A Probabilistic Inference Scaling Theory for LLM Self-Correction

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have demonstrated the capability to refine their generated answers through self-correction, enabling continuous performance improvement over multiple rounds. However, the mechanisms underlying how and why accuracy evolves during this iterative process remain unexplored. To fill this gap, we propose a probabilistic theory to model the dynamics of accuracy change and explain the performance improvements observed in multi-round self-correction. Through mathematical derivation, we establish that the accuracy after the t^{th} round of self-correction is given by: $Acc_t = Upp - \alpha^t(Upp - Acc_0)$, where Acc_0 denotes the initial accuracy, Upp represents the upper bound of accuracy convergence, and α determines the rate of convergence. Based on our theory, these parameters can be calculated and the predicted accuracy curve then can be obtained through only a single round of self-correction. Extensive experiments across diverse models and datasets demonstrate that our theoretical predictions align closely with empirical accuracy curves, validating the effectiveness of the theory. Additionally, we derive and experimentally verify three corollaries, further substantiating the theory. Finally, we discuss failure scenarios, bottlenecks, and the potential of self-correction from the perspective of our theory. Our work provides a theoretical foundation for understanding LLM self-correction, thus paving the way for further explorations.

1 Introduction

With the depletion of pre-training corpora, the training scaling law (Kaplan et al., 2020) reaches the saturation point, and an alternative way to further improve performance is introducing more computational cost at test time, also known as inference scaling (Snell et al., 2025; Hoffmann et al., 2022). Brown et al. (2024b) repeatedly sample multiple answers and select the optimal one with best-of-*n* (Li et al., 2023) or majority voting (Wang et al., 2023) strategy, and the curve of how accuracy changes in this process as inference costs increase is also experimentally recorded (Wu et al., 2024a). Another approach to inference scaling is self-correction (Kamoi et al., 2024; Pan et al., 2024), where LLMs can refine their answers based on intrinsic (Madaan et al., 2024) or external (Jiang et al., 2023b) feedback. Xi et al. (2023); Liu et al. (2024b) have empirically observed that model performance continuously improves and eventually converges during multi-round self-correction, but the underlying reasons and mechanisms remain poorly understood. To narrow this gap, we propose a probabilistic theory to model how accuracy evolves and explain why performance improves in multi-round self-correction.

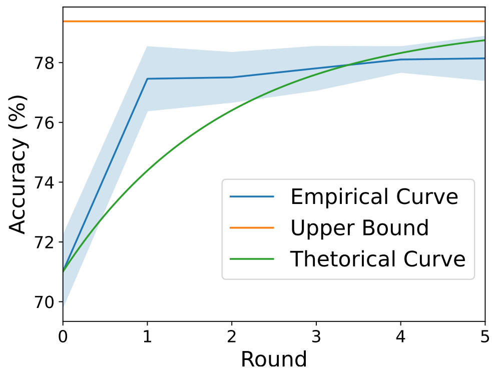

Figure 1: A demonstration of our theory on GSM8k with Llama3-8B-Instruct. The curves reflect how accuracy evolves in multi-round self-correction, and we depict the empirical curve, theoretical curve, and upper bound predicted by our theory in blue, green, and orange respectively. The theoretical curve fits the actual curve well, and both curves approach but do not exceed the upper bound.

In §3, we mathematically derive our theory from a probabilistic perspective. Yang et al. (2024b) decompose self-correction capabilities of LLMs into confidence capability and critique capability, introducing two metrics named Confidence Level (CL) and Critique Score (CS) to measure them, respectively. Based on their decomposition, we further discover a recursive relationship between the accuracy of successive rounds of self-correction: $Acc_t = (CL - CS)Acc_{t-1} + CS$, where Acc_t and Acc_{t-1} denote the accuracy after the t^{th} and $t-1^{th}$ round of self-correction, respectively. From this recursive relationship, we further find $Acc_t = Upp - \alpha^t(Upp - Acc_0)$, where $Upp = \frac{CS}{1-CL+CS}, \alpha = CL - CS$, and Acc_0 is the initial accuracy. This equation serves as the core part of our theory by describing how accuracy evolves in multi-round self-correction. To directly verify the theory, we compare the empirical accuracy curve with the theoretical curve given by our theory, and extensive experiments in §4 demonstrate that the theoretical curve fits the empirical curve well across various models and datasets (with an example illustrated in Figure 1).

Based on our theory and subsequent deduction, three corollaries can be derived (§5): (1). after infinite rounds of self-correction, the final accuracy converges to the upper bound Upp, which is solely determined by CL and CS and is independent of the initial accuracy Acc_0 (§5.1); (2). the speed of convergence depends $\alpha = CL - CS$, and accuracy converge faster when α is lower (§5.2); (3). in particular, under the ideal condition with an oracle verifier ($CL = 1$), the accuracy follows $Acc_t = 1 - (1 - CS)^t(1 - Acc_0)$, ultimately converging to 100% (§5.3). All these corollaries are then experimentally validated: (1). when manipulating initial accuracy Acc_0 to different values, we find final accuracy always converges to the same value, validating *corollary 1*; (2). by comparing the convergence rate of LLama3 and Qwen2.5, we find model with lower α converges faster, validating *corollary 2*; (3). after introducing an oracle verifier to make sure $CL = 1$, we find the theoretical curve still fits the empirical curve, validating *corollary 3*. These experiments directly verify three corollaries and provide further support for our theory.

Further discussions on self-correction based on our theory (§6). Huang et al. (2024); Jiang et al. (2024); Valmeekam et al. (2023); Zhang et al. (2024b) observe the failure of self-correction where accuracy can even decrease after self-correction. From the perspective of our theory, this failure can be explained as a special case when the converged upper bound Upp is lower than the initial accuracy Acc_0. We also discuss how far self-correction can go: the performance upper bound of self-correction has been given by our theory, which is empirically not that high, and this bottleneck can hardly be solved. In contrast, the great potential of external self-correction is showcased by our theory, and performance can be improved by a large margin when $CL = 1$. These discussions provide a theoretical perspective for a better understanding of self-correction, and bring more insights to further investigation.

Our contributions can be summarized as follows:

1. We propose a probabilistic theory to model how accuracy evolves in multi-round self-correction, along with 3 corollaries.

2. To validate our theory, we conduct extensive experiments and find that our theoretical curve fits empirical curve well. We also provide experimental validation of 3 corollaries as further support of the theory.

3. We discuss failure scenarios, bottlenecks, and the potential of self-correction based on the theory, bringing insights and a better understanding to further explorations.

2 Related Work

Inference Scaling Model performance can be improved by introducing more computational cost at test time, and this inference scaling (Snell et al., 2025; Hoffmann et al., 2022) can be achieved via various ways: Wei et al. (2022) find directly outputting the final answer limits model performance and propose a Chain-Of-Thought (COT) prompting strategy; Wu et al. (2024a) repeatedly sample multiple answers and choose the best one with best-of-n (Li et al., 2023; Brown et al., 2024b) or majority voting (Wang et al., 2023), and Zhang et al. (2023); Liu et al. (2024c) further substitute repeated sampling with Monte Carlo Tree Search (MCTS) for better efficiency; Liu et al. (2024b); Xi et al. (2023) utilize multi-round self-correction to obtain refined answers and Zhang et al. (2024a) refine answers with a search tree. while previous works have empirically observed that model performance improves with higher inference costs, deeper exploration into why these performance curves occur is still lacking. Our work partially fills this gap by providing a theoretical explanation and modeling how accuracy changes during multi-round self-correction.

LLM Self-Correction LLMs can generate feedback on their answer, revise this answer based on feedback, and output a refined answer that is considered to be better than the initial answer. This self-correction capability (Kamoi et al., 2024; Pan et al., 2024; Yang et al., 2024b) can be further improved through approaches: introducing external tools to provide more effective feedback (Jiang et al., 2023b), better prompting strategies (Li et al., 2024; Wu et al., 2024b), fine-tuning on a augmented critique dataset (Welleck et al., 2023), reinforcement learning with online feedback (Kumar et al., 2024) and iterative self-correction (Qu et al., 2024; Madaan et al., 2024). Different from previous works, we propose a theory to explain and model the accuracy curve for self-correction, providing theoretical support and beneficial insights for further investigations on LLM self-correction.

3 Theory

In this section, we introduce an inference scaling theory to model and explain how accuracy changes in multi-round self-correction. First, we formally define the multi-round self-correction process and provide mathematical notations in §3.1. Then we discuss a simple scenario where the test set consists of only one datum (§3.2), and further extend our analysis to the general case where the test set contains n questions (§3.3). According to our theory, the accuracy after t rounds of self-correction is given by $Acc_t = Upp - \alpha^t(Upp - Acc_0)$ and finally converges to Upp.

3.1 Problem Formulation and Notations

Initially, we have a set comprising of n questions denoted as $Q = \{q_1, q_2, ..., q_n\}$, and we utilize multi-round self-correction to boost model performance. For any given question q_i, we first directly query the model and generate an answer $a_{i,0}$. Then we utilize an appropriate prompt to encourage the model to self-correct $a_{i,0}$ and get a refined answer $a_{i,1}$ and subsequently self-correct $a_{i,1}$ to get $a_{i,2}$, and so on. This process is conducted iteratively, yielding a sequence of answers $a_{i,0}, a_{i,1}, ..., a_{i,k}$ after k rounds of self-correction. It is worth noting that during the t^{th} self-correction, only the $(t-1)^{th}$ answer $a_{i,t-1}$ is provided as input to the model, rather than the entire sequence $a_{i,0}, ..., a_{i,t-1}$, which ensures that the computational cost per self-correction round remains approximately constant, rather than scaling linearly with t. For the answer $a_{i,t}$ from the t^{th} self-correction, we denote the probability that the model generates a correct answer through a single temperature-based sampling as $P(a_{i,t})$. The initial accuracy is defined as $Acc_0 = \frac{\sum_{i=1}^{n} P(a_{i,0})}{n}$, and the accuracy after the t^{th} self-correction round is defined as $Acc_t = \frac{\sum_{i=1}^{n} P(a_{i,t})}{n}$. For clarity, all notations and their corresponding definitions are summarized in Appendix A.

3.2 Question-Level Theory

Rather than exploring the change of accuracy on the whole dataset, we first discuss a simpler problem: how the probability of generating a correct answer for a single question q_i evolves as the number of self-correction rounds increases.

For answer $a_{i,t}$ generated in the t^{th} self-correction, the answer before self-correction $a_{i,t-1}$ may be either correct or wrong, so by the Law of Total Probability we have:

$$\begin{aligned} P(a_{i,t}) =& P(a_{i,t-1})P(a_{i,t}|a_{i,t-1}) \\ &+ [1 - P(a_{i,t-1})]P(a_{i,t}|\neg a_{i,t-1}), \end{aligned} \tag{1}$$

where $P(a_{i,t}|a_{i,t-1})$ and $P(a_{i,t}|\neg a_{i,t-1})$ denote the conditional probabilities that $a_{i,t}$ is correct given that $a_{i,t-1}$ is correct or incorrect, respectively. During the t^{th} self-correction round, only $a_{i,t-1}$ is fed into the model, rather than the whole sequence $a_{i,0}, ..., a_{i,t-1}$. Consequently, these two probabilities depend solely on the question index i and are independent of the current self-correction round t. We denote these two probabilities as P_i^{con} and P_i^{cri} respectively, which represent the probability of generating a correct answer after self-correction, given the answer before self-correction is correct/wrong. P_i^{con} reflects model confidence in the correct answer and P_i^{cri} reflects the critique capability. For any $t \in N+$, we have $P(a_{i,t}|a_{i,t-1}) = P_i^{con}$ and $P(a_{i,t}|\neg a_{i,t-1}) = P_i^{cri}$, which we substitute into Equation 1 to obtain:

$$\begin{aligned} P(a_{i,t}) &= P(a_{i,t-1})P_i^{con} + [1 - P(a_{i,t-1})]P_i^{cri} \\ &= (P_i^{con} - P_i^{cri})P(a_{i,t-1}) + P_i^{cri} \end{aligned} \tag{2}$$

By subtracting $\frac{P_i^{cri}}{1-P_i^{con}+P_i^{cri}}$ from both sides of the Equation 2, we have:

$$\begin{aligned} &P(a_{i,t}) - \frac{P_i^{cri}}{1 - P_i^{con} + P_i^{cri}} \\ &= (P_i^{con} - P_i^{cri})(P(a_{i,t-1}) - \frac{P_i^{cri}}{1 - P_i^{con} + P_i^{cri}}) \end{aligned} \tag{3}$$

It is evident that $P(a_{i,t}) - P_i^{upp}$ forms a geometric progression with a common ratio of α_i, where $P_i^{upp} = \frac{P_i^{cri}}{1-P_i^{con}+P_i^{cri}}$ and $\alpha_i = P_i^{con} - P_i^{cri}$. By applying the general term formula of a geometric sequence, we obtain: $P(a_{i,t}) - P_i^{upp} = \alpha_i^t(P(a_{i,0}) - P_i^{upp})$. After k rounds of self-correction, the probability of the model correctly answering question q_i is expressed as:

$$P(a_{i,t}) = P_i^{upp} - \alpha_i^t(P_i^{upp} - P(a_{i,0})) \quad (4)$$

This equation characterizes the trajectory of the probability of correctly answering a single question q_i as the number of self-correction iterations t increases.

3.3 Dataset-Level Theory

In §3.2, we have demonstrated that the trajectory of correct probability for a single question (as shown in Equation 4) depends on three variables: $P(a_{i,0})$, P_i^{con} and P_i^{cri}. Further, we attempt to extend this finding to the dataset level, where we measure accuracy across the entire dataset. Specifically, we also use the initial accuracy, model confidence, and critique capability as three key indicators to characterize how accuracy evolves with the increase of self-correction rounds. Yang et al. (2024b) decompose the self-correction capability of a model into two components: confidence (the ability to maintain confidence in the correct answer) and critique (the ability to correct wrong answers), and propose two probabilistic metrics to measure these capabilities quantitatively, which we adopt directly:

- The **C**onfidence **L**evel (*CL*) measures the model confidence, defined as the probability that the model retains the correct answer after self-correction:

$$\begin{aligned} CL_t &= E[P(a_{\_,t+1}|a_{\_,t})] \\ &= \frac{\sum_{i=1}^{n} P(a_{i,t})P(a_{i,t+1}|a_{i,t})}{\sum_{i=1}^{n} P(a_{i,t})}, \end{aligned} \quad (5)$$

- The **C**ritique **S**core (*CS*) measures the capability to critique and reflect, defined as the probability that the model corrects a wrong answer to a right one after self-correction:

$$\begin{aligned} CS_t &= E[P(a_{\_,t+1}|\neg a_{\_,t})] \\ &= \frac{\sum_{i=1}^{n}[1 - P(a_{i,t})]P(a_{i,t+1}|\neg a_{i,t})}{\sum_{i=1}^{n}[1 - P(a_{i,t})]}, \end{aligned} \quad (6)$$

In the t^{th} round of self-correction, the relationship between accuracy before and after self-correction and the two metrics above is given by (with derivation details shown in Appendix B):

$$Acc_t = Acc_{t-1}CL_{t-1} + (1 - Acc_{t-1})CS_{t-1} \quad (7)$$

Assuming that *CL* and *CS* reflect the inherent confidence and critique capabilities of LLMs, so we treat these metrics as constants independent of the round number t, and this yields:

$$Acc_t = Acc_{t-1} * CL + (1 - Acc_{t-1}) * CS \quad (8)$$

Noticing that Equation 8 and Equation 2 are essentially the same recurrence relation, we can similarly derive that:

$$Acc_t = Upp - \alpha^t(Upp - Acc_0) \quad (9)$$

where $Upp = \frac{CS}{1-CL+CS}$, $\alpha = CL - CS$. Empirically we have $0 < \alpha < 1$, and as $t \to +\infty$, $Acc_t \to Upp$. This equation describes how accuracy changes in multi-round self-correction and provides a theoretical performance upper bound, serving as the core part of our theory.

4 Experiments

4.1 Experimental Setup

Models Similar to Yang et al. (2024b), experiments are conducted on both open-source and closed-source models. For the closed-source models, we assess Qwen-Max (Bai et al., 2023), GPT-3.5 Turbo, and GPT-4 Turbo (Achiam et al., 2023) by API calls. For the open-source models, we evaluate Llama3-8B (AI@Meta, 2024), Qwen2.5-7B (Yang et al., 2024a), DeepSeek-LLM-7B (DeepSeek-AI, 2024), Mistral-7B-v3 (Jiang et al., 2023a), and GLM4-9B (GLM et al., 2024), and parameters of these models are publicly available on HuggingFace [1]. For each open-source model (< 10B), we run the experiments on a single Nvidia A100 80G GPU, and utilize vllm [2] to accelerate generation.

Dataset We conduct experiments on both classification and generation tasks, including domains in mathematics, coding, instruction following, common-sense reasoning, and knowledge. To be specific, the dataset we utilized include GSM8k (Cobbe et al., 2021), Humaneval (Chen et al., 2021), IFEval (Zhou et al., 2023), MMLU (Hendrycks et al., 2021), BoolQ (Clark et al., 2019), CommonsenseQA (Talmor et al., 2019), PiQA (Bisk et al., 2019), and HotpotQA (Yang et al., 2018).

[1] https://huggingface.co/
[2] https://github.com/vllm-project/vllm

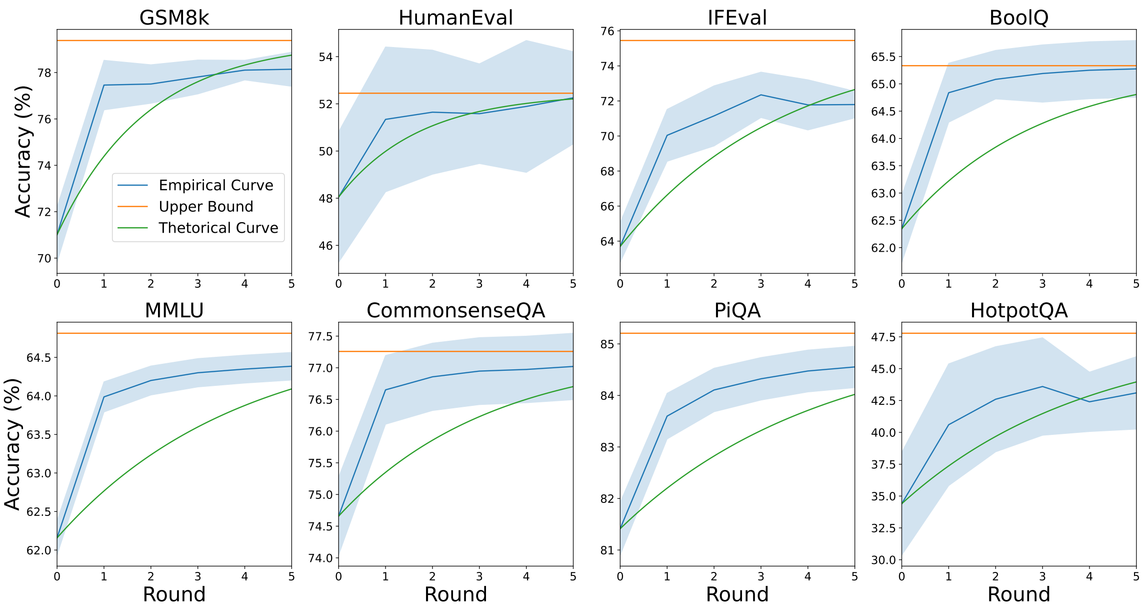

Figure 2: Experimental verification of our theory on Llama3-8B-Instruct. The empirical curve in mult-round self-correction, theoretical curve, and upper bound predicted by our theory are depicted in blue, green, and orange respectively. The theoretical curve fits the empirical curve well and accuracy approaches but never surpasses the upper bound.

4.2 Results

To validate our theory, we compare the empirical accuracy change curve with the theoretical curve predicted by our theory by visualizing them in the same figure and checking the alignment. The empirical curve is acquired from a 5-round self-correction process across multiple models and datasets, during which we track accuracy and variance changes and plot the results. To enhance the numerical stability of experimental results, we sample five responses independently for each question and use the average accuracy for analysis. For the theoretical curve, we compute three key parameters: initial accuracy (Acc_0), confidence level (CL), and critique score (CS). Using these values and Equation 9, we generate the theoretical curve and its upper bound, which are then plotted alongside the empirical curve. Since the calculation of CL and CS relies on probability, we utilize the probability estimation methods provided by Yang et al. (2024b), and more details are shown in the Appendix C.

The experimental results of Llama3-8B-Instruct are presented in Figure 2, with more results of other models provided in Appendix D. For better visualization, the empirical curve, theoretical curve, and upper bound are depicted in blue, green, and orange respectively. The results demonstrate that the theoretical curve closely aligns with the empirical curve across various datasets, suggesting that the proposed theory effectively models and explains the variations in accuracy during self-correction. Furthermore, the upper bound derived from the theory holds practical relevance, as the accuracy curve consistently approaches but does not exceed it, further validating the effectiveness of our theory.

5 Corollaries

Based on the theory in §3, three corollaries can be further derived: (1). the final converged accuracy is independent of the initial accuracy (§5.1); (2). the convergence rate of accuracy increases as α decreases (§5.2); (3). a special case of the theory where $CL = 1$ (§5.3). We provide both mathematical derivation and experimental verification of these corollaries, which can also serve as further validation of our theory.

5.1 Corollary 1

Corollary 1: The final converged accuracy is exclusively determined by the confidence and critique capabilities (i.e., CL and CS), and remains independent of the initial accuracy Acc_0.

Derivation of Corollary 1 Intuitively, when the model is provided with an initial correct or incor-

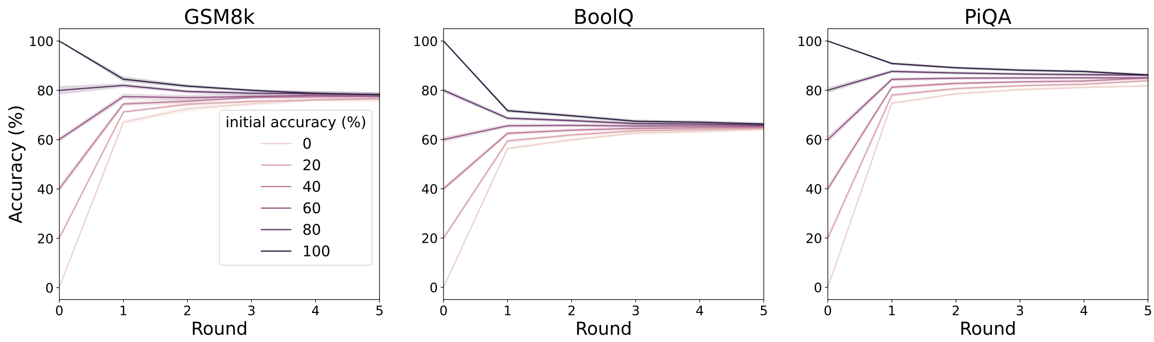

Figure 3: The accuracy convergency results with different initial accuracy Acc_0 for Llama3-8B-Instruct: the accuracy consistently converges to the same final value regardless of the initial accuracy.

rect answer to self-correct, it has a higher probability of reaching the correct answer when the initial answer is correct. This implies that $CL > CS$, which is also empirically demonstrated by Yang et al. (2024b). Given that $CL, CS \in (0, 1)$, it follows that $0 < \alpha = CL - CS < 1$. Based on Equation 9, as $t \to +\infty$, $\alpha^t \to 0$, and thus $Acc_t \to Upp$. This indicates after sufficient rounds of self-correction the final accuracy converges to $Upp = \frac{CS}{1-CL+CS}$. Notably, Upp is entirely determined by CL and CS and is independent of the initial accuracy Acc_0.

Verification of Corollary 1 To validate this corollary and investigate whether the initial accuracy influences the final converged accuracy after infinite rounds of self-correction, we systematically manipulate the initial accuracy to various target values and observe its impact on the final accuracy. Unlike the experiments described in §4, where the initial answer $a_{i,0}$ is generated by feeding the question q_i to the model, we directly control the initial accuracy to achieve a desired value Acc_{target} by carefully setting the initial answers. For a K-class classification task, we assign the initial probability of the correct class to Acc_{target} and distribute the remaining probability uniformly among the incorrect classes, ensuring that each incorrect class has a probability of $\frac{1-Acc_{target}}{K-1}$. This guarantees that the initial accuracy $Acc_0 = Acc_{target}$. For generation tasks with n items in the dataset, we first sample multiple answers for each question q_i to obtain both correct and incorrect answers. We then randomly select $\lfloor Acc_{target} \times n \rfloor$ items to use correct answers as initial answers, while assigning incorrect answers to the remaining items, which ensures that the initial accuracy $Acc_0 \approx Acc_{target}$. In cases where no correct answer is sampled for a question, we use the standard correct answer from the dataset. Conversely, if no incorrect answers are sampled, we truncate a correct answer to create an incorrect one. As the results of Llama3-8B-Instruct illustrated in Figure 3, the final accuracy consistently converges to the same value regardless of whether the initial accuracy is set to 0%, 20%, 40%, 60%, 80%, or 100%, which experimentally verifies *Corollary 1*.

5.2 Corollary 2

> *Corollary 2*: The convergence rate of accuracy is determined by the parameter $\alpha = CL - CS$. Specifically, a model with a lower value of α exhibits faster convergence in accuracy.

Derivation of Corollary 2 As discussed in §5.1, as $t \to +\infty$, $\alpha^t \to 0$, and consequently $Acc_t \to Upp$. The convergence rate of α^t is decided by the value α, and the closer the value of α is to 0, the faster α^t will converge to 0. To better illustrate this difference in convergence speed, consider the following example: when $\alpha = 0.9$, $\alpha^{10} \approx 0.35$; whereas when $\alpha = 0.2$, $\alpha^{10} \approx 10^{-7}$.

Verification of Corollary 2 To validate this corollary, we compare the convergence rates of models with distinct α values. Given the difficulty in discerning convergence speed differences between models with similar α values, we select two models with significantly differing α values for comparison. As experimentally demonstrated in §4, the Llama3-8B-Instruct model exhibits a lower α value, while the Qwen2.5-7B-Chat model has a higher α value, so we choose these two models for comparison and analysis. The experimental results are shown in Figure 4, with more results provided in Appendix D. Llama3-8B-Instruct (lower α) converges noticeably faster and its accuracy gets closer to the upper bound after 5 rounds of self-correction

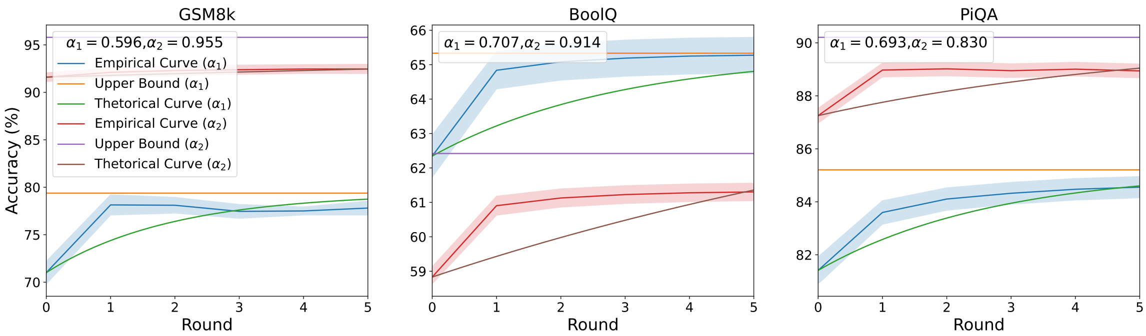

Figure 4: Convergence speed comparison of Llama3-8B-Instruct (α_1) and Qwen2.5-7B-Chat (α_2): we have $\alpha_1 < \alpha_2$, and Llama3-8B-Instruct (α_1) converges noticeably faster and its accuracy gets closer to the upper bound after 5 rounds of self-correction than Qwen2.5-7B-Chat (α_2).

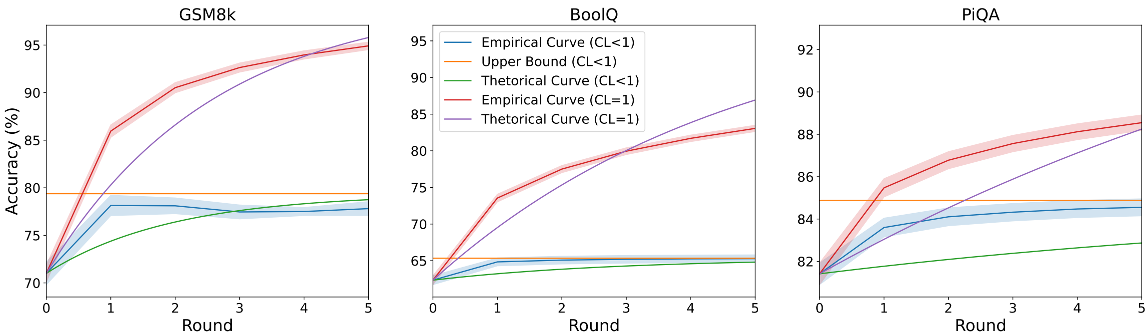

Figure 5: Curves for the special case ($CL = 1$) on Llama3-8B-Instruct. The theoretical curve fits the actual curve well when $CL = 1$, and exceeds the standard intrinsic self-correction ($CL < 1$) by a large margin.

than Qwen2.5-7B-Chat (higher α), which experimentally verifies *Corollary 2*.

5.3 Corollary 3

> *Corollary 3*: A special case where CL=1, we have $Acc_t = 1 - (1 - CS)^t(1 - Acc_0)$, and $Acc_t \to 1$ as $t \to +\infty$.

Derivation of Corollary 3 For intrinsic self-correction, LLMs need to independently evaluate the correctness of their generated answers (Zhang et al., 2024d), and errors in this process are almost inevitable (Stechly et al., 2023; Tyen et al., 2024). In cases where LLMs incorrectly identify a correct initial answer as erroneous and subsequently generate an incorrect answer after self-correction (✓ → ✗), we have $CL < 1$ instead of $CL = 1$. In contrast, external self-correction helps LLMs determine the correctness of their answers through external feedback, leading to a higher CL. For instance, Zhang et al. (2023); Kim et al. (2023) employ an oracle verifier to evaluate answer correctness, while Brown et al. (2024a) investigate inference scaling laws under the best-of-n metric, which can be considered as a special case in our theory when $CL = 1$. Specifically, when $CL = 1$, we have $Upp = \frac{CS}{1-CL+CS} = 1, \alpha = 1 - CS$, yielding:

$$Acc_t = 1 - (1 - CS)^t(1 - Acc_0) \tag{10}$$

As $t \to +\infty$, $\alpha^t \to 0$, and thus $Acc_t \to 1$, which aligns with the idea proposed in Brown et al. (2024a) that with sufficient times of sampling, the correct answer will always be encountered.

Verification of Corollary 3 To validate this corollary, we compare whether the accuracy change curve derived from our theory for the ideal scenario ($CL = 1$) aligns with the actual experiment curve. To simulate this special case ($CL = 1$) and equip the model with an oracle verifier, once a correct answer is generated in generation tasks, we halt subsequent rounds of self-correction and directly treat the following answers as correct. For classification tasks, we set the conditional probability of selecting the correct/incorrect answer after self-correction given the answer before self-correction is correct to 1/0 (i.e. setting $P(a_{i,t+1}|a_{i,t}) = 1, P(a_{i,t+1}|\neg a_{i,t}) = 0$). As the experimental results illustrated in Figure 5 and Appendix D, we show the experimental curve and theoretical curve for the special case ($CL = 1$), along with the curves for standard intrinsic self-correction ($CL < 1$) for comparison. The results demonstrate that the theoretical curve can still align well with the empirical curve in this special case ($CL = 1$), which experimentally verifies *Corollary 3*. Besides, we also find the accuracy of $CL = 1$ is improved by a large margin compared to that of $CL < 1$ and can exceed the upper bound of $CL < 1$, which shows a promising direction for further optimization of self-correction.

6 Discussion

The Failure of Self-Correction Though Madaan et al. (2024); Liu et al. (2024a) have found LLMs can achieve better performance after self-correction, there is still a debate on the effectiveness of self-correction and Huang et al. (2024); Jiang et al. (2024); Valmeekam et al. (2023) observe accuracy can even decrease after self-correction with poor prompts. For instance, Xie et al. (2024); Zhang et al. (2024b) find adding "Are you sure?" to the prompt will significantly reduce model confidence, causing it to change correct answers to incorrect ones after self-correction. Our theory can provide a new perspective to understand how self-correction fails: poor prompts can disrupt the balance between the confidence and critique capabilities of LLMs (CL and CS), thereby reducing the upper bound (Upp) to which the accuracy converges, ultimately resulting in $Upp < Acc_0$, and in this scenario accuracy will decrease after self-correction. Figure 6 shows a failure case of Llama3-8B-Instruct on GSM8k under the poor prompt of "Are you sure?", where accuracy converges to the bound in a descending fashion. For a given model and test set, different prompts correspond to different Upp values, suggesting that we should choose better prompts to avoid the failure of self-correction. A simple approach inspired by our theory could be testing various prompts and selecting the one with the highest Upp, and we leave further explorations in avoiding this failure to future work.

How Far Can LLM Self-Correction Go? Although previous works (Li et al., 2024; Zhang et al., 2024c; Wu et al., 2024b) have utilized and optimized self-correction for better performance, the extent of performance improvements achievable through self-correction under different settings and methods is still not thoroughly explored, and our theory partially fills this gap by providing a theoretical upper bound of accuracy. Our theory almost announces the death of intrinsic self-correction (Xi et al., 2023; Madaan et al., 2024), as it demonstrates that intrinsic self-correction cannot surpass the upper bound (Upp), which is empirically shown to be not that high in §4. A more promising direction lies in external self-correction (Jiang et al., 2023b; Chen et al., 2024), as we have discussed in §5.3 the great performance improvement brought by an oracle verifier (i.e. $CL = 1$), and external feedback can be viewed as an approximation of oracle verifier. Similarly, Kamoi et al. (2024) also discuss this problem and point out future directions for self-correction, and our work provides theoretical support to these discussions.

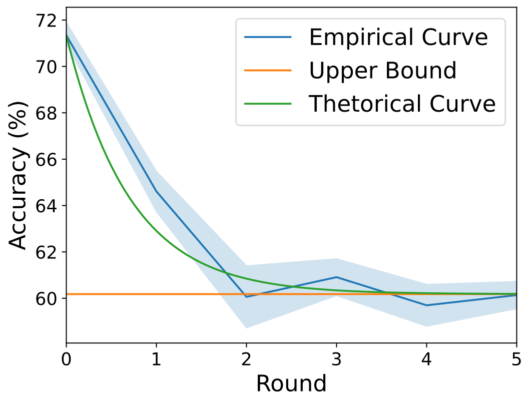

Figure 6: The failure of self-correction of Llama3-8B-Instruct on GSM8k under prompt of "Are you sure?". The accuracy decreases after self-correction and converges to the bound in a descending fashion.

7 Conclusion

We propose a probabilistic theory to model and explain how accuracy evolves in multi-round self-correction. Based on our theory and further deduction, we acquire 3 corollaries about convergence upper bound, the rate of convergence, and a special scenario. Extensive experiments validate the theory by showing the alignment between our theoretical curves and empirical curves, and empirical verification of 3 corollaries also futher supports the theory. Finally, from the perspective of our theory, we explain why self-correction can fail sometimes and discuss the bottle and potential direction of self-correction. Our theory provides theoretical support and a better understanding of LLM self-correction, thus paving the way for further explorations.

Limitations

The calculation of our theoretical curve relies on probability estimation, which necessitates repeated sampling for the same question, and the simulation of multi-round self-correction (i.e. actual curve) also generates multiple answers for the same question. These can be more computationally expensive than traditional experiments where only one answer is generated for a question. We only experimentally validate our theory on 8 models and 8 datasets, leaving more verification experiments on more models and datasets for future work.

Though our theoretical curve can fit the actual curve to some extent, what happens in self-correction and how accuracy changes can be much more complex than our theory. Our theory can only describe how accuracy changes in multi-round self-correction, but how performance improves in other inference scaling settings (e.g. long COT, MCTS) is still unknown, and we leave it to future work.

Ethical Considerations

The data we utilized are open for research, and evaluated LLMs are all publicly available by either parameters or API calls. Therefore, we do not anticipate any ethical concerns in our research.

References

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

AI@Meta. 2024. Llama 3 model card.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. *Preprint*, arXiv:1911.11641.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. 2024a. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. 2024b. Large language monkeys: Scaling inference compute with repeated sampling. *Preprint*, arXiv:2407.21787.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2024. Teaching large language models to self-debug. In *The Twelfth International Conference on Learning Representations*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

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

DeepSeek-AI. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023a. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Dongwei Jiang, Jingyu Zhang, Orion Weller, Nathaniel Weir, Benjamin Van Durme, and Daniel Khashabi. 2024. Self-[in] correct: Llms struggle with refining self-generated responses. *CoRR*.

Shuyang Jiang, Yuhao Wang, and Yu Wang. 2023b. Selfevolve: A code evolution framework via large language models. *arXiv preprint arXiv:2306.02907*.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can LLMs actually correct their own mistakes? a critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics*, 12:1417–1440.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Geunwoo Kim, Pierre Baldi, and Stephen Marcus McAleer. 2023. Language models can solve computer tasks. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. 2024. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*.

Loka Li, Zhenhao Chen, Guangyi Chen, Yixuan Zhang, Yusheng Su, Eric Xing, and Kun Zhang. 2024. Confidence matters: Revisiting intrinsic self-correction capabilities of large language models. *arXiv preprint arXiv:2402.12563*.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333, Toronto, Canada. Association for Computational Linguistics.

Dancheng Liu, Amir Nassereldine, Ziming Yang, Chenhui Xu, Yuting Hu, Jiajie Li, Utkarsh Kumar, Changjae Lee, and Jinjun Xiong. 2024a. Large language models have intrinsic self-correction ability. *CoRR*.

Guangliang Liu, Haitao Mao, Bochuan Cao, Zhiyu Xue, Xitong Zhang, Rongrong Wang, Jiliang Tang, and Kristen Johnson. 2024b. On the intrinsic self-correction capability of llms: Uncertainty and latent concept. *Preprint*, arXiv:2406.02378.

Jiacheng Liu, Andrew Cohen, Ramakanth Pasunuru, Yejin Choi, Hannaneh Hajishirzi, and Asli Celikyilmaz. 2024c. Don't throw away your value model! generating more preferable text with value-guided monte-carlo tree search decoding. In *First Conference on Language Modeling*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2024. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics*, 12:484–506.

Yuxiao Qu, Tianjun Zhang, Naman Garg, and Aviral Kumar. 2024. Recursive introspection: Teaching language model agents how to self-improve. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling test-time compute optimally can be more effective than scaling LLM parameters. In *The Thirteenth International Conference on Learning Representations*.

Kaya Stechly, Matthew Marquez, and Subbarao Kambhampati. 2023. Gpt-4 doesn't know it's wrong: An analysis of iterative prompting for reasoning problems. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Gladys Tyen, Hassan Mansoor, Victor Carbune, Peter Chen, and Tony Mak. 2024. LLMs cannot find reasoning errors, but can correct them given the error location. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13894–13908, Bangkok, Thailand. Association for Computational Linguistics.

Karthik Valmeekam, Matthew Marquez, and Subbarao Kambhampati. 2023. Can large language models really improve by self-critiquing their own plans?

In *NeurIPS 2023 Foundation Models for Decision Making Workshop*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2023. Generating sequences by learning to self-correct. In *The Eleventh International Conference on Learning Representations*.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2024a. Scaling inference computation: Compute-optimal inference for problem-solving with language models. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Zhenyu Wu, Qingkai Zeng, Zhihan Zhang, Zhaoxuan Tan, Chao Shen, and Meng Jiang. 2024b. Large language models can self-correct with key condition verification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12846–12867, Miami, Florida, USA. Association for Computational Linguistics.

Zhiheng Xi, Senjie Jin, Yuhao Zhou, Rui Zheng, Songyang Gao, Jia Liu, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. Self-Polish: Enhance reasoning in large language models via problem refinement. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11383–11406, Singapore. Association for Computational Linguistics.

Qiming Xie, Zengzhi Wang, Yi Feng, and Rui Xia. 2024. Ask again, then fail: Large language models' vacillations in judgment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10709–10745, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Zhe Yang, Yichang Zhang, Yudong Wang, Ziyao Xu, Junyang Lin, and Zhifang Sui. 2024b. Confidence v.s. critique: A decomposition of self-correction capability for llms. *Preprint*, arXiv:2412.19513.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Di Zhang, Xiaoshui Huang, Dongzhan Zhou, Yuqiang Li, and Wanli Ouyang. 2024a. Accessing gpt-4 level mathematical olympiad solutions via monte carlo tree self-refine with llama-3 8b. *Preprint*, arXiv:2406.07394.

Qingjie Zhang, Han Qiu, Di Wang, Haoting Qian, Yiming Li, Tianwei Zhang, and Minlie Huang. 2024b. Understanding the dark side of llms' intrinsic self-correction. *Preprint*, arXiv:2412.14959.

Shun Zhang, Zhenfang Chen, Yikang Shen, Mingyu Ding, Joshua B. Tenenbaum, and Chuang Gan. 2023. Planning with large language models for code generation. In *The Eleventh International Conference on Learning Representations*.

Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yueting Zhuang, and Weiming Lu. 2024c. Self-contrast: Better reflection through inconsistent solving perspectives. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3602–3622, Bangkok, Thailand. Association for Computational Linguistics.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Jaekyeom Kim, Moontae Lee, Honglak Lee, and Lu Wang. 2024d. Small language models need strong verifiers to self-correct reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15637–15653, Bangkok, Thailand. Association for Computational Linguistics.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

Appendix

A Mathematical Notations

This section shows all of the mathematical notations used in our theory. If you forget the meaning of any notation, please refer to Table 1. We leverage ˆ to symbolize estimates (e.g. $\hat{P}(a_i)$ represents the estimate of the true value $P(a_i)$).

B Mathematical Derivations

The detailed derivation of Equation 7 is show as follows:

$$
\begin{aligned}
& Acc_t \\
&= \frac{\sum_{i=1}^{n} P(a_{i,t})}{n} \\
&= \frac{\sum_{i=1}^{n} P(a_{i,t}|a_{i,t-1})P(a_{i,t-1})}{n} \\
&+ \frac{P(a_{i,t}|\neg a_{i,t-1})P(\neg a_{i,t-1})}{n} \\
&= \frac{\sum_{i=1}^{n} P(a_{i,t-1})}{n} \frac{\sum_{i=1}^{n} P(a_{i,t-1})P(a_{i,t}|a_{i,t-1})}{\sum_{i=1}^{n} P(a_{i,t-1})} \\
&+ \frac{\sum_{i=1}^{n} [1 - P(a_{i,t-1})]}{n} \\
&* \frac{\sum_{i=1}^{n} P(\neg a_{i,t-1})P(a_{i,t}|\neg a_{i,t-1})}{\sum_{i=1}^{n} [1 - P(a_{i,t-1})]} \\
&= Acc_{t-1} * CL_{t-1} + (1 - Acc_{t-1}) * CS_{t-1}
\end{aligned}
$$

C Probability Estimation

The metrics CL and CS discussed in §3 are derived from a probabilistic perspective and the calculation depends on three key probability values for each question q_i: $P(a_{i,t})$, $P(a_{i,t+1}|a_{i,t})$, and $P(a_{i,t+1}|\neg a_{i,t})$. However, these probabilities are not directly observable. Therefore, we employ statistical methods proposed by Yang et al. (2024b) to estimate these probabilities as $\hat{P}(a_{i,t})$, $\hat{P}(a_{i,t+1}|a_{i,t})$, and $\hat{P}(a_{i,t+1}|\neg a_{i,t})$ for metric computation. Natural Language Processing (NLP) tasks are generally divided into classification and generation tasks, and we will separately discuss the probability estimation methods applicable to each type of task.

Probability Estimation for Classification Tasks. In a K-class classification task, let the set of all candidate labels be denoted by $L = \{l_0, l_1, \ldots, l_{K-1}\}$ (e.g., the candidate set for MMLU is $\{A, B, C, D\}$). A question q_i is input into the model, which outputs a predicted label. During next-token prediction, the model generates a logit vector $(o_0, o_1, \ldots, o_{|V|-1})$, where each element corresponds to a token in the vocabulary V, whose size is $|V|$. The logits are then passed through a softmax function to compute the probability distribution for the next token across the entire vocabulary. For classification tasks, we focus only on probabilities over the candidate label set L, not the whole vocabulary V. Thus, we discard most logits, retaining only those corresponding to candidate labels, producing a reduced logit vector $(o_0^{'}, o_1^{'}, \ldots, o_{K-1}^{'})$. After applying the softmax function, the model predicts the probabilities for each label $P(l_0), P(l_1), \ldots, P(l_{K-1})$.

(1) Assuming without loss of generality that the correct label is l_0, then $\hat{P}(a_{i,t}) = P(l_0)$.

(2) By feeding the correct answer l_0 back into the model for self-correction, it outputs a probability distribution over candidate labels, denoted as $P(l_0|l_0), P(l_1|l_0), \ldots, P(l_{K-1}|l_0)$, leading to $\hat{P}(a_{i,t+1}|a_{i,t}) = P(l_0|l_0)$.

(3) The computation of $\hat{P}(a_{i,t+1}|\neg a_{i,t})$ is more complex. For each incorrect label l_j ($j \neq 0$), we input it to the model, allowing for self-correction, yielding the probability of correcting to the correct label $P(l_0|l_j)$. Using the law of total probability, we have $\hat{P}(a_{i,t+1}|\neg a_{i,t}) = \sum_{j=1}^{K-1} P(l_0|l_j)P(l_j)$.

Probability Estimation for Generation Tasks. We utilize multiple sampling to estimate probabilities by observing the frequency of correct and incorrect answers. Given a question q_i, we input it to the model to obtain an initial answer, which the model then attempts to self-correct to produce a refined answer. This process is independently repeated M times, and each pair of initial and refine answers is evaluated for correctness, yielding a sequence of results $(a_{i,t}^0, a_{i,t+1}^0), (a_{i,t}^1, a_{i,t+1}^1), \ldots, (a_{i,t}^{M-1}, a_{i,t+1}^{M-1})$, where $(a_{i,t}^m, a_{i,t+1}^m)$ denotes the outcome of the m^{th} repetition. Specifically, $P(a_{i,t}^m)$ and $P(a_{i,t+1}^m)$ indicate the correctness of the initial and refined answers, respectively. For a correct initial answer $a_{i,t}^m$, $P(a_{i,t}^m) = 1$; otherwise, $P(a_{i,t}^m) = 0$. The same logic applies to $a_{i,t+1}^t$. Using these frequencies, we estimate the probabilities as follows:

(1) $\hat{P}(a_{i,t}) = \frac{\sum_{m=0}^{M-1} P(a_{i,t}^m)}{M}$;

(2) $\hat{P}(a_{i,t+1}|a_{i,t}) = \frac{\sum_{m=0}^{M-1} P(a_{i,t}^m)P(a_{i,t+1}^m)}{\sum_{m=0}^{M-1} P(a_{i,t}^m)}$;

(3) $\hat{P}(a_{i,t+1}|\neg a_{i,t}) = \frac{\sum_{m=0}^{M-1} (1-P(a_{i,t}^m))P(a_{i,t+1}^m)}{\sum_{m=0}^{M-1} (1-P(a_{i,t}^m))}$.

Notations	Meanings
Q	a dataset with n questions
q_i	the i^{th} question in Q
$a_{i,t}$	the answer to question q_i generated in the t^{th} round of self-correction
$P(a_{i,t})$	the probability of generating a correct answer for question q_i through a single temperature-based sampling in the t^{th} round of self-correction
$P(a_{i,t}\|a_{i,t-1})$	the conditional probability of $a_{i,t}$ is correct given $a_{i,t-1}$ is correct
$P(a_{i,t}\|\neg a_{i,t-1})$	the conditional probability of $a_{i,t}$ is correct given $a_{i,t-1}$ is incorrect
P_i^{con}	model confidence in question q_i: for any $t \in N+$, we have $P(a_{i,t}\|a_{i,t-1}) = P_i^{con}$
P_i^{cri}	critique capability in question q_i: for any $t \in N+$, we have $P(a_{i,t}\|\neg a_{i,t-1}) = P_i^{cri}$,
P_i^{upp}	the upper bound of $P(a_{i,t})$, and we have $P_i^{upp} = \frac{P_i^{cri}}{1-P_i^{con}+P_i^{cri}}$
α_i	the convergence rate of $P(a_{i,t})$, and we have $\alpha_i = P_i^{con} - P_i^{cri}$
Acc_0	the initial accuracy
Acc_t	accuracy after the t^{th} round of self-correction
CL	the conditional probability of getting a correct answer after self-correction, given the answer before self-correction is correct. (defined in Equation 5)
CS	the conditional probability of getting a correct answer after self-correction, given the answer before self-correction is incorrect. (defined in Equation 6)
Upp	the upper bound of Acc_t, and we have $Upp = \frac{CS}{1-CL+CS}$
α	the convergence rate of Acc_t, and we have $\alpha = CL - CS$

Table 1: Mathematical notations and their meanings.

D More Experiment Results

We try to verify on 8 models and 8 datasets in §4, but full experiments include $8 * 8 = 64$ groups, which is extremely expensive. So we only do a part of them and we believe that is sufficient to validate our theory. We show the results of 8 datasets on GLM4-9B-Chat in Figure 7, and we also show the results of 8 models on BoolQ in Figure 8, leaving more validation experiments on other models and datasets to further work.

Except for the main experiments, we also provide more results on the validation of corollaries (§5). More results on convergence rate (§5.2) are shown in Figure 9, and more results on a special case where $CL = 1$ (§5.3) are illustrated in Figure 10.

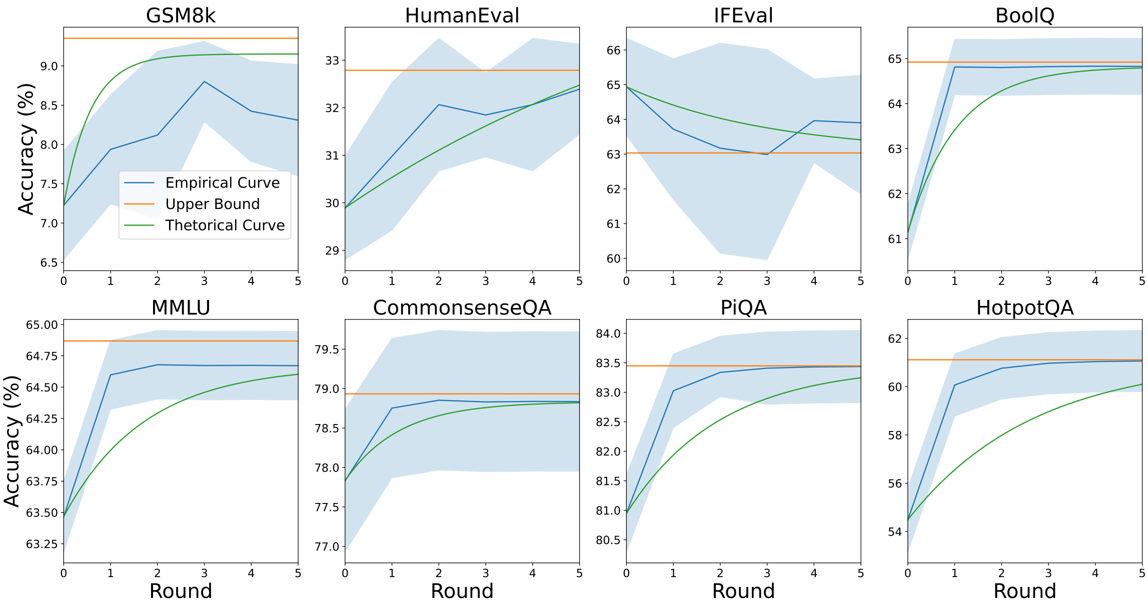

Figure 7: Experimental verification of our theory on BoolQ. The actual curve in mult-round self-correction, theoretical curve, and upper bound predicted by our theory are shown in blue, green, and orange respectively. The theoretical curve fits the actual curve well.

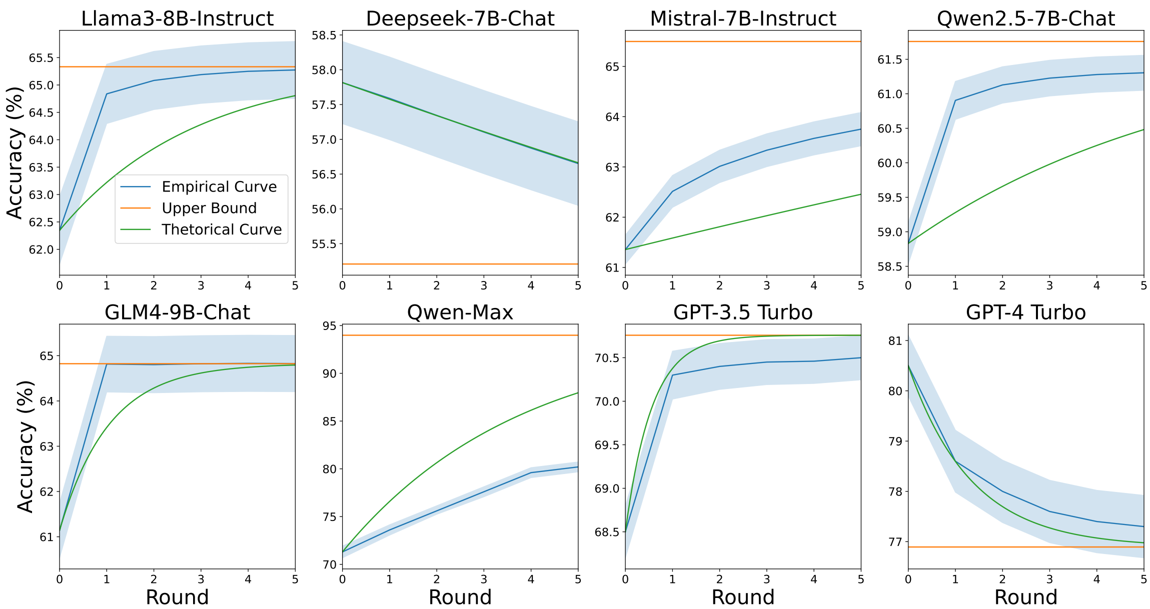

Figure 8: Experimental verification of our theory on GLM4-9B-Chat. The actual curve in mult-round self-correction, theoretical curve, and upper bound predicted by our theory are shown in blue, green, and orange respectively. The theoretical curve fits the actual curve well.

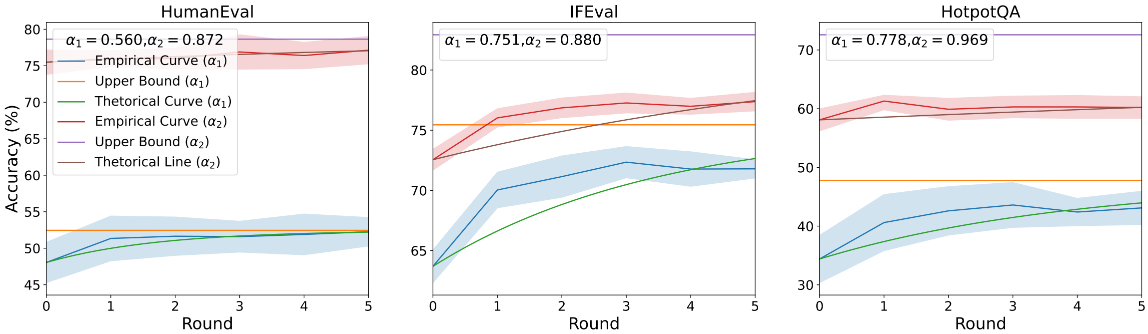

Figure 9: Convergence speed comparison of Llama3-8B-Instruct (α_1) and Qwen2.5-7B-Chat (α_2): we have $\alpha_1 < \alpha_2$, so Llama3-8B-Instruct (α_1) converges noticeably faster and its accuracy gets closer to the upper bound after 5 rounds of self-correction than Qwen2.5-7B-Chat (α_2).

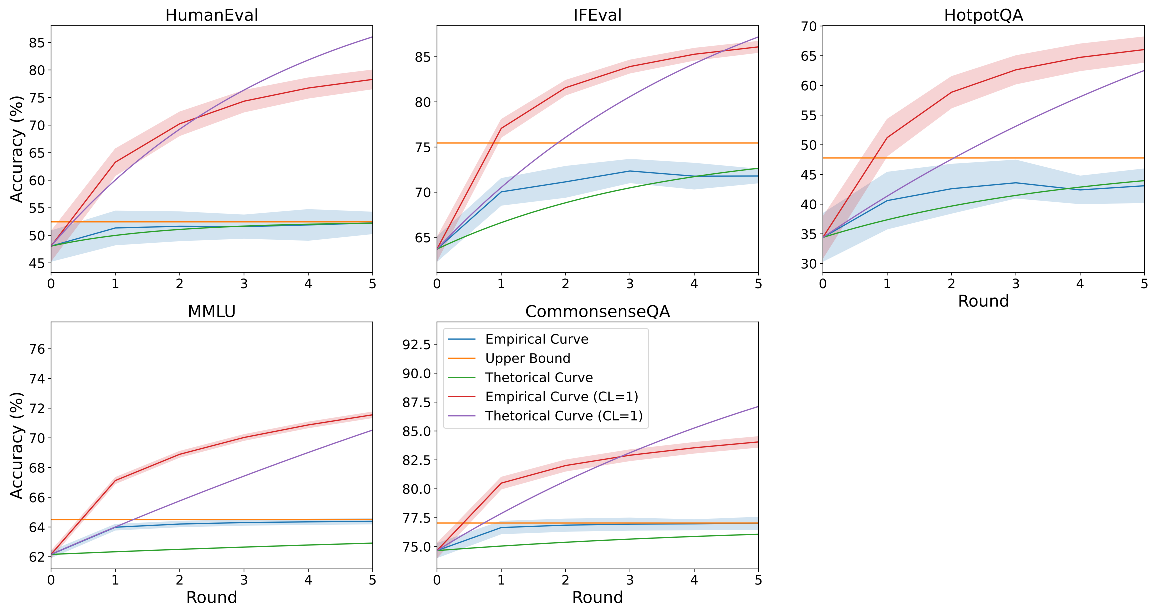

Figure 10: Curves for the special case ($CL = 1$) on Llama3-8B-Instruct. The theoretical curve fits the actual curve well when $CL = 1$, and exceeds the standard intrinsic self-correction ($CL < 1$) by a large margin.