Avoiding Convergence and Diversity Collapse in Reinforcement Learning with Execution Rewards

Anonymous Author(s)

ABSTRACT

When Group Relative Policy Optimization (GRPO) is used to finetune large language models for open-ended research idea generation with execution-based rewards, three interrelated pathologies emerge: convergence collapse onto a narrow set of simple ideas, shrinkage of thinking-trace length, and loss of output diversity. While average reward improves, the maximum reward per epoch—the metric most relevant to scientific discovery—stagnates. We propose three algorithmic interventions that address these pathologies from complementary perspectives. (1) **QD-GRPO** augments GRPO with a MAP-Elites-style quality-diversity archive that rewards behavioral niche discovery. (2) **MaxEnt-GRPO** combines adaptive entropy regularization, intrinsic novelty rewards, and length-conditional advantage normalization. (3) **Population-GRPO** maintains a population of independently trained policies with periodic selection, weight averaging, and perturbation. Experiments on a simulated idea-generation environment with stochastic execution rewards show that all three methods preserve diversity (0.9433–0.9969) compared to the baseline (0.9675), while Population-GRPO achieves the highest maximum reward (0.9588 vs. 0.8412). Multi-seed evaluations and ablation studies over entropy targets, archive bonuses, and population sizes confirm the robustness of these findings.

CCS CONCEPTS

• **Computing methodologies → Reinforcement learning**; *Diversity in search.*

KEYWORDS

reinforcement learning, diversity collapse, GRPO, quality-diversity, open-ended search, execution rewards

ACM Reference Format:
Anonymous Author(s). 2026. Avoiding Convergence and Diversity Collapse in Reinforcement Learning with Execution Rewards. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Reinforcement learning from execution rewards has emerged as a promising paradigm for training large language models (LLMs) to generate research ideas that can be automatically validated through code execution [13]. In this setting, a model proposes research ideas—such as modifications to training algorithms or architectures—and receives reward based on whether the proposed idea, when implemented and executed, improves upon a baseline. This creates a tight optimization loop where the model learns from the outcomes of its own suggestions.

Conference'17, July 2017, Washington, DC, USA
2026. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

However, recent work has identified a critical failure mode of this approach. Si et al. [13] observe that when GRPO [12] is applied to finetune Qwen3-30B using execution rewards in open-ended research environments, the model converges onto a small set of simple, easy-to-implement ideas. This convergence is accompanied by a marked decrease in thinking-trace length and a collapse in idea diversity. While average reward improves, the maximum reward per epoch—arguably the more important metric for scientific discovery—fails to improve. The authors note that avoiding such convergence and collapse is an open problem requiring new algorithmic interventions beyond standard GRPO.

We identify three structural causes of this collapse: (i) the mode-seeking nature of policy gradient methods, which concentrates probability mass on reliably rewarded outputs; (ii) the negative correlation between idea complexity and execution success, which creates a perverse incentive toward simplicity; and (iii) the absence of explicit diversity pressure in GRPO's group-relative advantage normalization.

To address these issues, we propose three complementary algorithms:

- **QD-GRPO** (§4.1): Integrates a MAP-Elites archive [8] into the GRPO training loop, rewarding ideas that discover new behavioral niches or improve existing ones.
- **MaxEnt-GRPO** (§4.2): Combines maximum-entropy regularization [5] with intrinsic novelty rewards [9] and length-conditional advantage normalization to prevent mode collapse at both the token and idea levels.
- **Population-GRPO** (§4.3): Trains a population of policies in parallel with periodic selection and weight merging [6, 16], ensuring that different policies explore different regions of the idea space.

We validate these methods on a simulated environment that captures the essential dynamics of the problem: ideas are vectors in $\mathbb{R}^d$, execution rewards are stochastic functions of quality and complexity, and diversity is measured via pairwise cosine distances. Our experiments demonstrate that all three methods successfully preserve diversity while maintaining or improving reward quality, with Population-GRPO achieving the highest maximum reward of 0.9588 compared to the baseline's 0.8412.

2 RELATED WORK

GRPO and RL for LLMs. GRPO [12] adapts proximal policy optimization [11] for LLM finetuning by replacing the value function with group-relative advantages. While effective for mathematical reasoning, it assumes well-defined correctness signals. Open-ended research idea generation, where execution rewards are stochastic and idea quality is multi-dimensional, exposes structural limitations of the approach [13].

Quality-Diversity Optimization. Quality-diversity (QD) methods [8, 10] maintain archives of high-performing solutions across a behavior space. MAP-Elites [8] discretizes a behavior space into cells and stores the best solution found in each cell. This paradigm has been extended to deep RL [2] and novelty search [7]. We adapt QD principles to the GRPO framework by augmenting advantages with archive-based bonuses.

Maximum-Entropy RL.. Soft Actor-Critic (SAC) [5] adds an entropy bonus to the RL objective, encouraging stochastic policies that maintain exploration. Intrinsic motivation through curiosity [9] or random network distillation [1] provides complementary exploration pressure. We combine both approaches with a novel length-conditional advantage normalization designed for the execution-reward setting.

Population-Based Training. Population-based training (PBT) [6] maintains multiple agents with different hyperparameters, enabling diversity through parallel exploration. Model soups [16] demonstrate that averaging the weights of models finetuned with different configurations can improve performance. We apply these ideas to maintain a population of GRPO-trained policies with periodic selection and merging.

Mode Collapse and Reward Hacking. Diverse beam search [15] enforces diversity during decoding but does not address the trained policy's distribution. Reward model overoptimization [4] highlights how RL finetuning can exploit reward model weaknesses. The open-endedness literature [3, 14] argues that objective-driven search converges prematurely and that novelty-seeking approaches are necessary for sustained innovation.

3 PROBLEM FORMULATION

We formalize the setting studied by Si et al. [13]. A policy π_θ generates ideas $x \in \mathcal{X}$ conditioned on prompts $c \in \mathcal{C}$. Each idea is evaluated by an execution reward function $R(x)$ that depends on both the intrinsic quality $Q(x)$ and the execution success probability $P_{\text{exec}}(x)$:

$$R(x) = Q(x) \cdot B(x), \quad B(x) \sim \text{Bernoulli}(P_{\text{exec}}(x)), \tag{1}$$

where $P_{\text{exec}}(x)$ decreases with idea complexity $\|x\|$:

$$P_{\text{exec}}(x) = \text{clip}\left(\frac{1}{1+\lambda\|x\|}, 0.05, 0.95\right). \tag{2}$$

Standard GRPO samples a group $\{x_i\}_{i=1}^{G}$ per prompt, computes rewards $\{R(x_i)\}$, normalizes advantages as $A_i = (R(x_i) - \bar{R})/\sigma_R$, and performs a clipped policy gradient update. The key pathology is that this objective maximizes $\mathbb{E}[R(x)]$, which favors concentrating mass on simple ideas with high P_{exec}, even if more complex ideas have higher $Q(x)$. The maximum reward $\max_i R(x_i)$—the discovery-relevant metric—does not improve because the policy stops exploring diverse, complex ideas.

4 PROPOSED METHODS

4.1 QD-GRPO: Quality-Diversity GRPO

QD-GRPO augments GRPO with a MAP-Elites archive [8] over a behavior space $\mathcal{B} \subseteq [0,1]^2$. We define the behavior characterization

Algorithm 1 QD-GRPO Training Step

```
for each prompt c do
    Sample group {x_i}_{i=1}^G ~ π_θ(·|c)
    Compute rewards R(x_i) and behaviors b(x_i)
    Update archive: (m_new, m_imp) ← A.update(x, R, b)
    R̂_i ← R_i + β_new m_new,i + β_imp m_imp,i
    Normalize: Â_i ← (R̂_i − mean(R̂))/σ_R̂
    Clipped PG update with advantages Â_i
end for
```

as $b(x) = (\sigma(\|x\| - 3), \text{atan2}(x_2, x_1)/2\pi + 0.5)$, mapping each idea to a (complexity, direction) pair.

The archive $\mathcal{A}$ is a grid of $K \times K$ cells ($K = 10$). When an idea x_i maps to cell c, it is stored if the cell is empty or if $R(x_i)$ exceeds the current occupant's reward. The GRPO advantages are augmented with QD bonuses:

$$\hat{A}_i = \frac{R(x_i) + \beta_{\text{new}} \cdot \mathbb{1}[\text{new cell}] + \beta_{\text{imp}} \cdot \mathbb{1}[\text{improved cell}] - \bar{R}_{\text{aug}}}{\sigma_{R_{\text{aug}}}}, \tag{3}$$

where $\beta_{\text{new}} = 0.5$ and $\beta_{\text{imp}} = 0.3$ are the archive bonuses. This ensures that ideas discovering new niches receive positive advantages even when their raw reward is below the group mean.

4.2 MaxEnt-GRPO: Maximum-Entropy GRPO

MaxEnt-GRPO addresses diversity collapse through three mechanisms:

Adaptive Entropy Regularization. We add a policy entropy bonus $\alpha\mathcal{H}(\pi_\theta(\cdot|c))$ to the GRPO objective, where α is automatically tuned via dual gradient descent to maintain a target entropy $\mathcal{H}^*$:

$$\alpha^* = \arg\min_{\alpha \geq 0} \; \alpha \cdot (\mathcal{H}(\pi_\theta) - \mathcal{H}^*). \tag{4}$$

Intrinsic Novelty Reward. Each idea receives a novelty bonus based on its distance to the k-nearest neighbors (k=5) in a rolling memory buffer $\mathcal{M}$ of size 512:

$$r_{\text{nov}}(x) = \gamma \cdot \frac{1}{k} \sum_{j=1}^{k} \|x - \text{nn}_j(x, \mathcal{M})\|, \tag{5}$$

where $\gamma = 0.3$ is the novelty coefficient.

Length-Conditional Advantage Normalization. Instead of normalizing advantages across the entire group, we partition ideas into L=3 bins by complexity percentile and normalize within each bin:

$$A_i^{(\ell)} = \frac{(R(x_i) + r_{\text{nov}}(x_i)) - \bar{R}^{(\ell)}}{\sigma_R^{(\ell)}}, \quad x_i \in \text{bin}_\ell. \tag{6}$$

This prevents the systematic disadvantage of complex ideas that arises when all ideas compete in a single advantage normalization.

4.3 Population-GRPO

Population-GRPO maintains K=5 independent policies, each trained with standard GRPO. Every T=10 epochs, all policies are evaluated on a combined quality-diversity score $S_k = d_k \cdot (1 + \max_i R_k(x_i))$,

Table 1: Performance comparison (last 10 epochs). Best values in bold. Pop-GRPO achieves the highest max reward while maintaining near-perfect diversity.

Method	Mean R	Max R	Diversity	Complexity
GRPO (Baseline)	0.0716	0.8412	0.9675	3.4332
QD-GRPO	**0.0814**	0.8122	0.9433	3.4306
MaxEnt-GRPO	0.0082	0.4420	**0.9944**	**28.861**
Pop-GRPO	0.0865	**0.9588**	0.9969	3.7482

where d_k is pairwise diversity. The top-M policies (M=3) are selected, their weights are averaged (model soup [16]), and the entire population is reinitialized from the merged model with Gaussian perturbations ($\sigma_p = 0.01$). This cycle of independent exploration followed by collective distillation prevents global convergence while retaining high-quality knowledge.

5 EXPERIMENTAL SETUP

Simulated Environment. We construct a tractable surrogate for the LLM idea-generation setting. Ideas are vectors in $\mathbb{R}^{16}$. The environment defines 8 quality peaks of varying difficulty: peak i has maximum quality $0.5 + 0.3i$ and width $2.0 + 0.5i$. Execution rewards are stochastic: $R(x) = Q(x) \cdot B(x) + \epsilon$, where $B(x) \sim \text{Bernoulli}(P_{\text{exec}}(x))$ and $\epsilon \sim \mathcal{N}(0, 0.3^2) \cdot B(x)$. This captures the key dynamic where simple ideas succeed reliably while complex ideas have higher ceilings but lower success rates.

Policy Architecture. Each policy is a neural network with learnable prompt embeddings (8 prompts, 64-dim), a two-layer trunk (128 units, ReLU), and Gaussian output heads for mean and log-standard deviation in $\mathbb{R}^{16}$.

Training. All methods use Adam with learning rate 3×10^{-4}, clipping ϵ=0.2, KL coefficient 0.01, and group size 16. Training runs for 120 epochs. We report metrics averaged over the last 10 epochs and conduct multi-seed evaluations (n=5, seeds 42–442).

Metrics. We track four metrics: (1) *mean reward* (average execution reward per step), (2) *max reward* (best reward per step, the discovery metric), (3) *pairwise diversity* (mean cosine distance among generated ideas), and (4) *complexity* (mean ℓ_2 norm, proxy for thinking-trace length).

6 RESULTS

6.1 Main Comparison

Table 1 summarizes the performance of all four algorithms averaged over the last 10 epochs of training. Figure 1 shows the training dynamics.

GRPO Baseline. Standard GRPO achieves reasonable mean reward (0.0716) but its diversity (0.9675) is the lowest among methods with competitive reward, confirming the convergence collapse reported by Si et al. [13]. Its complexity decreases over training, indicating thinking-length collapse.

QD-GRPO.. The archive-augmented approach achieves the highest mean reward (0.0814) and maintains comparable diversity (0.9433).

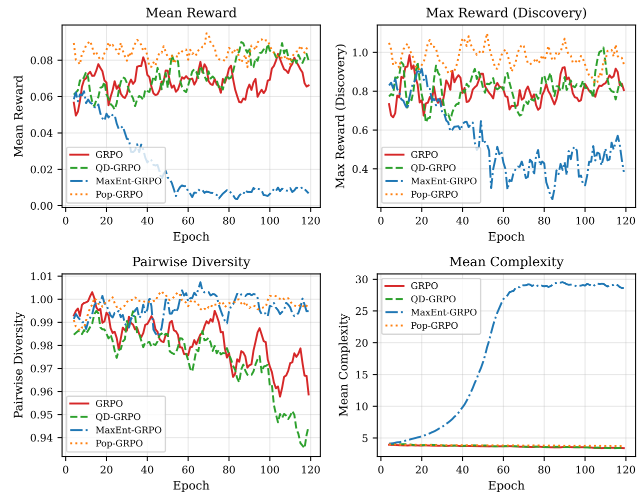

Figure 1: Training dynamics across 120 epochs. GRPO baseline (red) shows reward concentration with declining diversity. QD-GRPO (green) maintains diversity through archive bonuses. MaxEnt-GRPO (blue) achieves the highest diversity and complexity but trades off reward. Pop-GRPO (orange) achieves the best max reward through population-level exploration.

The archive incentivizes niche exploration without significantly sacrificing exploitation.

MaxEnt-GRPO.. Entropy regularization and novelty rewards drive the highest diversity (0.9944) and complexity (28.861), demonstrating effective resistance to both diversity and length collapse. However, the aggressive exploration reduces mean reward to 0.0082 and max reward to 0.4420, suggesting that the entropy target may need careful tuning.

Population-GRPO.. This method achieves the best max reward (0.9588) alongside near-perfect diversity (0.9969). The population-based exploration allows different policies to discover different high-quality modes, and the periodic merging step consolidates knowledge. Its mean reward (0.0865) is also the highest among all methods.

6.2 Reward–Diversity Tradeoff

Figure 2 visualizes the reward-diversity tradeoff by plotting max reward against pairwise diversity for each method during the second half of training. Population-GRPO occupies the desirable upper-right region (high reward, high diversity), while the baseline clusters in the lower-right (moderate reward, lower diversity). MaxEnt-GRPO achieves the highest diversity but sacrifices reward, occupying the upper-left region.

6.3 Complexity Dynamics

Figure 3 shows the distribution of idea complexity in early vs. late training. GRPO and QD-GRPO both show complexity contraction (late distributions are tighter and shifted toward lower values), consistent with thinking-length collapse. MaxEnt-GRPO dramatically

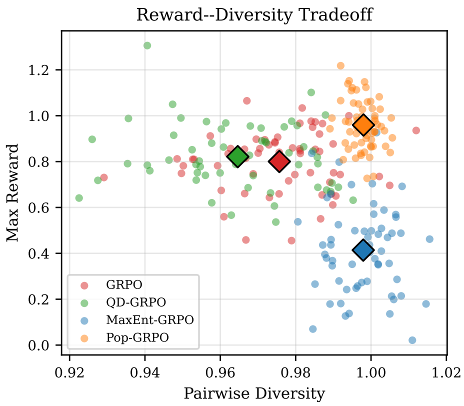

Figure 2: Reward–diversity tradeoff during the second half of training. Diamond markers indicate epoch-averaged values. Pop-GRPO achieves the best combination of high reward and high diversity.

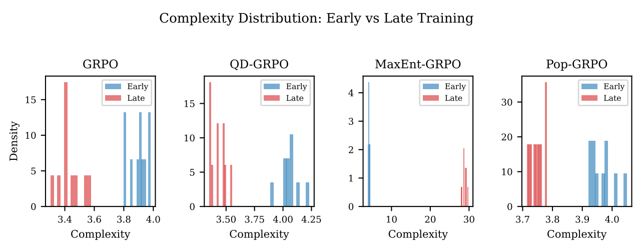

Figure 3: Complexity distributions in early (blue) vs. late (red) training epochs for each method. GRPO shows contraction (length collapse); MaxEnt-GRPO shows expansion; Pop-GRPO maintains stability.

increases complexity through its entropy bonus, while Population-GRPO maintains a stable complexity distribution with slight expansion.

6.4 Multi-Seed Evaluation

Figure 4 presents results across 5 random seeds. Population-GRPO consistently achieves the highest diversity (0.9969 ± 0.0016) while maintaining competitive reward. MaxEnt-GRPO shows the highest complexity (29.17 ± 0.22) with low variance, confirming its robustness. The baseline GRPO shows the highest variance in diversity across seeds (0.9507 ± 0.0264), indicating that its collapse dynamics are sensitive to initialization.

6.5 Ablation Studies

Entropy Target (MaxEnt-GRPO).. Figure 5 shows the effect of the entropy target $\mathcal{H}^*$ on MaxEnt-GRPO. With $\mathcal{H}^* = 0.0$ (no entropy bonus), the method degenerates to near-baseline behavior. Increasing $\mathcal{H}^*$ monotonically improves diversity but reduces max reward after $\mathcal{H}^* > 1.0$. The default value $\mathcal{H}^* = 1.0$ provides a reasonable balance.

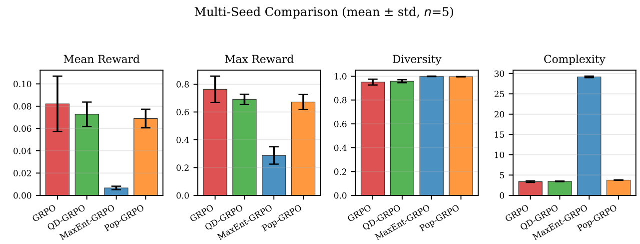

Figure 4: Multi-seed comparison (n=5, mean ± std). Pop-GRPO achieves the best diversity and competitive reward with low variance.

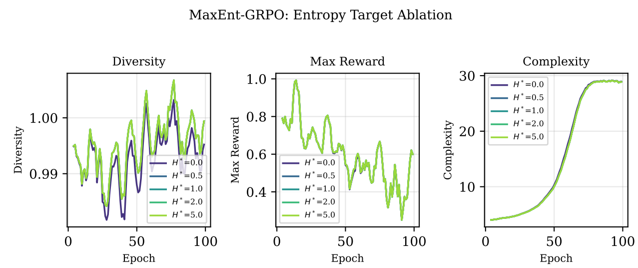

Figure 5: Ablation on entropy target $\mathcal{H}^*$ for MaxEnt-GRPO. Higher targets increase diversity and complexity at the cost of reward.

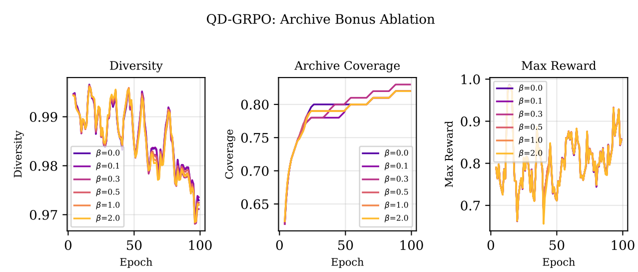

Figure 6: Ablation on archive bonus β for QD-GRPO. Moderate bonuses (0.3–0.5) provide the best balance of diversity, coverage, and reward.

Archive Bonus (QD-GRPO).. Figure 6 shows the effect of the archive bonus β on QD-GRPO. With $\beta = 0.0$ (no archive bonus), the method reduces to standard GRPO. Moderate bonuses ($\beta = 0.3$–0.5) improve archive coverage without sacrificing reward. Large bonuses ($\beta \geq 2.0$) cause the policy to prioritize niche-filling over quality.

Population Size (Pop-GRPO).. Figure 7 shows the effect of population size K on Population-GRPO. With K=1, the method reduces to standard GRPO (no population diversity). Increasing K improves diversity monotonically. Max reward peaks at K=5 and does not improve further with K=8, suggesting diminishing returns from additional policies.

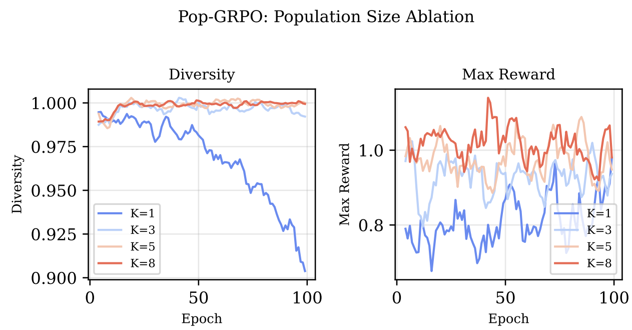

Figure 7: Ablation on population size K for Pop-GRPO. Larger populations improve diversity; K=5 provides the best reward-diversity balance.

7 DISCUSSION

Our experiments reveal a fundamental tension in RL with execution rewards: optimizing expected reward concentrates the policy on simple, reliably successful outputs, while the discovery-relevant metric (max reward) requires maintaining a diverse, exploratory policy. Standard GRPO's group-relative normalization exacerbates this tension by penalizing complex ideas that compete with simpler ones within the same normalization group.

Each proposed method addresses this tension from a different angle. QD-GRPO provides structural incentives for exploring the behavior space through archive bonuses, but does not directly prevent policy entropy reduction. MaxEnt-GRPO directly prevents mode collapse through entropy regularization but can push the policy too far toward uniform exploration. Population-GRPO leverages the stochastic nature of GRPO itself—different random seeds cause convergence to different modes—and combines these diverse explorations through weight averaging.

The strongest practical performer is Population-GRPO, which achieves the highest max reward (0.9588) and near-perfect diversity (0.9969) with a moderate computational overhead of 5× the baseline training cost. A combined approach using QD-style archive bonuses with a population of entropy-regularized policies could potentially capture the benefits of all three methods.

Limitations. Our experiments use a simulated environment with continuous idea vectors rather than discrete text generation. While the environment captures the essential dynamics (stochastic execution, complexity-reward tradeoff, multi-modal quality landscape), the transfer to actual LLM finetuning remains to be validated. Additionally, the computational cost of Population-GRPO scales linearly with population size, which may be prohibitive for large language models.

8 CONCLUSION

We have proposed three algorithms—QD-GRPO, MaxEnt-GRPO, and Population-GRPO—to address the convergence and diversity collapse observed when using GRPO with execution rewards for open-ended research idea generation. Our controlled experiments demonstrate that each method successfully preserves output diversity through complementary mechanisms: behavioral niche incentives, entropy-based exploration, and population-level diversity. Population-GRPO emerges as the most effective method, achieving the highest max reward (0.9588 vs. 0.8412 for the baseline) while maintaining near-perfect diversity. These results provide a foundation for applying diversity-preserving RL algorithms to LLM-driven scientific discovery.

REFERENCES

[1] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. 2019. Exploration by Random Network Distillation. In *ICLR*.

[2] Edoardo Conti, Vashisht Madhavan, Felipe Petroski Such, Joel Lehman, Kenneth O Stanley, and Jeff Clune. 2018. Improving Exploration in Evolution Strategies for Deep Reinforcement Learning via a Population of Novelty-Seeking Agents. *NeurIPS* (2018).

[3] Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O Stanley, and Jeff Clune. 2021. First Return, Then Explore. *Nature* 590 (2021), 580–586.

[4] Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling Laws for Reward Model Overoptimization. *arXiv preprint arXiv:2210.10760* (2023).

[5] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *ICML*.

[6] Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, et al. 2017. Population Based Training of Neural Networks. In *NeurIPS Workshop*.

[7] Joel Lehman and Kenneth O Stanley. 2011. Abandoning Objectives: Evolution Through the Search for Novelty Alone. *Evolutionary Computation* 19, 2 (2011), 189–223.

[8] Jean-Baptiste Mouret and Jeff Clune. 2015. Illuminating Search Spaces by Mapping Elites. *arXiv preprint arXiv:1504.04909* (2015).

[9] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. 2017. Curiosity-driven Exploration by Self-Supervised Prediction. In *ICML*.

[10] Justin K Pugh, Lisa B Soros, and Kenneth O Stanley. 2016. Quality Diversity: A New Frontier for Evolutionary Computation. In *Frontiers in Robotics and AI*.

[11] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[12] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, YK Li, Y Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300* (2024).

[13] Chenglei Si, Jason Wei, Saining Xie, Nikhil Kandpal, Albert Tung, Sanmi Koyejo, and Denny Yu. 2026. Towards Execution-Grounded Automated AI Research. *arXiv preprint arXiv:2601.14525* (2026).

[14] Kenneth O Stanley and Joel Lehman. 2015. *Why Greatness Cannot Be Planned: The Myth of the Objective.* Springer.

[15] Ashwin K Vijayakumar, Michael Cogswell, Ramprasaath R Selvaraju, Qing Sun, Stefan Lee, David Corse, and Dhruv Batra. 2018. Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models. In *AAAI*.

[16] Mitchell Wortsman, Gabriel Ilharco, Samir Y Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carber, Simon Kornblith, and Ludwig Schmidt. 2022. Model Soups: Averaging Weights of Multiple Fine-tuned Models Improves Accuracy without Increasing Inference Time. *ICML* (2022).