On the Optimal Reasoning Length for RL-Trained Language Models

paper

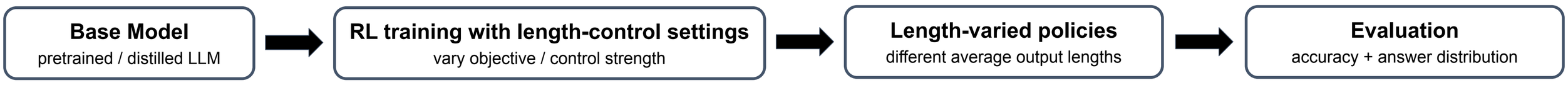
Daisuke Nohara, Taishi Nakamura, Rio Yokota
Institute of Science Tokyo

Motivation

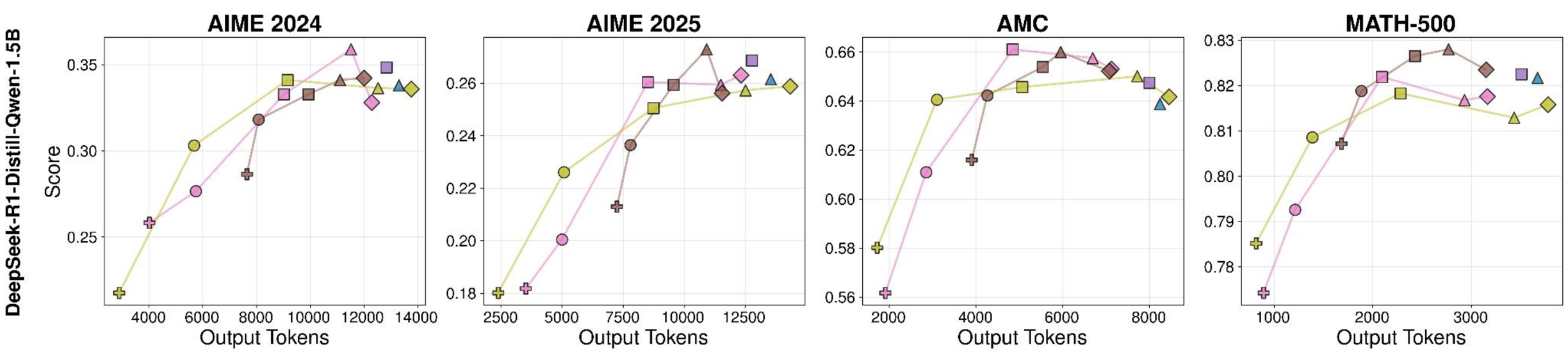
- RL post-training has produced strong reasoning models, but it also tends to increase output length and inference cost.
- Recent work suggests that simply extending test-time thinking can lead to non-monotonic gains and overthinking [1].
- During RL training, we can encourage shorter or longer outputs, but it is unclear how this changes downstream accuracy.
- We ask: do RL-trained language models have a reasoning-length regime that maximizes single-sample accuracy?

Experimental Design

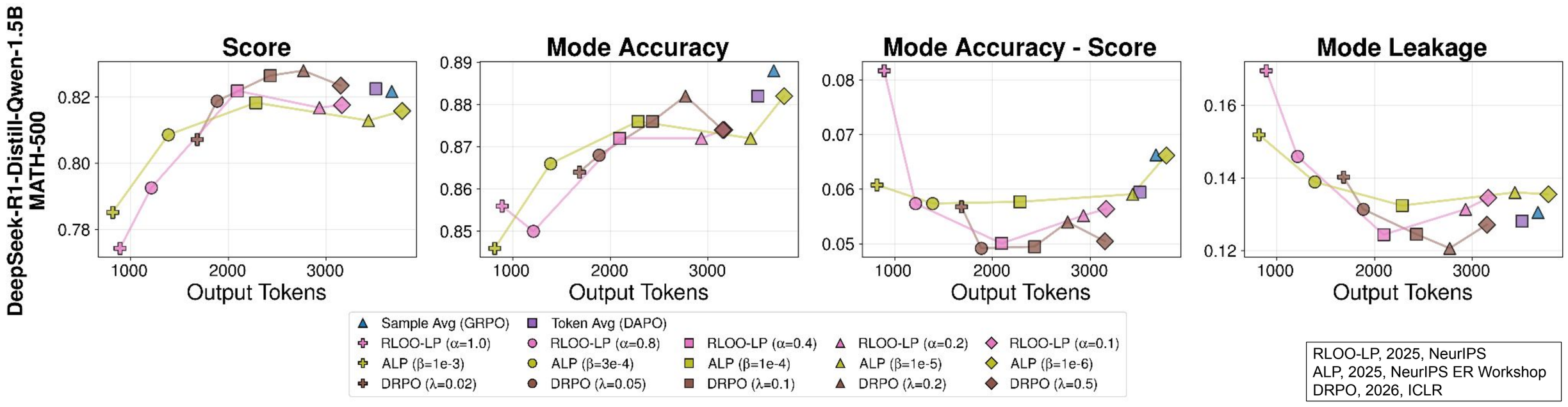
- Train multiple RL policies with different length-control settings.
- **Vary:** length-control objective / strength
- **Measure:** average output length, sample accuracy, mode accuracy, mode leakage
- **Models:** DeepSeek-R1-Distill-Qwen-1.5B, Qwen3-1.7B, Qwen3-4B-Base
- **Benchmarks:** AIME, AMC, MATH-500, HumanEval, MBPP+
- **Length-control methods:** Sample Avg, Token Avg, ALP, DRPO, RLOO-LP



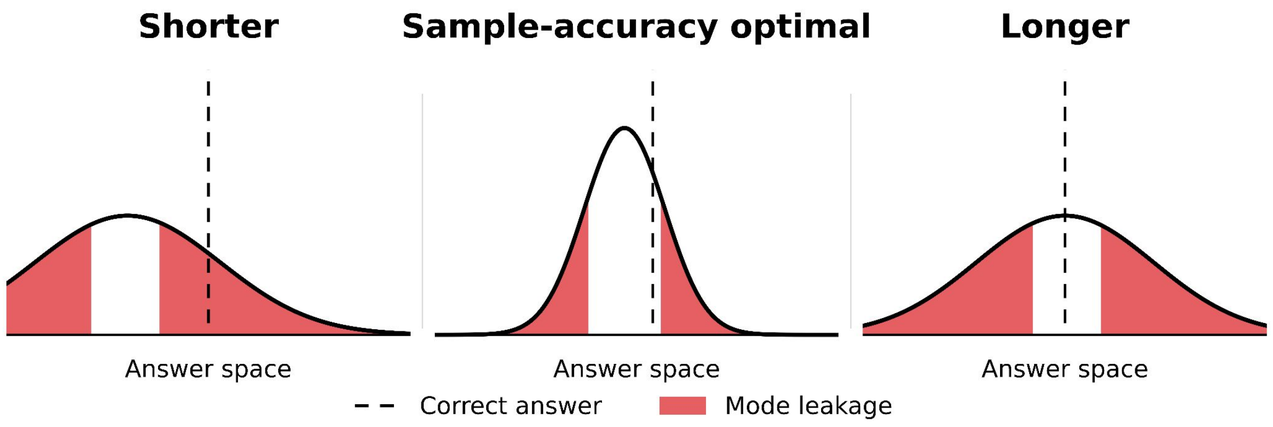
Result 1 — Sample accuracy is non-monotonic and peaks at an intermediate reasoning length



Result 2 — Longer outputs improve the modal answer but increase answer dispersion



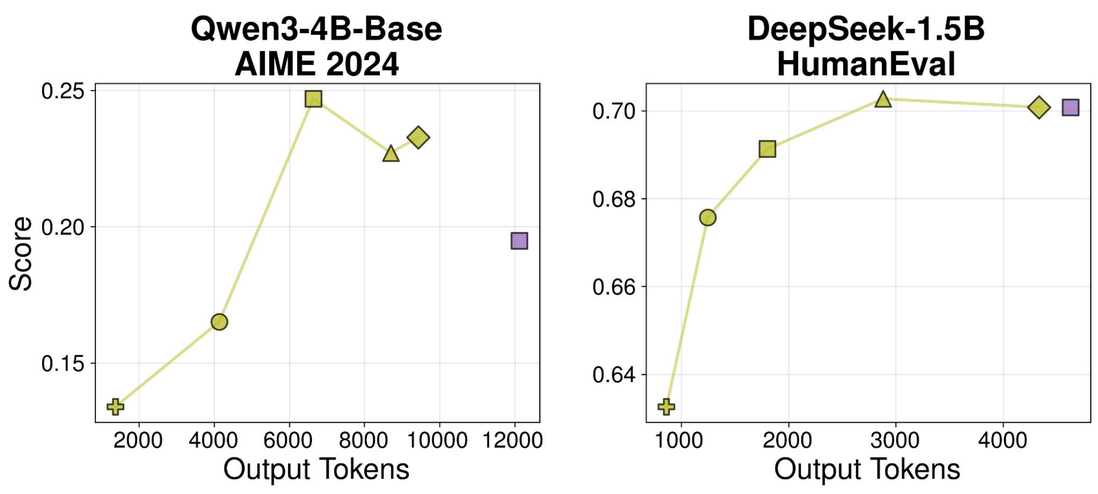
RLOO-LP, 2025, NeurIPS
ALP, 2025, NeurIPS ER Workshop
DRPO, 2026, ICLR



Mechanism

- **Mode accuracy ↑:** the most frequent answer becomes more often correct
- **Mode leakage ↑:** more samples fall outside the modal answer group
- **Sample score:** improves when mode gains dominate, but drops when leakage dominates

Result 3 — Generality



- Same non-monotonic trend across model families and code generation

Summary & Implications

- Single-sample accuracy peaks at intermediate reasoning lengths.
- Long reasoning can improve the modal answer while increasing leakage.
- Stronger length control may improve the accuracy–cost tradeoff when leakage dominates.
- High mode accuracy can benefit test-time aggregation, but mode leakage determines the required number of samples.

References

[1] Ghosal et al. Does Thinking More Always Help? Mirage of Test-Time Scaling in Reasoning Models. NeurIPS 2025.