RQ1	RQ2	RQ3	Requirements for Verifying the Target RQs		
✓	✓	✓	Clearly stating the target RQ and the category of self-correction framework discussed.	(§3.2)	Required
✓	✓	✓	Not using oracle information, such as ground-truth answers.	(§4)	Required
✓	✓	✓	*When using fine-tuning*, reporting the detailed settings, including the number of annotations and computational cost required to achieve the reported performance.	(§5.2)	Required
✓	✓	✓	Evaluating the quality of feedback directly (e.g., error detection accuracy).	(§7)	Recommended
✓	✓		Using sufficiently strong prompts for generating initial responses.	(§4)	Required
✓			Using intrinsic self-correction.	(§3.2)	Required
			When using external tools or knowledge,		
	✓		Using external tools or knowledge to improve initial response generation as much as possible.	(§5.1)	Required
			When using fine-tuning for self-correction,		
	✓		Fine-tuning initial response generators as well, as much as possible.	(§5.2)	Required
	✓		Evaluating the minimum required size of training data that enables self-correction.	(§5.2)	Recommended
	✓		Evaluating cross-model correction setting that refines mistakes in responses from stronger LLMs.	(§3.2)	Recommended
		✓	Comparing with strong baselines using comparable computational cost.	(§6)	Required

Table 7: Checklist for self-correction research for different target research questions.

Clearly stating the RQ that is refuted by the reported results and the category of the framework discussed.	(§3.2)	Required
Using strong prompts for self-correction (e.g., state-of-the-art reference-free metrics).	(§11)	Required
When not using external tools or knowledge available in real-world applications, explicitly reporting that the evaluation is done under weak conditions.	(§5.1)	Required
Evaluating with external tools or knowledge available in real-world applications.	(§5.1)	Recommended

Table 8: Checklist for reporting negative results of self-correction.

decoding generates multiple responses and selects the best response for each reasoning step using generation probability (Hao et al., 2023; Tyen et al., 2024), prompted self-evaluation (Jung et al., 2022; Creswell and Shanahan, 2022; Xie et al., 2023; Yao et al., 2023; Miao et al., 2024), or fine-tuned verifiers (Uesato et al., 2022; Tafjord et al., 2022; Yang et al., 2022a; Asai et al., 2024).

7 Summary of Our Analysis

Bottleneck is in Feedback Generation. Prior studies widely agree that LLMs can *refine* their responses given reliable feedback (§5). However, generating reliable *feedback* on their own responses is still observed to be challenging for LLMs without using additional information (§4). In other words, for the current LLMs, the hypothesis that *recognizing errors is easier than avoiding them* (Saunders et al., 2022) is only true for certain tasks whose verification is exceptionally easy, according to our analysis of the experiments in prior studies. We recommend that self-correction research analyze the quality of generated feedback in more detail, not only evaluate the downstream performance of the refined responses.

Tasks Suitable for Self-Correction. Our analysis identifies the properties of tasks that are suitable for self-correction under different conditions.

- Intrinsic Self-Correction (§4)
  - Tasks whose verification tasks are much easier than the original tasks (e.g., tasks whose responses are decomposable)
- Self-Correction with External Information (§5.1)
  - Tasks for which external tools that provide reliable feedback exist (e.g., code generation)
  - Tasks for which responses can be utilized to obtain useful information that is difficult to obtain before generating initial responses (e.g., generate queries from responses to retrieve documents for verifying information)
- Self-Correction with Fine-tuning (§5.2)
  - Self-correction works in many tasks when large training data for feedback generation is available
  - Tasks that can use reinforcement learning or self-corrective learning (Welleck et al., 2023), i.e., tasks whose responses can be easily evalu-

ated given ground-truth answers

8 Checklist for Self-Correction Research

Our analysis shows that many studies do not clearly define their research questions and fail to conduct appropriate experiments (§3.1, 4). To tackle these issues, we provide a checklist for self-correction research that provides requirements for designing appropriate experiments for verifying target RQs and recommended experiments for comprehensive analysis. Table 7 provides a checklist for verifying different RQs identified in Section 3.1. Table 8 provides a checklist for reporting negative results.

9 Differences from Other Survey

Pan et al. (2024) provide a comprehensive survey on broad topics related to self-correction, including training strategies. Our work specifically focuses on (inference-time) self-correction and provides a more detailed and critical analysis of prior work. Huang et al. (2024a) provide an analysis of problems in the evaluation settings of self-correction research, which motivates our work. They focus on analyzing a few papers on intrinsic self-correction in reasoning tasks. We provide a more comprehensive analysis of self-correction with in-context learning, external tools, and fine-tuning.

10 Related Work of Self-Correction

Self-Detection of mistakes in LLM responses using LLMs (possibly with external information) has been studied in various domains, including misinformation detection (Zhang et al., 2024b; Chern et al., 2023; Chen and Shu, 2024; Mishra et al., 2024), context-faithfulness (Wang et al., 2020; Durmus et al., 2020; Scialom et al., 2021), harmful content detection (Rauh et al., 2022), and bias detection (Blodgett et al., 2020; Feng et al., 2023). However, recent studies (Tyen et al., 2024; Kamoi et al., 2024) show that even strong LLMs often cannot detect their own mistakes in various tasks.

Editing Human-Written Text by using language models has been studied in various domains, including information update (Shah et al., 2020; Iv et al., 2022; Schick et al., 2023), grammatical error correction (Ng et al., 2014; Lichtarge et al., 2019), factual error correction (Cao et al., 2020; Thorne and Vlachos, 2021), and code repair (Gupta et al., 2017; Mesbah et al., 2019; Bader et al., 2019; Chen et al., 2021; Yasunaga and Liang, 2020, 2021).

Self-Training or self-improvement is an approach to train models using their own responses. Some studies use self-evaluation or self-correction for creating training data (Bai et al., 2022; Gulcehre et al., 2023) or use self-evaluation as training signals (Pang et al., 2024). Another approach improves the reasoning of LLMs using LLM-generated reasoning by selecting high-quality outputs using ground-truth final answers (Zelikman et al., 2022) or self-consistency (Huang et al., 2023). As another direction, Meng et al. (2022) use sentences generated by LLMs with high confidence for training classifiers.

11 Future Directions

Improving Feedback. Prior studies indicate that it is difficult for LLMs to generate feedback on their own responses with in-context learning (§4, 7). However, most studies in intrinsic self-correction (Madaan et al., 2023; Huang et al., 2024a) use simple prompts for generating feedback, and there is room for improvement. A possible direction to improve feedback is to apply (reference-free and point-wise) **LLM-based evaluation metrics**. Recent approaches for improving the model-based evaluation include using human-written evaluation criteria (Chiang and Lee, 2023; Liu et al., 2023) and decomposing responses (Saha et al., 2024; Min et al., 2023). As another direction, recent studies in self-correction propose frameworks using the **confidence** in their responses, estimated by generation probabilities (Varshney et al., 2023; Jiang et al., 2023b), prompting (Li et al., 2024a), or generating new questions from their answers to evaluate logical consistency (Jung et al., 2022; Tafjord et al., 2022; Wu et al., 2024).

Unexplored Tasks. The difficulty of self-evaluation differs from task to task (§4), while many studies assume that verification is consistently easier than generation. We expect that there are unexplored tasks in which intrinsic self-correction works well, although self-correction research mostly focuses on reasoning tasks such as math reasoning and coding (Madaan et al., 2023; Gou et al., 2024; Huang et al., 2024a). For example, LLM-based evaluation is often studied in open-ended text generation, such as dialogue generation and text summarization (Fu et al., 2024; Liu et al., 2023), suggesting that reasonable model-based feedback is available for these tasks.

Fine-tuning on Small Training Data. Fine-tuning of feedback generation often relies on large training data, which requires large-scale human annotations (§5.2). We expect future work to explore self-correction with smaller training data. Although reinforcement learning (Akyurek et al., 2023) or self-corrective learning (Welleck et al., 2023) do not require human feedback, they require reasonable reward functions for evaluating LLM responses, which are not available in many tasks. For example, RL4F (Akyurek et al., 2023) uses ROUGE as a reward function for text summarization and action planning, which is sub-optimal.

Pre-training for Improving Self-Correction. Prior studies show that large-scale fine-tuning on reference feedback improves the self-correction capability of LLMs (§5.2). This observation suggests that the current approach or datasets for pre-training LLMs are insufficient to make LLMs acquire self-correction capability. We expect future work to explore pre-training strategies to improve the intrinsic self-correction capability of LLMs.

12 Emerging Trends and Recent Developments

This survey, originally published in June 2024, primarily focuses on papers published before May 2024. However, to provide a broader perspective, this section briefly highlights emerging trends and recent advancements from June 2024 onward.

A recent trend of self-correction involves employing reinforcement learning (Kumar et al., 2024; Qu et al., 2024). Specifically, OpenAI has published o1 (OpenAI, 2024), a model for reasoning tasks trained with reinforcement learning to explore different strategies, recognize their own mistakes, and refine their thinking process. OpenAI o1 has been reported to outperform state-of-the-art LLMs in various reasoning tasks, including Math Olympiad, PhD-level academic problems, competitive programming, and Kaggle (Chan et al., 2024b).

13 Conclusion

We provide a critical survey of self-correction to identify in which conditions LLMs can self-correct their mistakes. Our analysis reveals that many studies fail to define their research questions clearly or design experiments appropriately. To tackle these issues, we categorize research questions and frameworks in self-correction research and provide a checklist for conducting appropriate experiments.

Acknowledgments

This work was supported by a Cisco Research Grant. We appreciate valuable suggestions from the action editor and anonymous reviewers.

References

Afra Feyza Akyurek, Ekin Akyurek, Ashwin Kalyan, Peter Clark, Derry Tanti Wijaya, and Niket Tandon. 2023. RL4F: Generating natural language feedback with reinforcement learning for repairing model outputs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7716–7733, Toronto, Canada. Association for Computational Linguistics.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Johannes Bader, Andrew Scott, Michael Pradel, and Satish Chandra. 2019. Getafix: learning to fix bugs automatically. *Proc. ACM Program. Lang.*, 3(OOPSLA).

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP.