

We found that the remaining chains of thought (27 of 50; 54%) would require substantial edits to make into a correct chain of thought. Almost all cases here involved some error in semantic understanding (see Table 11: “semantic understanding error”), and 8 of the 27 also had incoherent chain of thoughts, meaning that some statements in the generated chain of thought did not follow from prior ones or violated basic world knowledge (see Table 11: “incoherent chain of thought error”).

Table 11: Example of incorrect chains of thought, categorized as described in Appendix D.2. These chains of thought had multiple mistakes that could not be fixed with minor edits.

<p><b>QUESTION:</b> Gretchen has 110 coins. There are 30 more gold coins than silver coins. How many gold coins does Gretchen have?</p> <p><b>MODEL ANSWER (INCORRECT; SEMANTIC UNDERSTANDING ERROR):</b> Gretchen has 110 coins. There are 30 more gold coins than silver coins. So there are <math>110 - 30 = 80</math> silver coins. So there are 80 silver coins and <math>110 - 80 = 30</math> gold coins. The answer is 30. ✗</p> <p><b>EXPLANATION FOR ERROR CATEGORY:</b> The model does not understand that 110 coins are in total and that it must solve an equation to get the number of gold and silver coins.</p>
<p><b>QUESTION:</b> In a dance class of 20 students, 20% enrolled in contemporary dance, 25% of the remaining enrolled in jazz dance, and the rest enrolled in hip-hop dance. What percentage of the entire students enrolled in hip-hop dance?</p> <p><b>MODEL ANSWER (INCORRECT; INCOHERENT CHAIN OF THOUGHT ERROR):</b> 20% of the students enrolled in contemporary dance. 25% of the remaining students enrolled in jazz dance. The rest enrolled in hip-hop dance. So the percentage of the entire students enrolled in hip-hop dance is the percentage of the students enrolled in hip-hop dance, minus the percentage of the students enrolled in contemporary dance, minus the percentage of the students enrolled in jazz dance. So the percentage of the entire students enrolled in hip-hop dance is <math>(25 + 20) - (25 + 20) = 100\%</math>. The answer is 100%. ✗</p> <p><b>EXPLANATION FOR ERROR CATEGORY:</b> This chain of thought is incoherent in that the percent of entire students enrolled in hip-hop dance cannot be the percent of student enrolled in hip-hop dance minus another term.</p>

Overall, there are no guarantees that the reasoning processes generated by large language models are coherent or factually correct, as underscored by the recent work evaluating the factuality of language model generations and explanations (Maynez et al., 2020; Rashkin et al., 2021; Ye and Durrett, 2022; Marasović et al., 2022; Wiegrefe et al., 2022). Incorrect reasoning processes can lead to both incorrect final answers as well as accidentally correct final answers (with accidentally correct final answers being more likely for tasks such as binary classification as opposed to free response). Improving the factuality of language model generations with respect to context and world knowledge is an important direction open problems in language model research and could also be expected to potentially improve multi-step reasoning abilities of language models. One potential method for improving the quality of decoding could involve generating multiple reasoning paths and scoring each of them with a verifier, though this requires training the verifier (Cobbe et al., 2021; Shen et al., 2021; Thoppilan et al., 2022).

### D.3 Additional Robustness Analysis

As the experiments in the main paper use a fixed number of few-shot exemplars (8; as constrained by the input length of 1024 tokens), we verify that the chain-of-thought prompting is robust to various numbers of few-shot exemplars. We run experiments for LaMDA 137B, comparing chain-of-thought prompting with standard prompting for the five datasets where standard prompting had a mostly flat scaling curve (the largest model did not achieve high performance). As shown in Figure 11, the improvement of chain-of-thought prompting over standard prompting remains robust to varying the number of few-shot exemplars in the prompt.

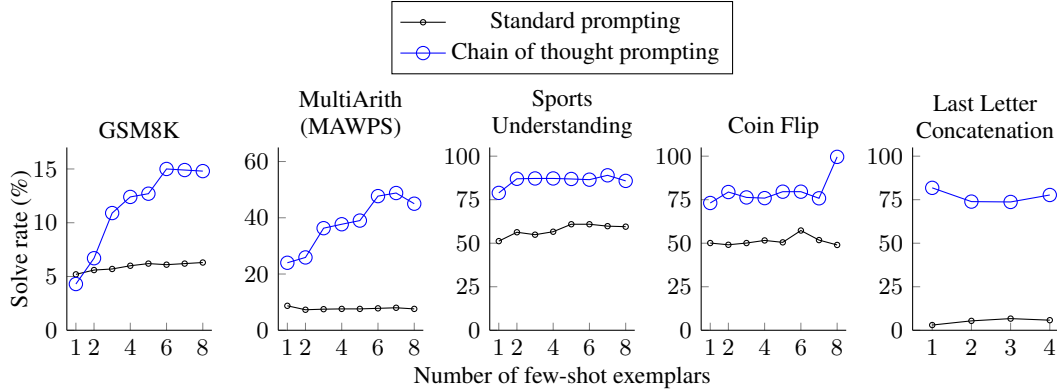


Figure 11: The improvement of chain of thought prompting over standard prompting appears robust to varying the number of few-shot exemplars in the prompt.

Table 12: Summary of math word problem benchmarks we use in this paper with examples.  $N$ : number of evaluation examples.

Dataset	$N$	Example problem
GSM8K	1,319	Josh decides to try flipping a house. He buys a house for \$80,000 and then puts in \$50,000 in repairs. This increased the value of the house by 150%. How much profit did he make?
SVAMP	1,000	Each pack of dvds costs 76 dollars. If there is a discount of 25 dollars on each pack. How much do you have to pay to buy each pack?
ASDiv	2,096	Ellen has six more balls than Marin. Marin has nine balls. How many balls does Ellen have?
AQuA	254	A car is being driven, in a straight line and at a uniform speed, towards the base of a vertical tower. The top of the tower is observed from the car and, in the process, it takes 10 minutes for the angle of elevation to change from $45^\circ$ to $60^\circ$ . After how much more time will this car reach the base of the tower? Answer Choices: (a) $5\sqrt{3} + 1$ (b) $6\sqrt{3} + \sqrt{2}$ (c) $7\sqrt{3} - 1$ (d) $8\sqrt{3} - 2$ (e) None of these
MAWPS: SingleOp	562	If there are 7 bottle caps in a box and Linda puts 7 more bottle caps inside, how many bottle caps are in the box?
MAWPS: SingleEq	508	Benny bought a soft drink for 2 dollars and 5 candy bars. He spent a total of 27 dollars. How much did each candy bar cost?
MAWPS: AddSub	395	There were 6 roses in the vase. Mary cut some roses from her flower garden. There are now 16 roses in the vase. How many roses did she cut?
MAWPS: MultiArith	600	The school cafeteria ordered 42 red apples and 7 green apples for students lunches. But, if only 9 students wanted fruit, how many extra did the cafeteria end up with?

## E Additional Details

### Version Control

**V5** → **V6**. Fixed minor typo in Figure 3.

**V4** → **V5**. Added Codex and UL2 results. Small changes to writing and style of paper.

**V3** → **V4**. Fixed typo in Figure 3 and added a couple citations.

**V2** → **V3**. Added GPT-3 results. Added SVAMP and AQuA eval datasets for math. Added SayCan eval for commonsense. Added Extended Related Work section (Appendix C). Added ablations for Commonsense and Symbolic Reasoning (Table 7). Added FAQ section (Appendix A). Added raw results in Appendix B.

**V1** → **V2**. Added PaLM results (V1 only had LaMDA).

### E.1 Reproducibility Statement

As our results make use of two sets of large language models that is not publicly available, we take the following actions to facilitate reproducibility. First, we provide the exact input prompts for all tasks in Table 20–Table 27 in Appendix G (and emphasize that we do not perform any finetuning and only apply prompting to off-the-shelf language models). Second, we conduct experiments using the publicly available GPT-3 API for four model scales text-ada-001, text-babbage-001, text-curie-001, text-davinci-002). Finally, we make exact inputs, targets, and predictions for LaMDA 137B for each task available as a zip file in the supplementary material.

### E.2 Computational Resources

For all three language models we evaluated, we did prompting-based inference only. No finetuning was done for this paper. For inference on LaMDA 137B we use TPU v3 (8x8 configuration, 64 chips / 128 cores), and for inference on PaLM 540B we use TPU v4 (4x4x12 configuration, 192 chips / 384 cores). GPT-3 experiments were done using the public API.<sup>5</sup>

### E.3 Dataset Details and Licenses

We list the details and licenses for all arithmetic and commonsense datasets used in this paper. The symbolic reasoning datasets were created synthetically, as described in Section 4.

#### Arithmetic reasoning

- Math Word Problem Repository (Koncel-Kedziorski et al., 2016): AddSub (Hosseini et al., 2014): <https://www.cs.washington.edu/nlp/arithmetic>; MultiArith (Roy and Roth, 2015), license: CC BY 4.0.
- ASDiv (Miao et al., 2020): <https://github.com/chaochun/nlu-asdiv-dataset>.
- AQuA (Ling et al., 2017): <https://github.com/deepmind/AQuA>, license: <https://github.com/deepmind/AQuA/blob/master/LICENSE>.
- GSM8K (Cobbe et al., 2021): <https://github.com/openai/grade-school-math>, MIT license: <https://github.com/openai/grade-school-math/blob/master/LICENSE>.
- SVAMP (Patel et al., 2021): <https://github.com/arkilpatel/SVAMP>, MIT license: <https://github.com/arkilpatel/SVAMP/blob/main/LICENSE>.

#### Commonsense reasoning

- CSQA (Talmor et al., 2019): <https://www.tau-nlp.org/commonsenseqa>, <https://github.com/jonathanherzig/commonsenseqa>.

---

<sup>5</sup><https://beta.openai.com/docs/api-reference/making-requests>