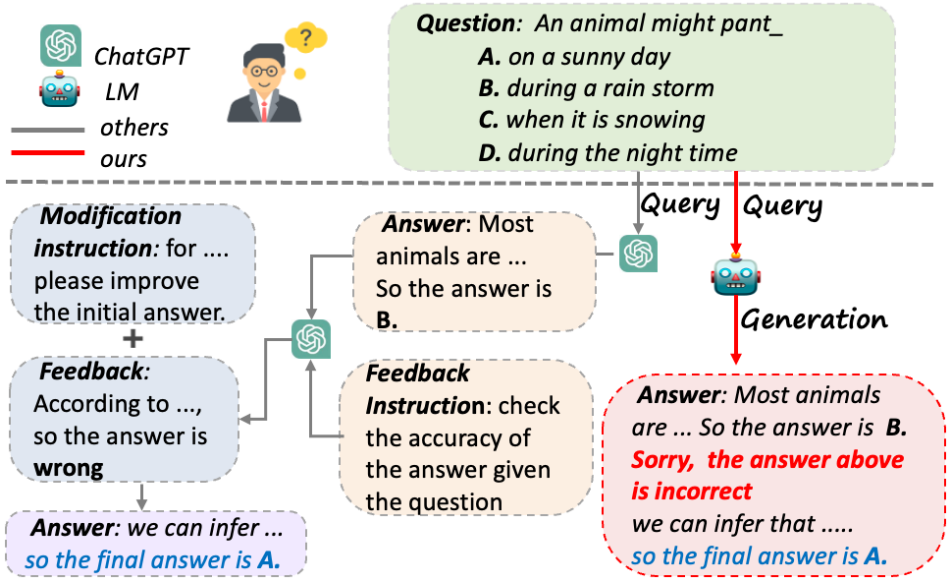
Small Language Model Can Self-Correct

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

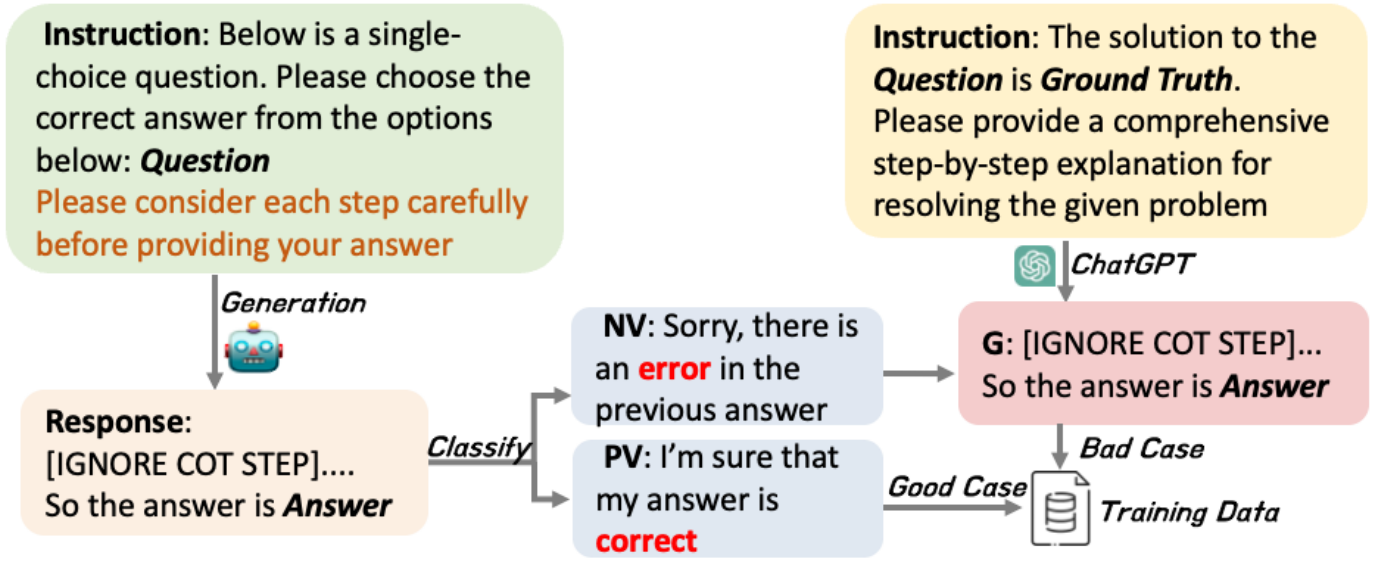
Motivation

- 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 their 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, even for those small LMs with 6 billion parameters.



Method: Build Self-correction Data

- **Question preparation**: we use two methods to enhance the model's understanding of various self-correction instructions.
  - Rewrite self-correction instruction using ChatGPT
  - Rearrang the different prompt: TaskP, COTP and SCP
- **Answer preparation**: through string matching to check the accuracy. We categorize the generated answers of *M* into two groups: good cases and bad cases
- **Data format**: $A_n^1 - COT \parallel A_n^1 \parallel \text{NV} \parallel A_n^2 - COT \parallel A_n^2 \cdots A_n^n \parallel \text{PV}$



Method: Partial Answer Mask

- Good case： only the loss associated with the answer component of the training data is used for gradient back propagation
- Bad case: we refrain from computing the loss corresponding to the incorrect answer part in the output during training. Instead, we calculate the loss from the output on self-verification

Good case sample

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**

Bad case sample

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**.

Main Results

- **Baselines：** CuteGPT-7B、CuteGPT-13B、Llama2-7B、ChatGLM-6B、Vicuna-7B、Vicuna-13B
- This intrinsic correction ability allows the small model to modify the answer when it detects an error in the initial response generation.
- ChatGLM-6B, correcting answers yields a notable improvement of **5.6%** in accuracy on the OpenBookQA dataset, improving it from **37%** to **42.6%**

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

- ***Discovery 1：*** When base models lack strong inherent capabilities but exhibit a high degree of confidence in their generated outcomes, the accuracy of self-verification tends to be notably lower. The performance gains derived from self-correction remain constrained
- ***Discovery 2:*** Observing the values in the W2R column, we find that if an LM can recognize errors in its responses and attempt to modify the initial answers

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

Quantitative Analysis

- ***Discovery 3:*** The values in the W2W column also constitute a significant portion of the total modifications. However, despite undergoing self-correction, the models have not achieved successful revisions.
- ***Discovery 4:*** *From R2R,* we note that the models, through self-correction, could identify inadequacies in the analytical process and subsequently provide supplementary 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

Ablation Analysis

- Nearly all results indicate that employing the PAM leads to higher accuracy in generating answers. The improvement in answer quality through self-correction is most prominent after utilizing the PAM
- Training without PAM appears to reduce the accuracy of self-verification, which is particularly evident in the CuteGPT family models.

Table 4: The impact of Partial Answer Masking on capability for self-correction.

Base Models	OpenBookQA				CommonsenseQA			
	ACC-First	ACC	Confidence	EvalACC	ACC-First	ACC	Confidence	EvalACC
CuteGPT-7B w/ PAM	25.2	29.0	70.6	36.6	23.2	26.9	87.7	26.8
CuteGPT-7B w/o PAM	26.0	28.2	2.0	68.6	18.8	31.2	2.2	76.2
CuteGPT-13B w/ PAM	37.2	42.0	40.8	52.0	36.6	37.9	42.2	53.2
CuteGPT-13B w/o PAM	28.6	32.4	1.6	68.2	20.5	34.2	1.9	77.8
ChatGLM-6B w/ PAM	37.0	42.6	60.4	27.2	34.3	38.7	52.7	52.3
ChatGLM-6B w/o PLM	30.2	36.0	56.6	32.2	26.0	32.9	50.0	47.2
Vicuna-7B w/ PAM	28.6	28.8	98.6	28.8	25.9	26.2	98.9	26.0
Vicuna-7B w/o PAM	23.4	23.6	97.4	25.0	13.8	13.9	39.8	14.0
Vicuna-13B w/ PAM	33.8	34.0	97.2	34.2	32.4	32.6	97.5	33.9
Vicuna-13B w/o PAM	32.0	31.2	89.6	32.6	31.5	32.4	97.5	33.1

Zero-shot Performance on New Task

- We discover that ISC remains effective for the new task. After the second round of correction, the accuracy of all LMs improve.

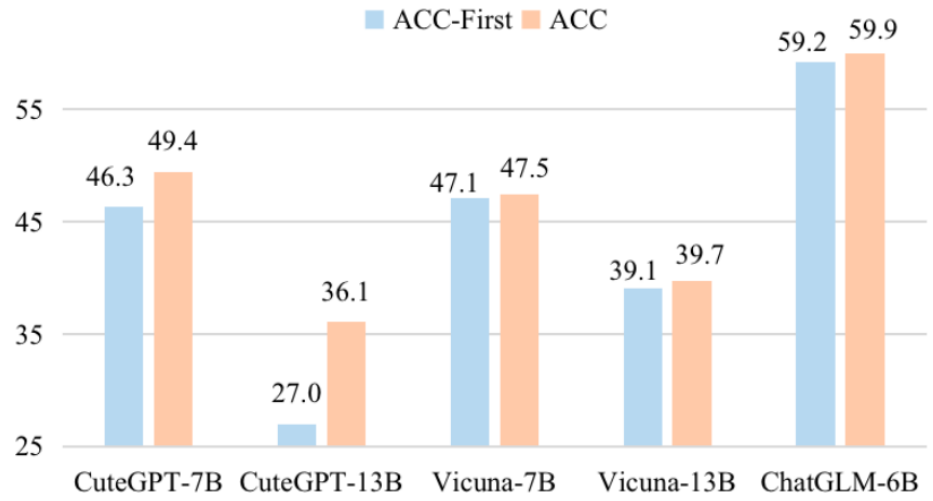


Figure 3: Zero-shot performance of Intrinsic Self-Correction on the StrategyQA test data. After the second round of correction, the accuracy improves obviously.

Conclusion

- We introduce Intrinsic Self-Correction (ISC) in LMs, an approach that utilizes models' own capabilities to identify and further modify their initial responses autonomously. This strong capability can even be applied to smaller LMs.
- We first devise a general process for constructing self-correction data and introduce a novel fine-tuning method named PAM to instruct LMs to self-correct.
- The experimental results on two distinct tasks consistently demonstrate that the utilization of ISC empowers the models with the capability for self-correction, and improves the accuracy of generated answers. In the best case, the accuracy enhancement reaches up to 5.6%.

Thank You !