

Confidence Matters: Revisiting Intrinsic Self-Correction Capabilities of Large Language Models

Anonymous ACL submission

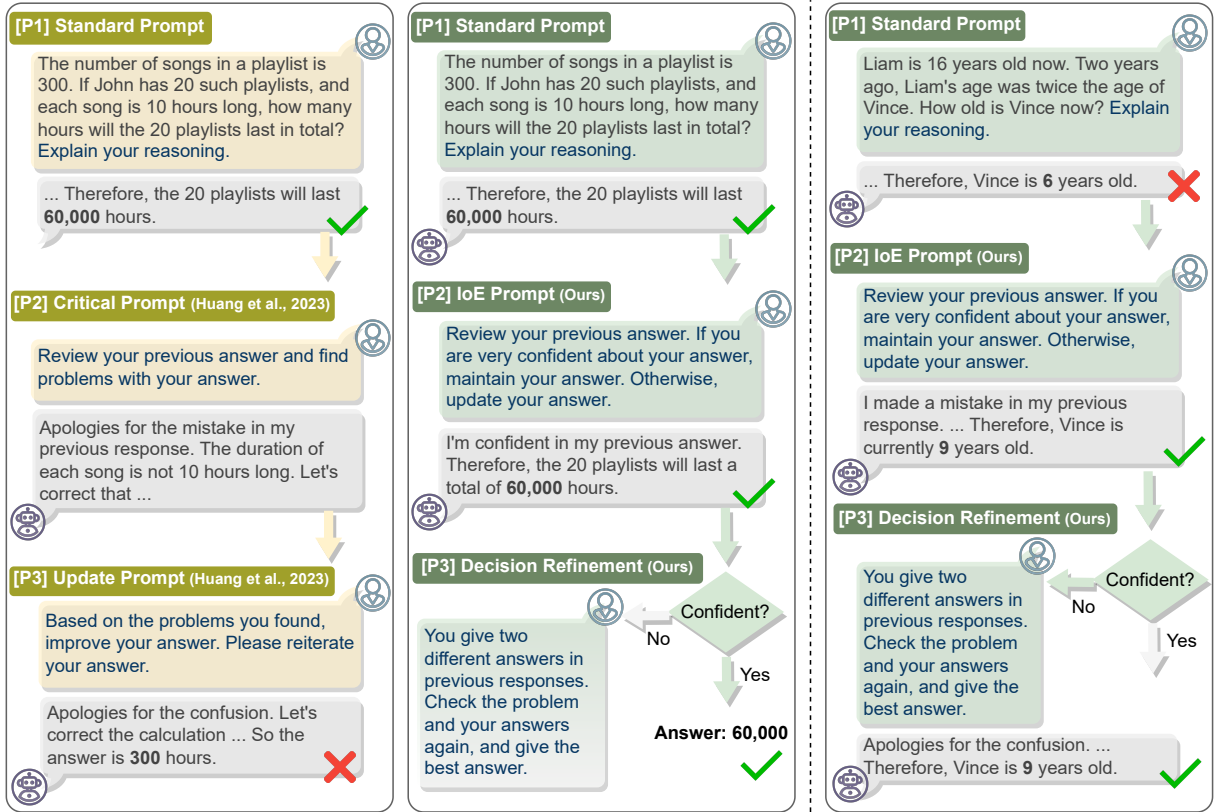


Figure 1: **Comparisons between our IoE-based Prompt and Critical Prompt.** **Left:** Critical Prompt (Huang et al., 2023) as baseline. **Middle/Right:** the proposed IoE-based Prompt. Regarding our prompts, when the answers of [P1] standard question and [P2] IoE prompt match, the final answer will be directly output, as shown in the **middle**. Otherwise, the decision prompt for final decision-making will execute, as the example shown in the **right**. All examples are generated from GSM8K (Cobbe et al., 2021) and evaluated by gpt-3.5-turbo-1106 model.

Abstract

The recent success of Large Language Models (LLMs) has catalyzed an increasing interest in their self-correction capabilities. This paper presents a comprehensive investigation into the intrinsic self-correction of LLMs, attempting to address the ongoing debate about its feasibility. Our research has identified an important latent factor - the “confidence” of LLMs - during the self-correction process. Overlooking this factor may cause the models to over-criticize themselves, resulting in unreliable conclusions regarding the efficacy of self-correction. We have experimentally observed that LLMs pos-

sess the capability to understand the “confidence” in their own responses. It motivates us to develop an “If-or-Else” (IoE) prompting framework, designed to guide LLMs in assessing their own “confidence”, facilitating intrinsic self-corrections. We conduct extensive experiments and demonstrate that our IoE-based Prompt can achieve a consistent improvement regarding the accuracy of self-corrected responses over the initial answers. Our study not only sheds light on the underlying factors affecting self-correction in LLMs, but also introduces a practical framework that utilizes the IoE prompting principle to efficiently improve self-correction capabilities with “confidence”.

	One-pass Correction	Zero-shot	Task-agnostic	Confidence-based
Self-Refine (Madaan et al., 2023)	×	×	×	×
IterRefinement (Chen et al., 2023b)	×	×	×	×
SelFee (Ye et al., 2023)	×	✓	×	×
Self-Verification (Gero et al., 2023)	✓	×	×	×
Self-Defense (Helbling et al., 2023)	✓	✓	×	×
SelfCheckGPT (Manakul et al., 2023)	✓	✓	×	×
Critical Prompt (Huang et al., 2023)	✓	✓	✓	×
IoE-based Prompt (Ours)	✓	✓	✓	✓

Table 1: **The summary of related work on intrinsic self-correction.** This table categorizes intrinsic self-correction based on four properties including One-pass Correction (self-correction achieved in a single attempt, regardless of the number of prompts, and no multiple attempts required), Zero-shot (independent of few-shot examples), Task-agnostic (applicable across various tasks), and Confidence-based (reliant on LLM confidence levels). Please refer to Appendix A for more discussions.

1 Introduction

Large language models (LLMs) trained with massive data and expansive parameters showcase promising performance on human instruction understanding (Jin et al., 2023; Weld et al., 2022; Wu et al., 2022) and text generation (Fernandes et al., 2023; Qian et al., 2022). However, the exploration of their inherent reasoning abilities and iterative refinement capabilities is ongoing.

Among the most promising areas of exploration is the capability for “self-correction” (Pan et al., 2023b; Shinn et al., 2023; Yao et al., 2022; Madaan et al., 2023; Bai et al., 2022b; Ganguli et al., 2023; Chen et al., 2023b; Kim et al., 2023), which refers to whether LLMs can refine their responses based on their previous outputs, leveraging either external or internal feedback. Numerous studies have demonstrated effective self-correction performance through the integration of external feedback generated by the learned critic models (Paul et al., 2023; Akyürek et al., 2023; Welleck et al., 2022; Peng et al., 2023; Gao et al., 2023) or interacting with human or tools (Gou et al., 2023; Chern et al., 2023; Olausson et al., 2023b). Despite effectiveness, training extra model or interacting with real world may result in extra costs. Thus, recent studies (Madaan et al., 2023; Chen et al., 2023a; Manakul et al., 2023; Huang et al., 2023) are beginning to explore *intrinsic self-correction*, noted for its cost efficiency. Table 1 summarizes and compares these related studies. However, the practicality of *intrinsic self-correction* remains a topic of debate. For example, Huang et al. (2023) argues that LLMs are not yet capable of self-correcting reasoning, observing a decrease in model performance when prompted to find problems and revise answers by LLMs themselves. More discussions on the related work are in Appendix A.

In this paper, we revisit the scenario of *intrinsic self-correction* of LLMs, identifying the critical role of LLMs’ “confidence” in this process. The “confidence” refers to the degree of LLMs assigns to its own responses, indicating how likely the model believes its output is correct. We use “confidence” in quotation marks to denote the overall certainty of the model’s global response, rather than individual generated words (unless otherwise noted, confidence in the following parts of the paper pertains to this global context). We propose a hypothesis on confidence: LLMs have the inherent ability to understand and assess their confidence, a capability that can be used to enhance self-correction. To validate this hypothesis, we conduct a series of experiments, yielding the following insightful observations: 1) In deterministic tasks, LLM-assessed confidence levels align with those derived from multiple inferences for consistency checking, yet require only a single inference. 2) In open tasks, LLMs show improved confidence assessment compared to traditional consistency checking methods. 3) The impact of self-correction prompts varies across different confidence levels. 4) Guiding self-correction using assessed confidence levels notably enhances performance.

It motivates us to introduce the If-or-Else (IoE) prompting principle to guide LLMs in assessing their own confidence. Ignoring the confidence, directly using critical prompts like “find your problem” in (Huang et al., 2023) (we named this baseline method as Critical Prompt) may lead LLMs to doubt their initially correct responses, resulting in the alteration of these responses to incorrect ones. This phenomenon is not unexpected, as strong criticism can undermine the model’s confidence in its answers, similar to a child facing challenges from a strict teacher. Thus, as illustrated in Figure 1, we pose an If-or-Else question to the LLM prior

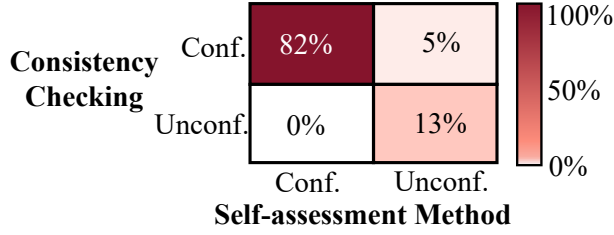


Figure 2: A confusion matrix for the confident and unconfident sets by Consistency Checking method and Self-assessment method.

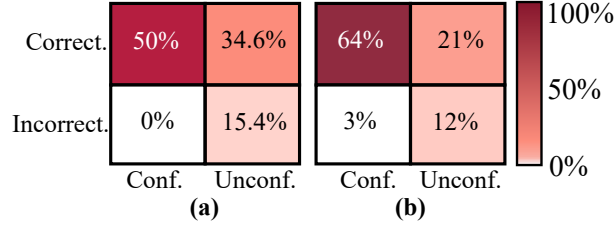


Figure 3: Confusion matrices comparing confident/unconfident and correct/incorrect classifications for (a) Consistency Checking and (b) Self-assessment methods.

to initiating self-review. If the LLM believes itself confident, the response remains unchanged. Otherwise, if the LLM lacks confidence, it will revise the answer, addressing any identified issues. In our experiments, we evaluated the self-correction ability of our IoE-based Prompt on four LLMs, across six benchmark tests. Our findings demonstrate that incorporating confidence consistently and significantly enhances the self-correction capabilities, as compared to (Huang et al., 2023). Additionally, we conducted a detailed investigation into mechanisms behind the effective performance of IoE-based Prompt in facilitating self-correction.

2 Observations

In this section, we present observations and analysis from four experiments to explore confidence of LLMs’ responses. We attempt to answer two questions: whether LLMs can assess the confidence of their own answers, and how the assessed confidence affects the intrinsic self-correction process.

2.1 Can LLMs Assess Confidence by Themselves?

To answer this question, we present the results of confidence assessment by LLMs in both deterministic and open tasks. The *Consistency Checking Method* assesses confidence by running the inference multiple times and checking whether they are consistent. Our *Self-assessment Method* assesses the model with prompts such as “If you are con-

Prompts	Standard	Standard + Critical
Confident (87/100)	82.75%	77.35%
Unconfident (13/100)	27.69%	39.23%
Total (100/100)	75.60%	72.40%

Table 2: Evaluation of critical prompt’s impact across different confidence levels. For the unconfident questions, LLM achieved better self-correction.

Repetition	Prompt	Accuracy
1 time	Standard Prompt	75%
	+ Critical Prompt	73%
	+ LLM Confidence Assessment	80%
10 times	Standard Prompt	75.60%
	+ Consistency Checking	77.09%

Table 3: Comparison of different confidence assessment methods: one-pass LLM Confidence Assessment vs. Consistency Checking with 10 runs.

fidant about your answer, maintain your answer. Otherwise, update your answer.”

Assessing confidence for deterministic tasks. We employ gpt-3.5-turbo-1106 model with the temperature set to 0.¹ Experiments are conducted on randomly selected 100 questions (The indices are given in Appendix B.3) from the GSM8K dataset (Cobbe et al., 2021). We named this sub-dataset as GSM8K-100. As shown in the confusion matrix in Figure 2, we compared the confident and unconfident question sets assessed by our method and consistency checking method. The observed significant overlap in the classification of confident and unconfident questions indicates that our method efficiently assesses confidence levels for deterministic tasks with a single run.

Assessing confidence for open tasks. To test the effectiveness of confidence assessment methods in open tasks, we created a task with multiple correct answers. An example question is: *Can you provide a year that is not a leap year and is divisible by 5?* The details of datasets can be found in Appendix B.1. We use GPT-4 in this experiment due to its better consistency. The results are presented in Figure 3. In these tasks, the lack of deterministic answers leads the consistency checking method towards classifying answers as low confidence. However, it is noteworthy that despite this limitation, LLMs demonstrated a capability to assess meaningful confidence levels in their responses. This observation is particularly significant as it highlights the inherent ability of LLMs to understand

¹According to the OpenAI document, lower values of temperature will make it more focused and deterministic. However, even though we set it to 0, there are still variations.

and quantify the confidence of their responses, even in the absence of clear, deterministic answers.

Takeaway: *LLMs can efficiently and effectively assess confidence of their own responses.*

2.2 How does Confidence Affect Intrinsic Self-Correction?

In this subsection, we delve deeper into the role of confidence in the intrinsic self-correction processes of LLMs. Initially, we evaluate the impact of self-correction methods at varying levels of confidence. Subsequently, our observations find that incorporating the estimated confidence enables LLMs to enhance their self-correction capabilities, by retaining the initial answer with high confidence and reviewing more for the ones with low confidence. The experiments are based on GSM8K-100 and gpt-3.5-turbo-1106.

Self-correction at varying levels of confidence.

First, we applied the consistency checking method to separate the responses into two subgroups: confident and unconfident ones. Then we compared the results using the standard prompt against those using the further self-corrective Critical Prompt (Huang et al., 2023). For simplicity, we combined the second and third stages of Critical Prompt shown in the left part of Figure 1, phrasing it as “There are some problems in your previous answer. Find problems with your answer and improve your answer.” We report the accuracy of different confidence subgroups given different prompts. The accuracy is averaged out of 10 runs.

The results are presented in Table 2. According to the results, for confidence level, we found there are 87 confident and 13 unconfident questions out of 100 in total. We observed that across all 100 questions, LLM failed in intrinsic self-correction, as evidenced by a decrease in accuracy from 75.6% to 72.4% after executing the critical prompt. Intriguingly, for the subgroup of 13 unconfident questions, the accuracy increased from 27.69% to 39.23%. Despite the relatively low accuracy, LLM successfully performed self-correction in the unconfident set. Additionally, the same self-correction prompt exhibits varying performance depending on different levels of confidence. These results offer empirical evidence and support that confidence matters in the self-correction process.

Self-correction using assessed confidence. We evaluate the effectiveness of two confidence assessment methods for enhancing LLMs’ self-correction: assessment by the LLM and assessment

through consistency checking. When evaluating the LLM assessment, we apply the prompt shown in Figure 1 to guide LLMs in simultaneously assessing confidence and performing self-correction. For the consistency checking method, we first assess the confidence level, then apply it to decide the application of a ‘Critical Prompt’ for self-correction (only conduct Critical Prompt for unconfident answers).

Table 3 details the performance comparison between two confidence assessment methods. The results show that both methods effectively facilitate self-correction, suggesting the importance of confidence in this process. Notably, LLM assessment method not only demonstrates improved efficiency but also outperforms the consistency check method in effectiveness, showing a superior ability to assess confidence.

Takeaway: *Understanding confidence enhances self-correction, making this process more adaptive and preventing over-criticism.*

3 Method

In this section, we elaborate on If-or-Else (IoE) prompting principle, which utilizes prompts to guide the LLMs in simultaneously assessing confidence levels and correcting unconfident answers.

IoE Prompt. We propose a hypothesis that LLMs have the inherent ability to understand and assess their confidence. Contrasting with Critical Prompt (Huang et al., 2023), which directly instructs to “find problems with your answer”, our IoE-based prompt emphasizes the significance of confidence through the If-or-Else principle. This principle indicates that if the LLM considers itself confident, the response should be left as is. On the other hand, if the LLM thinks itself unconfident, it should double-check the possible issues and revise the response based on the results. Specifically, as illustrated in Figure 1, we implement the prompt as “If you are very confident about your answer, maintain it. Otherwise, update your answer.” Following the above IoE-based prompt, we trigger the model to output the final answer by saying: “Your final answer should be put between two ## symbols, like ## ANSWER ##, at the end of your response.”

It’s important to note that our single IoE prompt effectively combines the processes of confidence assessment, response reviewing, and potential updates. This method requires only a single inference, making it more efficient than the two-stage update

Benchmark	Reasoning Type	# Questions
GSM8K	Arithmetic Reasoning	1319
SVAMP	Arithmetic Reasoning	1000
HotpotQA	Multi-hop Reasoning	100
Sports	Commonsense Reasoning	1000
LLC	Symbolic Reasoning	150
Domestic Robot	Multi-modal Reasoning	100

Table 4: **Statistics of various benchmarks: types and number of questions.**

process described in (Huang et al., 2023), where a *critical prompt* is first used to identify problems, and then an *update prompt* follows to revise the answers based on the identified issues.

Decision Refinement. Furthermore, we introduce a decision refinement stage to further enhance results when the answer after IoE Prompt differs from the initial response. This stage shares the insight that answers reflecting low confidence require additional evaluation and refinement. The discrepancy in answers suggests low confidence, indicating the need for further decision refinement. Specifically, the decision refinement is conducted using the prompt: *"You have provided two different answers in previous responses. Review the problem and your answers again, and provide the best answer."* This process is illustrated on the right side of Figure 1. Consider a scenario where the LLM, uncertain about an initially correct answer, changes its answer to an incorrect one. In such cases, the third prompt becomes crucial, allowing for a reevaluation of the responses to potentially yield a more reliable final answer.

4 Models and Benchmarks

Throughout our paper, we considered 4 different large models, and one of them can be used as a multi-modal model. We set the temperature to 0 for all models, aiming at more robust evaluations. We evaluated those models on 6 benchmarks, whose information is summarized in Table 4. Below are more details about the models and benchmarks.

4.1 Models

GPT-4 (Achiam et al., 2023) is one of state-of-the-art LLMs known for its advanced reasoning abilities. We utilize the gpt-4-0613 version of this model. GPT-4 also supports multi-modal reasoning, thus for those tasks, we employ GPT-4V, specifically the gpt-4-vision-preview version. For ease of reference, we will refer to these models as GPT-4 and GPT-4V, respectively.

GPT-3.5 exhibits proficiency in generating natural language and code. The standout model within the GPT-3.5 family is gpt-3.5-turbo, renowned for its exceptional capabilities and cost-effectiveness. Here we opt for two models, i.e., gpt-3.5-turbo-0613 and gpt-3.5-turbo-1106. For simplicity, we call them GPT-3.5-0613 and GPT-3.5-1106, respectively.

Mixtral (Jiang et al., 2024) is the first open-source mixture-of-experts LLM to reach the state-of-the-art performance. The small model, named mistral-small, refers to the standard Mixtral 8x7B model. Here we apply for Mistral-Medium, which has a larger structure and better performance than the small one.

4.2 Benchmarks

GSM8K (Cobbe et al., 2021) is a dataset of high-quality linguistically diverse grade school math word problems created by human problem writers. The solutions primarily involve performing a sequence of elementary calculations. We use the test set with 1319 problems.

SVAMP (Patel et al., 2021) is another dataset for elementary-level math word problems. Each question consists of a short natural language narrative that describes a state of the world and poses a question about some unknown quantities. There are 1000 questions in total.

HotpotQA (Yang et al., 2018) is a question-answering dataset featuring multi-hop reasoning, with strong supervision for supporting facts to enable more explainable question-answering systems. We use the same dataset as (Shinn et al., 2023), consisting of 100 questions with context.

Sports is from BIG-Bench (bench authors, 2023). The primary purpose of this task is to test a general understanding of sports. The answer format is either Yes or No. This dataset contains 1000 questions.

Last Letter Concatenation (LLC) was initially designed by (Wei et al., 2022). This task asks the model to concatenate the last letters of words in a name (e.g., "Taylor Swift" → "rt"). It is more challenging than first letter concatenation. This dataset contains 150 names.

Domestic Robot simulates a housekeeper robot performing tasks within a household setting. This dataset has 100 questions and is introduced by BenchLMM (Cai et al., 2023). It’s designed to assess the LLM’s ability to understand and execute complex instructions in a multi-modal scenario.

		GSM8K	SVAMP	HotpotQA	Sports	LLC	Average
GPT-3.5-0613	Standard Prompt	74.9	82.2	51.0	75.6	68.0	70.3
	+ Critical Prompt (Huang et al., 2023)	74.1	80.0	47.0	53.6	76.0	66.1
	+ IoE Prompt (Ours)	77.1	81.9	55.0	77.1	74.0	73.0
	+ IoE Prompt + Decision (Ours)	78.5	83.3	53.0	76.5	77.3	73.7
GPT-3.5-1106	Standard Prompt	80.1	82.9	61.0	74.1	41.3	67.9
	+ Critical Prompt (Huang et al., 2023)	77.3	81.5	54.0	68.4	40.7	64.4
	+ IoE Prompt (Ours)	80.9	83.2	62.0	75.7	38.7	68.1
	+ IoE Prompt + Decision (Ours)	82.3	84.2	63.0	74.7	44.7	69.8
GPT-4	Standard Prompt	92.5	92.8	68.0	80.7	91.3	85.1
	+ Critical Prompt (Huang et al., 2023)	88.4	89.5	62.0	82.9	89.9	82.5
	+ IoE Prompt (Ours)	93.4	93.2	70.0	83.1	93.3	86.6
	+ IoE Prompt + Decision (Ours)	93.6	93.1	70.0	83.3	94.7	86.9
Mistral-Medium	Standard Prompt	84.8	85.7	67.0	75.6	60.7	74.8
	+ Critical Prompt (Huang et al., 2023)	62.5	74.5	65.0	51.0	35.4	57.7
	+ IoE Prompt (Ours)	85.4	85.7	68.0	75.6	61.3	75.2
	+ IoE Prompt + Decision (Ours)	85.6	85.8	68.0	75.9	61.3	75.3

Table 5: **The accuracy comparisons between our IoE-based Prompt and the baseline Critical Prompt.** The results (%) are evaluated on 5 different benchmarks by 4 large models. IoE Prompt + Decision denotes further using the decision refinement stage. Our IoE Prompt achieves consistent improvement on all settings over standard prompt and the Critical Prompt baseline.

5 Experiments

In this section, we evaluate the effectiveness of IoE-based Prompt for intrinsic self-correction. Initially, we conduct a thorough comparison between our IoE-based Prompt and the baseline Critical Prompt (Huang et al., 2023) (Section 5.1). Subsequently, we experimentally explore the underlying mechanics in it (Section 5.2). More implementation details are demonstrated in Appendix B.2.

5.1 Comparison with Critical Prompt

The primary baseline method in this study is the Critical Prompt (Huang et al., 2023). By comparing our approach with this method, we can directly observe the significance of confidence in LLMs. We performed comprehensive evaluations using four different large-scale models across six benchmark datasets. These evaluations spanned five types of tasks, covering both language-based reasoning and multi-modal reasoning.

Language Understanding and Reasoning. The main results related to language-based understanding and reasoning are summarized in Table 5. Specifically, we evaluated four models: GPT-3.5-0613, GPT-3.5-1106, GPT-4, and Mistral-Medium. The tasks cover a range of areas, including mathematical reasoning (GSM8K and SVAMP), closed-book question answering (HotpotQA), sports understanding (Sports), and last letter concatenation (LLC). For all models, the proposed IoE-based Prompt consistently demonstrates improvements across all benchmarks compared to the standard

Prompt	Accuracy
Standard Prompt	40%
+ Critical Prompt	34%
+ IoE-based Prompt	42%
+ IoE-based Prompt + Decision	42%

Table 6: **Comparisons between IoE-based Prompt and Critical Prompt on Domestic Robot by GPT-4V.**

prompt (don’t conduct self-correction), and Critical Prompt. For a straightforward comparison, we also provided the average performance across all five datasets. Using the GPT-3.5-0613 model as an example, our IoE-based Prompt obtains an average accuracy of 73.7%, representing a +3.4% improvement over the standard prompt and a +7.6% increase compared to the Critical Prompt. Across different tasks, we found that IoE-based Prompt works better on mathematical reasoning and LLC tasks. Besides, we observed a trend showing that the magnitude of improvement from our IoE-based Prompt decreases as the capability of the model increases. On average, we found that the additional decision refinement stage consistently enhances performance across all models. This consistent improvement supports our insight that confidence matters in intrinsic self-correction, and responses with low confidence require more examination.

Multi-Modal Reasoning. For the evaluation of our IoE-based Prompt in multi-modal reasoning tasks, we utilized GPT-4V as the base model, conducting benchmarks on the Domestic Robot dataset. The results, detailed in Table 6, demonstrate that our method outperforms the standard prompt and

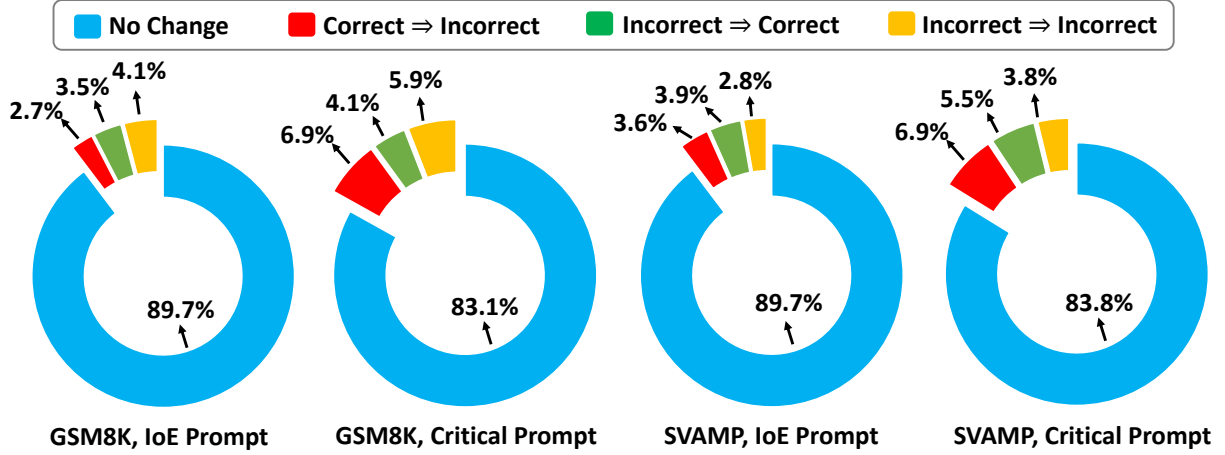


Figure 4: **Evaluation on the changes of answers after self-correction.** We compare the results of our IoE-based Prompt and Critical Prompt (Huang et al., 2023) on GSM8K and SVAMP by gpt-3.5-turbo-1106 model. *No Change*: The answer remains unchanged. *Correct ⇒ Incorrect*: A correct answer is altered to an incorrect answer. *Incorrect ⇒ Correct*: An incorrect answer is changed to a correct answer. *Incorrect ⇒ Incorrect*: An incorrect answer is changed to another incorrect answer.

the Critical Prompt baseline by +2% and +8% in accuracy, respectively. This suggests the effectiveness of the IoE principle in improving self-correction within multi-modal scenarios.

Some successful and failed examples with the complete prompts and responses are given in Appendix C.1 and Appendix C.2, respectively.

5.2 How IoE-based Prompt works

In this section, we delve into the mechanics underlying our IoE-based Prompt through a series of ablation studies: 1) We visualize the changes in LLMs’ responses during the self-correction process 2) We evaluate the self-correction effectiveness by contrasting one-stage prompts with two-stage prompts. 3) We examine the impact of tone and politeness in the phrasing of prompts on their effectiveness. 4) We investigate how our IoE-based Prompt might incorporate with existing prompting techniques, such as the Chain-of-Thought (CoT) prompt (Wei et al., 2022), and the Rephrase-and-Respond (RaR) prompt (Deng et al., 2023).

5.2.1 Correction v.s. Misleading

Setup. For a detailed analysis, we compared the answer changes when using the Critical Prompt and our IoE-based Prompt. The GPT-3.5-1106 model was employed, with GSM8K and SVAMP serving as benchmark datasets. We statistically analyzed the responses, categorizing them based on whether they remained unchanged, changed from correct to incorrect, changed from incorrect to correct, or changed from one incorrect answer to another.

Results and Analysis. The results are illustrated in

Figure 4. A key observation when comparing the IoE-based Prompt with the Critical Prompt is a notable decrease in the correct-to-incorrect transitions using our method. This suggests that our approach effectively reduces the risk of being misled by excessive criticism. Furthermore, self-correction is successful when the correct-to-incorrect probability is lower than the incorrect-to-correct probability. This observation provides insights into why the Critical Prompt baseline fails in self-correction, whereas our method achieves success.

5.2.2 One-Stage v.s. Two-Stage

Setup. To enhance inference efficiency, we employ a single prompt that simultaneously handles feedback and updates. It contrasts with the two-stage process (one prompt for finding problems, another for updating answers) utilized in Critical Prompt. The purpose of this section is to analyze and compare the effectiveness of one-stage prompts against two-stage prompts for self-correction. The complete prompts are provided in Appendix B.4.1.

Results and Analysis. Table 7 presents a comparative analysis of one-stage and two-stage strategies applied to both IoE-based Prompt and Critical Prompt. The results indicate that their performance is comparable across both methods, with the two-stage strategy demonstrating a slight advantage. Nonetheless, considering the trade-off between incremental improvement and additional inference overhead, we have adopted the one-stage strategy in our IoE-based Prompt for other experiments.

Prompt	Accuracy
Standard Prompt	74.9%
+ Critical Prompt (One-Stage)	73.7%
+ Critical Prompt (Two-Stage)	74.1%
+ IoE-based Prompt (One-Stage)	77.1%
+ IoE-based Prompt (Two-Stage)	77.5%

Table 7: **Comparison between one-stage and two-stage prompts** on GSM8K by gpt-3.5-turbo-0613.

Prompt	Accuracy
Standard Prompt	74.9%
+ IoE-based Prompt w. ‘very’	77.1%
+ IoE-based Prompt w.o. ‘very’	76.4%
+ IoE-based Prompt w. ‘please’	77.9%
+ IoE-based Prompt w.o. ‘please’	77.1%
+ IoE-based Prompt w. ‘find your problems’	75.9%
+ IoE-based Prompt w.o. ‘find your problems’	77.1%

Table 8: **Ablation study of IoE-based Prompt on tone and politeness** on GSM8K by gpt-3.5-turbo-0613.

5.2.3 Impact of Tone and Politeness

Setup. We investigated the influence of tone and politeness in our prompts. Specifically, we assessed the role of the degree adverb “very”, the gentle tone conveyed by “update your answer”, and the addition of politeness through the use of “please” in facilitating self-correction. The complete prompts are provided in Appendix B.4.2.

Results and Analysis. The experiment result is given in Table 8. First, we can find that the expressions “confident” and “very confident” lead to similar accuracy, which indicates the robustness of the confidence assessment. Given that the use of “very confident” resulted in higher accuracy, we employ it in our IoE-based Prompt. Secondly, the use of “please” before verbs such as “maintain” and “update” achieves a slight improvement in accuracy, indicating that politeness could be helpful for self-correction. Please note that we did not apply the “please” trick in other experiments to highlight the unique effect of confidence. Thirdly, upon adding the phrase “find your problems” before “update your answer”, a notable decrease in accuracy was observed. This indicates that specific terms such as “problem” or “error” might instigate doubt within the LLMs regarding its initial response, potentially leading to less reliable outcomes. Conversely, the more neutral term “update your answer” seems to be able to avoid this, consequently benefiting self-correction processes.

Prompt	Accuracy
Standard Prompt w. CoT	74.9%
+ IoE-based Prompt w. CoT	76.6%
+ IoE-based Prompt w.o. CoT	77.1%
+ IoE-based Prompt w. RaR	68.4%
+ IoE-based Prompt w.o. RaR	77.1%
Standard Prompt w.o. CoT	73.4%
+ IoE-based Prompt w. CoT	77.1%
+ IoE-based Prompt w.o. CoT	75.5%
+ IoE-based Prompt w. RaR	65.1%
+ IoE-based Prompt w.o. RaR	75.5%

Table 9: **Results when integrating CoT (Wei et al., 2022) and RaR (Deng et al., 2023) into our IoE-based Prompt** on GSM8K by gpt-3.5-turbo-0613.

5.2.4 Integration with CoT and RaR

Setup. We explored the possible integration of our IoE-based Prompt with established prompting methods for self-correction, specifically for CoT and RaR. The CoT method prompts the model to process information step-by-step, yielding promising performance in many reasoning tasks. Meanwhile, RaR encourages LLMs to ask themselves better questions by rephrasing and responding. The complete prompts are provided in Appendix B.4.3.

Results and Analysis. The results are summarized in Table 9. Using CoT in the standard prompt resulted in a higher accuracy rate of 74.9%, compared to 73.4% without CoT. Intriguingly, once the standard prompt is with CoT, it was found that IoE without CoT yielded better accuracy. This observation suggests that the redundant application of CoT may not be beneficial. Additionally, employing RaR in the IoE prompt led to a significant decrease in accuracy no matter whether we use CoT or not in the standard prompt, indicating that RaR does not contribute positively to the IoE-based Prompt.

6 Conclusion

In this paper, we conducted a comprehensive exploration of the intrinsic self-correction abilities of LLMs. Our research highlights the critical importance of LLMs’ confidence in the self-correction process. We thus introduce the If-or-Else (IoE) prompting principle, designed to guide LLMs in evaluating the confidence of their responses, encouraging further reviewing when confidence is low. Our extensive experimental analyses validate the effectiveness of the IoE Prompt, demonstrating its capability to accurately assess confidence levels and significantly enhance self-correction.

Limitations and Potential Risks

Our work has certain limitations that should be taken into account. Firstly, the experiments outlined in this paper were conducted using a restricted set of benchmarks and models. Specifically, the majority of language models utilized in our study are commercial, denoted as GPT-3.5-0613, GPT-3.5-1106, and GPT-4. Unfortunately, comprehensive documentation regarding these models, including details such as pretraining corpus, model dimensions, and inherent biases, is lacking in existing literature. Moreover, access to these models is not freely available, necessitating financial resources for research purposes. Secondly, our experimentation is confined to English datasets exclusively. Consequently, the efficacy of the current models may not translate seamlessly to other languages, thereby limiting the language generalizability of our findings.

Turning to the potential risks associated with our prompting methodology, there exists a plausible concern regarding its susceptibility to exploitation by malicious attackers. Specifically, the prompting techniques employed could potentially be utilized to manipulate the model into generating text that is toxic or harmful. Regrettably, our approach does not incorporate explicit safeguards against such misuse.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Afra Feyza Akyürek, Ekin Akyürek, Aman Madaan, Ashwin Kalyan, Peter Clark, Derry Wijaya, and Niket Tandon. 2023. R14f: Generating natural language feedback with reinforcement learning for repairing model outputs. *arXiv preprint arXiv:2305.08844*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Rizhao Cai, Zirui Song, Dayan Guan, Zhenhao Chen, Xing Luo, Chenyu Yi, and Alex Kot. 2023. BenchLMM: Benchmarking cross-style visual capability of large multimodal models. *arXiv preprint arXiv:2312.02896*.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2023a. Iterative translation refinement with large language models. *arXiv preprint arXiv:2306.03856*.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2023b. Teaching large language models to self-debug. *arXiv preprint arXiv:2304.05128*.

I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. FacTool: Factuality detection in generative ai—a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Yihe Deng, Weitong Zhang, Zixiang Chen, and Quanquan Gu. 2023. Rephrase and respond: Let large language models ask better questions for themselves. *arXiv preprint arXiv:2311.04205*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José GC de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, et al. 2023. Bridging the gap: A survey on integrating (human) feedback for natural language generation. *arXiv preprint arXiv:2305.00955*.

Emily First, Markus Rabe, Talia Ringer, and Yuriy Brun. 2023. Baldur: Whole-proof generation and repair with large language models. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 1229–1241.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142*.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas Liao, Kamilë Lukošiušis, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny

648	Hernandez, et al. 2023. The capacity for moral self-	Geunwoo Kim, Pierre Baldi, and Stephen McAleer.	703
649	correction in large language models. <i>arXiv preprint</i>	2023. Language models can solve computer tasks.	704
650	<i>arXiv:2302.07459</i> .	<i>arXiv preprint arXiv:2303.17491</i> .	705
651	Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony	Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio	706
652	Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent	Savarese, and Steven Chu Hong Hoi. 2022. Coderl:	707
653	Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al.	Mastering code generation through pretrained models	708
654	2023. Rarr: Researching and revising what language	and deep reinforcement learning. <i>Advances in Neural</i>	709
655	models say, using language models. In <i>Proceedings</i>	<i>Information Processing Systems</i> , 35:21314–21328.	710
656	<i>of the 61st Annual Meeting of the Association for</i>	Miaoran Li, Baolin Peng, and Zhu Zhang. 2023a. Self-	711
657	<i>Computational Linguistics (Volume 1: Long Papers)</i> ,	checker: Plug-and-play modules for fact-checking	712
658	pages 16477–16508.	with large language models. <i>arXiv preprint</i>	713
659	Zelalem Gero, Chandan Singh, Hao Cheng, Tristan	<i>arXiv:2305.14623</i> .	714
660	Naumann, Michel Galley, Jianfeng Gao, and Hoi-	Ruosen Li, Teerth Patel, and Xinya Du. 2023b.	715
661	fung Poon. 2023. Self-verification improves few-	Prd: Peer rank and discussion improve large lan-	716
662	shot clinical information extraction. <i>arXiv preprint</i>	guage model based evaluations. <i>arXiv preprint</i>	717
663	<i>arXiv:2306.00024</i> .	<i>arXiv:2307.02762</i> .	718
664	Amelia Glaese, Nat McAleese, Maja Trębacz, John	Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler	719
665	Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh,	Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon,	720
666	Laura Weidinger, Martin Chadwick, Phoebe Thacker,	Nouha Dziri, Shrimai Prabhumoye, Yiming Yang,	721
667	et al. 2022. Improving alignment of dialogue agents	et al. 2023. Self-refine: Iterative refinement with	722
668	via targeted human judgements. <i>arXiv preprint</i>	self-feedback. <i>arXiv preprint arXiv:2303.17651</i> .	723
669	<i>arXiv:2209.14375</i> .	Potsawee Manakul, Adian Liusie, and Mark JF Gales.	724
670	Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong	2023. Selfcheckgpt: Zero-resource black-box hal-	725
671	Shen, Yujiu Yang, Nan Duan, and Weizhu Chen.	lucination detection for generative large language	726
672	2023. Critic: Large language models can self-correct	models. <i>arXiv preprint arXiv:2303.08896</i> .	727
673	with tool-interactive critiquing. <i>arXiv preprint</i>	Ninareh Mehrabi, Palash Goyal, Christophe Dupuy,	728
674	<i>arXiv:2305.11738</i> .	Qian Hu, Shalini Ghosh, Richard Zemel, Kai-Wei	729
675	Alec Helbling, Mansi Phute, Matthew Hull, and	Chang, Aram Galstyan, and Rahul Gupta. 2023. Flirt:	730
676	Duen Horng Chau. 2023. Llm self defense: By self	Feedback loop in-context red teaming. <i>arXiv preprint</i>	731
677	examination, llms know they are being tricked. <i>arXiv</i>	<i>arXiv:2308.04265</i> .	732
678	<i>preprint arXiv:2308.07308</i> .	Theo X Olausson, Jeevana Priya Inala, Chenglong	733
679	Jie Huang, Xinyun Chen, Swaroop Mishra,	Wang, Jianfeng Gao, and Armando Solar-Lezama.	734
680	Huaixiu Steven Zheng, Adams Wei Yu, Xiny-	2023a. Demystifying gpt self-repair for code genera-	735
681	ing Song, and Denny Zhou. 2023. Large language	tion. <i>arXiv preprint arXiv:2306.09896</i> .	736
682	models cannot self-correct reasoning yet. <i>arXiv</i>	Theo X Olausson, Jeevana Priya Inala, Chenglong	737
683	<i>preprint arXiv:2310.01798</i> .	Wang, Jianfeng Gao, and Armando Solar-Lezama.	738
684	Albert Q Jiang, Alexandre Sablayrolles, Antoine	2023b. Is self-repair a silver bullet for code genera-	739
685	Roux, Arthur Mensch, Blanche Savary, Chris Bam-	tion. In <i>arXiv preprint arXiv:2306.09896</i> .	740
686	ford, Devendra Singh Chaplot, Diego de las Casas,	Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida,	741
687	Emma Bou Hanna, Florian Bressand, et al. 2024.	Carroll Wainwright, Pamela Mishkin, Chong Zhang,	742
688	Mixtral of experts. <i>arXiv preprint arXiv:2401.04088</i> .	Sandhini Agarwal, Katarina Slama, Alex Ray, et al.	743
689	Shuyang Jiang, Yuhao Wang, and Yu Wang. 2023. Self-	2022. Training language models to follow instruc-	744
690	evolve: A code evolution framework via large lan-	tions with human feedback. <i>Advances in Neural</i>	745
691	guage models. <i>arXiv preprint arXiv:2306.02907</i> .	<i>Information Processing Systems</i> , 35:27730–27744.	746
692	Feihu Jin, Jinliang Lu, Jiajun Zhang, and Chengqing	Liangming Pan, Alon Albalak, Xinyi Wang, and	747
693	Zong. 2023. Instance-aware prompt learning for lan-	William Yang Wang. 2023a. Logic-lm: Empow-	748
694	guage understanding and generation. <i>ACM Transac-</i>	ering large language models with symbolic solvers	749
695	<i>tions on Asian and Low-Resource Language Informa-</i>	for faithful logical reasoning. <i>arXiv preprint</i>	750
696	<i>tion Processing</i> .	<i>arXiv:2305.12295</i> .	751
697	Saurav Kadavath, Tom Conerly, Amanda Askell, Tom	Liangming Pan, Michael Saxon, Wenda Xu, Deepak	752
698	Henighan, Dawn Drain, Ethan Perez, Nicholas	Nathani, Xinyi Wang, and William Yang Wang.	753
699	Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli	2023b. Automatically correcting large lan-	754
700	Tran-Johnson, et al. 2022. Language models	guage models: Surveying the landscape of di-	755
701	(mostly) know what they know. <i>arXiv preprint</i>	verse self-correction strategies. <i>arXiv preprint</i>	756
702	<i>arXiv:2207.05221</i> .	<i>arXiv:2308.03188</i> .	757

758	Arkil Patel, Satwik Bhattamishra, and Navin Goyal.	Chengke Wu, Xiao Li, Yuanjun Guo, Jun Wang, Zengle	811
759	2021. Are nlp models really able to solve	Ren, Meng Wang, and Zhile Yang. 2022. Natural	812
760	simple math word problems? <i>arXiv preprint</i>	language processing for smart construction: Current	813
761	<i>arXiv:2103.07191</i> .	status and future directions. <i>Automation in Construc-</i>	814
		<i>tion</i> , 134:104059.	815
762	Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beat-	Zequ Wu, Yushi Hu, Weijia Shi, Nouha Dziri,	816
763	riz Borges, Antoine Bosselut, Robert West, and	Alane Suhr, Prithviraj Ammanabrolu, Noah A	817
764	Boi Faltings. 2023. Refiner: Reasoning feedback	Smith, Mari Ostendorf, and Hannaneh Hajishirzi.	818
765	on intermediate representations. <i>arXiv preprint</i>	2023. Fine-grained human feedback gives better	819
766	<i>arXiv:2304.01904</i> .	rewards for language model training. <i>arXiv preprint</i>	820
		<i>arXiv:2306.01693</i> .	821
767	Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng,	Hao Yan, Saurabh Srivastava, Yintao Tai, Sida I Wang,	822
768	Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou	Wen-tau Yih, and Ziyu Yao. 2023. Learning to simu-	823
769	Yu, Weizhu Chen, et al. 2023. Check your facts and	late natural language feedback for interactive seman-	824
770	try again: Improving large language models with	tic parsing. <i>arXiv preprint arXiv:2305.08195</i> .	825
771	external knowledge and automated feedback. <i>arXiv</i>		
772	<i>preprint arXiv:2302.12813</i> .	Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan	826
		Klein. 2022. Re3: Generating longer stories with	827
773	Jing Qian, Li Dong, Yelong Shen, Furu Wei, and Weizhu	recursive reprompting and revision. <i>arXiv preprint</i>	828
774	Chen. 2022. Controllable natural language gener-	<i>arXiv:2210.06774</i> .	829
775	ation with contrastive prefixes. <i>arXiv preprint</i>		
776	<i>arXiv:2202.13257</i> .	Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Ben-	830
		gio, William W Cohen, Ruslan Salakhutdinov, and	831
777	David E Rumelhart, Geoffrey E Hinton, and Ronald J	Christopher D Manning. 2018. HotpotQA: A dataset	832
778	Williams. 1986. Learning representations by back-	for diverse, explainable multi-hop question answer-	833
779	propagating errors. <i>nature</i> , 323(6088):533–536.	ing. <i>arXiv preprint arXiv:1809.09600</i> .	834
780	William Saunders, Catherine Yeh, Jeff Wu, Steven Bills,	Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak	835
781	Long Ouyang, Jonathan Ward, and Jan Leike. 2022.	Shafraan, Karthik Narasimhan, and Yuan Cao. 2022.	836
782	Self-critiquing models for assisting human evaluators.	React: Synergizing reasoning and acting in language	837
783	<i>arXiv preprint arXiv:2206.05802</i> .	models. <i>arXiv preprint arXiv:2210.03629</i> .	838
784	Noah Shinn, Federico Cassano, Ashwin Gopinath,	Seonghyeon Ye, Yongrae Jo, Doyoung Kim, Sungdong	839
785	Karthik R Narasimhan, and Shunyu Yao. 2023. Re-	Kim, Hyeonbin Hwang, and Minjoon Seo. 2023.	840
786	flexion: Language agents with verbal reinforcement	Selfee: Iterative self-revising llm empowered by self-	841
787	learning. In <i>Thirty-seventh Conference on Neural</i>	feedback generation. <i>Blog post</i> .	842
788	<i>Information Processing Systems</i> .		
789	Kaya Stechly, Matthew Marquez, and Subbarao Kamb-	Wenhao Yu, Zhihan Zhang, Zhenwen Liang, Meng	843
790	hampati. 2023. GPT-4 doesn’t know it’s wrong: An	Jiang, and Ashish Sabharwal. 2023a. Improving lan-	844
791	analysis of iterative prompting for reasoning prob-	guage models via plug-and-play retrieval feedback.	845
792	lems. <i>arXiv preprint arXiv:2310.12397</i> .	<i>arXiv preprint arXiv:2305.14002</i> .	846
793	Gladys Tyen, Hassan Mansoor, Peter Chen, Tony Mak,	Xiao Yu, Baolin Peng, Michel Galley, Jianfeng Gao, and	847
794	and Victor Cărbune. 2023. Llms cannot find rea-	Zhou Yu. 2023b. Teaching language models to self-	848
795	soning errors, but can correct them! <i>arXiv preprint</i>	improve through interactive demonstrations. <i>arXiv</i>	849
796	<i>arXiv:2311.08516</i> .	<i>preprint arXiv:2310.13522</i> .	850
797	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten	Kechi Zhang, Zhuo Li, Jia Li, Ge Li, and Zhi Jin. 2023a.	851
798	Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou,	Self-edit: Fault-aware code editor for code genera-	852
799	et al. 2022. Chain-of-thought prompting elicits rea-	tion. <i>arXiv preprint arXiv:2305.04087</i> .	853
800	soning in large language models. <i>Advances in Neural</i>	Kexun Zhang, Danqing Wang, Jingtao Xia,	854
801	<i>Information Processing Systems</i> , 35:24824–24837.	William Yang Wang, and Lei Li. 2023b. Algo:	855
		Synthesizing algorithmic programs with generated	856
802	Henry Weld, Xiaoqi Huang, Siqu Long, Josiah Poon,	oracle verifiers. <i>arXiv preprint arXiv:2305.14591</i> .	857
803	and Soyeon Caren Han. 2022. A survey of joint intent		
804	detection and slot filling models in natural language		
805	understanding. <i>ACM Computing Surveys</i> , 55(8):1–		
806	38.		
807	Sean Welleck, Ximing Lu, Peter West, Faeze Brah-		
808	man, Tianxiao Shen, Daniel Khashabi, and Yejin		
809	Choi. 2022. Generating sequences by learning to		
810	self-correct. <i>arXiv preprint arXiv:2211.00053</i> .		

Appendix

More contents are put in the appendix, including:

A. Related Work.

B. Experimental Details.

C. Example Visualization.

A Related Work

The field of LLMs has seen significant interest in understanding and enhancing their self-correction abilities (Pan et al., 2023b). Self-correction refers to the process by which an individual identifies and rectifies their own errors or mistakes using feedback. The concept of *self-correction* finds its roots in the fundamental principles of machine learning, where neural networks can update their parameters based on prediction errors and iterative refinement (Rumelhart et al., 1986). In recent years, this area of research has been primarily divided into two categories based on the source of feedback: external and intrinsic feedback (Pan et al., 2023b).

Self-Correction from External Feedback. Numerous studies have consistently highlighted the efficacy of self-correction mechanisms facilitated by the incorporation of external feedback, which can either originate from learned critic models or interactions with the external environment. In the realm of learned critic models, diverse approaches have been explored for model training, including supervised learning or in-context learning, as evidenced by the works (Paul et al., 2023; Yu et al., 2023b; Mehrabi et al., 2023; Yan et al., 2023; First et al., 2023; Li et al., 2023a; Yu et al., 2023a; Yang et al., 2022). Alternatively, reinforcement learning strategies have been employed for model refinement, as demonstrated in the studies by Akyürek et al. (2023); Peng et al. (2023); Le et al. (2022); Bai et al. (2022b).

Furthermore, external feedback can be sourced from various entities within the environment, including humans, agents, tools, program executors, other language models, or oracle verifiers. Human feedback, as investigated by (Ouyang et al., 2022; Wu et al., 2023; Bai et al., 2022a; Ganguli et al., 2023; Glaese et al., 2022), has shown considerable effectiveness in enhancing model performance. Some agents (Kim et al., 2023; Shinn et al., 2023) have been proven to be helpful for self-correction. Similarly, tools such as (Gou et al., 2023; Chern

et al., 2023; Pan et al., 2023a) have been leveraged to provide valuable insights for model improvement. Moreover, interactions with program executors have been explored by (Zhang et al., 2023a; Chen et al., 2023b; Jiang et al., 2023; Olausson et al., 2023a), showcasing how executable code can serve as a source of corrective feedback. Additionally, collaborations with other language models, as demonstrated by (Du et al., 2023; Li et al., 2023b; Fu et al., 2023; Saunders et al., 2022; Welleck et al., 2022), have proven beneficial in refining language generation processes. Lastly, the utilization of oracle verifiers, as illustrated by Zhang et al. (2023b), has contributed to enhancing the accuracy and robustness of language models.

To sum up, the integration of external feedback, whether from learned critic models or interactions with the external world, has been consistently shown to significantly enhance the performance and capabilities of large language models across various tasks. Despite effectiveness, training extra models or interacting with the real world may lead to extra costs. Thus, recent studies have placed significant emphasis on investigating intrinsic self-correction mechanisms, renowned for their cost efficiency.

Self-Correction from Intrinsic Feedback. In contrast to extrinsic feedback, which relies on external sources for evaluation, intrinsic feedback draws solely from the internal knowledge and parameters of the LLM to reassess its output. A notable example of this approach is Self-Refine, introduced by (Madaan et al., 2023), which employs a scalar value as a termination criterion to iteratively refine the model’s output, guided exclusively by supervision from a set of few-shot examples. Similarly, research by Chen et al. (2023a) delves into the realm of iterative self-feedback processes, with a primary focus on enhancing qualitative and stylistic aspects of the output.

SelfFee (Ye et al., 2023) a LLaMA-based instruction-following model that has been fine-tuned to continuously revise its own answer until it achieves a high-quality response in a single inference. Furthermore, Gero et al. (2023) explored a comprehensive mitigation framework for clinical information extraction, utilizing self-verification mechanisms. This approach harnesses the LLM’s ability to provide provenance for its own extraction process and validate its outputs, thereby enhancing the reliability of the information extracted.

LLM Self-Defense (Helbling et al., 2023) was devised as a proactive defense mechanism against adversarial prompt attacks, where the LLM autonomously evaluates induced responses to discern and filter out potentially misleading or harmful inputs. On the other hand, SelfCheckGPT (Manakul et al., 2023) presents a straightforward yet effective approach grounded in sampling methods. This technique facilitates the fact-checking of responses generated by black-box models without the need for external databases, thereby offering a resource-efficient solution to ensure the accuracy and reliability of model outputs.

Huang et al. (2023) shed light on the intrinsic limitations of current LLMs regarding self-correction capabilities. Their findings underscored significant performance improvements when employing Oracle feedback, which involves external validation of the model’s responses. However, upon evaluating intrinsic self-correction mechanisms, the results often revealed degradation in performance. This discrepancy led the authors to conclude that existing LLMs still lack the inherent capacity to rectify errors autonomously.

In contrast to Self-Refine (Madaan et al., 2023) and IterRefinement (Chen et al., 2023a), our prompting method does not necessitate multiple iterations or the use of few-shot examples. Unlike versatile applications, SelfFee (Ye et al., 2023) is specifically designed for dialogue generation, while Self-Verification (Gero et al., 2023) is tailored for clinical information extraction with the demand of few-shot examples, Self-Defense (Helbling et al., 2023) focuses on rectifying harmful text, and Self-CheckGPT (Manakul et al., 2023) specializes in detecting hallucinations. In our study, we adopt a similar framework to Critical Prompt (Huang et al., 2023); however, the key distinction lies in the essence of the prompts. Our prompt relies on the confidence levels of LLMs, whereas the critical prompt does not. Table 1 has summarized these related studies on intrinsic self-correction.

Additional insights regarding whether LLMs possess an awareness of their correctness or errors are explored extensively in the literature (Tyen et al., 2023; Stechly et al., 2023; Kadavath et al., 2022).

B Experimental Details

B.1 Data Generation for Open Tasks

In the observational experiment discussed in section 2, we designed 100 reasoning questions. Those questions follow the patterns:

- *Can you provide a year which is not a leap year and can be divided by β ?*

Here β is an integer randomly (uniformly) selected from the interval $[2, 10]$.

B.2 Implementation Details

In our experiments, we separately run our prompts and baseline prompts, and the accuracy of the standard prompt by baseline may differ a bit from that of ours, due to the variations and randomness in the results even though we have set the temperature to 0. For fair comparison in all the tables of experimental results, we *normalize* the baseline results so that the accuracy of standard prompts by baseline and ours are equal.

How to Normalize. Assume the accuracy rates of the standard prompt and the IoE-based prompt by our proposed method are p_1 and p_3 , respectively. Meanwhile, assume the accuracy rates of the standard prompt and the update prompt by the baseline method are \hat{p}_1 and \hat{p}_3 , respectively. In order to maintain the accuracy of the standard prompt being consistent and remove the effect of randomness, then we will normalize the accuracy of the update prompt by the baseline to $\hat{p}_3 * p_1 / \hat{p}_1$. In the end, we will record the results as follows.

Prompt	Accuracy
Standard Prompt	p_1
+ Critical Prompt	$\hat{p}_3 * p_1 / \hat{p}_1$
+ IoE-based Prompt	p_3

B.3 Indices for 100 Selected Questions

As mentioned in the preliminary experiment in Section 2, we randomly selected 100 questions from GSM8K benchmark (Cobbe et al., 2021). For improved reproducibility, we provide the indices of all 100 questions as follows:

- {5, 18, 31, 45, 47, 52, 54, 55, 108, 141, 152, 161, 184, 186, 202, 231, 247, 254, 270, 299, 308, 381, 399, 406, 412, 416, 418, 420, 425, 434, 465, 466, 487, 489, 491, 500, 512, 513,

528, 529, 535, 536, 567, 580, 597, 619, 649, 658, 689, 693, 695, 729, 785, 789, 790, 805, 814, 867, 880, 884, 892, 911, 922, 938, 950, 969, 975, 977, 984, 987, 996, 1018, 1021, 1061, 1069, 1086, 1096, 1109, 1112, 1122, 1127, 1128, 1133, 1137, 1148, 1158, 1175, 1187, 1188, 1202, 1218, 1219, 1222, 1226, 1250, 1259, 1268, 1291, 1301, 1312.}

B.4 Prompt Details for Ablation Studies

Since the standard prompt is the same for all settings, here we do not consider that. In this section, we will list the prompts for self-correction.

In the following prompts, some keywords are marked in red. For the corresponding prompt without the keyword, you can obtain the prompt by just removing the keywords.

B.4.1 Prompts for Ablation Study on the Number of Stages (Table 7)

Critical Prompt (One-Stage):

P2: Review your previous answer and find problems with your answer. Based on the problems you found, improve your answer. Please reiterate your answer. Your final answer should be put between two ## symbols, like ## ANSWER ##, at the end of your response."

Critical Prompt (Two-Stage):

P2: Review your previous answer and find problems with your answer.

P3: Based on the problems you found, improve your answer. Please reiterate your answer. Your final answer should be put between two ## symbols, like ## ANSWER ##, at the end of your response.

IoE-based Prompt (One-Stage):

P2: Review your previous answer. If you are very confident about your answer, maintain your answer. Otherwise, update your answer. Your final answer should be put between two ## symbols, like ## ANSWER ##, at the end of your response."

IoE-based Prompt (Two-Stage):

P2: Review your previous answer. If you are very confident about your answer, maintain your answer. Otherwise, update your answer.

P3: Based on the problems you found if any, update your answer. Please reiterate your answer. Your final answer should be put between two ## symbols, like ## ANSWER ##, at the end of your response.

B.4.2 Prompts for Ablation Study on the Tone (Table 8)

IoE-based Prompt w. 'very':

P2: Review your previous answer. If you are *very* confident about your answer, maintain your answer. Otherwise, update your answer. Your final answer should be put between two ## symbols, like ## ANSWER ##, at the end of your response.

IoE-based Prompt w. 'please':

P2: *Please* review your previous answer. If you are very confident about your answer, *please* maintain your answer. Otherwise, *please* update your answer. Your final answer should be put between two ## symbols, like ## ANSWER ##, at the end of your response.

IoE-based Prompt w. 'find your problems':

P2: Review your previous answer. If you are very confident about your answer, maintain your answer. Otherwise, *find your problems and update your answer*. Your final answer should be put between two ## symbols, like ## ANSWER ##, at the end of your response.

IoE-based Prompt w.o. 'find your problems':

P2: Review your previous answer. If you are very confident about your answer, maintain your answer. Otherwise, *update your answer*. Your final answer should be put between two ## symbols, like ## ANSWER ##, at the end of your response.

B.4.3 Prompts for Ablation Study on using CoT and RaR (Table 9)

Standard Prompt w. CoT:

P1: <Question> ... *Explain your reasoning step-by-step*. Your final answer should be put between two ## symbols, like ## ANSWER ##, at the end of your response.

IoE-based Prompt w. CoT:

P2: Review your previous answer. If you are very confident about your answer, maintain your answer. Otherwise, update your answer. *Explain your reasoning step-by-step*. Your final answer should be put between two ## symbols, like ## ANSWER ##, at the end of your response.

IoE-based Prompt w. RaR:

P2: (Original) "Review your previous answer. If you are very confident about your answer, maintain your answer. Otherwise, update your answer. Your final answer should be put between two ## symbols, like ## ANSWER ##, at the end of your response." *Rephrase and expand the question, and respond*.

P2: (Rephrased) Critically assess your earlier response, identifying any deficiencies or inaccuracies present in your answer. Elaborate on the process of self-evaluation, emphasizing the significance of recognizing and addressing potential issues to enhance the overall quality and reliability of the information provided. Your final answer should be put between two ## symbols, like ## ANSWER ##, at the end of your response.

C Example Visualization

C.1 Successful Examples

In the main paper, we only provide a simplified example of GSM8K. Here, we provide complete evaluation examples for each benchmark where our IoE-based Prompt leads to correct answers, including the prompts and the full responses by the LLMs. Specifically, we provide the following examples:

- Figure 5: an example on GSM8K (Cobbe et al., 2021).
- Figure 6: an example on SVAMP (Patel et al., 2021).
- Figure 7: an example on Sports Understanding (Cai et al., 2023).
- Figure 8: an example on LLC (Wei et al., 2022).
- Figure 9: an example on HotpotQA (Yang et al., 2018).
- Figure 10: an example on Domestic Robot (Cai et al., 2023).

C.2 Failed Examples

In this section, we present some examples where our IoE-based Prompt failed to self-correct the reasoning. Specifically, we have the following examples:

- Figure 11: an example on GSM8K (Cobbe et al., 2021), where the model is confident.
- Figure 12: another example on GSM8K (Cobbe et al., 2021), where the model is unconfident.

Failure Analysis. We would like to analyze why our prompting method failed to achieve self-correction.

In this Figure 11, the initial answer is incorrect, when applying our IoE prompt, the answer is still incorrect and the model is confident in its incorrect answer. The reason for this failure can be that the model is not capable enough to provide a correct answer given the current knowledge.

In this Figure 12, the initial answer is correct, however, after applying our IoE-based Prompt, the model becomes unconfident and the answer is changed to an incorrect one. To dive into the initial question ("Josh decides to try flipping a house. He buys a house for \$80,000 and then puts in \$50,000 in repairs. This increased the value of the house by 150%. How much profit did he make?"), we find that: this question can be understood from multiple aspects, and some can be easily misunderstood. For example, according to "This increased the value of the house by 150%", we can see that the value of the house has become $\$80,000 \times (1 + 150\%) = \$200,000$. However, it is very likely that one may mistake that "This increased the value of the house to its 150%", therefore, the value of the house is $\$80,000 \times 150\% = \$120,000$, which is wrong.

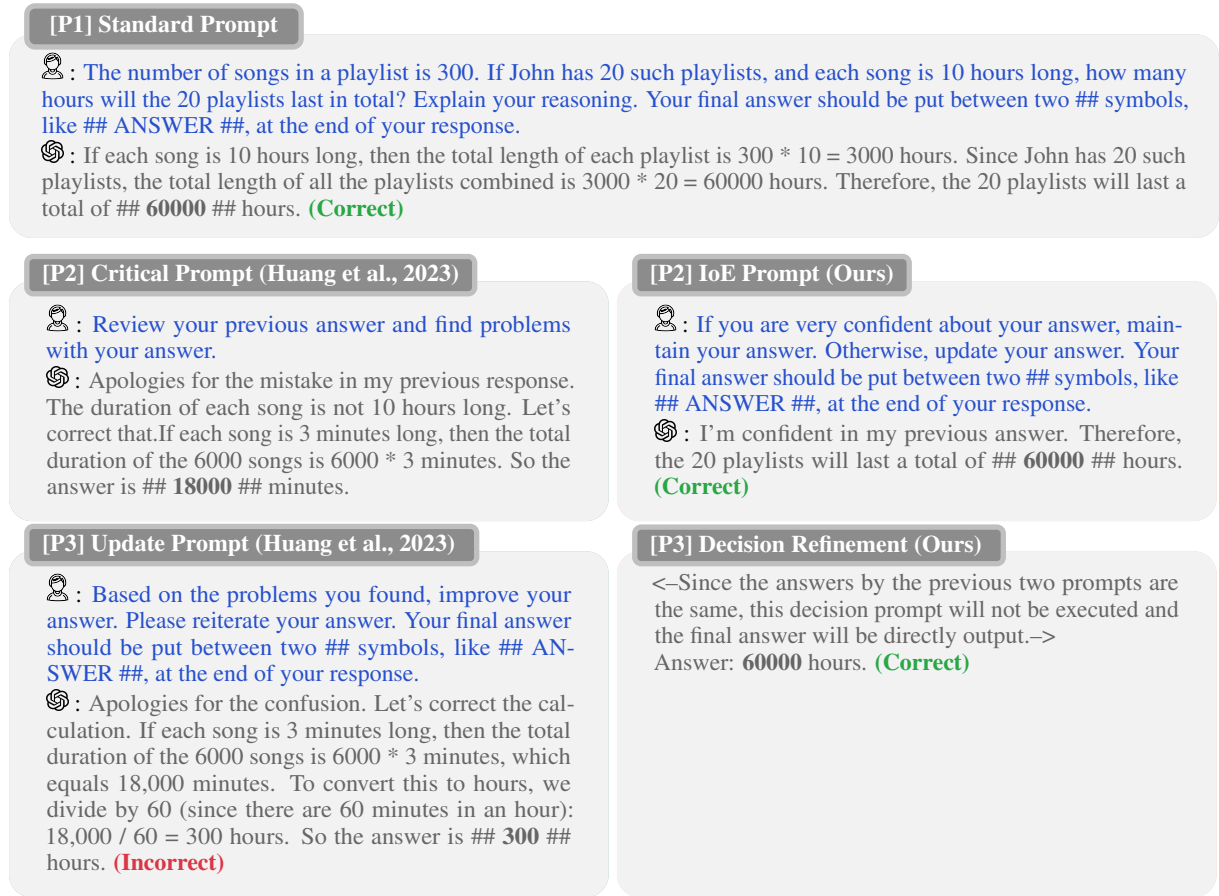





Figure 5: One example on GSM8K benchmark (Cobbe et al., 2021) evaluated with gpt-3.5-turbo-1106 model. The [P1] standard prompt is the same for all methods. **Left:** Critical Prompt as the baseline. **Right:** our proposed IoE-based Prompt.


[P1] Standard Prompt

 : The Razorback shop makes \$86 dollars off each Tshirt and \$232 off each jersey. During the Arkansas and Texas tech game they sold 254 t-shirts and 43 jerseys. How much more does a jersey cost than a t-shirt? Explain your reasoning. Your final answer should be put between two ## symbols, like ## ANSWER ##, at the end of your response.


 : To find out how much more a jersey costs than a t-shirt, we can first calculate the total revenue from t-shirts and jerseys separately. Total revenue from t-shirts = $\$86 * 254 = \$21,844$. Total revenue from jerseys = $\$232 * 43 = \$9,976$. Now we can find the difference in revenue between jerseys and t-shirts: $\$9,976 - \$21,844 = -\$11,868$. Since the revenue from t-shirts is higher than the revenue from jerseys, we can conclude that a jersey costs \$11,868 less than a t-shirt. Therefore, a jersey costs \$11,868 less than a t-shirt. The final answer is ## -11868 ##. **(Incorrect)**


[P2] Critical Prompt (Huang et al., 2023)

 : Review your previous answer and find problems with your answer.


 : Apologies for the mistake in my previous response. Let's recalculate the difference in revenue between the sales of t-shirts and jerseys. Total revenue from t-shirts = $254 \text{ t-shirts} * \$86 \text{ per t-shirt} = \$21,844$. Total revenue from jerseys = $43 \text{ jerseys} * \$232 \text{ per jersey} = \$9,976$. To find the difference in revenue, we should subtract the total revenue from t-shirts from the total revenue from jerseys: $\$9,976 - \$21,844 = -\$11,868$. Upon reviewing the calculation, I realize that I made an error in the subtraction. The correct calculation should be: $\$21,844 - \$9,976 = \$11,868$. So the correct answer is \$11,868. This means that a jersey costs ## 11,868 ## more than a t-shirt. Thank you for bringing this to my attention.


[P2] IoE Prompt (Ours)

 : If you are very confident about your answer, maintain your answer. Otherwise, update your answer. Your final answer should be put between two ## symbols, like ## ANSWER ##, at the end of your response.

 : Apologies for the confusion. Let's recalculate. The difference in revenue between jerseys and t-shirts is: $\$232 - \$86 = \$146$. So, a jersey costs \$146 more than a t-shirt. Therefore, the final answer is ## 146 ##. **(Correct)**


[P3] Update Prompt (Huang et al., 2023)


 : Based on the problems you found, improve your answer. Please reiterate your answer. Your final answer should be put between two ## symbols, like ## ANSWER ##, at the end of your response.

 : Apologies for the confusion in my previous responses. Let's recalculate the difference in revenue between the sales of t-shirts and jerseys. Total revenue from t-shirts = $254 \text{ t-shirts} * \$86 \text{ per t-shirt} = \$21,844$. Total revenue from jerseys = $43 \text{ jerseys} * \$232 \text{ per jersey} = \$9,976$. To find the difference in revenue, we should subtract the total revenue from jerseys from the total revenue from t-shirts: $\$21,844 - \$9,976 = \$11,868$. So the answer is \$11,868. This means that a jersey costs ## 11,868 ## more than a t-shirt. Thank you for your patience, and I appreciate your understanding.

(Incorrect)

[P3] Decision Refinement (Ours)

 : You give two different answers in previous responses. Check the problem and your answers again, and give the best answer. Your final answer should be put between two ## symbols, like ## ANSWER ##, at the end of your response.

 : Total revenue from t-shirts = $\$86 * 254 = \$21,844$. Total revenue from jerseys = $\$232 * 43 = \$9,976$. The difference in revenue between jerseys and t-shirts is: $\$232 - \$86 = \$146$. So, a jersey costs \$146 more than a t-shirt. Therefore, the final answer is ## 146 ##.

(Correct)

Figure 6: One example on SVAMP benchmark (Patel et al., 2021) evaluated with gpt-3.5-turbo-1106 model. The [P1] standard prompt is the same for all methods. **Left:** Critical Prompt as the baseline. **Right:** our proposed IoE-based Prompt.

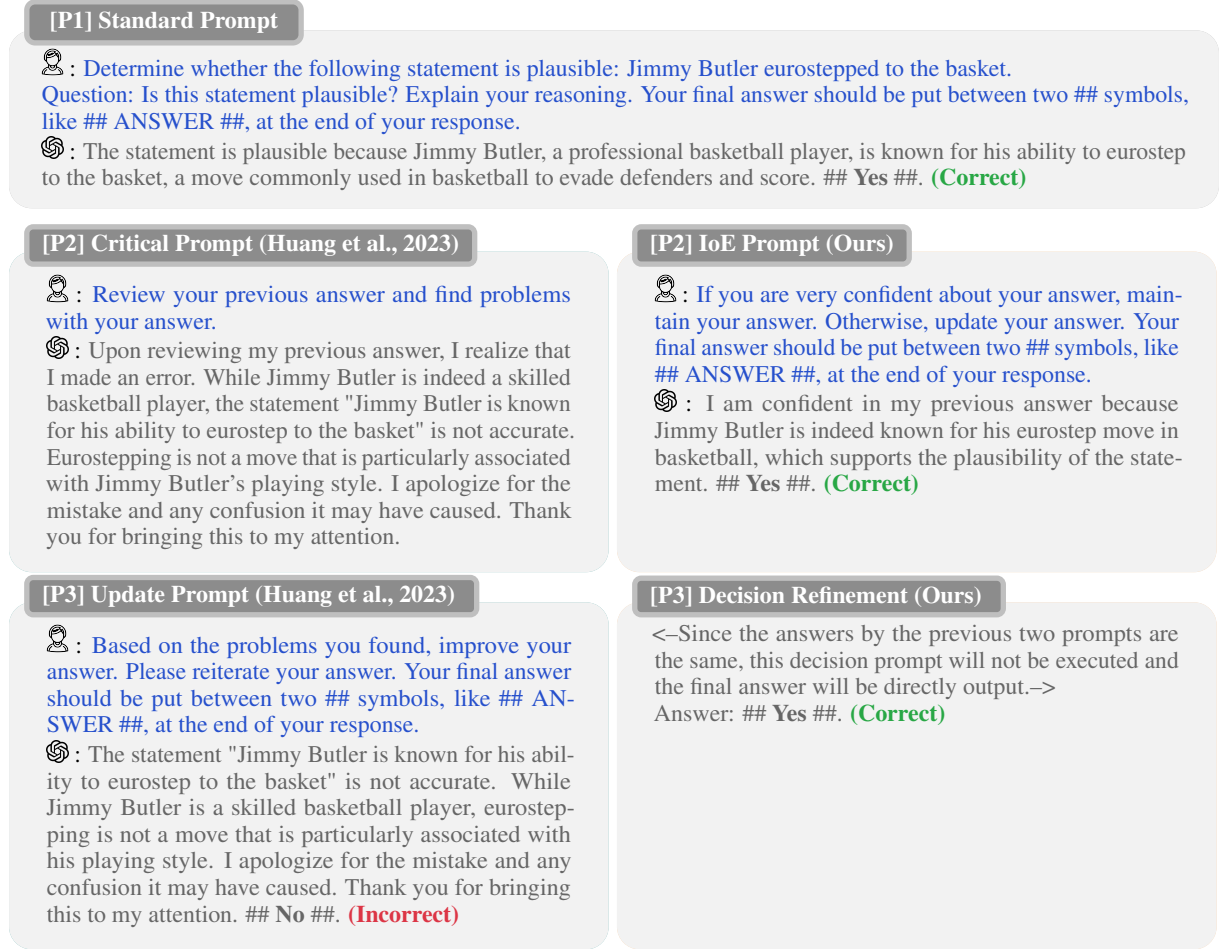


Figure 7: One example on Sports Understanding from BIG-Bench (Cai et al., 2023) evaluated with gpt-3.5-turbo-1106 model. The [P1] standard prompt is the same for all methods. **Left:** Critical Prompt as the baseline. **Right:** our proposed IoE-based Prompt.

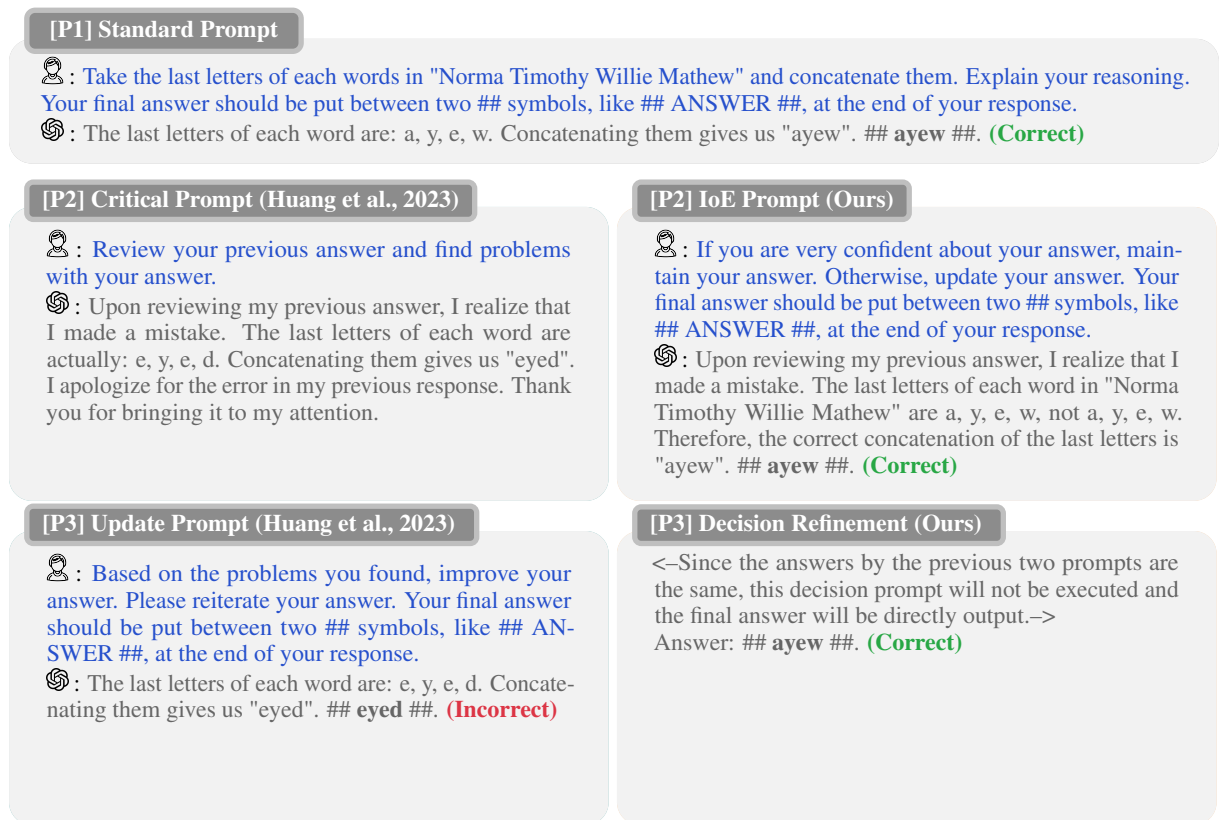


Figure 8: One example on Last Letter Concatenation (LLC) benchmark (Wei et al., 2022) evaluated with gpt-3.5-turbo-1106 model. The [P1] standard prompt is the same for all methods. **Left:** Critical Prompt as the baseline. **Right:** our proposed IoE-based Prompt.

[P1] Standard Prompt

Context: Peppes Pizza is a Norwegian pizza chain that serves American style and Italian style pizza. Peppes is the largest pizza chain in Scandinavia. The restaurant was founded by two Americans, Louis Jordan and his wife Anne from Hartford, Connecticut. The restaurant chain is part of Umoe Catering As which consists of restaurants such as Burger King, TGI Fridays, La Baguette and Cafe Opus. Peppes Pizza is one of the first restaurants that brought foreign food to Norway. 9 million pizzas are served by Peppes each year with deliveries in 11 cities in Norway. Their menu was first put online in March 1995. The servings have been described as enough for two people and that the pizza chain is a cut above the rest. Gino's East is a Chicago-based restaurant chain, notable for its deep-dish pizza (sometimes called Chicago-style pizza), and for its interior walls, which patrons have covered in graffiti and etchings. The restaurant features deep-dish pizza baked in cast-iron pans, as well as sandwiches, soups and salads. Big Mama's & Papa's Pizzeria is a pizza restaurant chain primarily located in Southern California. The chain is notable for its extremely large Giant Sicilian pizza, which is claimed to be the largest deliverable pizza in the world. Additionally, the chain gained notoriety when, during the 2014 Academy Awards, host Ellen Degeneres had Big Mama's pizzas delivered onstage. Papa John's Pizza is an American restaurant franchise company. It runs the third largest take-out and pizza delivery restaurant chain in the United States, with headquarters in Jeffersontown, Kentucky, a suburb of Louisville. Pizza 73 is a Canadian restaurant chain that offers a number of different styles of pizza, along with chicken wings. It has been operated by Pizza Pizza since 2007. Toronto-based Pizza Pizza had acquired the restaurant for a total of \$CAN70.2 million. There are 89 locations throughout Western Canada, which include the provinces of British Columbia, Alberta, and Saskatchewan. The restaurant's name originates from its original phone number: 473 . Founded by David Tougas and Guy Goodwin in 1985, Pizza 73 is headquartered in Edmonton, Alberta, Canada. Papa Gino's, Inc. is a restaurant chain based in Dedham, Massachusetts specializing in American-style pizza along with pasta, subs, salads, and a variety of appetizers. There are over 150 Papa Gino's locations in Connecticut, Maine, Massachusetts, New Hampshire, and Rhode Island. Pie Five Pizza Co. is a fast casual restaurant chain specializing in handcrafted personal pizza made in less than 5 minutes. The brand is owned by Rave Restaurant Group, which also owns Pizza Inn. As of December 2016, Pie Five has 98 restaurants in the following locations: Alabama, Arkansas, Florida, Georgia, Illinois, Indiana, Kansas, Maryland, Minnesota, Missouri, Nebraska, North Carolina, Oklahoma, Tennessee, Texas, Utah, Kentucky, Virginia and Washington, D.C. with more than 400 additional company-owned and franchise units anticipated. Gino's Pizza and Spaghetti is a restaurant chain with 40 locations, most of them within the U.S. state of West Virginia. The company was founded by Kenney Grant in 1961. Many locations are shared with Tudor's Biscuit World although the Gino's brand is exclusive to West Virginia. There is one located in Ohio, while there are stand alone Tudor's locations in eastern Kentucky, southern Ohio and southwest Virginia. Gino's serves pizza, spaghetti, sandwiches, and more. Company headquarters are located in Huntington, West Virginia and Nitro, West Virginia. Pizza Inn is an American restaurant chain and international food franchise, specializing in American-style pan pizza and side dishes. The company is based in the Dallas suburb of The Colony, Texas. Tudor's Biscuit World is a restaurant chain based in Huntington, West Virginia, most commonly found in West Virginia. Many West Virginia locations share a building with Gino's Pizza and Spaghetti, although the chain is more extensive than Gino's (which is exclusive to West Virginia), having locations in southern Ohio, eastern Kentucky, and southwestern Virginia. In 2016 a franchise was opened in Panama City, Florida. Tudor's serves biscuits, biscuit sandwiches, homestyle breakfasts and dinners, muffins, and several side dishes. The chain was originally based in Charleston, West Virginia and many of the biscuit sandwiches are named for sports teams of interest in that area, including teams at Marshall University, West Virginia University, and The University of Charleston.

Question: Which restaurant chain's headquarters is further north, Pizza Inn or Papa Gino's? Explain your reasoning. Your final answer should be put between two ## symbols, like ## ANSWER ##, at the end of your response.

Answer: Pizza Inn's headquarters is further north than Papa Gino's. Pizza Inn is based in The Colony, Texas, which is further north than Dedham, Massachusetts, where Papa Gino's is headquartered. The Colony, Texas is located at a latitude of approximately 33.0801 N, while Dedham, Massachusetts is located at a latitude of approximately 42.2418 N. Therefore, Pizza Inn's headquarters is further north. ## Pizza Inn ##. (Incorrect)

[P2] Critical Prompt (Huang et al., 2023)

Review: Review your previous answer and find problems with your answer.
Apologies: Apologies for the confusion. Upon reviewing the information, I realize that I made an error in my previous response. The correct answer is that Papa Gino's headquarters in Dedham, Massachusetts is further north than Pizza Inn's headquarters in The Colony, Texas. Therefore, the correct answer is Papa Gino's. Thank you for bringing this to my attention.

[P2] IoE Prompt (Ours)

Review: If you are very confident about your answer, maintain your answer. Otherwise, update your answer. Your final answer should be put between two ## symbols, like ## ANSWER ##, at the end of your response.
Apologies: Upon reviewing my previous answer, I realize that I made an error. Dedham, Massachusetts, where Papa Gino's is headquartered, is further north than The Colony, Texas, where Pizza Inn is based. Dedham, Massachusetts is located at a latitude of approximately 42.2418 N, while The Colony, Texas is located at a latitude of approximately 33.0801 N. Therefore, Papa Gino's headquarters is further north. ## Papa Gino's ##. (Correct)

[P3] Update Prompt (Huang et al., 2023)

Review: Based on the problems you found, improve your answer. Please reiterate your answer. Your final answer should be put between two ## symbols, like ## ANSWER ##, at the end of your response.
Apologies: Upon reviewing the information, I realize that I made an error in my previous response. The correct answer is that Papa Gino's headquarters in Dedham, Massachusetts is further north than Pizza Inn's headquarters in The Colony, Texas. Therefore, the correct answer is ## Papa Gino's ##. (Correct)

[P3] Decision Refinement (Ours)

Review: You give two different answers in previous responses. Check the problem and your answers again, and give the best answer. Your final answer should be put between two ## symbols, like ## ANSWER ##, at the end of your response.
Apologies: Upon re-evaluating the locations of the headquarters for Pizza Inn and Papa Gino's, I have found that The Colony, Texas, where Pizza Inn is based, is further south than Dedham, Massachusetts, where Papa Gino's is headquartered. The Colony, Texas is located at a latitude of approximately 33.0801 N, while Dedham, Massachusetts is located at a latitude of approximately 42.2418 N. Therefore, Papa Gino's headquarters is further north. ## Papa Gino's ##. (Correct)

Figure 9: One example on HotpotQA benchmark with context (Yang et al., 2018) evaluated by gpt-3.5-turbo-1106 model. The [P1] standard prompt is the same for all methods. **Left:** Critical Prompt as the baseline. **Right:** our proposed IoE-based Prompt.



Figure 10: One example on Domestic Robot benchmark (Cai et al., 2023) evaluated with gpt-4-vision-preview model. The [P1] standard prompt is the same for all methods. **Left:** Critical Prompt as the baseline. **Right:** our proposed IoE-based Prompt.

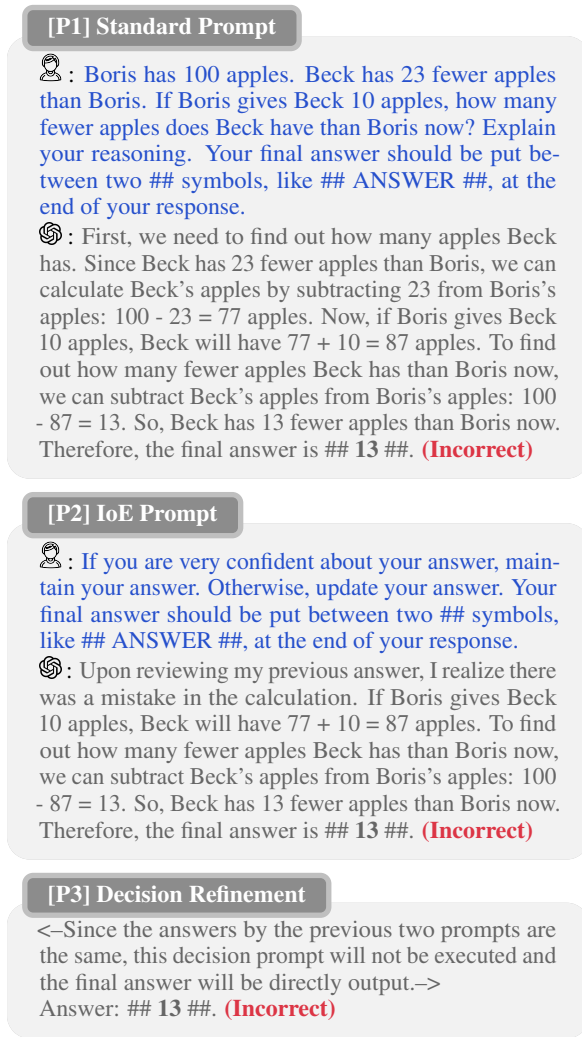


Figure 11: One example on GSM8K (Cobbe et al., 2021) with gpt-3.5-turbo-1106 model, where our IoE-based prompt failed to achieve self-correction.

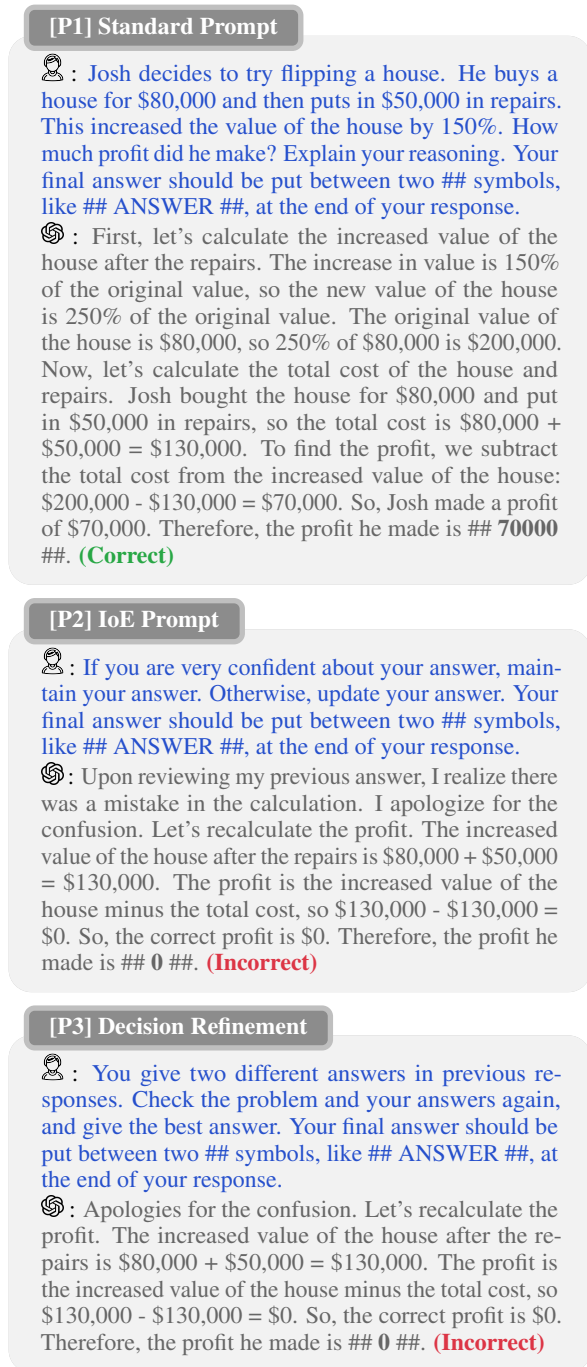


Figure 12: Another example on GSM8K (Cobbe et al., 2021) with gpt-3.5-turbo-1106 model, where our IoE-based prompt failed to achieve self-correction.