

B All Experimental Results

This section contains tables for experimental results for varying models and model sizes, on all benchmarks, for standard prompting vs. chain-of-thought prompting.

For the arithmetic reasoning benchmarks, some chains of thought (along with the equations produced) were correct, except the model performed an arithmetic operation incorrectly. A similar observation was made in Cobbe et al. (2021). Hence, we can further add a Python program as an external calculator (using the Python `eval` function) to all the equations in the generated chain of thought. When there are multiple equations in a chain of thought, we propagate the external calculator results from one equation to the following equations via string matching. As shown in Table 1, we see that adding a calculator significantly boosts performance of chain-of-thought prompting on most tasks.

Table 1: Chain of thought prompting outperforms standard prompting for various large language models on five arithmetic reasoning benchmarks. All metrics are accuracy (%). Ext. calc.: post-hoc external calculator for arithmetic computations only. Prior best numbers are from the following. *a*: Cobbe et al. (2021). *b* & *e*: Pi et al. (2022), *c*: Lan et al. (2021), *d*: Piękos et al. (2021).

	Prompting	GSM8K	SVAMP	ASDiv	AQuA	MAWPS
Prior best	N/A (finetuning)	55 ^a	57.4 ^b	75.3 ^c	37.9 ^d	88.4 ^e
UL2 20B	Standard	4.1	10.1	16.0	20.5	16.6
	Chain of thought	4.4 (+0.3)	12.5 (+2.4)	16.9 (+0.9)	23.6 (+3.1)	19.1 (+2.5)
	+ ext. calc	6.9	28.3	34.3	23.6	42.7
LaMDA 137B	Standard	6.5	29.5	40.1	25.5	43.2
	Chain of thought	14.3 (+7.8)	37.5 (+8.0)	46.6 (+6.5)	20.6 (-4.9)	57.9 (+14.7)
	+ ext. calc	17.8	42.1	53.4	20.6	69.3
GPT-3 175B (text-davinci-002)	Standard	15.6	65.7	70.3	24.8	72.7
	Chain of thought	46.9 (+31.3)	68.9 (+3.2)	71.3 (+1.0)	35.8 (+11.0)	87.1 (+14.4)
	+ ext. calc	49.6	70.3	71.1	35.8	87.5
Codex (code-davinci-002)	Standard	19.7	69.9	74.0	29.5	78.7
	Chain of thought	63.1 (+43.4)	76.4 (+6.5)	80.4 (+6.4)	45.3 (+15.8)	92.6 (+13.9)
	+ ext. calc	65.4	77.0	80.0	45.3	93.3
PaLM 540B	Standard	17.9	69.4	72.1	25.2	79.2
	Chain of thought	56.9 (+39.0)	79.0 (+9.6)	73.9 (+1.8)	35.8 (+10.6)	93.3 (+14.2)
	+ ext. calc	58.6	79.8	72.6	35.8	93.5