Reasoning Based LLMs from Scratch

VIZUARA AI LABS

July 8, 2025

1 Course Introduction

The course is titled "Reasoning Based LLMs from Scratch". It focuses on the process of answering questions that require complex multi-step reasoning with intermediate steps.

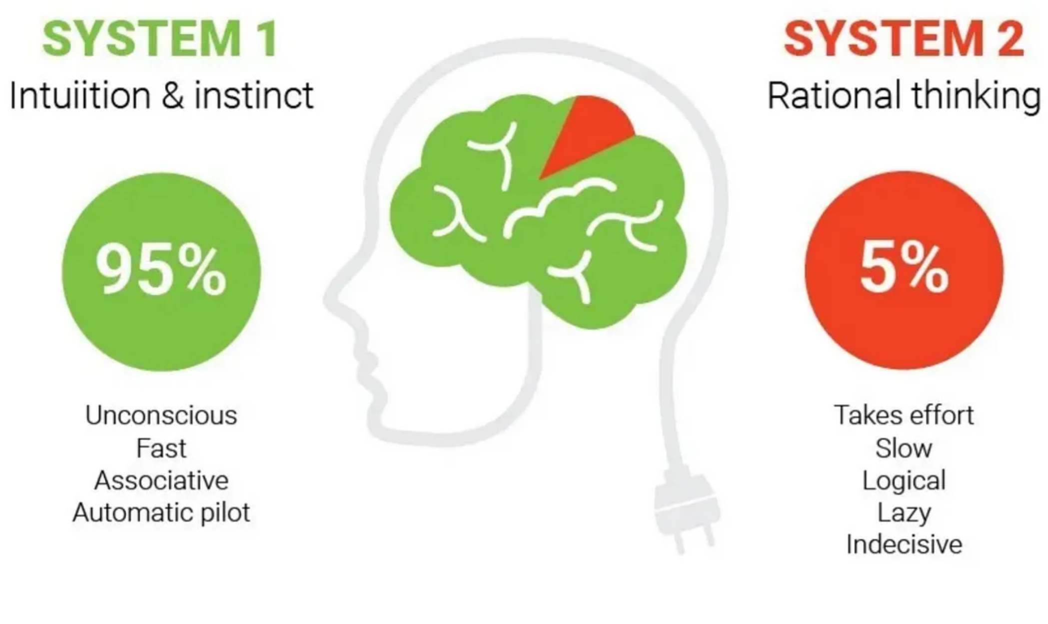
1.1 Instructor's Background and Philosophy

- **Education**: BTech, MTech from IIT Madras; PhD from Purdue University
- **Professional Background**: Co-Founder of VIZUARA AI LABS
- **Teaching Philosophy**: "If I cannot teach a subject to beginners, I haven't truly understood it myself." — R:D
- **Approach**: Simplify complex concepts to make them accessible.

2 Course Overview

2.1 Explanation of Reasoning in LLMs

Reasoning in LLMs refers to the ability to answer questions that require multi-step logical thinking, often involving intermediate steps to reach a conclusion.



2.2 System 1 vs. System 2 Reasoning

- **System 1 (Fast Reasoning)**: Intuitive, automatic, unconscious (e.g., ”What is the capital of India?”)
- **System 2 (Slow Reasoning)**: Rational, effortful, logical (e.g., ”Which movie to watch tonight?”)

2.3 Central Question: Can AI Reason?

AI is designed to answer like humans but not necessarily think like them. Early models like ChatGPT 3.5 excelled at System 1 tasks but struggled with System 2 tasks.

3 Progress in Reasoning AI

3.1 OpenAI’s o1 Model

Released on September 12, 2024, o1 marked a significant advancement in reasoning capabilities, evaluating preferences before making suggestions.

3.2 Industry Trends

All major LLM providers now offer reasoning models, trusted for their transparent thinking processes.

3.3 DeepSeek-R1 Model

Released on January 20, 2025, by DeepSeek, DeepSeek-R1 is open-source, comparable in accuracy to o1, and achieved reasoning through pure reinforcement learning.

4 DeepSeek-R1 Example

4.1 Problem and Solution

Problem: If $x > 1$, find the sum of the real solutions of $\sqrt{x\sqrt{x}} = x$.
Initial Steps:

1. Square both sides: $(\sqrt{x\sqrt{x}})^2 = x^2 \to x\sqrt{x} = x^2$
2. Rearrange: $\sqrt{x} = x$, then square again: $x = x^2$
3. Solve: $x^2 - x = 0 \to x = 0$ or $x = 1$

4.2 ”Aha Moment”

The model recognizes a need to reevaluate since $x > 1$, demonstrating an ”aha moment” where it allocates more thinking time to ensure correctness.

5 Course Objective

The course aims to teach students how to develop LLMs with reasoning capabilities, transitioning from a regular LLM to a reasoning LLM.

5.1 Flowchart Comparison

- **Regular LLM**: Question → Answer
- **Reasoning LLM**: Question → Thought 1 → Thought 2 → ... → Thought n → Answer

6 Example: Regular LLM vs. Reasoning LLM

Question: If a train travels at 60 mph for 3 hours, how far does it go?
Regular LLM Response: The train travels 180 miles.
Reasoning LLM Response:

1. Use the formula: Distance = Speed × Time
2. Calculate: $60 \times 3 = 180$ miles
3. Conclusion: The train travels 180 miles.

Key Difference: Reasoning LLMs show intermediate steps, demonstrating logical reasoning.

7 Course Structure

8 Module 1: Inference-Time Compute Scaling

Humans give better answers when they think for more time. Why? Because thinking step by step leads to a correct answer, especially for puzzles or complex math problems (e.g., Sudoku). For example, consider the question: Are there a finite or infinite number of prime numbers? A step-by-step reasoning process reveals the answer is infinite. Similarly, what happens if we make large language models (LLMs) think more before giving an answer?

8.1 Example: Tennis Balls Problem

Question: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 balls. How many tennis balls does he have now?
Regular LLM Response:

- Answer is 11
- Tokens used: 3

Reasoning LLM Response:

- Roger started with 5 balls (4 tokens)
- 2 cans of 3 tennis balls each is 6 tennis balls (11 tokens)
- $5 + 6 = 11$ (5 tokens)
- Answer is 11 (3 tokens)
- Total tokens used: 23

Insight: Forcing the model to think step-by-step uses more tokens (compute resources) during inference, potentially improving reasoning.

9 Test-Time Compute

Test-time compute refers to the computing resources used by the model during inference. For reasoning LLMs, this includes generating intermediate thoughts before the final answer.

9.1 Flowchart Representation

Regular LLM:

- Question $\rightarrow$ Regular LLM $\rightarrow$ Answer

Reasoning LLM:

- Question $\rightarrow$ Reasoning LLM $\rightarrow$ Thought 1 $\rightarrow$ Thought 2 $\rightarrow \cdots \rightarrow$ Thought n $\rightarrow$ Answer

Note: ”Test-time compute” includes the Reasoning LLM and the sequence of thoughts.

10 Inducing Reasoning During Inference

The first method to induce reasoning in LLMs is to allocate more computing resources during inference, without altering the model itself.

10.1 Comparison of Compute Resources

Regular LLMs:

- PRE-TRAINING $\rightarrow$ FINETUNING $\rightarrow$ INFERENCE (less compute resources)

Reasoning LLMs:

- PRE-TRAINING $\rightarrow$ FINETUNING $\rightarrow$ INFERENCE (more compute resources)

Insight: More ”thinking” during inference enhances reasoning.

11 Historical Context: Chain-of-Thought Prompting

In 2022, the Brain Team at Google Research introduced Chain-of-Thought Prompting to elicit reasoning in large language models.

11.1 Chain-of-Thought Prompting Example

Model Input:

- Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

- A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

- Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output:

- A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9.

11.2 Standard Prompting Example

Model Input:

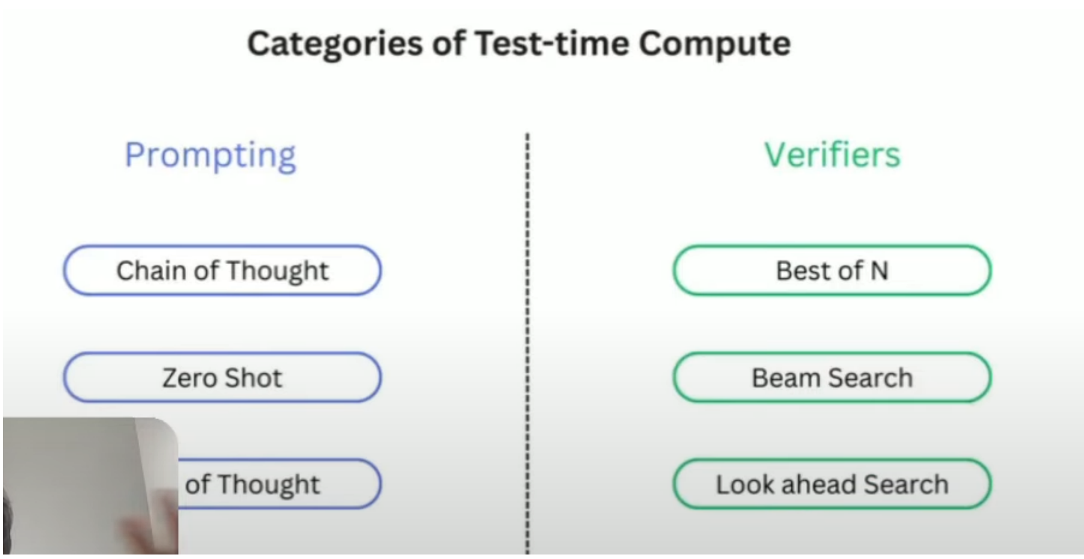
- Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
- A: The answer is 11.
- Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output:

- A: The answer is 27. (Incorrect)

12 Categories of Test-Time Compute Methods

Since Chain-of-Thought Prompting, several methods have been developed to induce reasoning during inference without modifying the model.



12.1 Prompting Methods

- Chain of Thought
- Zero Shot
- Tree of Thought

12.2 Verifiers

- Best of N
- Beam Search
- Look ahead Search

Note: These methods will be explored in detail throughout the course.

13 Module 2: Pure Reinforcement Learning

DeepSeek-R1's technical report highlighted that an intermediate version of DeepSeek-R1-Zero learned to "rethink" using pure reinforcement learning (RL). This module explores this capability through examples and structured learning approaches.

13.1 Example: Solving a Mathematical Equation

Question: If $a > 1$, find the sum of the real solutions of the equation

$$\sqrt{a - \sqrt{a + x}} = x.$$

Response:

$$\text{To solve } \sqrt{a - \sqrt{a + x}} = x, \text{ start by squaring both sides:}$$

$$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2.$$

Rearrange to isolate the inner square root:

$$\sqrt{a + x} = a - x^2.$$

Square both sides again:

$$\left(\sqrt{a + x}\right)^2 = \left(a - x^2\right)^2 \implies a + x = a^2 - 2ax^2 + x^4.$$

Rearrange into a polynomial:

$$x^4 - 2ax^2 - x + a^2 - a = 0.$$

Recognizing an "aha moment," reevaluate step-by-step:

$$\text{Return to } \sqrt{a - \sqrt{a + x}} = x, \text{ and square both sides:}$$

$$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2.$$

Proceed carefully, noting $a > 1$ and $x \geq 0$ (since $\sqrt{\cdot} \geq 0$).

Table 3: An interesting "aha moment" of an intermediate version of DeepSeek-R1-Zero. The model learns to rethink using an anthropomorphic tone, showcasing the power and beauty of reinforcement learning.

14 Understanding Reinforcement Learning with C-M-B Technique

We adopt a methodical approach called the C-M-B Technique: Classical, Modern, Build on your own, inspired by pioneers like Richard Sutton.

14.1 Classical RL Concepts

Reference: "Reinforcement Learning: An Introduction" by Richard S. Sutton and Andrew G. Barto, Second Edition—our bread and butter.

- What is RL? Elements of RL.
- Understanding agents, environment, rewards, returns
- Markov Decision Processes, Markov Property
- Value Functions, Optimal Value Functions
- Policy Evaluation, Policy Iterations, Policy and Value Iterations
- Monte Carlo Methods
- Moral Difference Learning
- Approximate-Solution Methods

14.2 Modern RL

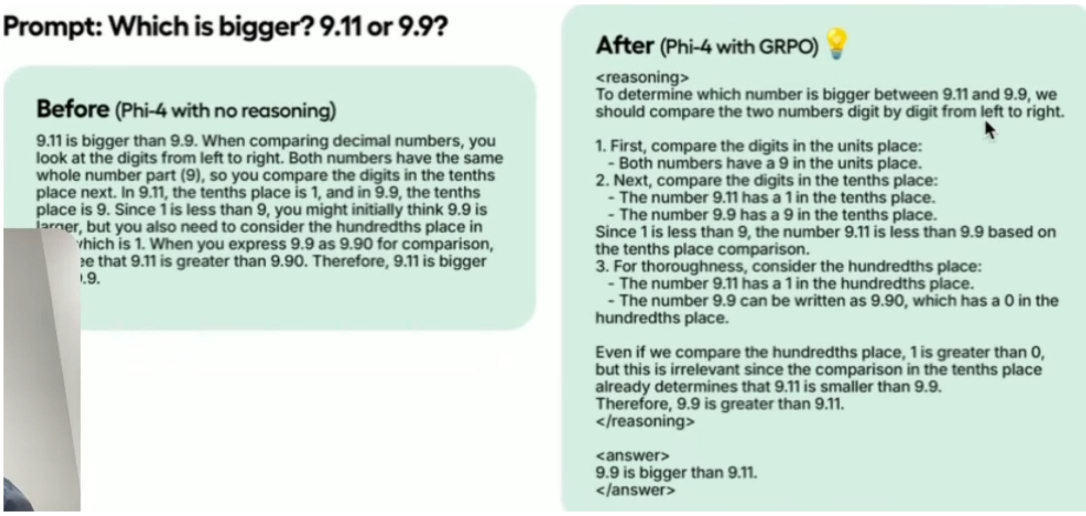
- Trust Region Policy Optimization (TRPO) – 2015
- Proximal Policy Optimization (PPO) – 2017
- Group Relative Policy Optimization (GRPO) – 2025

14.3 Build Your Own Reasoning Model

We will convert an open-source LLM (non-reasoning) into a reasoning model using pure reinforcement learning.

15 Example: Model Improvement with RL

Prompt: Which is bigger, 9.11 or 9.9?



Before (Phi-4 with no reasoning):

9.11 is bigger than 9.9. Both numbers share the same integer part, 9. Comparing the tenths place, 9.11 has a 1, while 9.9 has a 9. Since 1 is less than 9, one might initially consider 9.9 larger. However, examining the hundredths place, 9.11 has a 1, and 9.9, expressed as 9.90, has a 0. Thus, comparing 9.11 to 9.90, 9.11 appears greater.

After (Phi-4 with GRPO):

Reasoning: To compare 9.11 and 9.9, we examine digits step-by-step.

1. **Units place**: Both numbers have a 9.
2. **Tenths place**: 9.11 has a 1; 9.9 has a 9. Since $1 < 9$, 9.11 is less than 9.9.
3. **Hundredths place**: 9.11 has a 1; 9.9 (as 9.90) has a 0. This is irrelevant since the tenths place comparison is conclusive.

Answer: 9.9 is bigger than 9.11.

16 Method 3: Supervised Fine-Tuning and Reinforcement Learning

This method, likely used by OpenAI's o1 model (architecture undisclosed), serves as a blueprint for building reasoning models, as adopted by DeepSeek-R1.

16.1 Development Process of DeepSeek-R1

- Deepseek-V3 (87B)
- RL with accuracy & format rewards
- Deepseek-R1-Zero
- SFT ("cold start") data
- Train with "cold start" data
- RL with accuracy, format, and consistency rewards
- SFT (CoT) data
- SFT (knowledge) data
- Llama 3 & Qwen 2.5
- RL with rule-based verification (math, code) and human preference
- Deepseek-R1
- Deepseek-R1-Distill-Qwen (1.5B – 32B)
- Deepseek-R1-Distill-Llama (8B & 70B)

17 Method 4: Pure Supervised Fine-Tuning and Distillation

Distillation transfers knowledge from a large model to a smaller one to instill reasoning capabilities.

17.1 Distillation Process

- Large Model (with reasoning capabilities)
- Generate dataset (input-output pairs)
- Fine-tune
- Small Model (No reasoning capabilities)
- Small Model (with reasoning capabilities)

Analogy:

- Dataset from Large model with reasoning capability
- Small model without reasoning capability
- Small model with reasoning capability

In this section, we will train our own reasoning model using this distillation process.

18 Summary of Methods

This module covers four key methods for developing reasoning LLMs:

1. Inference-Time Compute Scaling
2. Pure Reinforcement Learning (Hands-on)
3. Supervised Fine-Tuning and Reinforcement Learning
4. Pure Supervised Fine-Tuning and Distillation (Hands-on)