Speculations on Test-Time Scaling

Sasha Rush Daniel Ritter

Cornell

Outline

Introduction

The Clues

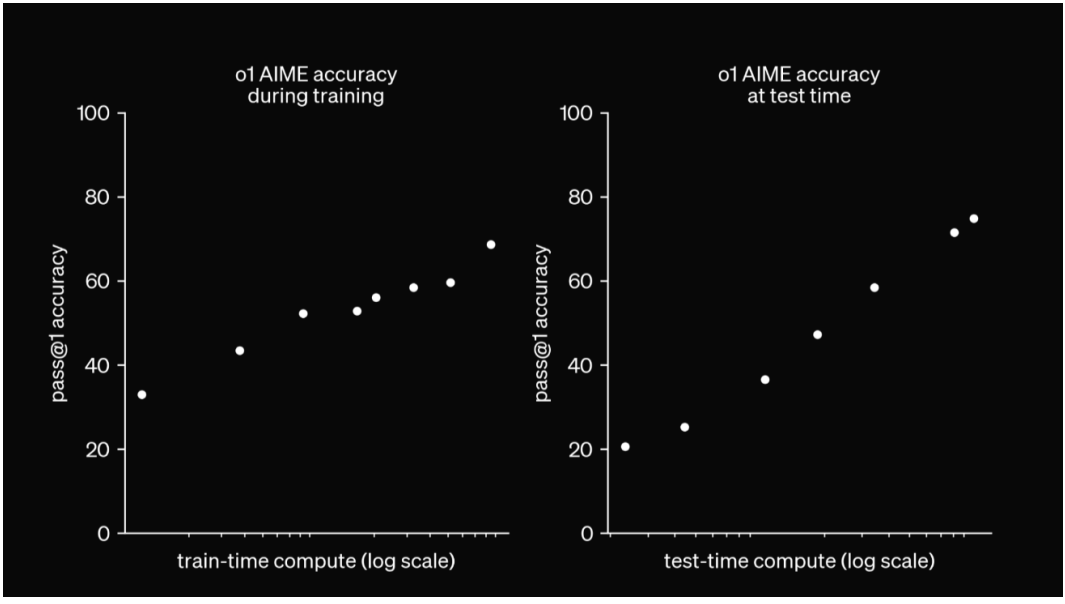
Guess and Check

Guided Search

Full AlphaZero

Learning to Search

Something Wild


o1 AIME accuracy
during training
100
80
60
40
20
0
pass@1 accuracy
train-time compute (log scale)
o1 AIME accuracy
at test time
100
80
60
40
20
0
pass@1 accuracy
test-time compute (log scale)

Context

- LLM (2018-2024) driven by training scaling
- Speculation: Benefit of static data running out

Implication

- Breakthrough in large-scale RL Training
-

What have we seen?

- Public demo model
- Strong result in constrained domains.

This Talk

- Survey of the public literature
- Synthesis of discussions with expert
- Gossip and hearsay

Thanks

Lewis Tunstall, Edward Beeching, Aviral Kumar, Charlie Snell, Michael Hassid, Yoav Artzi, Risab Agarwal, Kanishk Gandhi, Wenting Zhao, Yuntian Deng, Nathan Lambert

What we know

Our large-scale **reinforcement learning algorithm** teaches the model how to think productively using its **chain of thought** in a highly **data-efficient** training process.

What we know

- RL - Signal from verifiable problems
- CoT - “Thinking” occurs in token stream
- Data Efficient - Fixed set of good problems

From Gossip

- Single final model
- Not learned from expert examples
-

Chain of Thought

o1 learns to hone its chain of thought and refine the strategies it uses. It learns to recognize and **correct its mistakes**. It learns to **break down tricky steps** into simpler ones. It learns to try a **different approach** when the current one isn't working.

Review: Chain of Thought

-

Planning

Backtracking

Strategies

Summary

- Solves problems by very long CoT
- CoT includes “thinking” (search / planning)
- Core novelty: Inducing this behavior

The Suspects

- Guess + Check
- Guided Search
- AlphaZero
- Learn to Search
- Wildcard

A Note About Names

- Many different communities
- Names conflict and overlap with past methods
- This talk: First explain, then discuss names

Informal: Guess + Check

- Sample N CoTs
- Check if successful
- Train on good ones

Simple Formalization: EM

- Sample N CoTs
- Check if successful
- Train on good ones

Online Formalization: Policy Gradient

- Sample N CoTs
- Check if successful
- Train on good ones

Terminology

- STaR
- ReST
- ReST-EM
- Filtered Rejection Sampling
- Best-of-N

Why might this be right?

- Extremely simple and scalable
- Good baseline in past work

Why might this be wrong?

- No evidence this learns to correct, plan
- Well-explored in literature with marginal gains

Alternative

- Can we improve upon the process of finding adequate CoTs?

Informal: Guided Search

- Sample several next steps for CoT
- Check with a guide model for which to pursue
- Continue to the end
- Train on good ones

Where does the guide come from?

PRM/RollOuts

[Uesato et al., 2022]

- Point 1
- Point 2
- Point 3

Full AlphaZero

[Silver et al., 2017]

- Point 1
- Point 2
- Point 3

Learning to Search

[Gandhi et al., 2024]

- Point 1
- Point 2
- Point 3

Something Wild

[Brown et al., 2024]

- Point 1
- Point 2
- Point 3

Reference I

[Brown et al., 2024] Brown, B., Juravsky, J., Ehrlich, R., Clark, R., Le, Q. V., Ré, C., and Mirhoseini, A. (2024). Large language monkeys: Scaling inference compute with repeated sampling. *arXiv [cs.LG]*.

[Gandhi et al., 2024] Gandhi, K., Lee, D., Grand, G., Liu, M., Cheng, W., Sharma, A., and Goodman, N. D. (2024). Stream of search (SoS): Learning to search in language. *arXiv [cs.LG]*.

Reference II

[Silver et al., 2017] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. (2017).
Mastering chess and shogi by self-play with a general reinforcement learning algorithm.
arXiv [cs.AI].

Reference III

[Uesato et al., 2022] Uesato, J., Kushman, N., Kumar, R., Song, F., Siegel, N., Wang, L., Creswell, A., Irving, G., and Higgins, I. (2022).
Solving math word problems with process- and outcome-based feedback.
arXiv [cs.LG].