Method 1: Inference-Time Compute Scaling

VIZUARA AI LABS

July 9, 2025

1 Introduction

In this lecture, we explore the first method to import reasoning capabilities into large language models (LLMs): **Inference-Time Compute Scaling**. Humans provide more accurate answers when they allocate more time to think. Why? Because step-by-step reasoning often leads to correct solutions, especially for complex problems like puzzles (e.g., Sudoku) or mathematical questions (e.g., determining whether there are a finite or infinite number of prime numbers). This increased thinking time consumes more mental energy. Similarly, what happens if we make LLMs "think" more before answering?

1.1 Example: Tennis Balls Problem

Question: Roger has 5 tennis balls. He buys 2 more cans of tennis balls, each containing 3 balls. How many tennis balls does he have now?

Regular LLM Response:

- Answer: 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: 11 (3 tokens).
- Total tokens used: 23

Insight: Forcing the model to show step-by-step thinking increases token usage (i.e., compute resources) during inference, enhancing reasoning accuracy.

1.2 Flowchart: Regular vs. Reasoning LLM

2 Test-Time Compute

Test-time compute refers to the computing resources utilized by the model during inference. For reasoning LLMs, this includes generating intermediate thoughts before producing the final answer. The first method to induce reasoning is to allocate more computing resources during inference without altering the model itself.

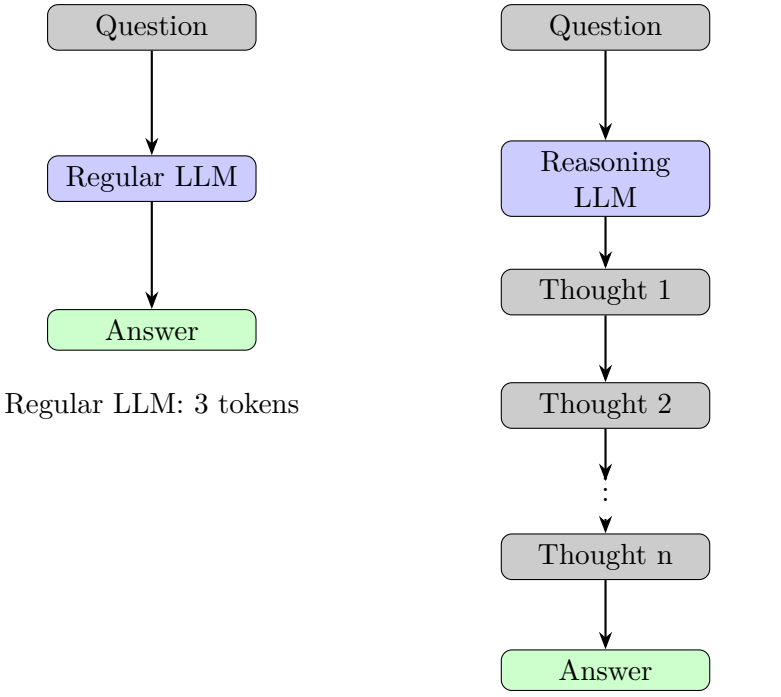

Figure 1: Regular LLM vs. Reasoning LLM: Test-time compute includes intermediate thoughts.

2.1 Pipeline Comparison

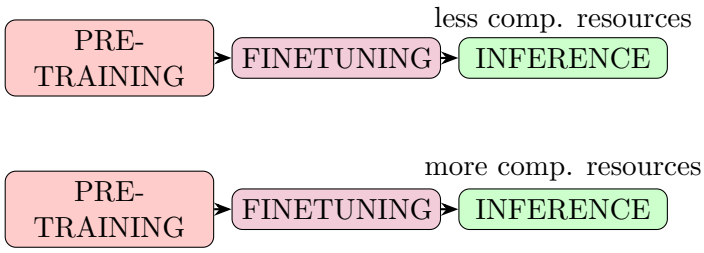

Figure 2: Regular LLMs use fewer compute resources during inference compared to Reasoning LLMs.

Insight: More "thinking" during inference induces reasoning, leading to increased model accuracy. This accuracy scales with test-time compute, hence the name *Inference-Time Compute Scaling*. Giving the model more time to think before answering results in better answers, aligning with intuitive human reasoning processes.

3 Historical Context: Chain-of-Thought Prompting

The idea of enhancing LLM reasoning gained traction in 2022 when the Google Research Brain Team introduced **Chain-of-Thought (CoT) Prompting** in the paper *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* by Jason Wei, Xuezhi Wang, Dale Schuurmans, Maurice Bosma, Brian Ichter, Fei Xia, Quoc V. Le, and Denny Zhou. The paper states:
"We explore how generating a chain of thought—a series of intermediate reasoning steps—significantly improves the ability of large language models to perform complex reasoning. In

particular, we show how such reasoning abilities emerge naturally in sufficiently large language models via a simple method called chain-of-thought prompting, where a few chain of thought demonstrations are provided as exemplars in prompting."

3.1 Key Ideas

- **Idea 1**: Arithmetic reasoning benefits from generating natural language rationales leading to the final answer.
- **Idea 2**: Prompting with input-output examples demonstrating the task can unlock reasoning abilities in LLMs.

3.2 Standard vs. Chain-of-Thought Prompting

Standard Prompting:

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?

Output:

- A: The answer is 27. (Incorrect)

Chain-of-Thought Prompting:

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?

Output:

- A: The cafeteria had 23 apples originally. They used 20, so $23 - 20 = 3$. They bought 6 more, so $3 + 6 = 9$. The answer is 9. (Correct)

3.3 Chain-of-Thought Reasoning Flow

3.4 Impact of the Paper

With over 12,000 citations, the paper demonstrated that sufficiently large models can generate chains of thought when provided with few-shot CoT demonstrations. It evaluated CoT for:

- Arithmetic reasoning
- Commonsense reasoning
- Symbolic reasoning

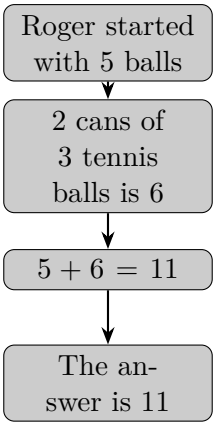

Figure 3: Chain-of-Thought Reasoning for the tennis balls problem.

Few-shot prompting existed prior, but this was the first study to incorporate chain-of-thought in prompting.

4 Examples from the Paper

The paper provides several examples of CoT prompting across different reasoning tasks:

4.1 Math Word Problems (Free Response)

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.

4.2 Math Word Problems (Multiple Choice)

Q: How many keystrokes are needed to type the numbers from 1 to 500?
Options: (a) 1156, (b) 1392, (c) 1480, (d) 1562, (e) 1788
A: There are 9 one-digit numbers (1 to 9). There are 90 two-digit numbers (10 to 99). There are 401 three-digit numbers (100 to 500). So, $9 \times 1 + 90 \times 2 + 401 \times 3 = 9 + 180 + 1203 = 1392$. The answer is (b).

4.3 Commonsense Reasoning (CSQA)

Q: Sammy wanted to go to where the people were. Where might he go?
Options: (a) race track, (b) populated areas, (c) desert, (d) apartment, (e) roadblock
A: Populated areas have a lot of people, unlike race tracks, deserts, apartments, or roadblocks. The answer is (b).

4.4 StrategyQA

Q: Yes or no: Would a pear sink in water?
A: The density of a pear is about 0.6 g/cm^3, which is less than water. Thus, a pear would float. The answer is no.

4.5 Date Understanding

Q: The concert was scheduled for 06/01/1943 but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY?
A: One day after 06/01/1943 is 06/02/1943. Ten days before 06/02/1943 is 05/23/1943. The answer is 05/23/1943.

4.6 Sports Understanding

Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship."
A: Joao Moutinho is a soccer player, not involved in American football's NFC championship. The answer is no.

5 Results and Observations

The paper's results highlight the effectiveness of CoT prompting:

5.1 GSM8K Dataset

- *Reference*: https://huggingface.co/datasets/openai/gsm8k (MIT License)
- *Description*: Contains 8.5K high-quality, linguistically diverse grade school math word problems.
- **Key Observations**:
  - Smaller models exhibit lower accuracy.
  - Larger models (e.g., PaLM) achieve accuracy comparable to supervised fine-tuning.
  - CoT prompting outperforms standard prompting on arithmetic datasets.
  - CoT reasoning is an emergent ability with increasing model scale.

5.2 Error Analysis

For a 62B model:

- Semantic understanding errors: 20 (540B fixes 6)
- One step missing errors: 18 (540B fixes 12)
- Other errors: 7 (540B fixes 4)

Larger models with CoT prompting fix more errors.

5.3 Example: Error Correction

Question: Tracy used a 4-foot-long wire to support tomato plants, cut into 6-inch pieces. How many pieces did she obtain?

62B Model Output: She cut the wire into 6-inch pieces. This means she got $4 \times 6 = 24$ pieces. The answer is 24. (Incorrect: semantic understanding error)
540B Model Output: The wire was 4 feet long, or $4 \times 12 = 48$ inches. Cut into 6-inch pieces, she obtained $48/6 = 8$ pieces. The answer is 8. (Correct)

Question: Tom's ship travels at 10 mph from 1 to 4 PM, then returns at 6 mph. How long does it take to travel back?

62B Model Output: He travels from 1 to 4 PM, or 3 hours. He travels back at 6 mph. Thus, $3 \times 6 = 18$ miles. The answer is 18. (Incorrect: semantic understanding error)
540B Model Output: He travels 3 hours at 10 mph, covering $3 \times 10 = 30$ miles. Returning at 6 mph, it takes $30/6 = 5$ hours. The answer is 5. (Correct)

5.4 Performance Graphs

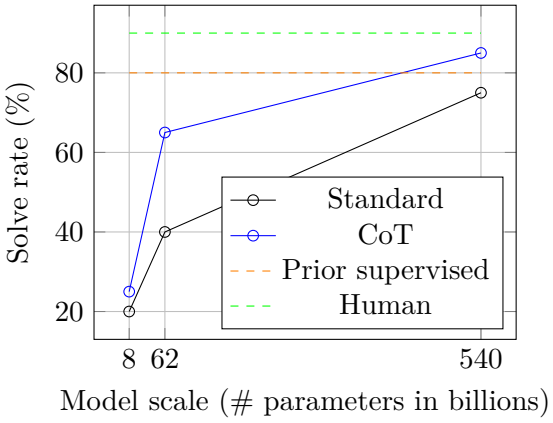

Figure 4: CSQA Dataset: CoT prompting outperforms standard prompting as model scale increases.

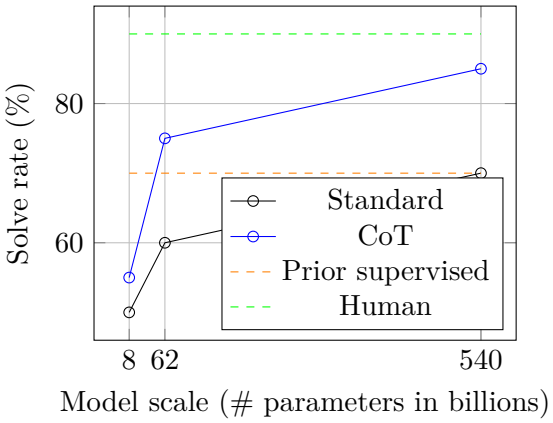

Figure 5: StrategyQA Dataset: CoT improves performance significantly.

6 Commonsense vs. Symbolic Reasoning

Commonsense Reasoning: Involves intuitive knowledge about the world.

> **Example**: Would a steel ball sink in water? (Answer: Yes, due to its high density compared to water.)

Symbolic Reasoning: Involves manipulating symbols or states according to rules.

> **Example**: A coin is heads up. Akshay flips it. Deepti does not flip it. Is it still heads up? (Answer: No, flipping changes the state.)

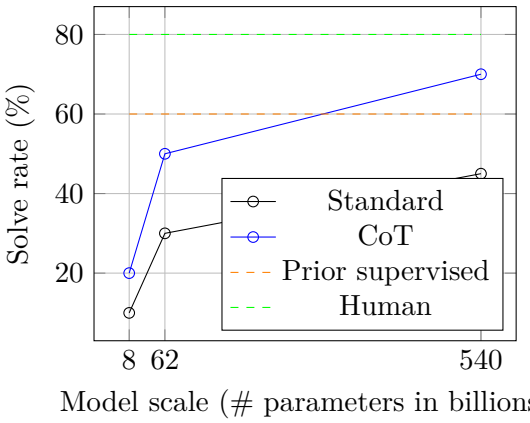

Figure 6: Date Dataset: CoT improves performance significantly.

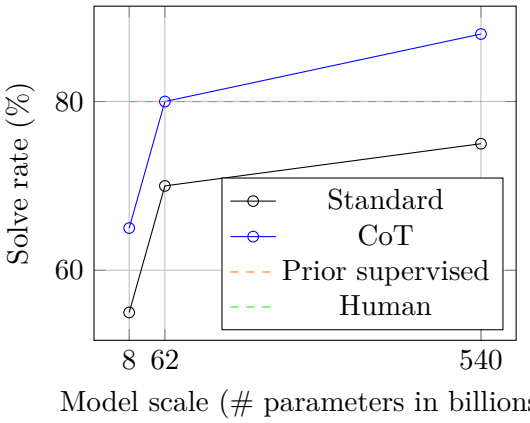

Figure 7: Sports Dataset: CoT improves performance significantly.

7 Main Result

Chain-of-Thought prompting significantly enhances performance in arithmetic, commonsense, and symbolic reasoning tasks, particularly as model scale increases, making it a powerful method for inducing reasoning in LLMs.

8 Additional Exploration: Zero-Shot CoT

A related paper, *Large Language Models are Zero-Shot Reasoners* by Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa, introduces **Zero-shot CoT**:

8.1 Example

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
Zero-shot: The answer is 8. (Incorrect)
Zero-shot CoT: Let's think step by step. There are 16 balls in total. Half are golf balls, so 8 golf balls. Half of those are blue, so 4 blue golf balls. The answer is 4. (Correct)

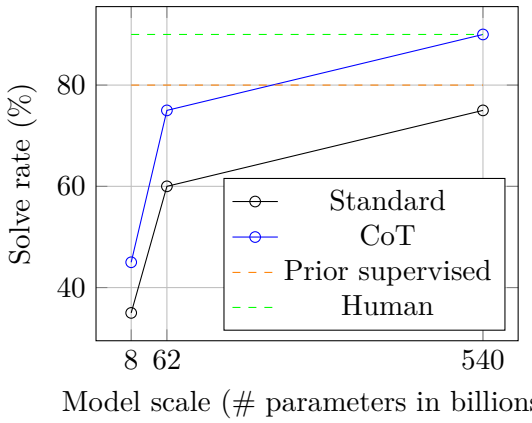

Figure 8: SayCan Dataset: CoT improves performance significantly.

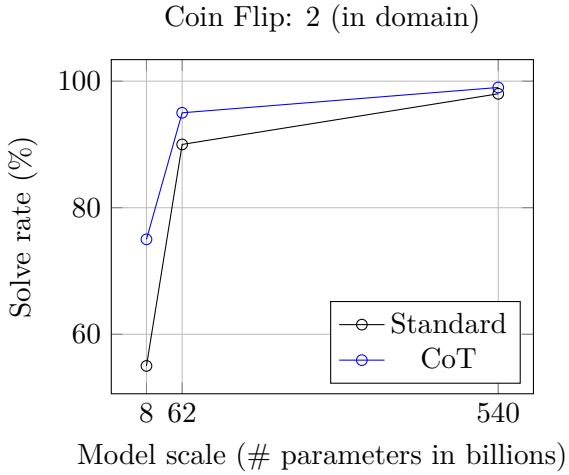

Figure 9: Coin Flip: 2 (in domain): CoT achieves near-perfect solve rates at larger scales.

8.2 Key Points

- **Why "Zero"**: No input-output examples are provided in the prompt.
- **Process**: Uses two prompts:
  1. "Let's think step by step" to extract reasoning.
  2. "Therefore, the answer is" to extract the final answer.
- **Example (Two-Step Prompting)**:
  - **Q**: On average, Joe throws 25 punches per minute. A fight lasts 5 rounds of 3 minutes. How many punches did he throw?
  - **First Prompt (Reasoning Extraction)**: Let's think step by step.
    - In one minute, Joe throws 25 punches.
    - In one round of 3 minutes, Joe throws $3 \times 25 = 75$ punches.
    - In five rounds, Joe throws $5 \times 75 = 375$ punches.
  - **Second Prompt (Answer Extraction)**: Therefore, the answer is 375.
- **Results**: Zero-shot CoT outperforms baseline zero-shot prompting, with larger models (e.g., PaLM) showing significantly higher accuracy on GSM8K (42–43% for Zero-shot CoT

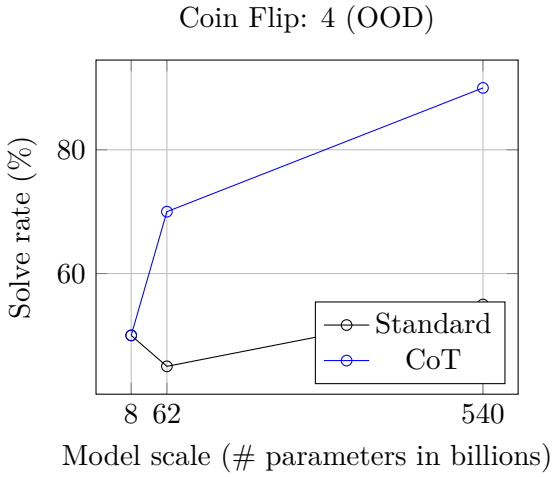

Figure 10: Coin Flip: 4 (OOD): CoT significantly improves performance on out-of-domain tasks.

vs. 12–13% for Zero-shot at 540B).

9 Hands-on Projects

9.1 Project 1: Evaluating the Effect of Model Size on CoT Reasoning

- **Models**: Flan-T5 Small (80M), Flan-T5 Base (250M), Flan-T5 Large (800M), Tiny Llama (1.1B), Phi-2 (2.7B), Hygr-7B (7B)
- **Dataset**: GSM8K (Arithmetic Reasoning)