When Can LLMs *Actually* Correct Their Own Mistakes? A Critical Survey of Self-Correction of LLMs

Ryo Kamoi[1] **Yusen Zhang**[1] **Nan Zhang**[1] **Jiawei Han**[2] **Rui Zhang**[1]
[1]Penn State University [2]University of Illinois Urbana-Champaign
{ryokamoi,rmz5227}@psu.edu

Abstract

Self-correction is an approach to improving responses from large language models (LLMs) by refining the responses using LLMs during inference. Prior work has proposed various self-correction frameworks using different sources of feedback, including self-evaluation and external feedback. However, there is still no consensus on the question of *when LLMs can correct their own mistakes*, as recent studies also report negative results. In this work, we critically survey broad papers and discuss the conditions required for successful self-correction. We first find that prior studies often do not define their research questions in detail and involve impractical frameworks or unfair evaluations that over-evaluate self-correction. To tackle these issues, we categorize research questions in self-correction research and provide a checklist for designing appropriate experiments. Our critical survey based on the newly categorized research questions shows that (1) no prior work demonstrates successful self-correction with feedback from prompted LLMs, except for studies in tasks that are exceptionally suited for self-correction, (2) self-correction works well in tasks that can use reliable external feedback, and (3) large-scale fine-tuning enables self-correction.

1 Introduction

Self-correction is a popular approach to improve responses from large language models (LLMs) by refining them using LLMs during inference (Bai et al., 2022; Madaan et al., 2023). Extensive studies on self-correction have been conducted in various tasks, including arithmetic reasoning, code generation, and question answering (Gao et al., 2023; Shinn et al., 2023). The simplest approach of self-correction prompts LLMs to provide feedback on their own responses and refine the responses using the feedback (Huang et al., 2024a), under the hypothesis that *recognizing errors is easier than avoiding them* (Saunders et al., 2022). As in Figure 1, self-correction has also been studied using additional information for improving feedback, including external tools such as code interpreters (Chen et al., 2024e; Gou et al., 2024), external knowledge retrieved via web search (Gao et al., 2023; Jiang et al., 2023b), or fine-tuning (Welleck et al., 2023; Ye et al., 2023). However, recent studies also report negative results indicating that LLMs cannot self-correct (Huang et al., 2024a; Gou et al., 2024; Li et al., 2024b; Chen et al., 2024f) or even self-detect (Chen and Shu, 2024; Tyen et al., 2024; Hong et al., 2024; Jiang et al., 2024; Kamoi et al., 2024) their own mistakes at least in certain conditions. These conflicting observations indicate that further analysis of self-correction is needed.

In this work, we provide a critical survey to investigate the conditions required for successful self-correction. First, our analysis finds that prior studies often do not define their research questions in detail. As a result, many papers fail to provide appropriate experiments to evaluate the research questions they implicitly target. To address this issue, we categorize research questions in self-correction research (§3.1) and discuss frameworks that should be used for verifying each research question (§3.2). Finally, we provide a checklist for designing appropriate experiments (§8).

Next, we analyze prior work to identify when LLMs can self-correct their mistakes, using the new definitions of the research questions. Our analysis highlights that the bottleneck is in the feedback generation (§7). Specifically, (1) no prior work shows successful self-correction with feedback from prompted LLMs in general tasks (§4), (2) self-correction works well in tasks where reliable external feedback is available (§5.1), (3) large-scale fine-tuning enables self-correction (§5.2), and (4) some tasks have properties exceptionally suit-

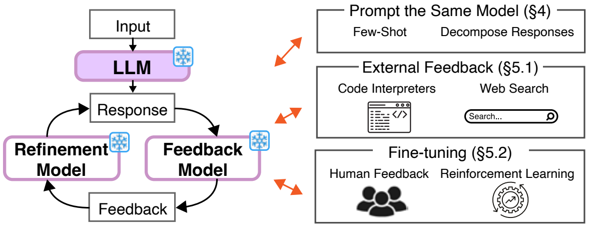

Figure 1: Self-correction in three stages: initial response generation, feedback, and refinement.

able for self-correction (§4). In summary, our analysis identifies the properties required for successful self-correction as follows:

[RQ1] When can LLMs self-correct *based solely on the inherent capabilities of LLMs?*

- In general tasks, no prior work shows reliable evidence of successful self-correction with in-context learning. (§4)
- In tasks with specific properties that are exceptionally favorable for self-correction (e.g., responses are decomposable), self-correction is effective even with in-context learning. (§4)

[RQ2] When can LLMs self-correct the best-possible initial responses *with external information?*

- Self-correction is effective in tasks where reliable external feedback is available. (§5.1)
- Fine-tuning enables self-correction when large training data is available but is unexplored for small training data. (§5.2)

[RQ3] When are the final outputs of self-correction *better than other approaches*?

- Self-correction is often not compared with sufficiently strong baselines, and it is still unclear whether it is better than other approaches. (§6)

This survey is organized as follows. Section 2 provides an overview of self-correction. Section 3 introduces a new approach to classify research questions and frameworks in self-correction research. Sections 4 and 5 analyze prior work in self-correction with in-context learning and external information (external tools, external knowledge, fine-tuning), respectively. Section 6 explains related approaches that should be compared with self-correction as baselines. Section 7 summarizes our findings from the analysis. Section 8 provides a checklist for self-correction research. Section 9 explains differences from other surveys. Section 10 provides studies related to self-correction. Section 11 provides future directions.

Timeframe. This survey was originally published in June 2024 and primarily includes research papers and studies published up to and including May 2024. While papers published after this date are not comprehensively analyzed, they are briefly discussed in Section 12.

2 Self-Correction of LLMs

The term "self-correction" is used in a wide range of scenarios, from a strict definition in which LLMs refine their own responses by themselves (Madaan et al., 2023; Huang et al., 2024a) to broader concepts that also involve feedback from external tools or knowledge (Shinn et al., 2023; Gou et al., 2024). In this work, we define self-correction as a framework that *refines* responses from LLMs using LLMs *during inference*, possibly with external tools or knowledge. As in Table 1, Figure 2, and Figure 3, self-correction has been studied in various frameworks with different sources of feedback.

2.1 Frameworks

Prior studies propose self-correction frameworks with various different architectures.

Explicit Feedback vs. Direct Refinement. Self-correction often consists of three stages including *feedback generation* (Kim et al., 2023; Madaan et al., 2023; Shinn et al., 2023; Huang et al., 2024a):

- **Initial Response Generation** is a stage of generating initial responses from an LLM.
- **Feedback** model generates feedback given the original input and initial response. This stage may use external tools or knowledge.
- **Refinement** model generates a refined response, given the input, initial response, and feedback.

Direct refinement is another approach that refines responses without generating feedback explicitly (Saunders et al., 2022; Bai et al., 2022; Welleck et al., 2023; Akyurek et al., 2023).

Post-hoc vs. Generation-time. *Post-hoc correction* refines responses after they are generated (Pan et al., 2024). *Generation-time correction* or step-level correction (Paul et al., 2024; Jiang et al., 2023b) improves step-by-step reasoning by providing feedback on intermediate reasoning steps. Post-hoc correction is more flexible and applicable to

Paper	Category	Main Models	Additional Feedback			Main Tasks				
			Oracle	External Tools	Fine-Tuning	Reasoning, Coding	Closed-book, Knowledge	Open-book, Context-based	Open-ended Text Gen	Decom-posable
Self-Correction with **In-context Learning** (*Intrinsic Self-Correction*)										
CoVe (2024)	Intrinsic	PaLM 540B	–	–	–	–	–	–	–	Multiple Answers
CAI Revisions (2022)♠	Intrinsic	52B (no details)	–	–	–	–	–	–	Detoxification	–
Self-Refine (2023)♠	Intrinsic	GPT-3.5, GPT-4	–	–	–	Math, Coding	–	Dialogue	–	–
RCI (2023, §3.1)	Oracle	GPT-3.5-T	✓	–	–	Computer Tasks	CSQA	–	–	–
Reflexion (2023, §4.2)	Oracle	GPT-4	✓	–	–	–	–	HotpotQA (GT Context)	–	–
Self-Correction with **External Tools or Knowledge**										
Reflexion (2023, §4.1, 4.3)	Fair-Asym.	GPT-4	–	Game Envs, Interpreter	–	Games, Coding	–	–	–	–
Self-Debug (2024e)	Fair-Asym.	GPT-3.5-T, GPT-4	–	Code Interpreter	–	Text-to-Code	–	–	–	–
CRITIC (2024)	Fair-Asym.	GPT-3, Llama 2 70B	–	Interpreter, Web Search	–	GSM8k, SVAMP	HotpotQA	–	Detoxification	–
RARR (2023)	Unfair-Asym.	Palm 540B	–	Web Search	–	–	NQ, SQA, QReCC	–	–	–
Reflexion (2023, §4.2)	Oracle	GPT-4	✓	Wikipedia API	–	–	HotpotQA	–	–	–
Self-Correction with **Fine-tuning**										
Self-Critique (2022)	Fair-Asym.	InstructGPT	–	–	Human Assessment	–	–	Topic-based Summarization	–	–
SelFee (2023)	Fair-Asym.	Llama 7B, 13B	–	–	ChatGPT Assessment	MT-Bench	MT-Bench	MT-Bench	MT-Bench	–
Baldur (2023)	Fair-Asym.	Minerva 8B ,62B	–	Proof Assistant	GT Answer	Proof Generation	–	–	–	–
REFINER (2024)	Cross-Model	GPT-3.5 (FB:T5-base)	–	–	Synthetic Data	Math, Logic	–	–	Moral Stories	–
RL4F (2023)	Cross-Model	GPT-3 (FB: T5-large)	–	–	Reinforcement Learning	Action Planning	–	Topic-based Summarization	–	–
Self-Correction (2023, §3.4)	Cross-Model	GPT-3 (FB: GPT-Neo)	–	–	GT Answer, External	GSM8k, SVAMP	–	–	Detoxification	–
Self-Correction (2023, §3.1-3.3)	Unfair-Asym.	GPT-Neo 1.3B, GPT-2	–	–	GT Answer, External	GSM8k, SVAMP	–	–	Detoxification, Const Gen	–
Negative Results of Self-Correction (i.e., LLMs cannot Self-Correct)										
RCI (Table 17) (2023)	Intrinsic	GPT-3.5-T	–	–	–	Computer Tasks	CSQA	–	–	–
CRITIC w/o Tool (2024)	Intrinsic	GPT-3, Llama 2 70B	–	–	–	GSM8k, SVAMP	Closed-book HotpotQA	–	Detoxification	–
Huang et al. (2024a)	Intrinsic	GPT-4-T, GPT-3.5-T	–	–	–	GSM8k	CSQA, HotpotQA	–	–	–

Table 1: Representative studies in self-correction of LLMs. Gray color represents unrealistic settings. ♠: Weak prompts for generating initial responses. FB: Feedback models for cross-model correction.

broader tasks, although generation-time correction is popular for reasoning tasks (Pan et al., 2024).

Same-model vs. Cross-model. *Cross-model correction* generates feedback or refines the responses using models different from the model that generates initial responses. Cross-model correction has been mostly studied in the settings of correcting mistakes of large proprietary LLMs using small fine-tuned models (Welleck et al., 2023; Akyurek et al., 2023; Paul et al., 2024) or multi-agent debate of multiple models with similar capabilities (Liang et al., 2024; Li et al., 2023; Cohen et al., 2023; Du et al., 2023; Zhang et al., 2024a; Chen et al., 2024b; Chan et al., 2024a; Wang et al., 2024a).

2.2 Sources of Feedback

Intrinsic (§4). Intrinsic self-correction prompts LLMs to generate feedback on their own responses. Prompting strategies include simple zero-shot or few-shot prompts (Madaan et al., 2023; Kim et al., 2023), decomposing the responses (Dhuliawala et al., 2024), and evaluating confidence (Varshney et al., 2023; Jiang et al., 2023b; Wu et al., 2024).

External Information (§5.1). Self-correction often relies on external information, including **external tools** such as code executors (Jiang et al., 2023a; Gou et al., 2024; Chen et al., 2024e; Stengel-Eskin et al., 2024), symbolic reasoners (Pan et al., 2023), proof assistant (First et al., 2023), or task-