Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Xuyuan Liu

December 7, 2023

DARTMOUTH

Chain of Thought

[Wei et al., 2022]

Contribution

+ Introduce intermediate steps to enhance reasoning capabilities in language models.

Weakness

- Small models' underperformance is understood, but the adverse effect of CoT is not explained

Model		GSM8K		SVAMP		ASDiv		AQuA		MAWPS	
		standard	CoT	standard	CoT	standard	CoT	standard	CoT	standard	CoT
UL2	20B	4.1	**4.4**	10.1	**12.5**	16.0	**16.9**	20.5	**23.6**	16.6	**19.1**
LaMDA	420M	2.6	0.4	2.5	1.6	3.2	0.8	23.5	8.3	3.2	0.9
	2B	3.6	1.9	3.3	2.4	4.1	3.8	22.9	17.7	3.9	3.1
	8B	3.2	1.6	4.3	3.4	5.9	5.0	22.8	18.6	5.3	4.8
	68B	5.7	**8.2**	13.6	**18.8**	21.8	**23.1**	22.3	20.2	21.6	**30.6**
	137B	6.5	**14.3**	29.5	**37.5**	40.1	**46.6**	25.5	20.6	43.2	**57.9**
GPT	350M	2.2	0.5	1.4	0.8	2.1	0.8	18.1	8.7	2.4	1.1
	1.3B	2.4	0.5	1.5	1.7	2.6	1.4	12.6	4.3	3.1	1.7
	6.7B	4.0	2.4	6.1	3.1	8.6	3.6	15.4	13.4	8.8	3.5
	175B	15.6	**46.9**	65.7	**68.9**	70.3	**71.3**	24.8	**35.8**	72.7	**87.1**
Codex	-	19.7	**63.1**	69.9	**76.4**	74.0	**80.4**	29.5	**45.3**	78.7	**92.6**
PaLM	8B	4.9	4.1	15.1	**16.8**	23.7	**25.2**	19.3	**21.7**	26.2	**30.5**
	62B	9.6	**29.9**	48.2	46.7	58.7	**61.9**	25.6	22.4	61.8	**80.3**
	540B	17.9	**56.9**	69.4	**79.0**	72.1	**73.9**	25.2	**35.8**	79.2	**93.3**

Model		SingleOp		SingleEq		AddSub		MultiArith	
		standard	CoT	standard	CoT	standard	CoT	standard	CoT
UL2	20B	24.9	**27.2**	18.0	**20.2**	18.5	18.2	5.0	**10.7**
LaMDA	420M	2.8	1.0	2.4	0.4	1.9	0.7	5.8	1.5
	2B	4.6	4.1	2.4	3.3	2.7	3.2	5.8	1.8
	8B	8.0	7.0	4.5	4.4	3.4	5.2	5.2	2.4
	68B	36.5	**40.8**	23.9	**26.0**	17.3	**23.2**	8.7	**32.4**
	137B	73.2	**76.2**	48.8	**58.7**	43.0	**51.9**	7.6	**44.9**
GPT	350M	3.2	1.8	2.0	0.2	2.0	1.5	2.3	0.8
	1.3B	5.3	3.0	2.4	1.6	2.3	1.5	2.2	0.5
	6.7B	13.5	3.9	8.7	4.9	8.6	2.5	4.5	2.8
	175B	90.9	88.8	82.7	**86.6**	83.3	81.3	33.8	**91.7**
Codex	-	93.1	91.8	86.8	**93.1**	90.9	89.1	44.0	**96.2**
PaLM	8B	41.8	**46.6**	29.5	28.2	29.4	**31.4**	4.2	**15.8**
	62B	87.9	85.6	77.2	**83.5**	74.7	**78.2**	7.3	**73.7**
	540B	94.1	94.1	86.5	**92.3**	93.9	91.9	42.2	**94.7**

Self-Consistency

[Wang et al., 2022]

Contribution

+ Employ a voting mechanism among various generated reasoning paths to determine the final answer.

Weakness

- Why Multiple Prompts Ensemble Approach is Ineffective?

	GSM8K	MultiArith	SVAMP	ARC-e	ARC-c
CoT (Wei et al., 2022)	17.1	51.8	38.9	75.3	55.1
Ensemble (3 sets of prompts)	18.6 ± 0.5	57.1 ± 0.7	42.1 ± 0.6	76.6 ± 0.1	57.0 ± 0.2
Ensemble (40 prompt permutations)	19.2 ± 0.1	60.9 ± 0.2	42.7 ± 0.1	76.9 ± 0.1	57.0 ± 0.1
Self-Consistency (40 sampled paths)	**27.7** ± 0.2	**75.7** ± 0.3	**53.3** ± 0.2	**79.3** ± 0.3	**59.8** ± 0.2

Table 7: Self-consistency outperforms prompt-order and multi-prompt ensembles on LaMDA-137B.

* A different prompt can generate a distinct path
* Sampling from the same decoder does not guarantee a divergent path

Self-Consistency

[Wang et al., 2022]

Weakness

- *"Language model can't distinguish correct solution"*, why still employ a conditional probability-based voting mechanism?

$$\mathcal{P}\left(\mathbf{r}_i, \mathbf{a}_i \mid \text{prompt,question}\right) = \exp^{\frac{1}{K}\sum_{k=1}^{K} \log P(t_k|\text{prompt,question},t_1,\ldots,t_{k-1})}$$

	GSM8K	MultiArith	AQuA	SVAMP	CSQA	ARC-c
Greedy decode	56.5	94.7	35.8	79.0	79.0	85.2
Weighted avg (unnormalized)	56.3 ± 0.0	90.5 ± 0.0	35.8 ± 0.0	73.0 ± 0.0	74.8 ± 0.0	82.3 ± 0.0
Weighted avg (normalized)	22.1 ± 0.0	59.7 ± 0.0	15.7 ± 0.0	40.5 ± 0.0	52.1 ± 0.0	51.7 ± 0.0
Weighted sum (unnormalized)	59.9 ± 0.0	92.2 ± 0.0	38.2 ± 0.0	76.2 ± 0.0	76.2 ± 0.0	83.5 ± 0.0
Weighted sum (normalized)	74.1 ± 0.0	99.3 ± 0.0	48.0 ± 0.0	86.8 ± 0.0	80.7 ± 0.0	88.7 ± 0.0
Unweighted sum (majority vote)	74.4 ± 0.1	99.3 ± 0.0	48.3 ± 0.5	86.6 ± 0.1	80.7 ± 0.1	88.7 ± 0.1

Table 1: Accuracy comparison of different answer aggregation strategies on PaLM-540B.

Least-to-Most Prompting

[Zhou et al., 2022]

Contribution

+ Decompose the question into sub-problems and solve each separately.

Weakness

- Lack generalizability due to the demands of task-specific prompt design .
- Lack empirical research examining the efficacy of Decomposition/Subproblem Solving methodologies.
  * Multiple decomposition strategies/granularity
  * How to ensure model compose the final solution correctly.

References

1

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E. H.-h., & Zhou, D. (2022).Self-consistency improves chain of thought reasoning in language models. *ArXiv*, *abs/2203.11171*. https://api.semanticscholar.org/CorpusID:247595263

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E. H.-h., Xia, F., Le, Q., & Zhou, D. (2022).Chain of thought prompting elicits reasoning in large language models. *ArXiv*, *abs/2201.11903*. https://api.semanticscholar.org/CorpusID:246411621

Zhou, D., Scharli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Bousquet, O., Le, Q., & Chi, E. H.-h. (2022).Least-to-most prompting enables complex reasoning in large language models. *ArXiv*, *abs/2205.10625*. https://api.semanticscholar.org/CorpusID:248986239