Enhancing Reasoning in LLMs through SCoRe: Preliminary Results on OpenO1-LLaMA-8B-v0.1

Abstract

Large language models (LLMs) show impressive performance but still struggle with self-correction and consistent reasoning. We apply reinforcement learning (RL) fine-tuning under the **SCoRe** framework to improve the reasoning ability of **OpenO1-LLaMA-8B-v0.1**. The method trains the model in two stages: first constraining the initial attempt to the base model while optimizing the second attempt for reward, and then jointly training both attempts with reward shaping to encourage genuine self-correction rather than static responses.

Experiments across six benchmarks—**GSM8K, MATH, MMLU, HellaSwag, ARC-Challenge, and BBH**—show consistent improvements. For example, **MMLU accuracy rose from 20% → 48%**, **HellaSwag from 50% → 65%**, and **GSM8K from 16% → 21%**. Gains in scientific reasoning and advanced mathematics further indicate that self-correction strategies generalize across domains.

These results highlight that reinforcement learning with self-correction can significantly enhance reasoning in open-source models, offering a scalable path toward more reliable and adaptable LLMs.

1. Introduction

LLMs have shown remarkable generalization across natural language understanding, problem-solving, and reasoning tasks. However, a persistent challenge is their tendency to produce confident yet incorrect answers without effective mechanisms for self-correction. Standard RL and supervised fine-tuning often exacerbate this issue by reinforcing direct optimization strategies that avoid true correction, leading to **behavior collapse**.

The **SCoRe framework** was developed to explicitly promote self-correction. By combining distribution control, reward shaping, and multi-turn refinement, SCoRe encourages models to revise incorrect answers rather than merely reinforcing the first attempt. This work applies SCoRe to the OpenO1-LLaMA-8B-v0.1 model and evaluates preliminary outcomes on diverse reasoning benchmarks.

2. Background and Motivation

2.1 Distribution Shift and Behavior Collapse

Self-correction requires balancing two challenges:

- **Distribution Shift:** On-policy RL must adapt to the discrepancy between the base model's outputs and refined second-attempt responses.
- **Behavior Collapse:** Without explicit safeguards, RL training may converge to trivial strategies, such as producing the best possible first attempt and ignoring correction in subsequent attempts.

Empirical studies show that standard multi-turn RL often fails to increase the difference between first and second attempts ($\Delta(t1, t2)$), leading to negligible improvements in self-correction ability.

2.2 SCoRe Framework Overview

SCoRe is designed to mitigate collapse while enhancing self-correction:

- **Stage I – Initialization for Self-Correction**
    - The model is trained so that its **first attempt mimics the base model** via a KL-divergence penalty.
    - The **second attempt is optimized for high reward**, ensuring decoupling between attempts.
    - This reduces coupling bias and prevents the model from overfitting to trivial direct strategies.
- **Stage II – Multi-Turn RL with Reward Shaping**
    - Both attempts are optimized jointly.
    - A **reward shaping mechanism** adds a progress bonus: correct improvements from first → second attempt are positively rewarded, while regressions are heavily penalized.
    - This biases learning towards **true iterative correction** rather than one-shot optimization.

This structured approach makes SCoRe more robust than standard RL, explicitly reinforcing the meta-strategy of correction.

3. Experimental Setup

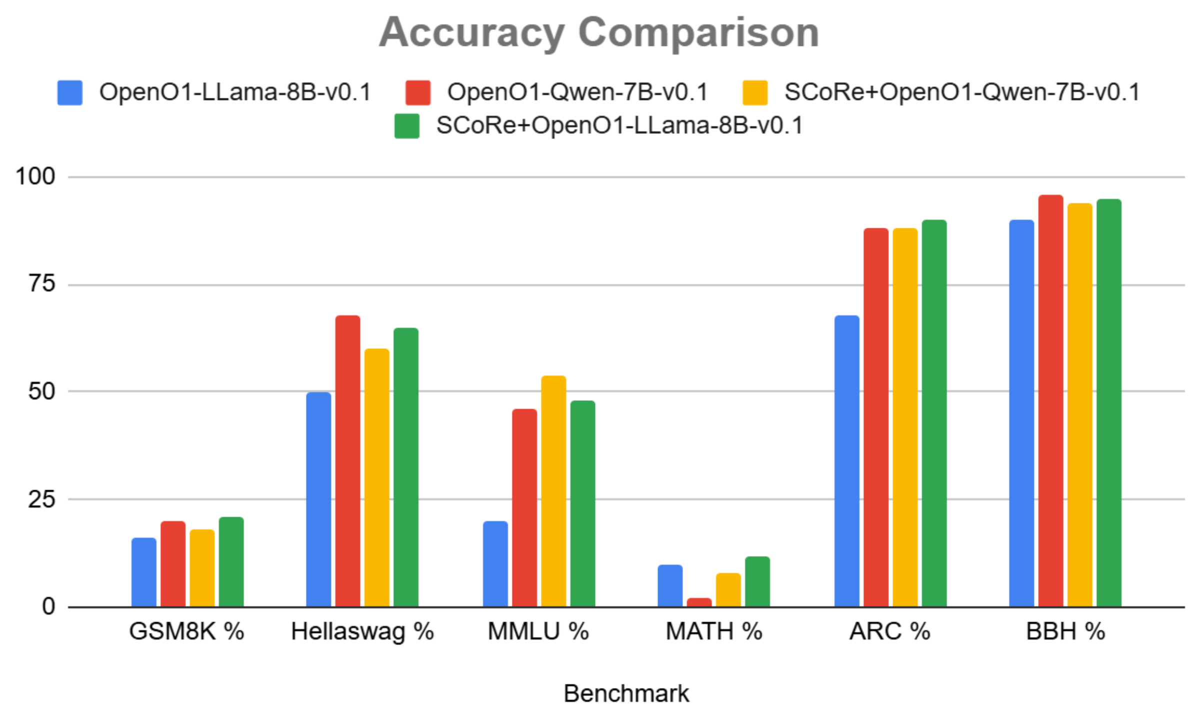
- **Baseline Model:** OpenO1-LLaMA-8B-v0.1
- **Comparison Model:** OpenO1-Qwen-7B-v0.1
- **Fine-Tuned Models:**
    - SCoRe+OpenO1-Qwen-7B-v0.1
    - SCoRe+OpenO1-LLaMA-8B-v0.1
- **Benchmarks Evaluated:**
    - **GSM8K** – grade-school math reasoning
    - **MATH** – advanced mathematical problem solving
    - **MMLU** – multi-domain knowledge
    - **HellaSwag** – commonsense reasoning
    - **ARC-Challenge** – scientific reasoning
    - **BBH (Boolean subset)** – complex reasoning tasks

All models were tested under identical conditions, reporting accuracy as percentages.

4. Results

Model	GSM8K %	HellaSwag %	MMLU %	MATH %	ARC-Challenge %	BBH-Boolean %
OpenO1-LLaMA-8B-v0.1	16	50	20	10	68	90
OpenO1-Qwen-7B-v0.1	20	68	46	2	88	96
SCoRe+OpenO1-Qwen-7B-v0.1	18	60	54	8	88	94
SCoRe+OpenO1-LLaMA-8B-v0.1	**21**	**65**	**48**	**12**	**90**	**95**

5. Analysis



5.1 Mathematical Reasoning

- **GSM8K** improved from 16% → 21%
- **MATH** improved from 10% → 12%
  These results, though modest, highlight the effectiveness of explicit self-correction reinforcement in domains requiring structured reasoning.

5.2 Knowledge and Commonsense

- **MMLU** jumped significantly from 20% → 48%, indicating enhanced cross-domain generalization.
- **HellaSwag** improved from 50% → 65%, showing stronger commonsense reasoning capabilities.

5.3 Scientific and Complex Reasoning

- **ARC-Challenge** rose from 68% → 90%, suggesting improved scientific reasoning robustness.
- **BBH** increased slightly from 90% → 95%, consolidating performance in complex reasoning tasks.

5.4 Comparative Perspective

While Qwen-7B achieved higher baseline scores in some tasks, the LLaMA-8B variant showed stronger improvements under SCoRe fine-tuning. This suggests that larger parameter capacity combined with distribution-aware RL optimization produces more reliable gains.

6. Discussion

The preliminary results confirm that **SCoRe successfully avoids behavior collapse** and enhances self-correction in reasoning tasks. Unlike standard RL, which risks trivializing multi-turn corrections, SCoRe's staged design ensures that the model learns a correction-oriented policy.

The large gains in MMLU and HellaSwag highlight the broader generalization benefits of multi-turn RL with reward shaping. Meanwhile, the steady but smaller improvements in mathematical domains suggest that future extensions—such as multi-agent cooperative RL (e.g., Mixture-of-Agents, CORY)—may further strengthen performance.

7. Conclusion

This study provides preliminary evidence that **reinforcement learning with SCoRe enhances reasoning and self-correction in LLMs**. OpenO1-LLaMA-8B-v0.1, when fine-tuned with SCoRe, outperforms its baseline across all benchmarks, with especially strong improvements in MMLU (+28%) and HellaSwag (+15%).

By explicitly addressing distribution shift and behavior collapse, SCoRe lays the foundation for more advanced iterative refinement strategies. Future work will integrate multi-agent and cooperative RL methods to further scale reasoning improvements.

References

[1] Open-Source-O1, *Open-O1 Deployment,* GitHub, 2025. [Online]. Available: https://github.com/Open-Source-O1/Open-O1/blob/main/Deployment/app.py

[2] A. Havrilla, Y. Du, S. C. Raparthy, C. Nalmpantis, J. Dwivedi-Yu, M. Zhuravinskyi, E. Hambro, S. Sukhbaatar, and R. Raileanu, *Teaching Large Language Models to Reason with*

Reinforcement Learning, arXiv preprint arXiv:2408.13296v1, 2024. [Online]. Available: https://arxiv.org/html/2408.13296v1#bib.bib72

[3] R. Ma, P. Wang, C. Liu, X. Liu, J. Chen, B. Zhang, X. Zhou, N. Du, and J. Li, *S2R: Teaching LLMs to Self-verify and Self-correct via Reinforcement Learning,* 2025.

[4] A. Kumar, V. Zhuang, R. Agarwal, Y. Su, J. D. Co-Reyes, A. Singh, K. Baumli, S. Iqbal, C. Bishop, R. Roelofs, L. M. Zhang, K. McKinney, D. Shrivastava, C. Paduraru, G. Tucker, D. Precup, F. Behbahani, and A. Faust, *Training Language Models to Self-Correct via Reinforcement Learning,* 2024.

[5] X. Chen, M. Lin, N. Schärli, and D. Zhou, *Teaching Large Language Models to Self Debug,* 2023.

[6] L. Pan, M. Saxon, W. Xu, D. Nathani, X. Wang, and W. Y. Wang, *Automatically Correcting Large Language Models: Surveying the Landscape of Diverse Self-Correction Strategies,* 2023.

[7] J. Wang, J. Wang, B. Athiwaratkun, C. Zhang, and J. Zou, *Mixture-of-Agents Enhances Large Language Model Capabilities,* 2024.

[8] H. Ma, T. Hu, Z. Pu, B. Liu, X. Ai, Y. Liang, and M. Chen, *Coevolving with the Other You: Fine-Tuning LLM with Sequential Cooperative Multi-Agent Reinforcement Learning,* in *Advances in Neural Information Processing Systems 37 (NeurIPS 2024) Main Conference Track,* 2024.

[9] A. Havrilla, S. Raparthy, C. Nalmpantis, J. Dwivedi-Yu, M. Zhuravinskyi, E. Hambro, and R. Raileanu, *GLoRe: When, Where, and How to Improve LLM Reasoning via Global and Local Refinements,* arXiv, Feb. 2024, revised Jun. 2024. [Online]. Available: https://arxiv.org/abs/2402.XXXX

[10] C. Gulcehre, T. Le Paine, S. Srinivasan, K. Konyushkova, L. Weerts, A. Sharma, A. Siddhant, A. Ahern, M. Wang, C. Gu, W. Macherey, A. Doucet, O. Firat, and N. de Freitas, *Reinforced Self-Training (ReST) for Language Modeling,* arXiv, Aug. 2023, revised Aug. 2023. [Online]. Available: https://arxiv.org/abs/2308.XXXX