Small Language Model Can Self-Correct

Haixia Han[1], Jiaqing Liang[2], Jie Shi[3], Qianyu He[3], Yanghua Xiao*[1,3]

[1]Shanghai Institute of AI for Education and School of Computer Science and Technology, East China Normal University

[2]School of Data Science, Fudan University

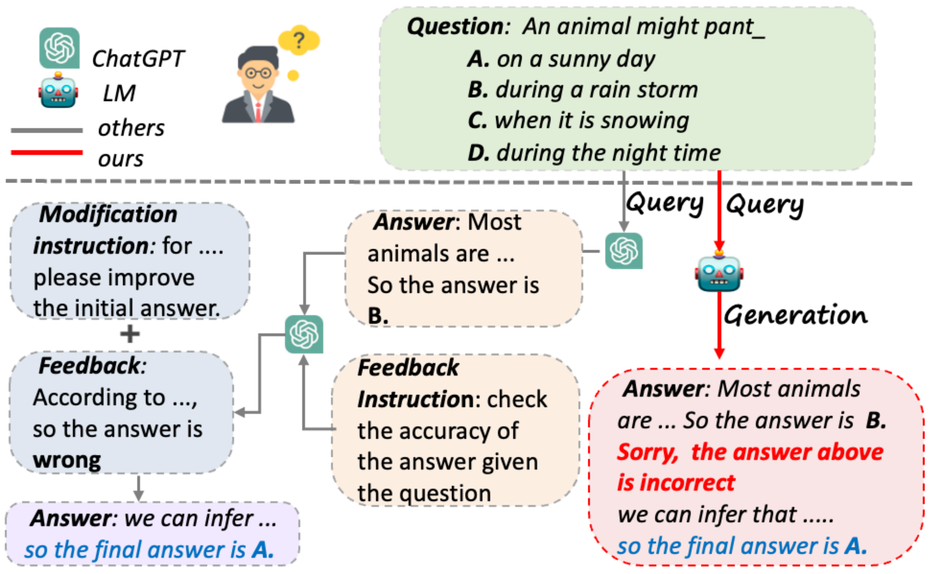
[3]Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University

Motivation

Previous studies have devised sophisticated pipelines and prompts to induce large LMs to exhibit the capability for self-correction.

- **Challenge 1**:large LMs are explicitly prompted to verify and modify its answers separately rather than completing all steps spontaneously like humans.
- **Challenge 2**: these complex prompts are extremely challenging for small LMs to follow.

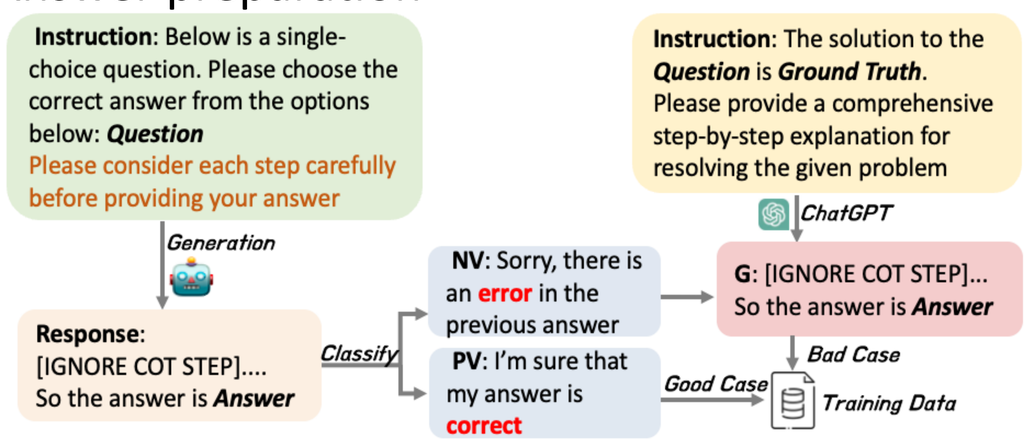
We introduce the *Intrinsic Self-Correction (ISC)* in generative language models, aiming to correct the initial output of LMs in a self-triggered manner, even for those small LMs with 6 billion parameters.



Methods

Build Self-correction Data

- Question preparation
- Answer preparation



Partial Answer Mask

⭐ Refrain from computing the loss corresponding to the incorrect answer part in the output during training

Question: *(TaskP)* Examine the following options carefully and select the correct one. *(COTP)* Before providing your final answer, give the analysis steps. *(SCP)* And you need double-check your response for accuracy before proceeding to submit.
(question) Where do you buy tickets at a ticket booth for games?
A. train station B. cathedral C. metro station D. fairgrounds E. amusement park
Answer: *(A_1^1- COT)* The question mentions the keywords "buying tickets" and "games", so we can guess that this is a · · ·
(A_1^1) Therefore, the correct answer is **E. amusement park**
(PV) Thinking about the correctness of the previous answer· · ·
Thinking result: I am sure that the earlier answer is **correct**

Question: *(TaskP)* Please choose the most appropriate one from the following options:
(question) what contributes more, though less than some believe, to the lung damage caused by smoking?
A. smoking less B. switching to chewing C. no filters D. switching to e-cigs
(COTP) Please give the detailed solving process and *(SCP)* verify your response before final submission.
Answer: *(A_2^1-COT)* Smoking causes less lung damage than people think, but it's not completely without effect.
So the answer is **A. smoking less**
(NV) Thinking about the correctness of the previous answer · · ·
Thinking result: Sorry, there is an error in the previous answer.
(SC-COT) Let's analyze each option:
A. smoking less: The question clearly mentions that it contributes less than some people think,· · · · · ·
C. no filters: filters it contributes to lung damage, and to a lesser extent than some believe. Therefore, the no filter option meets the requirement.
(A_2^2) So the correct option is **C. no filter**.

Experiments

Main Results

⭐A noticeable enhancement in accuracy is observed across all models for the two given tasks

Base Models	OpenBookQA		CommonsenseQA	
	ACC-First	ACC	ACC-First	ACC
CuteGPT-7B	25.2	29.0 (+3.8)	23.2	28.9 (+3.7)
CuteGPT-13B	37.2	42.0 (+4.8)	35.6	37.9 (+2.3)
Llama2-7B	52.2	52.2 (+0.0)	52.2	52.3 (+0.1)
ChatGLM-6B	37.0	42.6 (**+5.6**)	34.3	38.7 (**+4.4**)
Vicuna-7B	28.6	28.80 (+0.4)	25.9	26.2 (+0.3)
Vicuna-13B	33.8	34.0 (+0.2)	32.4	32.6 (+0.3)

Quantitative Analysis

Base Models	Confidence	EvalACC	R2R	R2W	W2W	W2R
OpenBookQA						
CuteGPT-7B	70.6	36.6	9	20	78	39
CuteGPT-13B	40.8	52.0	57	54	57	78
Llama2-7B	99.6	**52.2**	0	0	2	0
ChatGLM-6B	60.4	27.2	34	31	76	59
Vicuna-7B	98.6	28.8	0	0	6	1
Vicuna-13B	97.2	34.2	1	4	4	5
CommonsenseQA						
CuteGPT-7B	87.7	26.8	10	6	83	51
CuteGPT-13B	42.2	**53.8**	111	131	111	159
Llama2-7B	99.8	52.3	0	0	4	2
ChatGLM-6B	52.7	52.3	70	92	272	146
Vicuna-7B	98.9	26.0	1	2	13	9
Vicuna-13B	97.5	33.1	5	5	13	8

Performance on New Task