Inference Is All You Need

-

What did Ilya see?

rasdani/inference-is-all-you-need

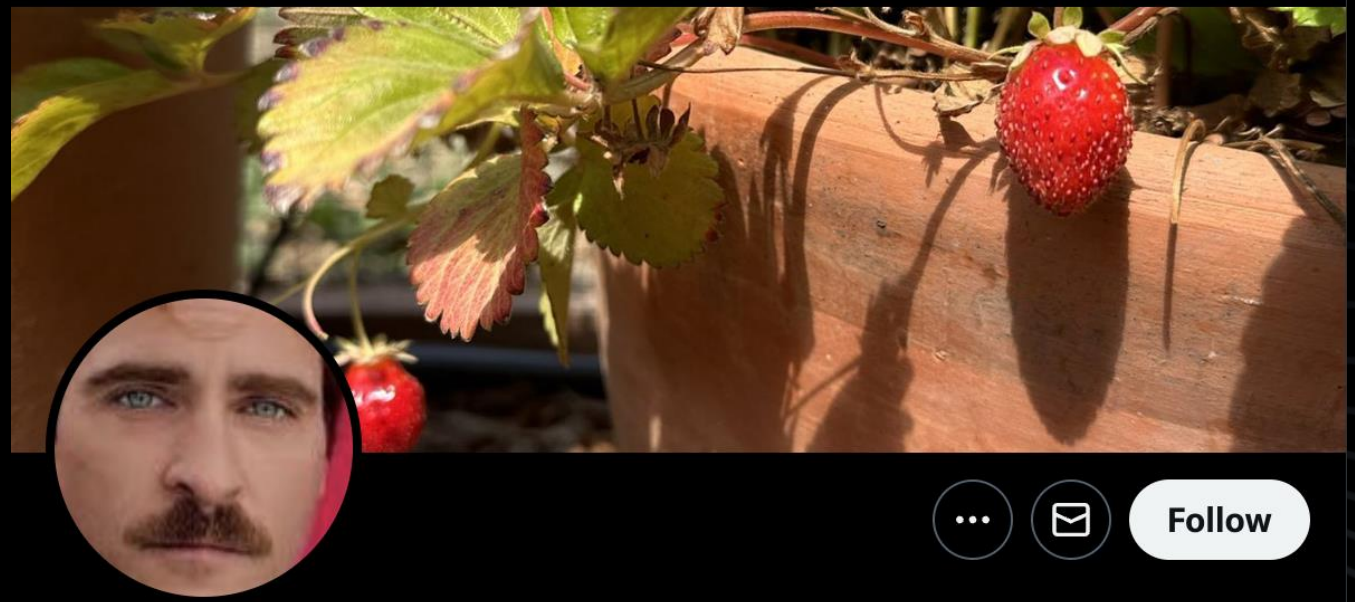
Sam Altman
@sama · Follow

i love summer in the garden

5:29 PM · Aug 7, 2024

12.5K Reply Share

Read 1.4K replies



Follow

@iruletheworldmo

purely theatre

Science & Technology Joined August 2022

1,673 Following **30K** Followers

ellamind

Large Language Monkeys: Scaling Inference Compute with Repeated Sampling

Bradley Brown*†‡, Jordan Juravsky*†, Ryan Ehrlich*†, Ronald Clark‡, Quoc V. Le§, Christopher Ré†, and Azalia Mirhoseini†§

†Department of Computer Science, Stanford University
‡University of Oxford
§Google DeepMind

bradley.brown@cs.ox.ac.uk, jbj@stanford.edu, ryanehrlich@cs.stanford.edu, ronald.clark@cs.ox.ac.uk, qvl@google.com, chrismre@stanford.edu, azalia@stanford.edu

July 30, 2024

Google DeepMind 2024-8-7

Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters

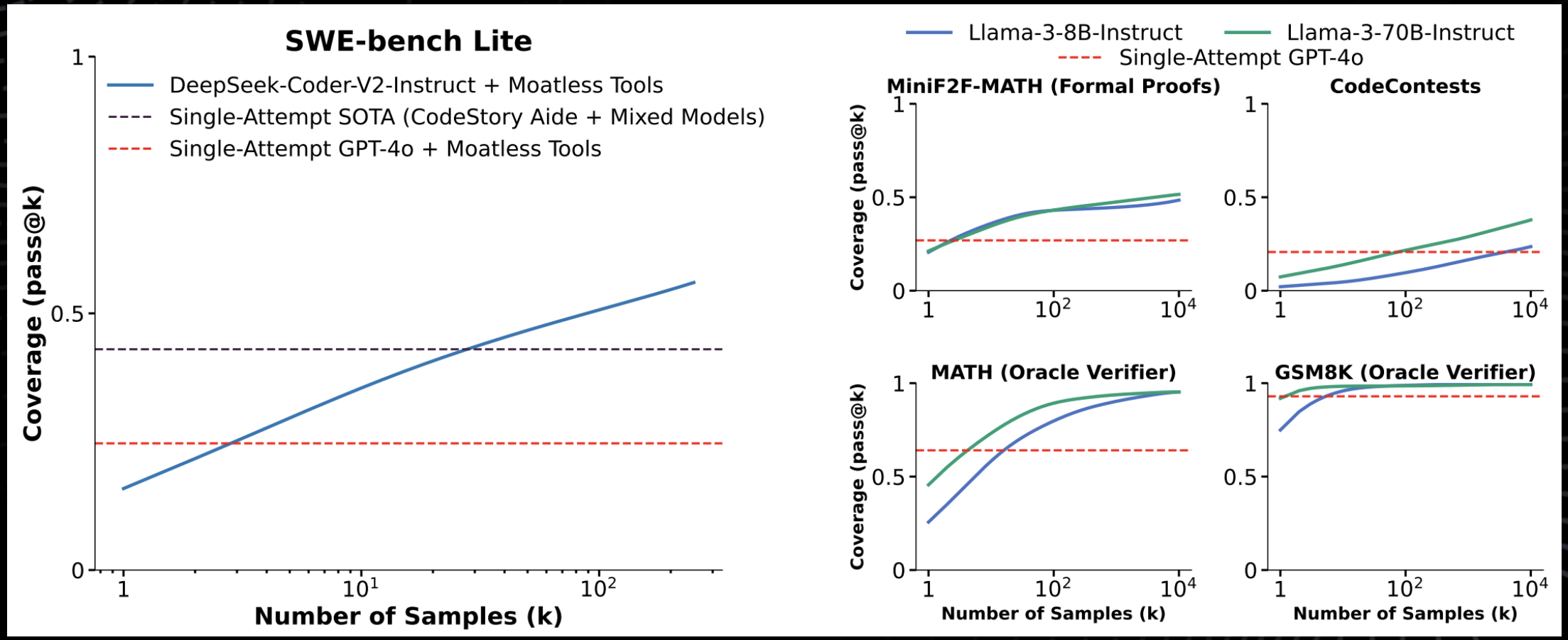
Charlie Snell[♦, 1], **Jaehoon Lee**[2], **Kelvin Xu**[♣, 2] **and Aviral Kumar**[♣, 2]

[♣]Equal advising, [1]UC Berkeley, [2]Google DeepMind, [♦]Work done during an internship at Google DeepMind

1. What does ”Scaling Test Time Compute” mean?

2. Why does it matter for AI Engineers?

“Ask the same question a couple of times“



“Ask the same question a couple of times“

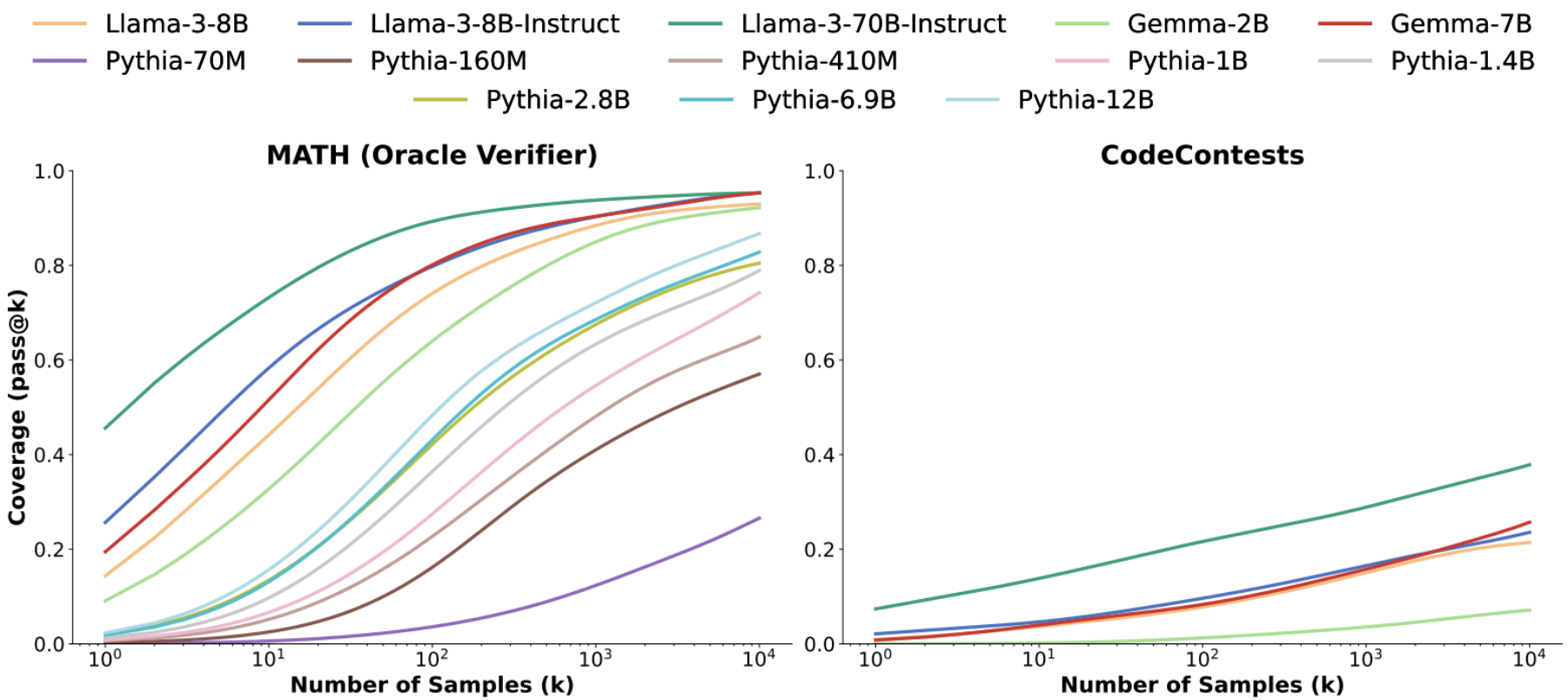


Figure 3: Scaling inference time compute via repeated sampling leads to consistent coverage gains across a variety of model sizes (70M-70B), families (Llama, Gemma and Pythia) and levels of post-training (Base and Instruct models).

Access to “Ground Truth”

Golden Answer
Unit Tests
Formal Verfication
Any strict test

Reward Models!

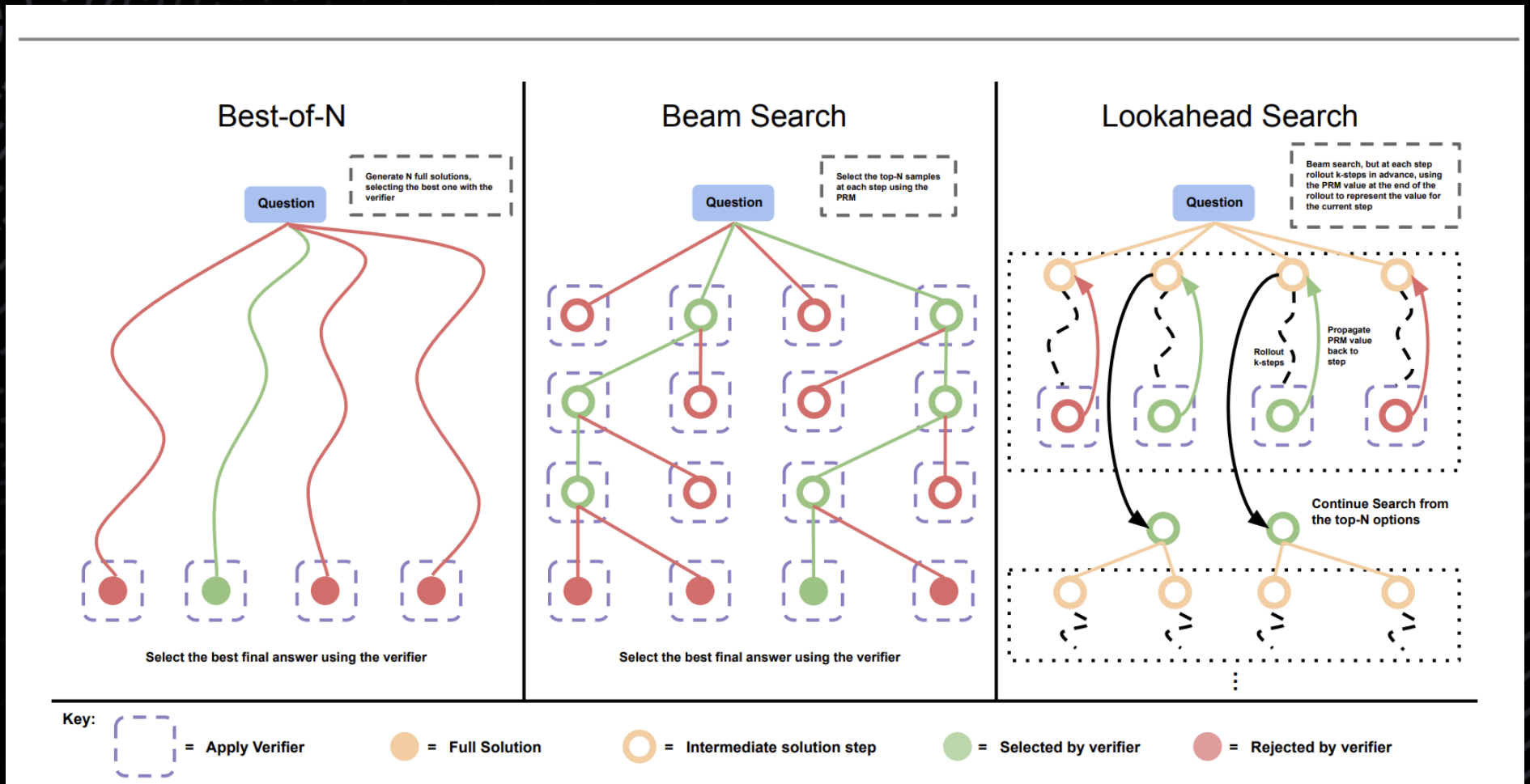


Figure 2 | ***Comparing different PRM search methods.*** **Left:** Best-of-N samples N full answers and then selects the best answer according to the PRM final score. **Center:** Beam search samples N candidates at each step, and selects the top M according to the PRM to continue the search from. **Right:** lookahead-search extends each step in beam-search to utilize a k-step lookahead while assessing which steps to retain and continue the search from. Thus lookahead-search needs more compute.

AGI achieved internally? 🍓

PRM800K

Let's Verify Step by Step

Hunter Lightman* Vineet Kosaraju* Yura Burda* Harri Edwards

Bowen Baker Teddy Lee Jan Leike John Schulman Ilya Sutskever

Karl Cobbe*

OpenAI

Abstract

peiyi9979/math-shepherd-mistral-7b-prm

```
If Buzz bought a pizza with 78 slices at a restaurant and then decided to share it

Step 1: The total ratio representing the pizza is 5+8 = <<5+8=13>>13. ки

Step 2: The waiter ate 13 x 8 / 13 = <<13*8/13=6>>6 slices of the pizza. ки

Step 3: Buzz ate 78 - 6 = <<78-6=72>>72 slices of the pizza. ки

Step 4: The waiter ate 20 less than the number of slices that Buzz ate which is 72

Step 5: The waiter ate 52 slices of the pizza. The answer is: 52 ки
```

2. "label": problem + step-by-step solution with automatic label, e.g.,

```
If Buzz bought a pizza with 78 slices at a restaurant and then decided to share it

Step 1: The total ratio representing the pizza is 5+8 = <<5+8=13>>13. +

Step 2: The waiter ate 13 x 8 / 13 = <<13*8/13=6>>6 slices of the pizza. -

Step 3: Buzz ate 78 - 6 = <<78-6=72>>72 slices of the pizza. -

Step 4: The waiter ate 20 less than the number of slices that Buzz ate which is 72

Step 5: The waiter ate 52 slices of the pizza. The answer is: 52 -
```

Why does this matter for an AI engineer?

- **Stronger capabilities with less memory**
- **Inference is getting more optimized**
  - Groq, Cerebras
  - Quantization
  - Speculative Decoding
- **Easier Data collection**

linktr.ee/rasdani

rasdani/inference-is-all-you-need

ellamind

jan@ellamind.com

https://ellamind.com

ellamind GmbH
Konsul-Smidt-Straße 8p
28217 Bremen