Mitigating Misleading Chain-of-Thought Reasoning with Selective Filtering

Yexin Wu, Zhuosheng Zhang, Hai Zhao

LREC-COLING 2024 | Shanghai Jiao Tong University

Abstract

- Proposed a novel approach called the selective filtering reasoner (SelF-Reasoner) that assesses the entailment relationship between the question and the candidate reasoning chain. Then, we proceed with CoT reasoning when the reasoning chain demonstrates confidence; otherwise, we opt to predict the answer directly.
- SelF-Reasoner improves the fine-tuned T5 baseline consistently over the ScienceQA, ECQA, and LastLetter tasks.

Motivation

- CoT(Few-shot, zero-shot, chatgpt…) has achieved success with LLM.
- CoT reasoning is considered one of the emergent abilities shown in “Scaling Laws”.
- In some situations (Li et al., 2022a; Magister et al., 2022; Ho et al., 2022; Li et al., 2022b; Wang et al., 2023), though small language models (SML) can not do few-shot or zero-shot CoT finetuning, they can still benefit from CoT finetuning.
- But small language models are easier to generate misleading CoTs compared with LLM.
- Ways to mitigate the effect of misleading CoTs

Q: departmental, dome, pressed, fascinating
Predicted rationale: The last letter of the first word 'departmental' is 'l'. The last letter of the second word 'dome' is 'e'. The last letter of the third word 'pressed' is 'd'. The last letter of the forth word 'funny' is 'y' .
Extracted Answer: The answer is ledy. (Incorrect, the rationale part is wrong)
Directly predicted Answer: The answer is ledg. (Correct)
Ground Truth Answer: The answer is ledg.

Question: What is the result of applying for a job? Choices: (a) anxiety and fear (b) increased workload (c) praise (d) less sleep (e) being employed
Ground Truth CoT: Applying for a job is followed by attending interview which results in being employed. Applying for a job may not result in the other options.
Answer: The answer is being employed.

Generated wrong CoT: People generally get praise when they get accepted for a job. When one applies for a job, they get praise. Anxiety and fear is not the result of applying for a job. Increased workload is not the result of applying for a job. Less sleep is not the result of applying for a job. Being employed is not the result of applying for a job.
Extracted incorrect answer: The answer is praise.

Figure 1: An example of an invalid CoT reasoning from ECQA (Aggarwal et al., 2021; Wang et al., 2023). The errors are highlighted in red. The generated CoT is wrong at the first step, and the error continues to the end. However, when altering to direct prediction, this one-step reasoning question is solved correctly.

Table 7: Dataset statistics used in our experiments.

Dataset	Train	Validation	Test
ScienceQA	12726	4241	4241
ECQA	8520	1221	1221
LastLetter	10000	5000	5000

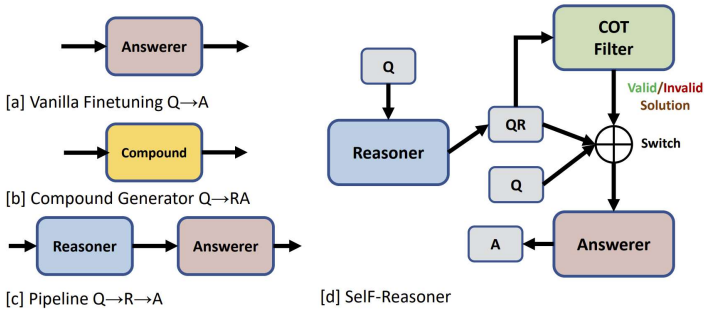
Methodology

Datasets

- Three Datasets:
  - ScienceQA(Lu et al., 2022a)
  - ECQA (Aggarwal et al., 2021; Talmor et al., 2019)
  - LastLetter

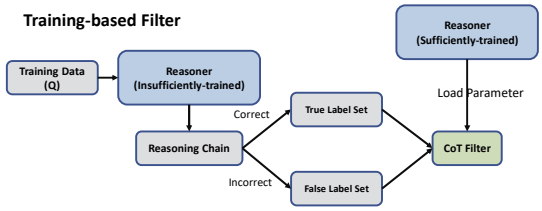
Model Architecture

- Three baselines:
  - Vanilla finetuning means to generate the answer only.
  - Compound generator generates the reasoning chain and the answer in one run.
  - Pipeline first generates the reasoning chain, and then use the question and the reasoning chain to generate the answer.
- SelF-Reasoner will judge the reasoning chain. If the filter thinks the reasoning chain is misleading, then it will adopt vanilla finetuning's method to predict the answer only using the question. Otherwise, it will go as the pipeline using the question and the reasoning chain to generate the answer.
- Our backbone model is T5.



Filter Design

- Training-based Filter.
  - Use an insufficiently trained reasoner to generate plausible reasoning chains and correct reasoning chains (correct means it will lead to correct answers).
  - T5 encoder is trained to be a filter.
  - Load finetuned T5 reasoner’s parameter to accelerate training and improve accuracy.
- Rule-based Filter.
- For the LastLetter task: **"The given word should appear in the valid CoT".**
- Observation: Many given words in the reasoning chain are different from the ones in the question due to tokenization and sampling, which will cause the incorrect answer extraction



Experiments

- Datasets:
  - ScienceQA, ECQA, LastLetter
- Metrics: Accuracy
- Result: New SOTA.

Method	Vanilla	Pipeline	SelF-Reasoner
Accuracy	58.07	54.95	**58.48 (+3.5)**

Table 3: Accuracy (%) on test split of ECQA. The backbone model is UnifiedQA-base. SelF-Reasoner outperforms the pipeline by 3.5%.

Method	Vanilla	Pipeline	SelF-Reasoner
Accuracy	64.22	76.80	**80.06 (+3.26)**

Table 4: Accuracy (%) on test split of LastLetter. The backbone model is UnifiedQA-base. SelF-Reasoner outperforms the pipeline by 3.26%.

Method	Model	Learning	Format	Accuracy
Lu et al. (2022a)	Human	-	*Q-A*	88.4
	GPT-3 (CoT)	In-Context Learning	*Q-ALE*	75.17
	$\text{UnifiedQA}_{\text{Base}}$	fine-tuning	*Q-A*	70.12
	$\text{UnifiedQA}_{\text{Base}}$	fine-tuning	*Q-AE*	73.33
	$\text{UnifiedQA}_{\text{Base}}$	fine-tuning	*Q-ALE*	74.11
Lu et al. (2023)	ChatGPT (CoT)	In-Context Learning	*Q-EA*	78.31
	GPT-4 (CoT)	In-Context Learning	*Q-EA*	83.99
	Chameleon (ChatGPT)	In-Context Learning	*Q-EA*	79.93
	Chameleon (GPT-4)	In-Context Learning	*Q-EA*	86.54
Vanilla	$\text{UnifiedQA}_{\text{Samll}}$	fine-tuning	*Q-A*	71.54
	$\text{UnifiedQA}_{\text{Base}}$	fine-tuning	*Q-A*	83.09
	$\text{UnifiedQA}_{\text{Large}}$	fine-tuning	*Q-A*	86.53
Compound	$\text{UnifiedQA}_{\text{Base}}$	fine-tuning	*Q-ALE*	76.13
	$\text{UnifiedQA}_{\text{Base}}$	fine-tuning	*Q-EA*	77.71
	$\text{UnifiedQA}_{\text{Base}}$	fine-tuning	*Q-LEA*	73.97
Pipeline	$\text{UnifiedQA}_{\text{Small}}$	fine-tuning	*Q-E→ QE-A*	66.37
	$\text{UnifiedQA}_{\text{Base}}$	fine-tuning	*Q-E→ QE-A*	79.32
	$\text{UnifiedQA}_{\text{Large}}$	fine-tuning	*Q-E→ QE-A*	84.98
SelF-Reasoner	$\text{UnifiedQA}_{\text{Small}}$	fine-tuning	SelF-Reasoner	69.55
	$\text{UnifiedQA}_{\text{Base}}$	fine-tuning	SelF-Reasoner	83.45
	$\text{UnifiedQA}_{\text{Large}}$	fine-tuning	SelF-Reasoner	**87.24**

Table 1: Accuracy (%) of each baseline on test split. In the format part, *Q = Question*, *A = Answer*, *E = Explanation*, *L = Lecture*. We list the results from ScienceQA (Lu et al., 2022a), ChatGPT, GPT-4 (Lu et al., 2023) for comparison. *L* and *E* can be treated as reasoning chain. So LEA/EA and ALE/AE correspond to the standard RA and AR as defined in Section 3.1, respectively. Our SelF-Reasoner (Large) is comparable in accuracy to a human's.

Analysis

Quality of CoT generated by small language models

- Metrics:
  - BLEU-1/4
  - ROUGE
  - Sentence Similarity
  - Human evaluation.
- Similar to the ground truth CoT in structure.
- Incorrect in some key parts, causing misleading results.

Model	Split	BLEU-1	BLEU-4	ROUGE-L	Similarity	Complete	Entailment	Correct
Base	Lead to Correct Answer	0.914	0.776	0.910	0.937	1.00	1.00	0.94
	Lead to Incorrect Answer	0.789	0.660	0.797	0.860	1.00	1.00	0.02
	All	0.892	0.756	0.891	0.924	-	-	-
Large	Lead to Correct Answer	0.937	0.810	0.929	0.949	1.00	0.98	0.96
	Lead to Incorrect Answer	0.775	0.642	0.784	0.847	1.00	1.00	0.02
	All	0.917	0.788	0.910	0.936	-	-	-

Table 2: Automatic metrics (BLEU-1/4, ROUGE-L, Similarity) and human evaluation of generated explanations. We evaluate these metrics on different splits of the produced CoT according to whether they can lead to the correct answer. Details of human evaluation are shown in Appendix A.8.

Filter’s performance

- Filter can predict whether the CoT will mislead the answer accurately.
- Introducing filter outperforms random assignment.

Model	Vanilla	Pipeline	Random	SelF-Reasoner
Small	71.54	66.37	68.76	69.55
Base	83.09	79.32	81.61	83.45
Large	86.53	84.98	86.09	87.24

Table 5: Ablation on the CoT filter on ScienceQA benchmark. Random refers to randomly choosing vanilla fine-tuning and pipeline to produce the answer.

Generator	Filter	Valid Acc	Invalid Acc	Acc	F1
Base	Base	76.96	76.39	76.84	0.841
	Large	81.30	81.64	81.37	0.874
Large	Base	74.97	75.03	74.98	0.836
	Large	80.07	78.17	79.78	0.871

Table 6: Accuracy and F1 score of the CoT filter on classifying the generated reasoning chain on ScienceQA benchmark. Valid/Invalid Acc refers to the filter's accuracy in discriminating valid/invalid reasoning chains. Acc is the overall accuracy.

Ablation Study: Filter&Pipeline size

- Larger Filter performs better
- The improvement in the larger pipeline is smaller because the reasoner’s ability to generate plausible CoT is improved, making it harder for filters to distinguish plausible CoT from correct CoT.

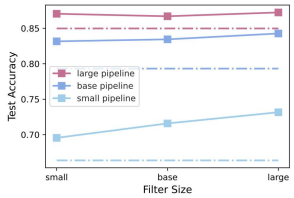


Figure 3: The "scaling law" in the size of the CoT filter and pipeline on ScienceQA benchmark. The dashed line presents the accuracy of the pipeline.

Conclusion

- We proposed a selective filtering reasoner (SelF-Reasoner) to perform CoT only as necessary and mitigate the detrimental effects of erroneous reasoning chains.
- Our SelF-Reasoner outperforms the finetuned CoT/vanilla baseline on ScienceQA, ECQA, and LastLetter datasets, advancing the effectiveness of CoT in small-scale language models.
- We analyze the obstructions of fine-tuning CoT on language models and conclude common types in invalid generated CoT.

Acknowledgement

[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2] Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University
[3] This paper was partially supported by Joint Research Project of Yangtze River Delta Science and Technology Innovation Community (No. 2022CSJGG1400) and Joint Funds of the National Natural Science Foundation of China (No. U21B2020)

Contact

wuyexin_libro_i131[AT]sjtu.edu.cn
zhangzs[AT]sjtu.edu.cn
Zhaohai[AT]cs.sjtu.edu.cn