CS234 Final Project Proposal, Winter 2025
`r1-arc-agi`: Saturating a Single Benchmark with Small Reasoning Models

Anonymous Authors[1]

Abstract

We propose to investigate whether the distilled DeepSeek-r1 model series can be fine-tuned with RL, free of a learned reward model, to deliberately achieve ARC-AGI-1 performance par with larger models (r1/o1). Success on this narrow task, which involves recognizing visuals patterns in nxm grid boxes, would suggest that even Semi-Private evaluations on any single domain-specific task is an insufficient test for AGI.

1. Introduction

This project examines if a distilled r1 model can match r1 on ARC-AGI-1 performance (or substantially improve base performance) by deliberately constructing a small, handcrafted dataset of reasoning tasks similar to ARC-AGI-1 (Chollet et al., 2024; DeepSeek, 2025). If a small model saturates performance on this single domain, it raises concerns regarding the validity of such isolated evaluations in testing AGI, implying small models well-optimized for domain-specific reasoning tasks may be misleading, and frontier models should be tested across an ensemble of reasoning tasks from very different domains.

2. Data

Our dataset begins with 400 public ARC-AGI (question, answer) tasks focused on visual pattern recognition in nxm grid boxes. We will augment this seed set by using a larger model (o3 or r1) to generate an additional 10–100$\times$ diverse QA pairs through permutations and repeated sampling, as described by the curriculum generation of phi-series papers. Each generated pair, along with its reasoning trace, will be verified by three human labellers; any disagreement will result in the pair's rejection, ensuring a correct reward.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

3. Method

To post-train a r1-distilled 7B model of Llama architecture, we will apply GRPO—a variant of PPO—with a rule-based reward that is positive when the model's answer matches the human/o3 reference, and zero otherwise (Schulman et al., 2017; Shao et al., 2024). If this reward signal is too sparse for policy model convergence, alternatively we can try a more continuous reward of percentage boxes correct in the visual grid of the ARC-AGI-1 task. We will also attempt s1's budget forcing method of test-time scaling: if the model's internal "thinking" sequence terminates before reaching a desired token count n, the premature "`</think>`" token is replaced with "Wait," and if it exceeds n, the sequence is truncated and ended with "`</think>` (Muennighoff et al., 2025)." This may amplify the DSL effect observed in R1-Zero: where language mixing and uninterpretable thinking token could be the policy optimizing its thinking for a narrow domain task.

4. Literature Review

The phi-series and recent s1 papers will inform our dataset curation, while R1, R1-Zero, and v3 are the best open-source reasoning models, particularly the rule-based reward of R1-Zero (Microsoft, 2024). DeepSeek-Math and Open-R1 offer more details to replicate the reasoning process, ARC-AGI's leaderboard papers will inform task-specific methods. GRPO and PPO learns our policy.

5. Evaluation

We will attempt to request the ARC foundation to evaluate on ARC-AGI-1 Semi-Private, or Public otherwise. We expect our smaller distilled model to compare with R1's metrics only on ARC-AGI-1, and to be likely much worse on all other benchmarks. Qualitatively, metrics should scale with generated data size and thinking token count. However, the possible failure modes are: the Semi-Private score is much lower due to overfitting; if so, our hypothesis that a small reasoning model can generalize to one well-defined task is false. It is also possible that the reward signal is too sparse to converge, or improvement is too slow at small model sizes.

References

Chollet, F., Knoop, M., Kamradt, G., and Landers, B. Arc prize 2024: Technical report. *arXiv preprint arXiv:2412.04604*, 2024.

DeepSeek, D. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Microsoft, M. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Muennighoff, N., Yang, Z., Shi, W., Li, X. L., Fei-Fei, L., Hajishirzi, H., Zettlemoyer, L., Liang, P., Candès, E., and Hashimoto, T. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y. K., Wu, Y., and Guo, D. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.