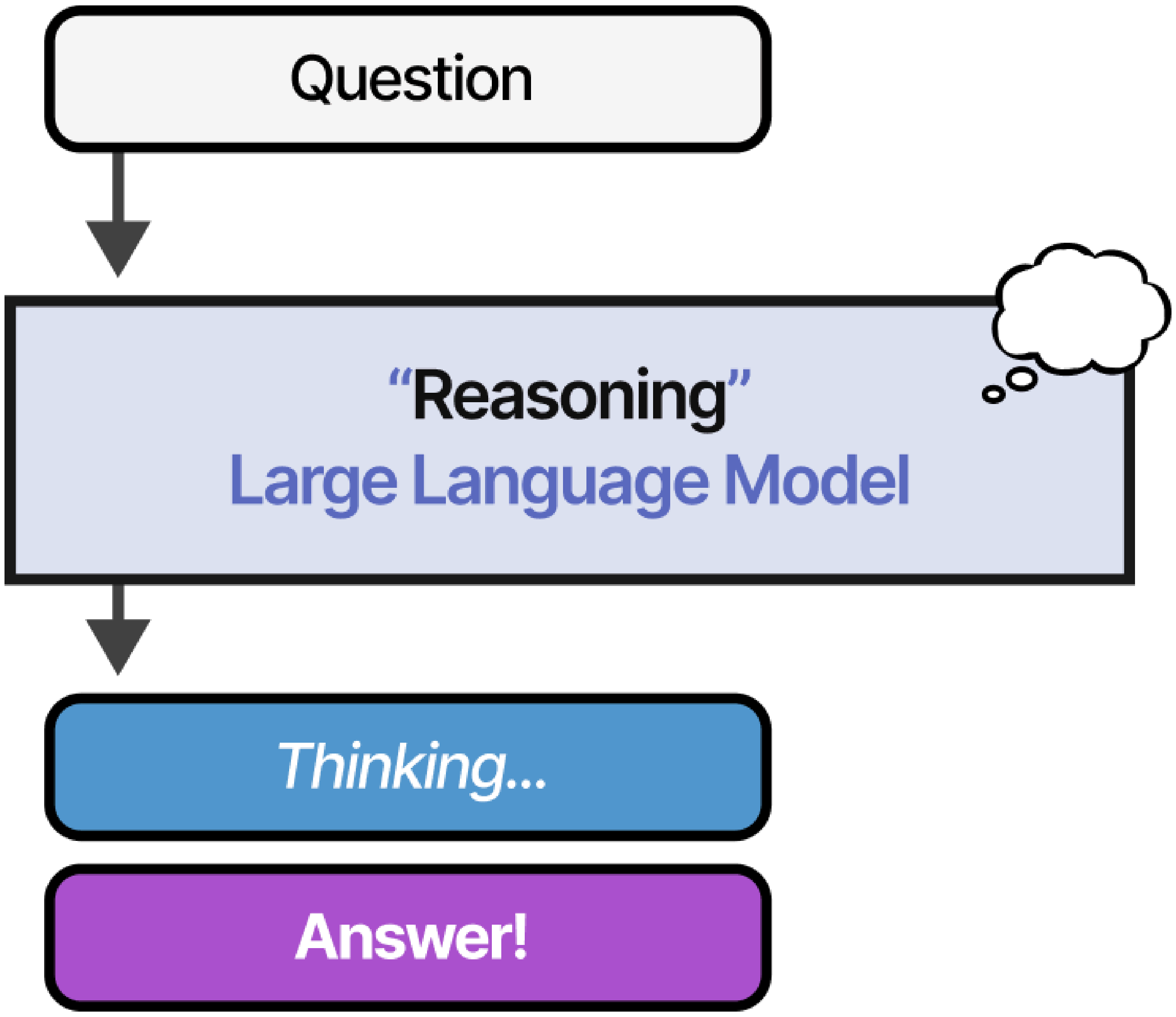
HOW ARE LLMS TRAINED TO REASON?

What’s behind the “on/off” reasoning mode?

Bhavishya Pandit

Introduction

We hear about AI models like **GPT-4, Claude, and Llama-Nemotron** solving complex math, coding, and even scientific reasoning, but how do they learn to reason step-by-step like humans?



OpenAI's O1 model was a pioneer in this space, and since then, many other reasoning models have emerged, each improving upon its predecessors.

DeepSeek-R1, for instance, has demonstrated significantly enhanced performance.

NVIDIA's Llama-Nemotron introduces a novel feature: a reasoning toggle that allows users to switch reasoning on or off, with the model adjusting its responses accordingly.

Let's dive into how reasoning models are trained and specifically how Nemotron's reasoning toggle works.

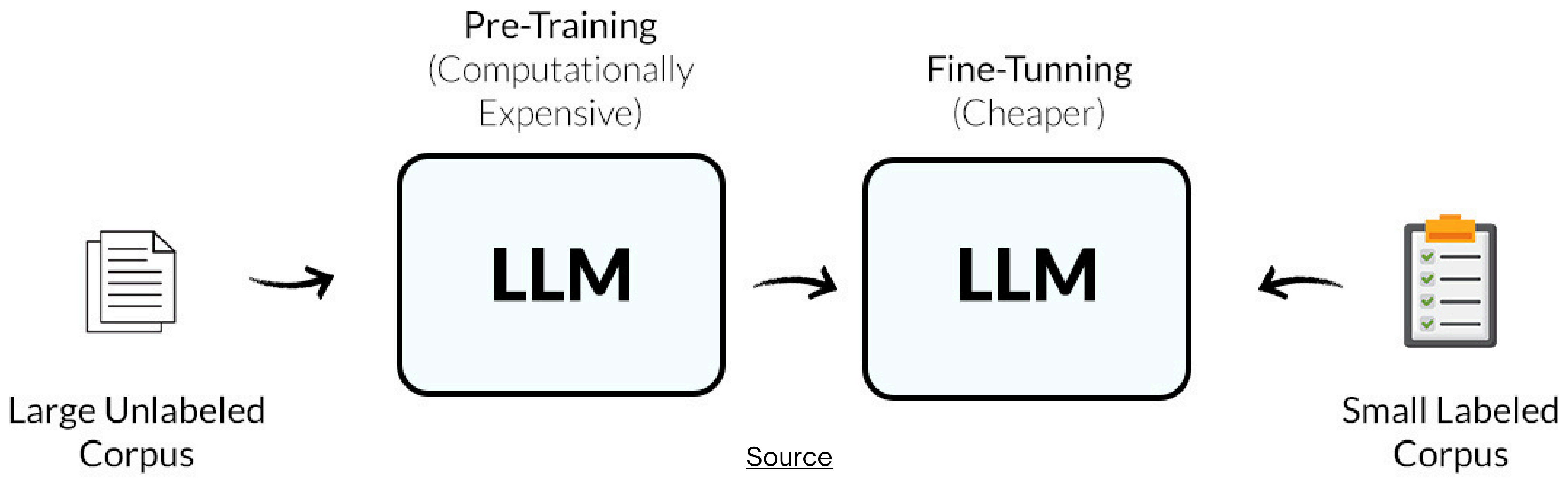
Source

Pre-Training & Fine-Tuning

Before reasoning, LLMs need knowledge. They start by analyzing trillions of words books, Wikipedia, code repositories, and research papers to build their knowledge base.

🔷 What they learn:

- Facts (e.g., "Paris is France’s capital")
- Certain logics (e.g., "If A=B and B=C, then A=C")
- Language patterns



Source

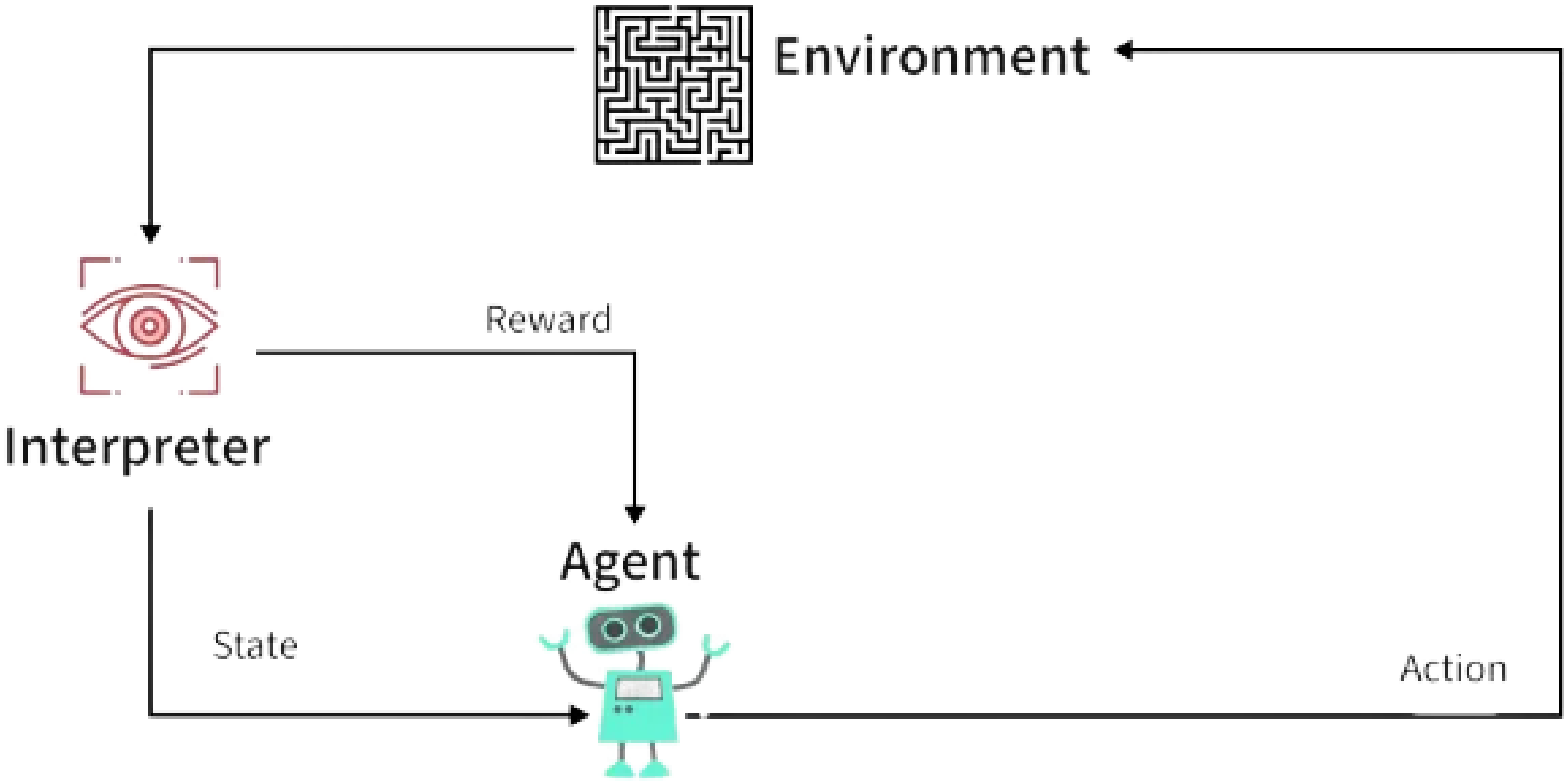
After pretraining, model is finetuned for its reasoning abilities from a teacher LLM or humans.

e.g. **Nvidia Nemotron** uses **Deepseek R1** as its teacher LLM.

Reinforcement Learning

Here’s how LLMs learn how to reason. They:

1. Generate multiple answers (e.g., 10 ways to solve a calculus problem).
2. Receive feedback (e.g., "Answer 3 is correct; however, steps 1 and 5 are flawed").
3. Optimise for better reasoning over time.



🚀 Why RL matters:

Source

- Encourages creative problem-solving (not just imitation).
- Fixes "lazy reasoning" (e.g., skipping steps).
- Models like **Llama-Nemotron** learn step by step: they start with easy tasks and move on to harder ones only after mastering the simpler examples.

Dynamic Reasoning Toggle

NVIDIA’s Llama-Nemotron is the first model for developers to introduce a reasoning toggle.

Here's how they did it:

✅ **Paired Training Data:** NVIDIA created a unique dataset where each query had two responses:

- A "reasoning" response (from models like DeepSeek-R1), if tagged with "**detailed thinking on**."
- A "non-reasoning" response (from models like Llama-3.1-Nemotron-70B-Instruct) if tagged with "**detailed thinking off**."

✅ **Supervised Fine-Tuning (SFT):** During SFT, the models learned to link these specific system prompts to the desired output.

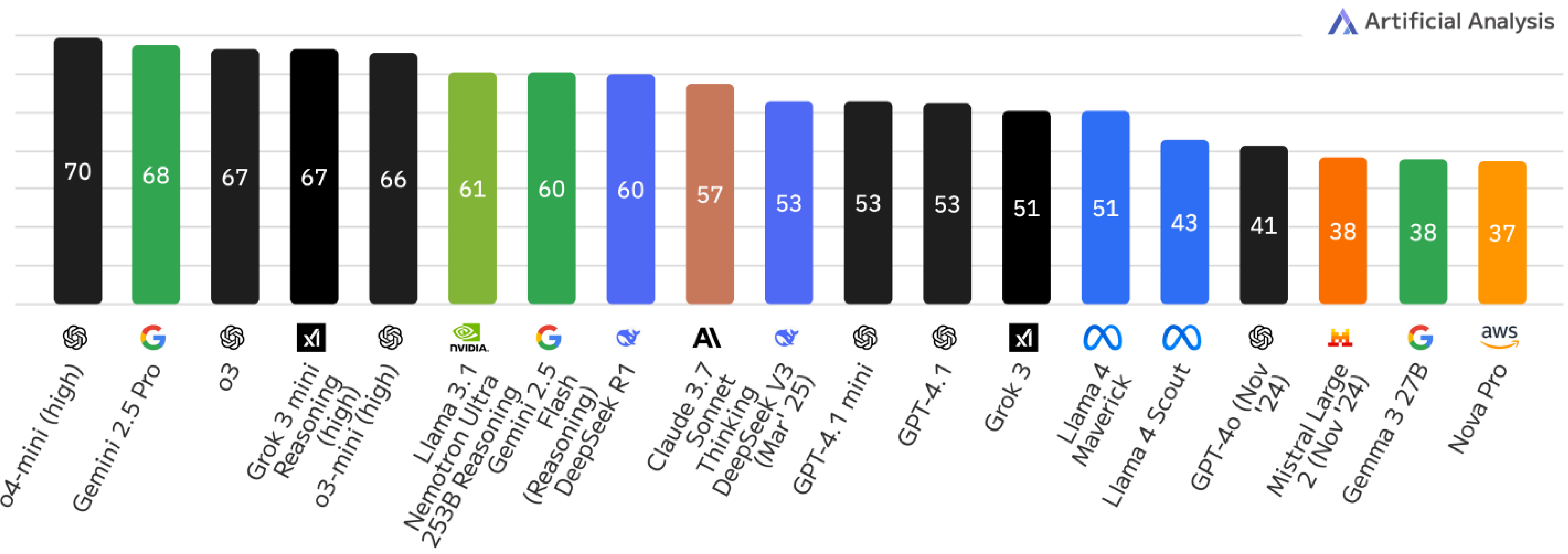
By mixing "on" and "off" examples, Llama-Nemotron learned to:

- Use full analytical capabilities when prompted with "detailed thinking on."
- Give brief answers when prompted with "detailed thinking off."

The Impact

As of April 2025, **o4-mini (high)** is the best model with a score of **70**, according to **Artificial Analysis Intelligence Index.**

However, **Nvidia's Nemotron (LN Ultra)** is the best open model in the list.



Nvidia LN-Ultra derived from **LLaMa 3.3** outperforms both **LLaMa 3.3** and **Deepseek R1** in reasoning capability while fitting on a **single 8xH100** node and achieving **higher inference** throughput.

This demonstrates how reasoning models continue to improve over time. With **DeepSeek's new R1 model** release, we may see even more significant advancements in the future.

Follow to stay updated on Generative AI

👍 LIKE

💬 COMMENT

🔁 REPOST