means there are $4 * 2 * 5 = 40$ players. Each team has a coach. That means there are $40 + 4 = 44$ coaches. The answer is 44. (Incorrect: one step missing error)</p>	<p>540B Model Output</p> <p>Each school has sent 2 teams, each with 5 players. This means each school has sent $2 * 5 = 10$ players. Each school has also sent 2 coaches. This means each school has sent $10 + 2 = 12$ people. There are 4 schools, so in total all of the schools have sent $4 * 12 = 48$ people. The answer is 48. (Correct)</p>

Figure 10: Examples of semantic understanding and one-step missing errors that were fixed by scaling PaLM from 62B to 540B.

A.2 What is the role of prompt engineering?

One of the key considerations of prompting is sensitivity to the exact prompt. There is no shortage of work showing that prompts affect language models in unexpected ways (Min et al., 2022). The general way that we created chain of thought annotations was by taking eight exemplars from the training set and decomposing the reasoning process into multiple steps leading to the final answer. Examples of chain of thought annotations are provided in Figure 3, with full prompts given in Appendix G. To analyze how sensitive chain of thought is to prompt engineering, we performed robustness experiments with respect to various factors.

- **Different annotators.** We first analyze robustness to three different annotators (Section 3.4 and Figure 6). Although there is notable variance in performance (which we will discuss later), chain of thought performed better than the baseline by a large margin for all three annotators on eight datasets in arithmetic, commonsense, and symbolic reasoning (Table 6 and Table 7). Similar to the annotation process in Cobbe et al. (2021), annotators were not given specific instructions about

how to write the chain of thought annotations other than to simply write the step-by-step reasoning process that led to the final answer. Thus, the annotations were written in each annotator’s own linguistic “chain of thought” writing style.

- **Annotators without machine learning background.** The GSM8K dataset (Cobbe et al., 2021) conveniently provides a training set with reasoning chains written by crowd compute workers, which enables us to investigate whether chain of thought still works with reasoning chains from an independent source without a background in machine learning. So we randomly sampled three sets of eight exemplars with chains of thought from GSM8K. These chain of thought annotations also outperformed the baseline by a large margin for all four arithmetic datasets (Table 6), indicating that chain of thought is not dependent on a particular set of annotators.
- **Different exemplars.** The different GSM8K exemplars experiment above (Table 6) also shows that chain-of-thought prompting works for different sets of exemplars. Notably, we test every set of exemplars on all four arithmetic datasets (instead of picking exemplars from the training set for each dataset), which suggests that the exemplars do not necessarily have to come from the same dataset distribution as the test examples.
- **Different order of exemplars.** Prior work has shown that in some cases (e.g., classification) even the order of prompts matter—varying the permutation of few-shot exemplars can cause the accuracy of GPT-3 on SST-2 to range from near chance (54.3%) to near SOTA (93.4%) (Zhao et al., 2021). We show the standard deviation of performance from different exemplars in Table 6 and Table 7. Standard deviations with respect to prompt order are relatively minimal in almost all cases. The one exception is the coin flip task, for which exemplar orders have high standard deviation, likely for the reason cited in Zhao et al. (2021)—for classification, many exemplars of the same category in a row biases the model outputs).
- **Different number of exemplars.** We also found that gains from chain-of-thought prompting generally still held when there was a varying number of few-shot exemplars. This is shown for five datasets in Figure 11 (we did not have the compute to run this for all datasets). We also found in preliminary experiments that further increasing the number of exemplars in standard prompting did not lead to significant gains (e.g., increasing from 8 to 16 exemplars did not improve the performance of standard prompting enough to catch up with chain-of-thought prompting).
- **Different language models.** Another interesting question is whether certain prompts that work better for one model work better for other large language models. We find that with the same prompts, chain-of-thought prompting improves performance across all three models (LaMDA, GPT-3, and PaLM) for all datasets except CSQA and StrategyQA for GPT-3 (Table 1, Table 4, Table 5). The fact that gains from chain of thought did not transfer perfectly among models is a limitation; further work could investigate why how different pre-training datasets and model architectures affect the performance gain from chain-of-thought prompting.

Prompt engineering still matters, though. Although the results are relatively robust to the prompt for arithmetic reasoning, we want to be clear that prompt engineering still does matter, and can improve performance significantly in many cases. Though most chain of thought annotations outperform standard prompting, there is large variation in many cases. For instance, for the coin flip task, the performance varied from 99.6% for Annotator A to 71.4% for Annotator C, though both were above standard prompting = 50.0% (see Table 7). There are even tasks where prompt engineering is a requirement for good performance. In preliminary experiments, we tried using chain of thought to enable language models to reverse the order of a list of 5 items. While two co-authors were not able to write chain of thought prompts that solved the task despite their best attempts, a third co-author was able to write a chain of thought that perfectly solved the task.

How to generate chain of thought annotations in a robust fashion could be an interesting direction for future work. For instance, an idea here could be to use a large language model to automatically generate chains of thought via prompting (and potentially optimize this over a validation set).

A.3 Will chain-of-thought prompting improve performance for my task of interest?

While chain-of-thought prompting is in principle applicable for any text-to-text task, it is more helpful for some tasks than others. Based on the experiments in this paper, our intuition is that chain of thought helps the most when three conditions are met: (1) the task is challenging and requires