International Research Journal of Multidisciplinary Sciences (Double - Blind Peer Review Journal)
VOL-2 ISSUE-4 April 2026 PP:27-33 ISSN:3107-930X(ONLINE)

Self Correction without External Feedback: Can LLMs Actually Verify Their Own Reasoning

[1]Dr.YazdaniHasan
[2]KavitaKumari
[1]NoidaInternationalUniversity
[2]HIMTGreaterNoida

Abstract

The ability of Large Language Models (LLMs) to self correct—detecting and rectifying errors in their own outputs without external feedback—has emerged as a cornerstone of autonomous AI systems. Yet the fundamental question of whether LLMs can genuinely verify their own reasoning remains deeply contested. This paper provides a comprehensive analysis of intrinsic self correction capabilities in LLMs, synthesizing recent empirical findings and theoretical frameworks. We decompose self correction into three distinct sub capabilities—error detection, error localization, and error correction—and examine the evidence for and against each. Our analysis reveals a consistent pattern: while LLMs possess robust error correction abilities when mistakes are explicitly identified, they fundamentally struggle with autonomous error detection and localization. We identify the "verification bottleneck" as the primary obstacle to effective self correction and survey emerging approaches including key condition verification, program driven verification, and reinforcement learning based training. We further analyze self correction through a feedback control theoretical lens, deriving stability conditions for iterative refinement. We conclude that intrinsic self correction without external feedback remains severely limited for complex reasoning tasks, though targeted prompting strategies and specialized training can partially mitigate these limitations.

Keywords: Large Language Models, Self Correction, Self Verification, Reasoning, Error Detection

1. Introduction

The capacity to recognize and correct one's own mistakes is a hallmark of intelligent reasoning. As Large Language Models (LLMs) are increasingly deployed in high stakes applications—from mathematical problem solving and code generation to scientific reasoning and medical diagnosis—understanding their capacity for autonomous self correction has become critically important. Self correction aims to enable LLMs to self verify and self refine their initial responses without external feedback, offering the promise of more reliable and autonomous AI systems.

The research community has witnessed a proliferation of claims about LLMs' ability to successfully verify or self critique their candidate solutions in reasoning problems. Self Refine (Madaan et al., 2023) and Reflexion (Shinn et al., 2023) demonstrated that iterative self refinement could improve outputs in terms of style and quality. However, recent attempts to self correct logical or

reasoning errors often cause correct answers to become incorrect, resulting in worse overall performance. A growing body of evidence suggests that intrinsic self correction—where models correct errors without external validation signals—is fundamentally limited.

This paper addresses a central research question: Can LLMs actually verify their own reasoning without

external feedback? We synthesize findings from over twenty recent studies to provide a nuanced answer. Our contributions are threefold: (1) we decompose self correction into its constituent sub capabilities, revealing where LLMs succeed and where they fail; (2) we survey emerging techniques that attempt to overcome these limitations; and (3) we provide a theoretical framework for understanding when and why self correction works or fails.

2. Decomposing Self Correction: Detection, Localization, and Correction

A critical insight emerging from recent research is that "self correction" conflates several distinct capabilities that deserve independent study. Following Li (2026), we decompose self correction into three measurable sub capabilities:

1. Error Detection : Can the model identify that its output contains an error?
2. Error Localization : Can it pinpoint where the error occurs?
3. Error Correction : Can it produce a corrected solution?

This decomposition reveals important and counterintuitive insights about the limitations of current LLMs.

2.1 The Verification Bottleneck

Tyen et al. (2024) conducted a landmark study demonstrating that poor self correction performance stems from LLMs' inability to find logical mistakes , rather than their inability to correct a known mistake. Benchmarking several state of the art LLMs on their mistake finding ability, they found that models generally struggle with error detection, even in highly objective, unambiguous cases. When provided with ground truth mistake location information through a backtracking setup, models showed robust correction abilities, boosting downstream task performance across five reasoning tasks.

This "verification bottleneck" has been corroborated by multiple studies. VeriCoT observed that while LLMs can perform multi step reasoning through Chain of Thought (CoT), they "cannot reliably verify their own logic". Even when models reach correct answers, the underlying reasoning may be flawed, undermining trust in high stakes scenarios.

2.2 The Accuracy Correction Paradox

Li (2026) uncovered a striking Accuracy Correction Paradox : weaker models achieve higher intrinsic correction rates than stronger models. Through cross model experiments on GSM8K Complex with three major LLMs, GPT 3.5 (66% accuracy) achieved a 26.8% intrinsic correction rate, while DeepSeek (94% accuracy) achieved only 16.7%. The proposed Error Depth Hypothesis explains this paradox: stronger models make fewer but "deeper" errors

that resist self correction, while weaker models make "shallower" errors that are easily fixable.

Error detection rates vary dramatically across architectures—from 10% (Claude) to 82% (GPT 3.5)—yet detection capability does not predict correction success. Claude detects only 10% of errors but corrects 29% intrinsically. Surprisingly, providing error location hints hurts all models. These findings challenge linear

assumptions about model capability and self improvement.

2.3 The Self Correction Blind Spot

Wu et al. (2024) identified a systematic "Self Correction Blind Spot": while LLMs can identify errors in user input, they fail to correct identical errors in their own outputs. This finding suggests that self correction failures are not merely about reasoning ability but involve a fundamental asymmetry in how models process external versus self generated content. Observing LLM self correction behavior in natural settings is challenging due to their inherent accuracy—the rarity of naturally occurring errors makes systematic evaluation difficult.

3. Empirical Evidence: When Self Correction Fails

The empirical literature on intrinsic self correction presents a sobering picture. Multiple studies have documented systematic failures.

3.1 Degradation Rather Than Improvement

Liu et al. (2026) recast self correction as a closed loop feedback control problem, analyzing error dynamics via a two state Markov model parameterized by Error Introduction Rate (EIR) and Error Correction Rate (ECR). Their empirical analysis across seven models and three datasets (GSM8K, MATH, StrategyQA) revealed a sharp near zero EIR boundary (< 0.5%) separating beneficial from harmful self correction. Only o3 mini (+3.4 pp), Claude Opus 4.6 (+0.6 pp), and o4 mini (+/ 0 pp) stayed non degrading, while GPT 5 and four others lost accuracy.

Valmeekam et al. (2023) found that self critiquing appears to diminish plan generation performance, especially when compared to systems with external, sound verifiers. LLM verifiers produce a notable number of false positives, compromising system reliability. The nature of feedback—whether binary or detailed—showed minimal impact on plan generation.

3.2 Introducing New Errors

Research on self reflection in medical question answering showed that self reflective loops can lead to error correction, error persistence, or the introduction of new errors. This "dark side" of intrinsic self correction has been further documented by studies showing that self correction can cause LLMs to waver on both intermediate and final answers, introducing prompt bias on simple factual questions and human like cognitive bias on complex tasks.

3.3 Reformulation Without Progress

Recursive self evaluation without external feedback frequently yields reformulation rather than progress. Models may produce revised outputs that are superficially different but do not address the underlying reasoning errors. This phenomenon underscores the distinction between surface level refinement and genuine error correction.

4. When Self Correction Succeeds: Emerging Approaches

Despite the sobering evidence, recent work has identified conditions under which self correction can succeed.

4.1 Key Condition Verification

Wu et al. (2024) demonstrated that a simple yet effective verification method can unleash the inherent capabilities of LLMs. The approach masks a key condition in the question, adds the current response to construct a verification question, and predicts the condition to verify the response. The condition can be an entity in an open domain question or a numeric value in a math question, requiring minimal effort to identify.

Their ProCo framework (Progressive Verify then Correct) yielded significant improvements: +6.8 exact match on open domain QA datasets, +14.1 accuracy on arithmetic reasoning datasets, and +9.6 accuracy on a commonsense reasoning dataset compared to Self Correct. This suggests that targeted verification prompts can overcome some limitations of intrinsic self correction.

4.2 Program Driven Verification

Song et al. (2025) proposed Program driven Self Correction (ProgCo), which addresses the failure of LLMs to effectively self verify, especially in complex reasoning tasks. Program driven verification (ProgVe) achieves complex verification logic and extensive validation through self generated, self executing verification pseudo programs. Program driven refinement (ProgRe) receives feedback from ProgVe and conducts dual reflection and refinement on both responses and verification programs. Experiments on instruction following and mathematical benchmarks demonstrated effective self correction.

4.3 Reinforcement Learning Approaches

Ma et al. (2025) introduced S2R, an efficient framework that enhances LLM reasoning by teaching models to self verify and self correct during inference. Using only 3.1k behavior initialization samples, Qwen2.5 math 7B achieved an accuracy improvement from 51.0% to 81.6%, outperforming models trained on equivalent amounts of long CoT distilled data. Self verification and self correction skills were strengthened through outcome level and process level reinforcement learning.

Xiong et al. (2025) proposed a two staged algorithmic framework for constructing self rewarding reasoning models using only self generated data. The first stage employs sequential rejection sampling to synthesize long chain of thought trajectories incorporating both self rewarding and self correction mechanisms. The second stage enhances models' ability to assess response accuracy and refine outputs through reinforcement learning with rule based signals. Experiments with Llama 3 and Qwen 2.5 demonstrated that this approach surpasses intrinsic

self correction capabilities and achieves performance comparable to systems relying on external reward models.

4.4 Test Time Scaling with Self Verification

Lee et al. (2025) proposed ReVISE (Refine via Intrinsic Self Verification), which enables LLMs to verify their reasoning processes and continually rethink reasoning trajectories based on verification. Through curriculum learning that collects both failed and successful reasoning paths to construct preference pairs, ReVISE achieves efficient self correction and significantly improves reasoning performance. During inference, the approach

enjoys natural test time scaling by integrating self verification and correction capabilities.

5. Theoretical Framework: Self Correction as Feedback Control

Liu et al. (2026) provide a rigorous theoretical framework for understanding self correction dynamics. By recasting self correction as a closed loop feedback control problem where the same model serves as both controller and plant, they derive a directly measurable stability threshold : iterate only when ECR/EIR > Acc/(1 Acc), where ECR is the Error Correction Rate and EIR is the Error Introduction Rate.

This framework yields several important insights:

1. EIR as stability margin : The Error Introduction Rate determines whether iterative refinement improves or degrades performance.
2. Prompt level interventions : A verify first prompt intervention drove GPT 4o mini's EIR from 2% to 0% and converted a 6.2 pp degradation into +0.2 pp improvement.
3. Two tier capability structure : Prompt level EIR suppression prevents degradation, whereas ECR enhancement—plausibly requiring training level interventions—is necessary for genuine gains.

The key prescription is that "self correction should thus be treated not as a default behavior but as a control decision governed by measurable error dynamics".

6. Discussion and Future Directions

6.1 The Fundamental Limitation

The cumulative evidence suggests that intrinsic self correction without external feedback is severely limited for complex reasoning tasks . The primary bottleneck is not correction ability but verification ability—LLMs cannot reliably detect their own errors. As Tyen et al. (2024) demonstrated, when provided with error locations, models can correct effectively. The challenge lies in autonomous error detection and localization.

6.2 Implications for System Design

These findings have important implications for deploying LLMs in high stakes applications:

External verification remains essential : Systems should not rely on LLMs' self verification alone but should incorporate external validation mechanisms.

Targeted prompting can help : Key condition verification and similar prompting strategies can improve self correction performance.

Training matters : Reinforcement learning approaches can teach models to self verify and self correct more effectively.

6.3 Open Questions

Several important questions remain for future research:

1. Can verification ability be learned? While current models struggle with autonomous verification, training approaches like S2R and ReVISE show promise.
2. What explains architectural differences? The dramatic variation in error detection rates across architectures (10% to 82%) warrants further investigation.
3. How can we measure error "depth"? The Error Depth Hypothesis needs empirical validation and operationalization.
4. What is the role of model scale? CriticBench findings suggest GQC knowledge inconsistencies decrease as model size increases, but the relationship between scale and self correction ability remains unclear.

7. Conclusion

This paper has critically examined whether LLMs can verify their own reasoning without external feedback. Our analysis, synthesizing evidence from over twenty recent studies, reveals a consistent pattern: while LLMs possess robust error correction abilities when mistakes are explicitly identified, they fundamentally struggle with autonomous error detection and localization. The "verification bottleneck" represents the primary obstacle to effective intrinsic self correction.

Emerging approaches—including key condition verification, program driven verification, and reinforcement learning based training—offer pathways to partially overcome these limitations. Theoretical frameworks, particularly the feedback control analysis of error dynamics, provide principled guidance for when and how to deploy self correction.

We conclude that intrinsic self correction without external feedback remains fundamentally limited for complex reasoning tasks . LLMs cannot reliably verify their own reasoning. However, with targeted prompting strategies, specialized training, and appropriate system design—including external verification mechanisms when feasible—self correction can be a valuable tool in the AI practitioner's arsenal. The path forward lies not in expecting LLMs to be perfect self verifiers, but in understanding their capabilities and limitations and designing systems that leverage their strengths while compensating for their weaknesses.

References

1. Huang, J., et al. (2024). On the Intrinsic Self Correction Capability of LLMs: Uncertainty and Latent Concept. arXiv preprint .
2. Li, Y. (2026). Decomposing LLM Self Correction: The Accuracy Correction Paradox and Error Depth Hypothesis. arXiv preprint .
3. Lin, Z., Gou, Z., Liang, T., Luo, R., Liu, H., & Yang, Y. (2024). CriticBench: Benchmarking LLMs for Critique Correct Reasoning. In Findings of ACL 2024 , pages 1552–1587.
4. Liu, A., et al. (2026). Self Correction as Feedback Control: Error Dynamics, Stability

Thresholds, and Prompt Interventions in LLMs. arXiv preprint .
5. Ma, R., Wang, P., Liu, C., Liu, X., Chen, J., Zhang, B., Zhou, X., Du, N., & Li, J. (2025). S2R: Teaching LLMs to Self verify and Self correct via Reinforcement Learning. In Proceedings of ACL 2025 (Volume 1: Long Papers) , pages 22632–22654.
6. Madaan, A., et al. (2023). Self Refine: Iterative Refinement with Self Feedback. arXiv preprint .
7. Shinn, N., et al. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning. arXiv preprint .
8. Song, X., Wu, Y., Wang, W., Liu, J., Su, W., & Zheng, B. (2025). ProgCo: Program Helps Self Correction of Large Language Models. In Proceedings of ACL 2025 (Volume 2: Short Papers) , pages 944–959.
9. Tyen, G., Mansoor, H., Carbune, V., Chen, P., & Mak, T. (2024). LLMs cannot find reasoning errors, but can correct them given the error location. In Findings of ACL 2024 , pages 13894–13908.
10. Valmeekam, K., Marquez, M., & Kambhampati, S. (2023). Can Large Language Models Really Improve by Self critiquing Their Own Plans? arXiv preprint .
11. Wu, Z., Zeng, Q., Zhang, Z., Tan, Z., Shen, C., & Jiang, M. (2024). Large Language Models Can Self Correct with Key Condition Verification. In Proceedings of EMNLP 2024 , pages 12846–12867.
12. Xiong, W., et al. (2025). Self rewarding correction for mathematical reasoning. arXiv preprint .
13. Lee, H., Oh, S., Kim, J., Shin, J., & Tack, J. (2025). ReVISE: Learning to Refine at Test Time via Intrinsic Self Verification. arXiv preprint .
14. When Can LLMs Actually Correct Their Own Mistakes? A Critical Survey of Self Correction of LLMs. (2024). Transactions of the Association for Computational Linguistics , 12:1417 1440.
15. Learning From Correctness Without Prompting Makes LLM Efficient Reasoner. (2024). arXiv preprint .

16. On the Intrinsic Self Correction Capability of LLMs: Uncertainty and Latent Concept. (2024). arXiv preprint .
17. Small Language Model Can Self correct. (2024). arXiv preprint .
18. Understanding the Dark Side of LLMs' Intrinsic Self Correction. (2024). ACL Anthology .
19. Can LLMs Learn from Previous Mistakes? Investigating LLMs' Errors to Boost for Reasoning. (2024). ACL Anthology .
20. When Does Verification Pay Off? A Closer Look at LLMs as Solution Verifiers. (2025). arXiv preprint .
International Research Journal of Multidisciplinary Sciences (Double - Blind Peer Review Journal)
VOL-2 ISSUE-4 April 2026 PP:27-33 ISSN:3107-930X(ONLINE)