DroneX AI

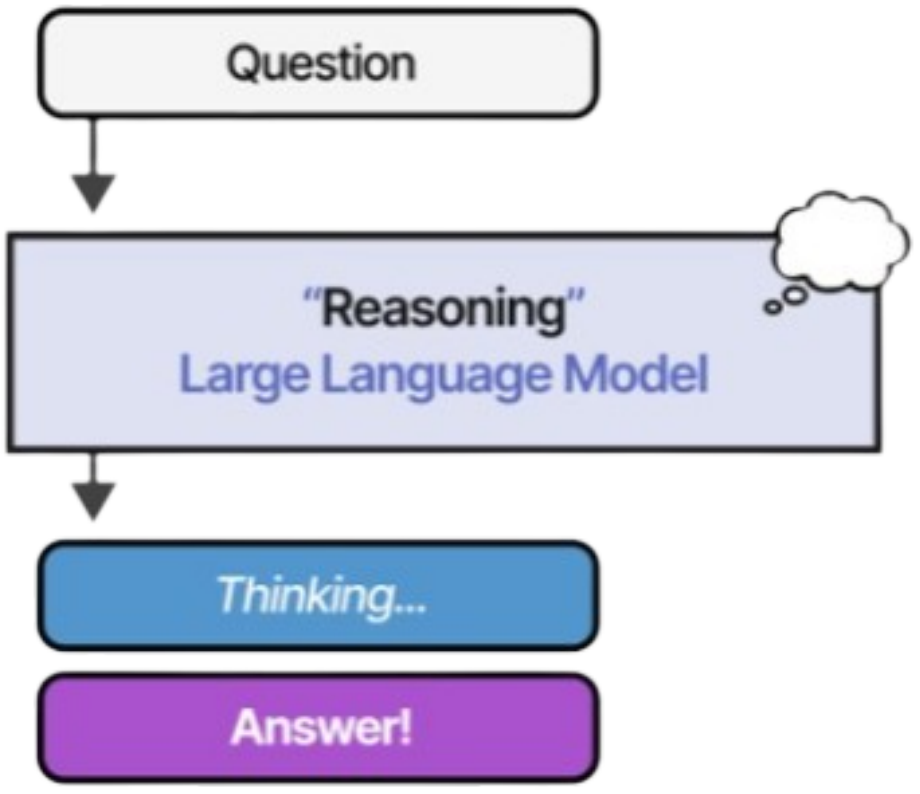
The Process Behind Reasoning in LLMs

What's behind the "on/off" reasoning mode?

Karn Singh

Introduction

Artificial intelligence models like **GPT-4, Claude**, and **Llama-Nemotron** are increasingly being used to solve complex tasks—from advanced mathematics and coding to scientific reasoning. But how do these models learn to reason in a step-by-step, human-like manner?



- **OpenAI's O1** model was a trailblazer in this field, setting the foundation for a new wave of AI models focused on logical reasoning. Since then, many successors have emerged, each building on and enhancing the capabilities of previous models.

- One such model, **DeepSeek-R1**, has demonstrated significantly improved reasoning performance and output quality.

- **NVIDIA's Llama-Nemotron** introduces an innovative feature: a reasoning toggle. This allows users to control whether the model engages in deep, step-by-step reasoning or provides faster, more direct answers—adapting its responses based on the user's needs.

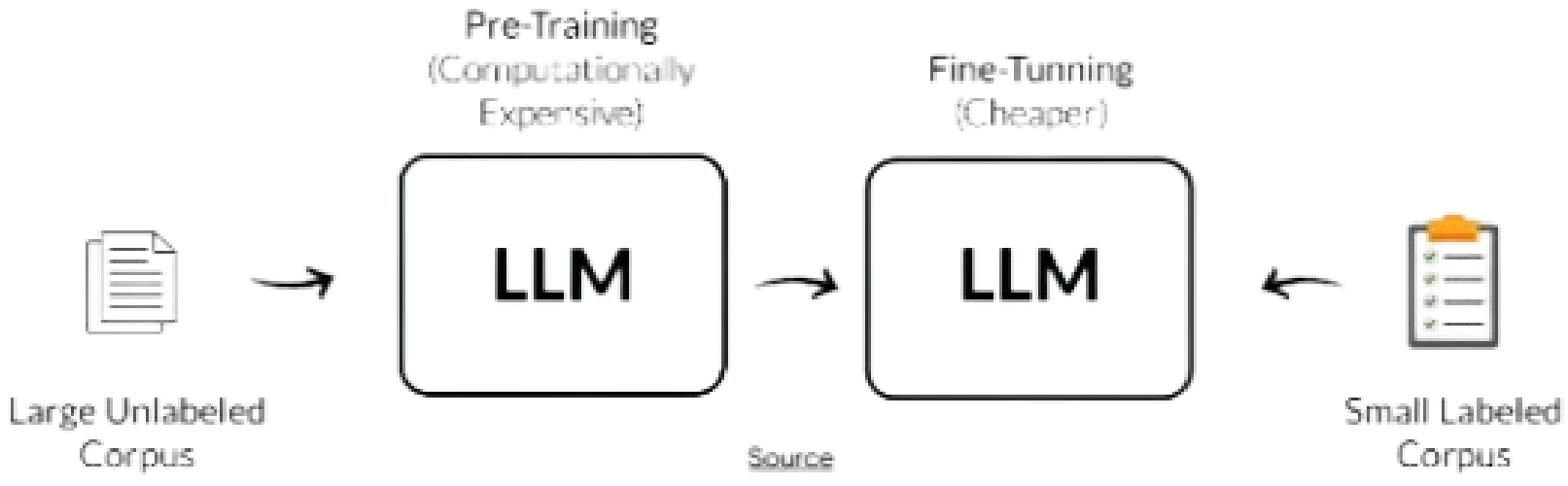
- In this guide, we'll explore how reasoning-focused AI models are trained and examine how Nemotron's reasoning toggle works under the hood to enhance flexibility and performance.

Knowledge Foundation: Pre-Training & Fine-Tuning

Before an AI model can reason, it must first acquire a vast amount of knowledge. Large Language Models (LLMs) achieve this by pre-training on trillions of words drawn from books, Wikipedia, programming code, and academic papers to understand the structure and meaning of language.

During this process, LLMs learn:

- Real-world facts (e.g., “Paris is the capital of France”)
- Logical relationships (e.g., “If A = B and B = C, then A = C”)
- Common language patterns and structures



Source

- Once pre-trained, the model undergoes fine-tuning, where it is further trained—often using labeled data or guidance from a more advanced model or human feedback—to specialize in tasks like reasoning, problem-solving, and nuanced conversation.

- For instance, NVIDIA's Llama-Nemotron enhances its reasoning capabilities by fine-tuning on guidance from a more capable teacher model, such as DeepSeek-R1.

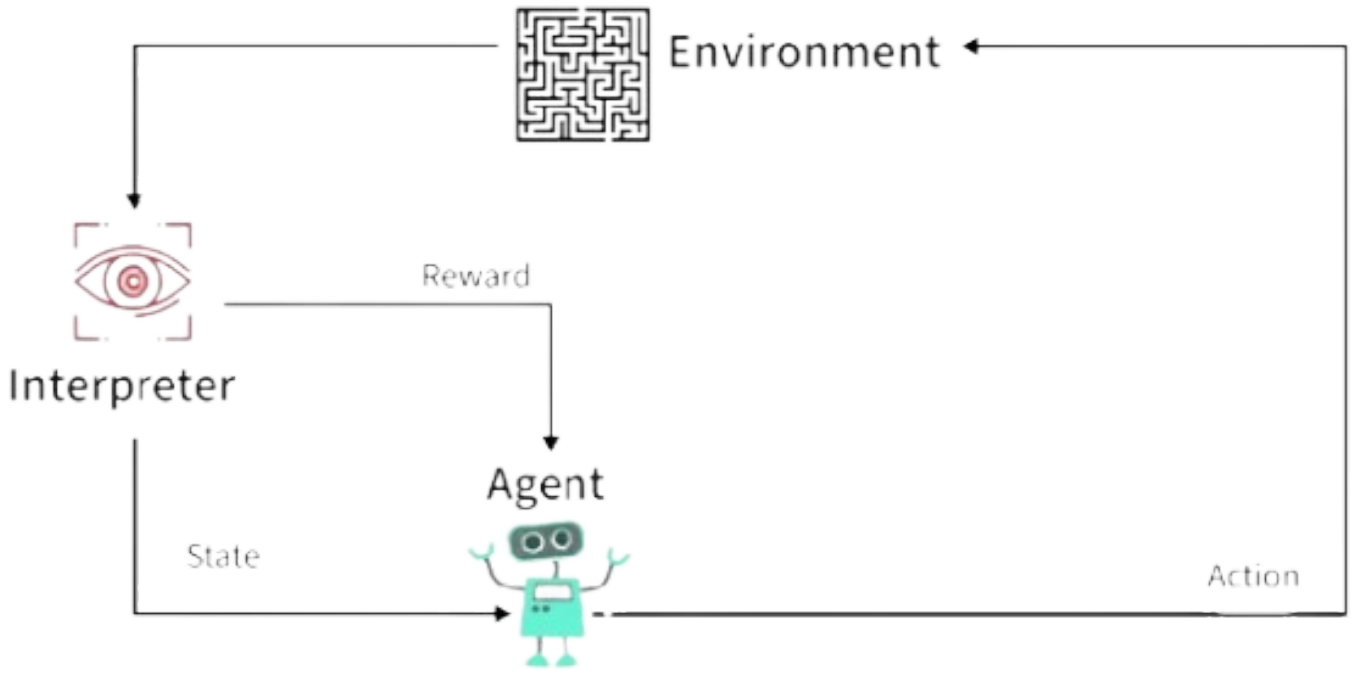
Reinforcement Learning: Teaching LLMs to Reason Better

- To develop strong reasoning abilities, large language models (LLMs) use **reinforcement learning (RL)**—a method where models learn from feedback and progressively improve. Here's how the process works:

- **Explore multiple solutions** (e.g., generate various methods to solve a calculus problem).

- **Receive detailed feedback** (e.g., "Solution 3 is mostly correct, but steps 1 and 5 need improvement").

- **Continuously refine** their reasoning strategy to get better over time



Why Reinforcement Learning is Crucial:

- Promotes **original thinking** rather than simple imitation.

- Helps eliminate **lazy shortcuts** in reasoning by addressing skipped or incorrect steps.

Dynamic Reasoning Toggle

NVIDIA's Llama-Nemotron is the first model for developers to introduce a reasoning toggle.
ON | REASONING

How it works:

1. Paired Training Data
NVIDIA built a specialized dataset where each query included two variations:

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

2. Supervised Fine-Tuning (SFT)

Through SFT, the model was trained to associate specific prompts (like "detailed thinking on/off") with the corresponding type of response.

What the toggle enables:

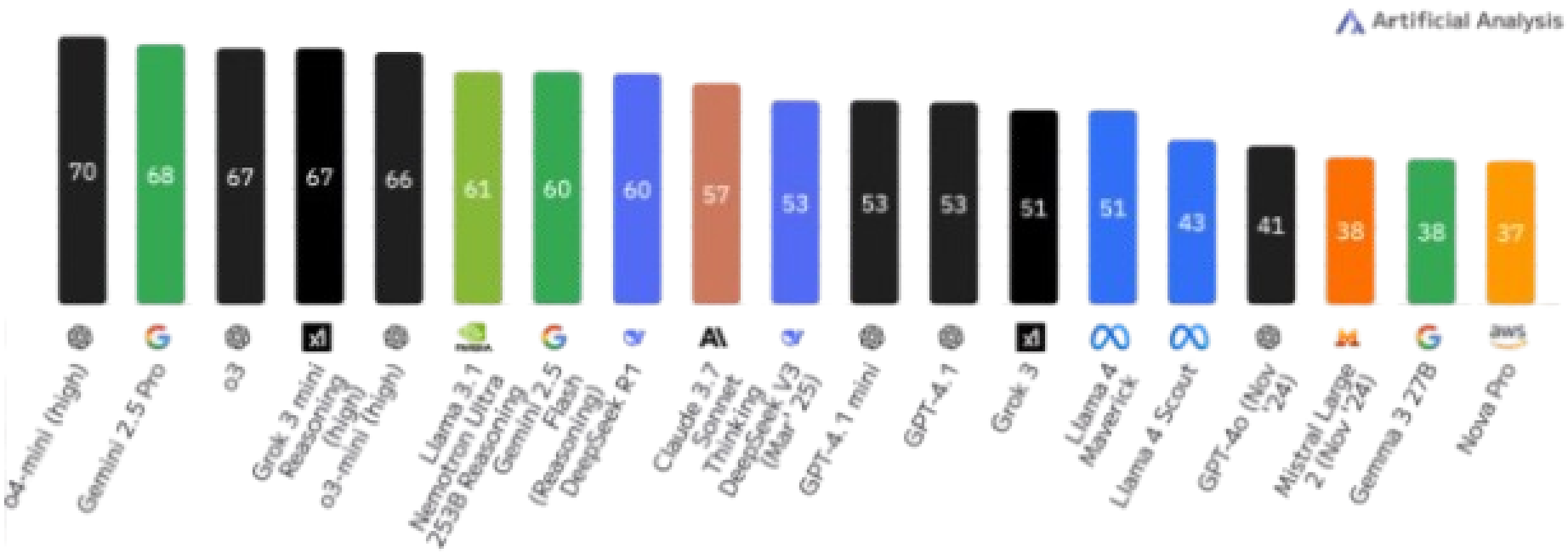
- When "**detailed thinking on**" is triggered, Llama-Nemotron activates its full reasoning abilities to provide thoughtful, step-by-step answers.

- When "**detailed thinking off**" is selected, it delivers quick, to-the-point responses with minimal reasoning.

The Impact of Modern LLMs

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

- Among open-source models, **Nvidia's Nemotron (LN Ultra)** ranks highest—positioning it as the most capable open model in the lineup.



Key Insights:

- **LN Ultra,** built on the **LLaMa 3.3** architecture, surpasses both its base model and **DeepSeek R1** in reasoning ability.

- It's optimized for performance and efficiency—running on a **single 8xH100 GPU node** while delivering **enhanced inference throughput.**

This trend highlights the rapid evolution of reasoning-based language models. With the arrival of **DeepSeek's R1**, and further fine-tuned architectures, we can expect even greater strides in model intelligence and efficiency in the near future.

DroneX AI

WAS THIS POST USEFUL?

FOLLOW FOR MORE!

REPOST