ICML
International Conference
On Machine Learning

Evolution Strategies at Scale: LLM Fine-Tuning Beyond Reinforcement Learning

Xin Qiu, Yulu Gan, Conor F. Hayes, Qiyao Liang, Yinggan Xu,
Roberto Dailey, Elliot Meyerson, Babak Hodjat, Risto Miikkulainen
Cognizant AI Lab

The LLM Training Pipeline: From Generalist to Specialist

1. Pre-training (General Language Understanding)

Learns grammar, facts, and reasoning from vast text data.
(e.g., The internet)

2. Fine-tuning (Task-Specific Specialization)

Adapts the model for specific tasks, instructions, or behaviors.

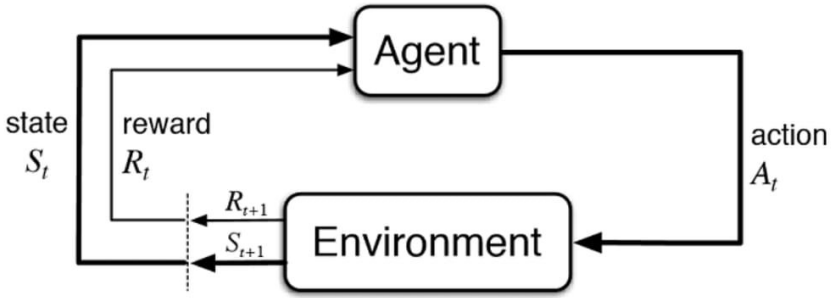
CRITICAL STEP: Unlocks performance on real-world applications (e.g., Coding, Chat, Summarization).

3. Inference (Deployment & Use)

Generates responses to user prompts in real-time applications.

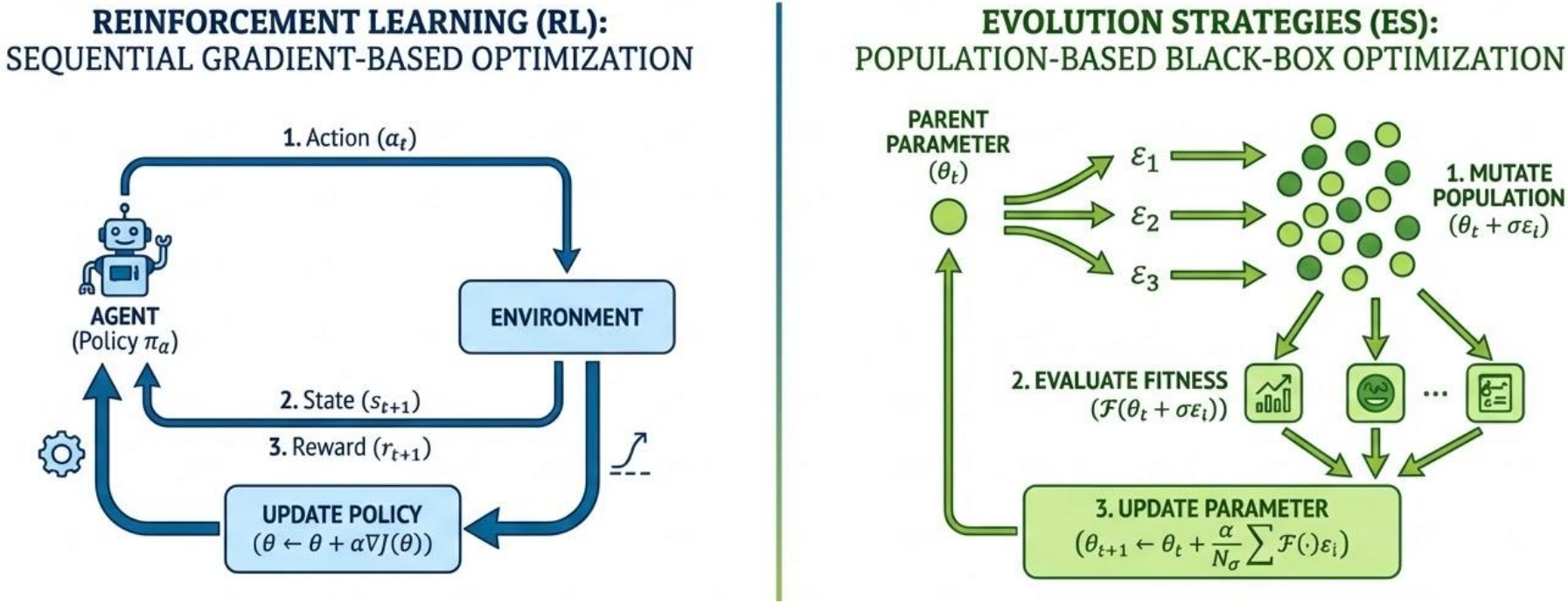
Reinforcement Learning (RL) in LLM fine-tuning

- Use cases:
  - Incentivizing reasoning capabilities, e.g., deepseek-R1
  - Improving alignment, e.g., RLHF
- Challenges:
  - Difficulty with long-horizon reward
  - High variance in performance
  - Reward hacking
  - Sensitivity to hyperparameters

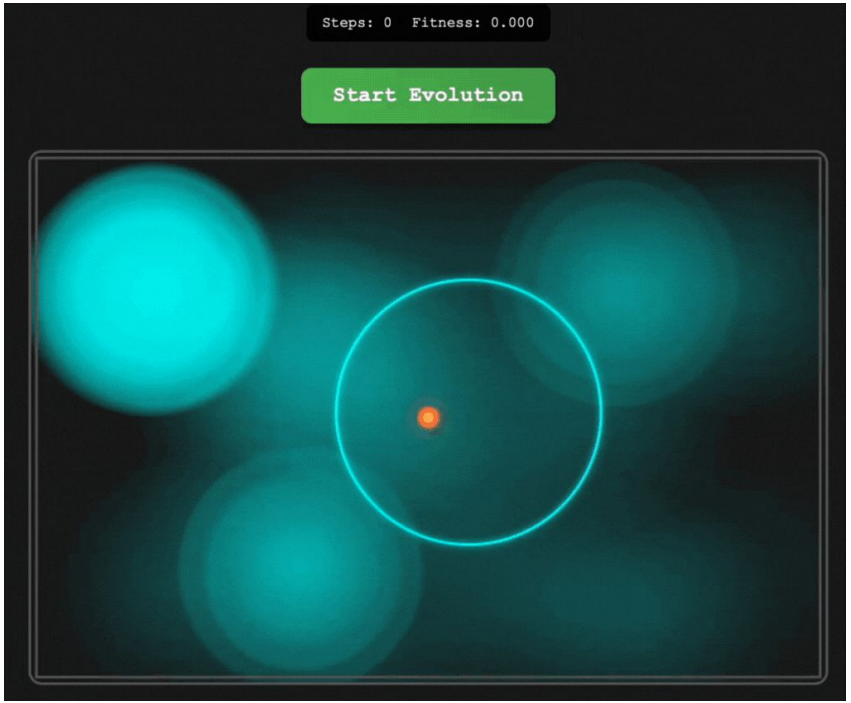


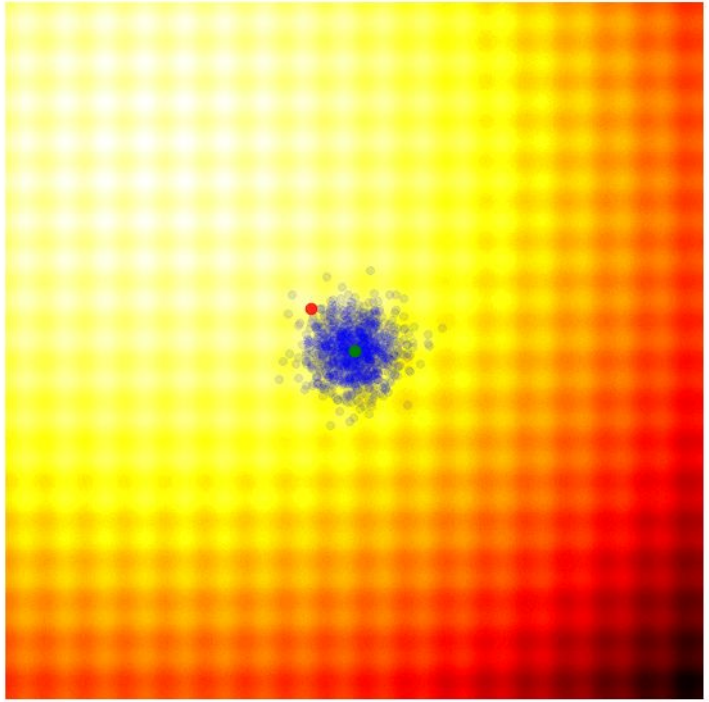
Source: Sutton, R. S. and Barto, A. G. Introduction to Reinforcement Learning

Reinforcement Learning (RL) vs. Evolution Strategies (ES)



Visualization of Searching Process in ES



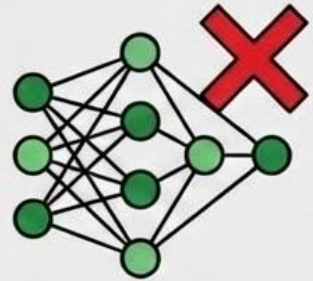


Source: David Ha, A Visual Guide to Evolution Strategies

Challenging the Status Quo with Evolution Strategies (ES)

Traditional Belief

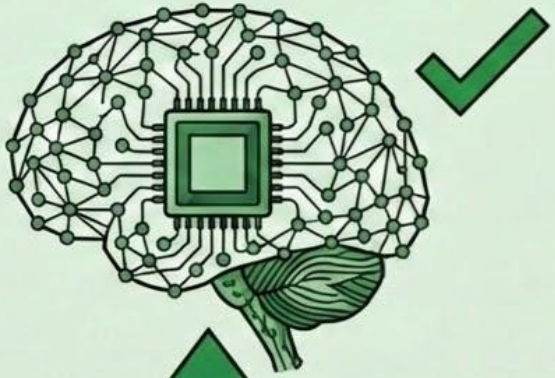
- Evolutionary search cannot scale to billion-parameter space.



Limited Scalability

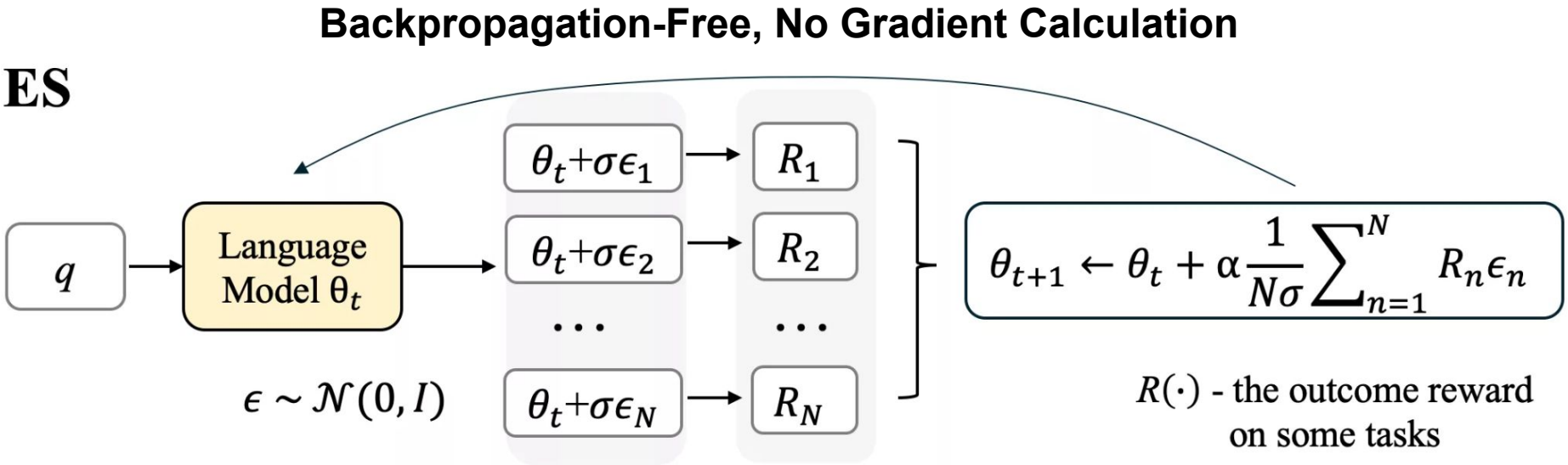
Our Breakthrough Contribution

- First successful application of ES to **full-parameter fine-tuning of LLMs at the billion-parameter scale**, without dimensionality reduction.



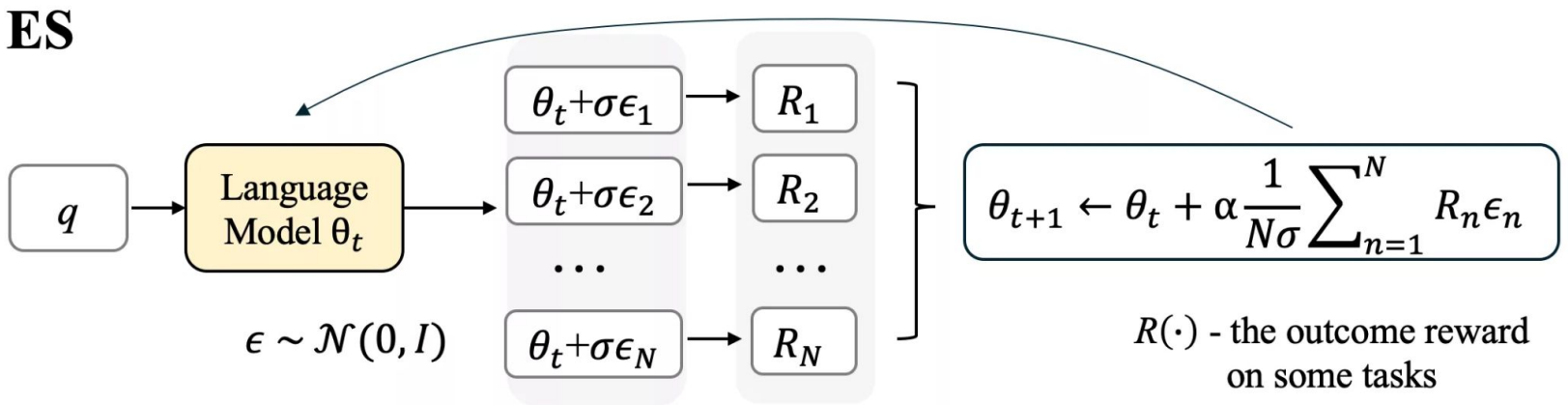
Billion-Parameter Scale

ES for LLM Fine-Tuning



Memory- and Compute-efficient Implementation

- Retrieval with random seed
- Parallel evaluation
- In-place weight perturbation and restoration
- Decomposition of the parameter update

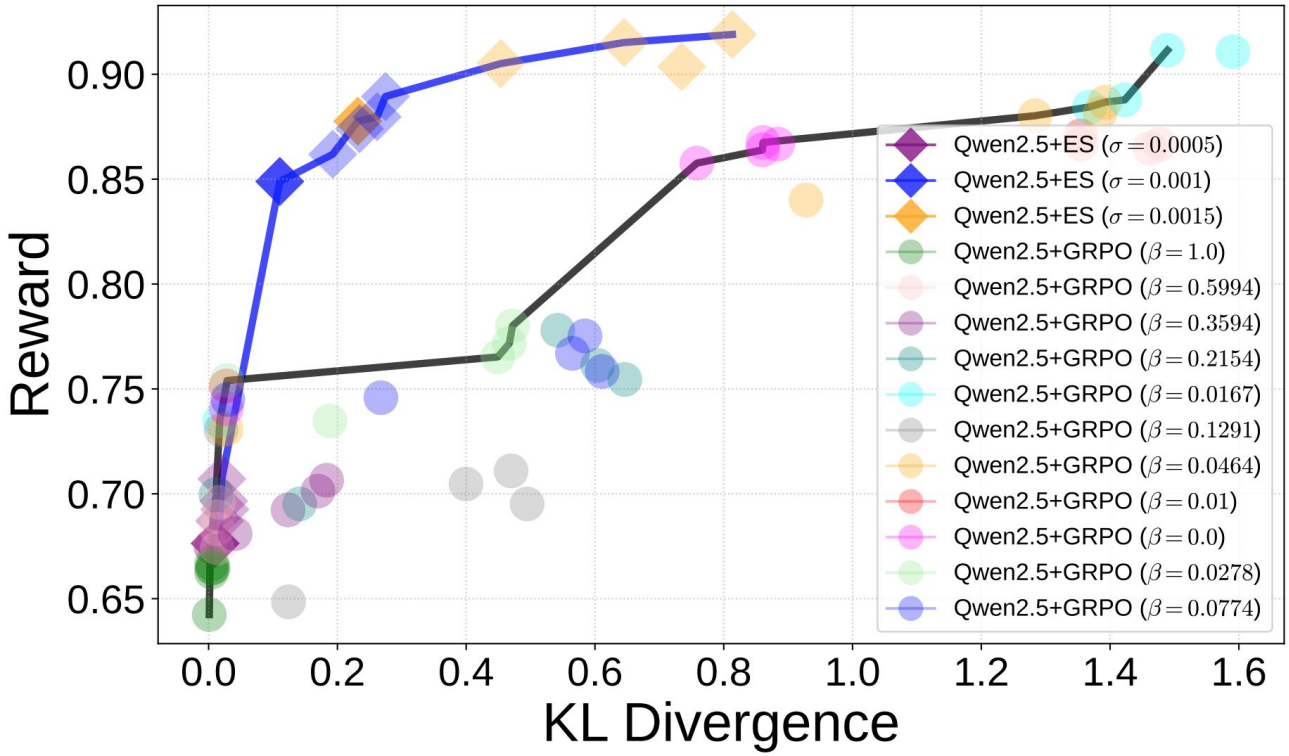


ES vs. RL in the Countdown Task

Base Model	Original	RL					ES (ours)
		PPO-z	GRPO-z (8)	GRPO-z (30)	GRPO-v	Dr.GRPO-v	
Qwen-2.5-0.5B-Instruct	0.1	0.3	0.3	0.5	13.0	13.5	**14.4**
Qwen-2.5-1.5B-Instruct	0.7	14.2	13.9	14.8	27.8	31.0	**37.3**
Qwen-2.5-3B-Instruct	10.0	20.1	30.9	32.5	37.8	43.8	**60.5**
Qwen-2.5-7B-Instruct	31.2	55.1	54.2	52.8	57.0	57.5	**66.8**
Llama-3.2-1B-Instruct	0.4	11.2	14.5	13.0	14.9	13.9	**16.8**
Llama-3.2-3B-Instruct	3.2	35.3	39.4	38.8	42.5	47.8	**51.6**
Llama-3.1-8B-Instruct	8.1	42.8	49.9	51.3	46.9	50.2	**61.2**

ES exhibits consistently better performance across different base models.

ES vs. RL in the Conciseness Task



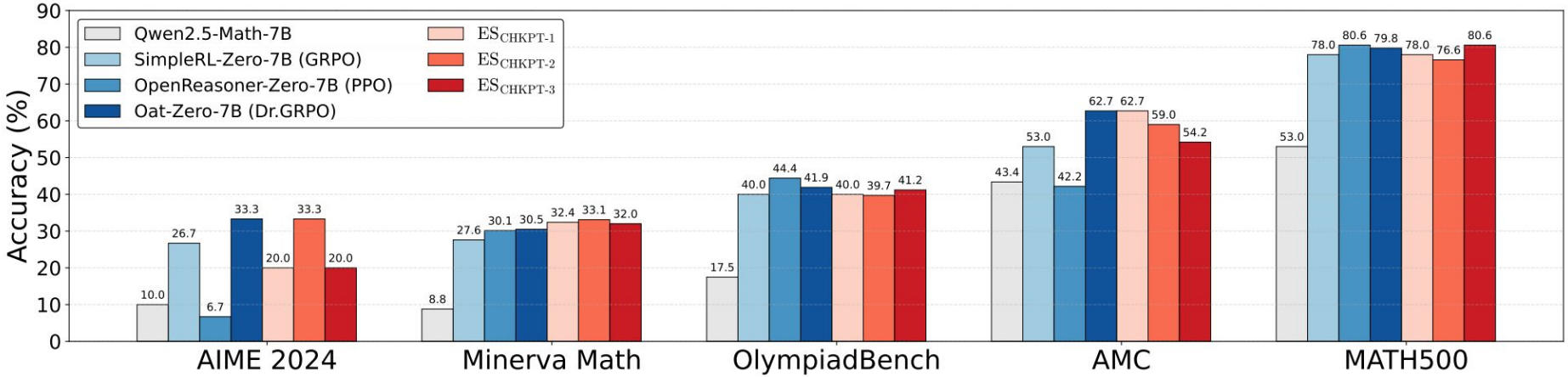
ES achieves better tradeoffs between performance and divergence from original LLMs, without explicit KL divergence penalty.

ES vs. RL in the Conciseness Task

Model	β	α	σ	Reward ↑	KL ↓
Qwen-2.5-7B+GRPO	0.0	5×10^{-6}	✗	$0.867 \pm 0.054^*$	$0.861 \pm 0.614^*$
Qwen-2.5-7B+GRPO	0.01	5×10^{-6}	✗	$0.871 \pm 0.060^*$	$1.354 \pm 0.873^*$
Qwen-2.5-7B+GRPO	0.0167	5×10^{-6}	✗	0.911 ± 0.038	1.591 ± 0.811
Qwen-2.5-7B+GRPO	0.0464	5×10^{-6}	✗	0.881 ± 0.062	1.384 ± 1.187
Qwen-2.5-7B+ES	✗	0.0005	0.001	$0.889 \pm \mathbf{0.004}$	$0.274 \pm \mathbf{0.096}$
Qwen-2.5-7B+ES	✗	0.00075	0.0015	$0.919 \pm \mathbf{0.008}$	$0.813 \pm \mathbf{0.212}$

ES shows lower variances in performances among different runs.

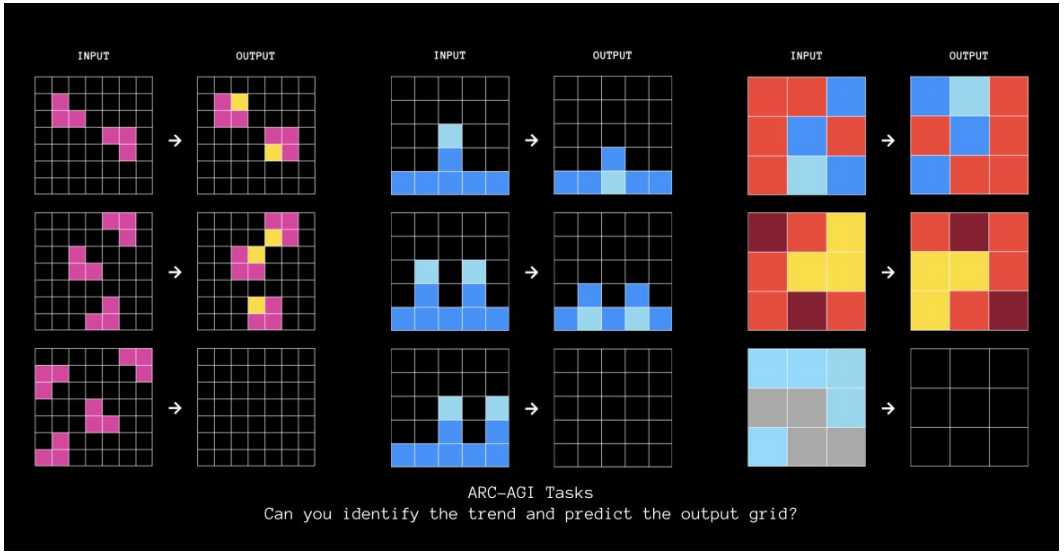
ES vs. RL in the Math Reasoning Benchmarks



ES performs competitively on SOTA math reasoning benchmarks.

Solving Puzzle Problems

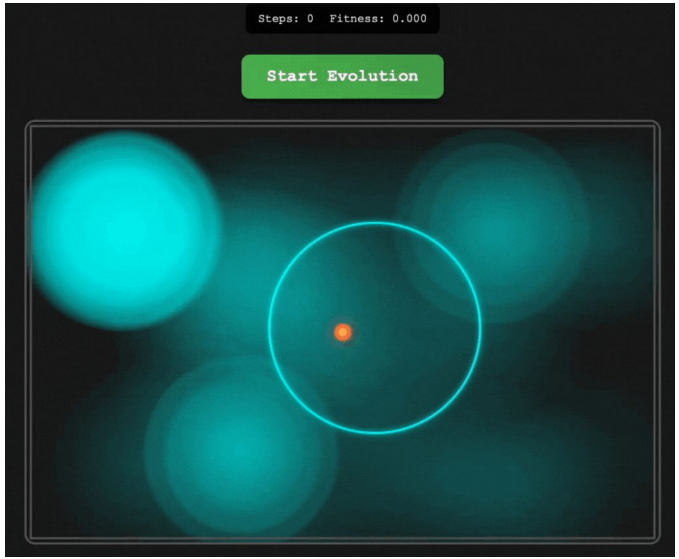
	3	1	2
2			4
3			1
		4	



ES generalizes to different types of problems.

Task	Base Model	Original	ES Fine-tuned
ARC-AGI	Qwen-2.5-14B	0.2	**29.5**
Sudoku	Qwen-2.5-3B	2.5	**69.5**

Summary



- Main surprise
  - ES can work in billion-parameter space
  - A small population size of 30 is enough
- Advantages of ES
  - Inference only, backpropagation-free
  - Robust performance across tasks and models
  - Insensitivity to hyperparameter
  - Less tendency to reward hacking
  - Stability across runs

Links & Contact

Email: **qiuxin.nju@gmail.com**

Paper

Code

Thank You