Synergistic Self-Correction: A Hierarchical Framework for Multi-Stage Reasoning and Error Recovery in Large Language Models

Pratham Patel[1]
Abhishek Jindal[1*]

[1]Department of Computer Science
Dhirubhai Ambani Institute of Information and Communication Technology
Gandhinagar, Gujarat, India
{prathambiren2618, abhishek_jindal}@daiict.ac.in
*Corresponding author

September 25, 2025

Abstract

Large Language Models (LLMs) have achieved remarkable success across diverse natural language processing tasks, yet they exhibit systematic failures in complex multi-step reasoning, particularly in mathematical domains where logical consistency and error recovery are paramount. The fundamental limitation stems from the autoregressive generation paradigm, where early reasoning errors propagate through subsequent steps, rendering final answers incorrect regardless of the overall approach validity. Existing solutions—external verification systems, ensemble methods, and process supervision—either require substantial computational overhead, fail to improve underlying model capabilities, or lack the sophistication needed for nuanced error identification and correction.

We introduce **S2C (Synergistic Self-Correction)**, a novel hierarchical framework that endows LLMs with metacognitive reasoning capabilities through a structured three-stage inference process. Our approach decomposes problem-solving into distinct computational personas: a **Generator** that produces initial solutions with explicit critical point identification, a **Critic** that systematically analyzes potential errors and logical inconsistencies, and a **Synthesizer** that integrates feedback to produce refined solutions. This decomposition enables targeted optimization of each reasoning stage while maintaining end-to-end differentiability.

Our training methodology, **CDT (Cognitive Dissonance Training)**, combines supervised fine-tuning on high-quality reasoning traces with reinforcement learning using a novel **HPBR (Hierarchical Process-Based Reward)** system. We introduce specialized reward models for critique quality evaluation and correction effectiveness assessment that provide fine-grained process supervision beyond traditional outcome-based metrics.

Comprehensive evaluation across multiple reasoning benchmarks demonstrates substantial improvements: S2C achieves 49.9% accuracy on GSM8K (60% relative improvement over 31.2% baseline), 21.3% on MATH (71% relative improvement), and consistent gains on commonsense reasoning tasks. Statistical significance testing confirms these improvements (p ¡ 0.001), with detailed error analysis revealing high success rates in correcting computational errors (78%) and missing reasoning steps (71%). Our work establishes a new paradigm for developing self-correcting AI systems with intrinsic metacognitive capabilities.

1 Introduction

The rapid advancement of Large Language Models (LLMs) has revolutionized natural language processing, demonstrating remarkable capabilities across diverse tasks from language generation to complex reasoning. However, despite these achievements, LLMs continue to exhibit systematic failures in multi-step reasoning tasks that require logical consistency, er-

ror detection, and self-correction capabilities. This limitation is particularly pronounced in mathematical reasoning domains, where a single computational or logical error can cascade through the entire solution process, rendering the final answer incorrect regardless of the soundness of the overall approach.

The fundamental challenge lies in the autoregressive nature of LLM generation: once an error is introduced into a reasoning chain, the model lacks intrinsic mechanisms to recognize, critique, and correct it. This represents a critical gap in current LLM architectures, which, while proficient at pattern matching and surface-level language understanding, lack the metacognitive capabilities necessary for systematic self-reflection and iterative improvement.

1.1 Motivation and Problem Statement

Current approaches to addressing reasoning failures in LLMs fall into three primary categories, each with significant limitations:

External Verification Systems rely on separate models or rule-based systems to validate LLM outputs, requiring additional computational resources and model maintenance overhead while failing to improve the underlying model's reasoning capabilities.

Ensemble Methods sample multiple solutions and select answers based on consistency or voting mechanisms. While sometimes effective, these approaches are computationally expensive and do not address the fundamental reasoning deficiencies of the base model.

Process Supervision approaches train separate verifier models to evaluate intermediate reasoning steps. However, these methods typically focus on binary correctness judgments rather than providing constructive feedback for improvement.

The core limitation across all these approaches is their failure to endow LLMs with intrinsic self-correction capabilities. They treat reasoning errors as external problems to be detected and filtered rather than internal cognitive processes to be improved through metacognitive skill development.

1.2 Our Contributions

This work introduces **Synergistic Self-Correction (S2C)**, a novel framework that addresses these limitations through the following key contributions:

1. **Hierarchical Reasoning Framework**: We formalize self-correction as a structured three-stage inference process with distinct computational personas (Generator, Critic, Synthesizer), each optimized for specific cognitive functions.

2. **Cognitive Dissonance Training**: We introduce a novel three-phase training methodology that combines supervised fine-tuning with process-based reinforcement learning using specialized reward models.

3. **Hierarchical Process-Based Rewards**: We develop reward models that evaluate critique quality and correction effectiveness, providing fine-grained process supervision beyond traditional outcome-based metrics.

4. **Comprehensive Evaluation**: We demonstrate significant improvements across multiple reasoning benchmarks with detailed analysis of correction patterns, computational efficiency, and statistical significance.

5. **Metacognitive Capabilities**: Our approach successfully teaches LLMs to develop intrinsic self-correction abilities without requiring external verification systems.

2 Related Work

2.1 Reasoning Enhancement in Large Language Models

The field of LLM reasoning enhancement has evolved through several paradigms, each contributing important insights while revealing fundamental limitations.

Chain-of-Thought (CoT) Prompting demonstrated that explicitly eliciting intermediate reasoning steps significantly improves performance on complex reasoning tasks. This foundational work established the importance of making reasoning processes transparent and sequential, providing the conceptual foundation for structured reasoning approaches.

Self-Consistency introduced ensemble approaches that sample multiple reasoning paths and select the most frequent answer. While effective in many scenarios, this method requires substantial computational overhead and doesn't improve the underlying model's reasoning capabilities—it merely filters outputs post-hoc.

Tree-of-Thoughts (ToT) extended reasoning paradigms by exploring solution spaces as tree structures, enabling backtracking and exploration of alternative solution paths. However, ToT requires manual specification of decomposition strategies and

doesn't scale effectively to arbitrary problem domains.

2.2 Self-Improvement and Iterative Refinement

Recent research has focused on training models to improve their own outputs through various self-improvement mechanisms.

Constitutional AI trains models to critique and revise their responses according to a set of principles or constitutions. While promising for alignment and safety applications, this approach relies on high-level behavioral guidelines rather than domain-specific reasoning skills.

Self-Taught Reasoner (STaR) employs iterative training loops where models generate rationales and fine-tune on successful examples. However, STaR only leverages correct solutions, missing opportunities to learn from errors and corrections—a critical limitation for developing error recovery capabilities.

Self-Refine enables models to iteratively improve their outputs based on self-generated feedback. While conceptually similar to our approach, Self-Refine lacks the structured decomposition and specialized training methodology that enables effective error identification and correction.

2.3 Process Supervision and Reward Modeling

Process supervision approaches have shown promise by providing more granular feedback than outcome-based methods.

Process Reward Models (PRMs) train separate models to evaluate reasoning steps, providing step-level feedback rather than just final answer correctness. However, traditional PRMs focus on binary correctness judgments rather than constructive feedback for improvement.

Reinforcement Learning from Human Feedback (RLHF) has demonstrated significant improvements in language model capabilities through human preference optimization. Our work extends these concepts by introducing specialized reward structures for metacognitive skill development.

Our approach builds on these foundations while addressing key limitations: we integrate process supervision directly into the generation model, provide constructive rather than binary feedback, and focus specifically on developing intrinsic self-correction capabilities.

3 Methodology

3.1 The Synergistic Self-Correction Framework

The S2C framework decomposes the reasoning process into three distinct computational stages, each optimized for specific cognitive functions and trained to work synergistically.

Stage 1: Generator receives an input problem and produces an initial solution attempt along with a set of Critical Points—key logical steps, assumptions, or calculations that are essential to the solution's validity. This stage is optimized for comprehensive problem analysis and solution generation.

Stage 2: Critic receives the complete context (problem, initial solution, and Critical Points) and generates a systematic evaluation of each critical point. The Critic is trained with an adversarial objective to identify potential errors, logical inconsistencies, or computational mistakes, providing specific and actionable feedback.

Stage 3: Synthesizer integrates all available information to produce a refined final solution. This stage is designed to address issues identified in the critique while preserving correct aspects of the original solution, demonstrating sophisticated error correction capabilities.

3.2 Cognitive Dissonance Training

We introduce a three-phase training methodology that progressively develops the model's self-correction capabilities:

Phase 1: Structural Alignment via Supervised Fine-Tuning establishes the structural foundation by training on high-quality examples of the complete S2C pipeline. We create datasets containing complete reasoning traces where solutions are generated by powerful teacher models and validated for correctness.

Phase 2: Specialized Reward Model Training develops two specialized reward models: an Insight Reward Model that evaluates critique quality by scoring error identification accuracy and feedback specificity, and a Correction Reward Model that assesses how effectively the Synthesizer addresses issues raised in the critique.

Phase 3: Hierarchical Process-Based Reward Optimization uses Proximal Policy Optimization (PPO) with a novel hierarchical reward structure that combines accuracy rewards, insight quality scores, correction effectiveness scores, and coherence penalties. This provides fine-grained opti-

mization signals for each component of the reasoning process.

4 Experimental Results

4.1 Main Results

We evaluate S2C across multiple reasoning benchmarks, demonstrating consistent improvements over strong baselines.

Mathematical Reasoning: On GSM8K, S2C achieves 49.9% accuracy compared to 31.2% for standard Chain-of-Thought prompting, representing a 60% relative improvement. On the more challenging MATH dataset, we achieve 21.3% accuracy (71% relative improvement over baseline).

Multi-hop Reasoning: On StrategyQA, S2C achieves 76.4% accuracy (11% relative improvement), demonstrating effectiveness beyond purely mathematical domains.

Commonsense Reasoning: On CommonsenseQA, we achieve 78.1% accuracy (8% relative improvement), showing that structured self-correction benefits extend to diverse reasoning tasks.

Statistical significance testing using McNemar's test confirms that all improvements are statistically significant (p ¡ 0.001).

4.2 Ablation Studies

Comprehensive ablation studies reveal the contribution of each framework component:

- Supervised Fine-Tuning alone provides 6.6 percentage point improvement over the base model
- Process-based rewards contribute an additional 7.5 percentage points over outcome-only optimization
- The three-stage architecture is crucial: removing the Critic stage results in 10.6 percentage point performance drop
- Critical Points contribute 5.2 percentage points to final performance

4.3 Error Analysis

Detailed analysis reveals S2C's effectiveness across different error types:

- **Computational Errors** (35% of errors): 78% correction success rate
- **Logical Errors** (28% of errors): 65% correction success rate
- **Missing Steps** (22% of errors): 71% correction success rate
- **Conceptual Errors** (15% of errors): 42% correction success rate

The framework demonstrates highest effectiveness in correcting computational errors and missing reasoning steps, while conceptual errors remain more challenging to address.

4.4 Computational Efficiency

Despite the multi-stage process, S2C achieves superior efficiency compared to ensemble methods:

- S2C uses 641 average tokens vs 2,470 for Self-Consistency
- Inference time: 3.8 seconds vs 12.1 seconds for Self-Consistency
- Accuracy improvement: 29% higher than Self-Consistency while using 74% fewer resources

5 Discussion

5.1 Theoretical Implications

Our results provide evidence for several important insights about LLM reasoning capabilities:

Metacognitive Development: The success of S2C demonstrates that LLMs can develop sophisticated metacognitive skills when provided with appropriate training signals and structured frameworks.

Process vs. Outcome Supervision: The superior performance of process-based rewards over outcome-only training validates educational psychology research showing that process-focused feedback leads to better learning outcomes.

Structured Reasoning Decomposition: The three-stage architecture provides a principled framework for decomposing complex reasoning tasks, potentially applicable beyond mathematical domains.

5.2 Limitations and Future Work

Several limitations suggest directions for future research:

Domain Specificity: While S2C shows strong performance on mathematical reasoning, generalization to other specialized domains requires further investigation.

Conceptual Error Correction: Our analysis reveals that conceptual errors remain challenging to correct, suggesting need for specialized techniques targeting fundamental misunderstandings.

Scalability: Experiments focus on 8B parameter models; investigating scaling behavior with larger models could provide insights into the relationship between model capacity and self-correction capabilities.

Real-time Applications: Current implementation requires multiple inference passes; developing more efficient architectures for real-time applications represents an important engineering challenge.

6 Conclusion

We have introduced Synergistic Self-Correction (S2C), a novel framework that enables Large Language Models to perform structured self-critique and iterative refinement through metacognitive skill development. Our approach addresses fundamental limitations in current LLM reasoning capabilities by teaching models to identify and correct their own errors through a principled three-stage process.

Key contributions include: (1) a theoretically grounded framework for multi-stage reasoning in LLMs, (2) a novel training methodology combining supervised learning with process-based reinforcement learning, (3) specialized reward models for evaluating critique quality and correction effectiveness, and (4) comprehensive experimental validation demonstrating significant improvements over existing methods.

Our results on mathematical reasoning benchmarks show that S2C achieves 60% relative improvement over standard approaches while maintaining computational efficiency. This work establishes a new paradigm for developing self-correcting AI systems with intrinsic metacognitive capabilities, representing a significant step toward more reliable and trustworthy artificial intelligence.

Future research directions include extending the framework to broader domains, improving correction of conceptual errors, and developing more efficient architectures for real-time deployment. The fundamental insight—that LLMs can learn to systematically critique and improve their own reasoning—opens new avenues for developing more sophisticated and reliable AI systems.

Acknowledgments

The authors thank the anonymous reviewers for their valuable feedback. This work was supported by the Student Research Initiative (SRI) program at DA-IICT. We acknowledge the computational resources provided by the institute's High Performance Computing facility.

References

[1] J. Wei, Y. Tay, R. Bommasani, et al., "Emergent abilities of large language models," *Transactions on Machine Learning Research*, 2022.

[2] K. Cobbe, V. Kosaraju, M. Bavarian, et al., "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168*, 2021.

[3] X. Wang, J. Wei, D. Schuurmans, et al., "Self-consistency improves chain of thought reasoning in language models," *International Conference on Learning Representations*, 2022.

[4] J. Wei, X. Wang, D. Schuurmans, et al., "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24824-24837, 2022.

[5] S. Yao, D. Yu, J. Zhao, et al., "Tree of thoughts: Deliberate problem solving with large language models," *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[6] Y. Bai, A. Jones, K. Ndousse, et al., "Constitutional AI: Harmlessness from AI feedback," *arXiv preprint arXiv:2212.08073*, 2022.

[7] E. Zelikman, Y. Wu, J. Mu, et al., "STaR: Bootstrapping reasoning with reasoning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 15476-15488, 2022.

[8] L. Ouyang, J. Wu, X. Jiang, et al., "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27730-27744, 2022.

[9] H. Lightman, V. Kosaraju, Y. Burda, et al., "Let's verify step by step," *arXiv preprint arXiv:2305.20050*, 2023.

[10] D. Hendrycks, C. Burns, S. Kadavath, et al., "Measuring mathematical problem solving with the MATH dataset," *NeurIPS*, 2021.