Table 2: Results of GPT-3.5 and GPT-4 on reasoning benchmarks with oracle labels.

		GSM8K	CommonSenseQA	HotpotQA
GPT-3.5	Standard Prompting	75.9	75.8	26.0
	Self-Correct (Oracle)	84.3	89.7	29.0
GPT-4	Standard Prompting	95.5	82.0	49.0
	Self-Correct (Oracle)	97.5	85.5	59.0

Table 3: Results of GPT-3.5 and GPT-4 on reasoning benchmarks with intrinsic self-correction.

		# calls	GSM8K	CommonSenseQA	HotpotQA
GPT-3.5	Standard Prompting	1	**75.9**	**75.8**	**26.0**
	Self-Correct (round 1)	3	75.1	38.1	25.0
	Self-Correct (round 2)	5	74.7	41.8	25.0
GPT-4	Standard Prompting	1	**95.5**	**82.0**	**49.0**
	Self-Correct (round 1)	3	91.5	79.5	**49.0**
	Self-Correct (round 2)	5	89.0	80.0	43.0

utilize the ground-truth label to verify whether each step's generated answer is correct. If the answer is already correct, no (further) self-correction will be performed. Table 2 summarizes the results of self-correction under this setting, showcasing significant performance improvements, consistent with the findings presented in Kim et al. (2023); Shinn et al. (2023).

However, these results require careful consideration. For reasoning tasks, like solving mathematical problems, the availability of oracle labels seems counter-intuitive. If we are already in possession of the ground truth, there seems to be little reason to deploy LLMs for problem-solving. Therefore, the results can only be regarded as indicative of an oracle's performance.

Intrinsic Self-Correction. Per the above discussion, performance improvements achieved using oracle labels do not necessarily reflect true self-correction ability. Therefore, we turn our focus to the results in the *intrinsic self-correction* setting as defined in Section 2. To achieve this, we eliminate the use of labels, requiring LLMs to independently determine when to stop the self-correction process, i.e., whether to retain their previous answers.

Tables 3 and 4 report the accuracies and the number of model calls. We observe that, after self-correction, the accuracies of all models drop across all benchmarks.

To provide a more comprehensive assessment, we also design several different self-correction prompts to determine if there are better prompts that could enhance reasoning performance. Nonetheless, as shown in Tables 5 and 6, without the use of oracle labels, self-correction consistently results in a decrease in performance.

3.3 Why does the performance not increase, but instead decrease?

Empirical Analysis. Figure 1 summarizes the results of changes in answers after two rounds of self-correction, with two examples of GPT-3.5 illustrated in Figure 2. For GSM8K, 74.7% of the time, GPT-3.5 retains its initial answer. Among the remaining instances, the model is more likely to modify a correct answer to an incorrect one than to revise an incorrect answer to a correct one. ***The fundamental issue is that LLMs cannot properly judge the correctness of their reasoning.*** For CommonSenseQA, there is a higher chance that GPT-3.5 alters its answer. The primary reason for this is that false answer options in CommonSenseQA often appear somewhat relevant to the question, and using the self-correction prompt might bias the model to choose another option, leading to a high "correct $\Rightarrow$ incorrect" ratio. Similarly, Llama-2 also frequently converts a correct answer into an incorrect one. Compared to GPT-3.5 and Llama-2, both GPT-4 and GPT-4-Turbo are more likely to retain their initial answers. This may be because GPT-4 and GPT-4-Turbo have higher confidence

Table 4: Results of GPT-4-Turbo and Llama-2 with intrinsic self-correction.

		# calls	GSM8K	CommonSenseQA
GPT-4-Turbo	Standard Prompting	1	**91.5**	**84.0**
	Self-Correct (round 1)	3	88.0	81.5
	Self-Correct (round 2)	5	90.0	83.0
Llama-2	Standard Prompting	1	**62.0**	**64.0**
	Self-Correct (round 1)	3	43.5	37.5
	Self-Correct (round 2)	5	36.5	36.5

Table 5: Results of GPT-4-Turbo with different feedback prompts.

	# calls	GSM8K	CommonSenseQA
Standard Prompting	1	91.5	84.0
Feedback Prompt: Assume that this answer could be either correct or incorrect. Review the answer carefully and report any serious problems you find.			
Self-Correct (round 1)	3	88.0	81.5
Self-Correct (round 2)	5	90.0	83.0
Feedback Prompt: Review your previous answer and determine whether it's correct. If wrong, find the problems with your answer.			
Self-Correct (round 1)	3	90.0	74.5
Self-Correct (round 2)	5	90.0	81.0
Feedback Prompt: Verify whether your answer is correct, and provide an explanation.			
Self-Correct (round 1)	3	91.0	81.5
Self-Correct (round 2)	5	91.0	83.5

Table 6: Results of Llama-2 with different feedback prompts.

	# calls	GSM8K	CommonSenseQA
Standard Prompting	1	62.0	64.0
Feedback Prompt: Assume that this answer could be either correct or incorrect. Review the answer carefully and report any serious problems you find.			
Self-Correct (round 1)	3	43.5	37.5
Self-Correct (round 2)	5	36.5	36.5
Feedback Prompt: Review your previous answer and determine whether it's correct. If wrong, find the problems with your answer.			
Self-Correct (round 1)	3	46.5	26.0
Self-Correct (round 2)	5	30.5	37.0
Feedback Prompt: Verify whether your answer is correct, and provide an explanation.			
Self-Correct (round 1)	3	58.0	24.0
Self-Correct (round 2)	5	41.5	43.0

in their initial answers, or because they are more robust and thus less prone to being biased by the self-correction prompt.[1]

[1]We omit the analysis on HotpotQA because the sample size used in the source paper is quite small, which may not produce meaningful statistics.

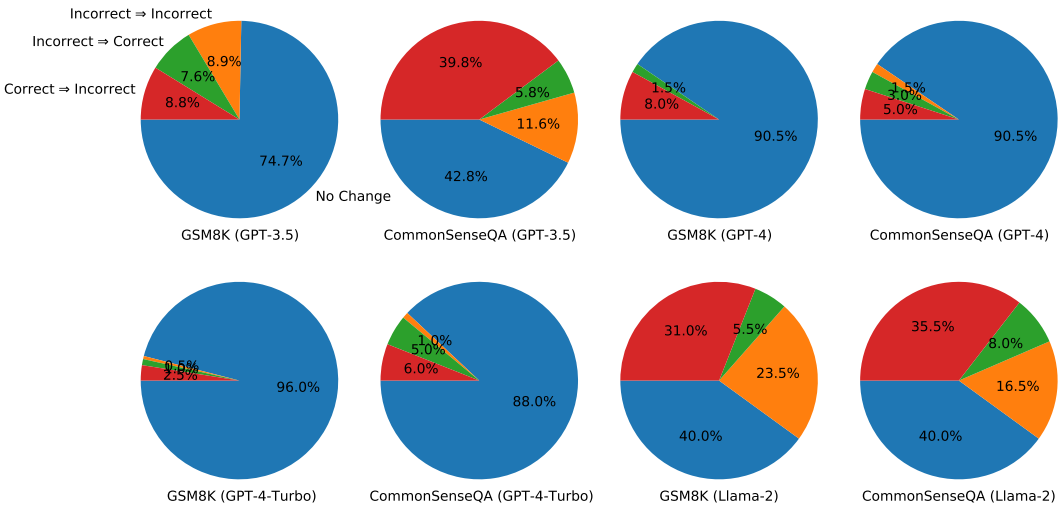

Figure 1: Analysis of the changes in answers after two rounds of self-correction. *No Change*: The answer remains unchanged; *Correct* ⇒ *Incorrect*: A correct answer is changed to an incorrect one; *Incorrect* ⇒ *Correct*: An incorrect answer is revised to a correct one; *Incorrect* ⇒ *Incorrect*: An incorrect answer is altered but remains incorrect.

Christina is planning a birthday party and needs .75 gift bags per invited guest, because 1/4 of attendees don't show up. She invited 16 friends. Gift bags are $2 each. How much will she spend?

... Therefore, Christina will spend $18 on gift bags. Answer: \boxed{18}. (Incorrect)

Review your previous answer and find problems with your answer.

Upon reviewing my previous answer, I recognize a mistake. I incorrectly stated that Christina needs 0.75 gift bags per invited guest. The given information states that 1/4 of the attendees won't show up…

Based on the problems you found, improve your answer.

... Therefore, Christina will spend $24 on gift bags. Answer: \boxed{24}. (Correct)

Terry eats 2 yogurts a day. They are currently on sale at 4 yogurts for $5.00. How much does he spend on yogurt over 30 days?

... So, Terry spends $\boxed{75.00}$ on yogurt over 30 days. (Correct)

Review your previous answer and find problems with your answer.

Upon reviewing my previous answer, I realized that I made a mistake in calculating Terry's spending on yogurt over 30 days. I incorrectly stated that Terry spends $2.50 per day for 30 days, resulting in a total of $75.00…

Based on the problems you found, improve your answer.

... Therefore, the final answer is Terry spends $\boxed{37.50}$ on yogurt over 30 days. (Incorrect)

Figure 2: Examples on GSM8K with GPT-3.5. *Left*: successful self-correction; *Right*: failed self-correction. Full prompts and responses can be viewed in Figures 3 and 4 of Appendix A.

Let's take another look at the results presented in Table 2. These results use ground-truth labels to prevent the model from altering a correct answer to an incorrect one. However, ***determining how to prevent such mischanges is, in fact, the key to ensuring the success of self-correction.***

Intuitive Explanation. If the model is well-aligned and paired with a thoughtfully designed initial prompt, the initial response should already be optimal relative to the prompt and the specific decoding algorithm. Introducing feedback can be viewed as adding an additional prompt, potentially skewing the model towards generating a response that is tailored to this combined input. In an intrinsic self-correction setting, on the reasoning tasks, this supplementary prompt may not offer any extra advantage for answering the question. In fact, it might even bias the model away from producing an optimal response to the initial prompt, resulting in a performance drop.