Towards Reasoning in Large Language Models

Jie Huang **Kevin C.C. Chang**

University of Illinois at Urbana-Champaign

61 ACL 2023

COMPUTER SCIENCE

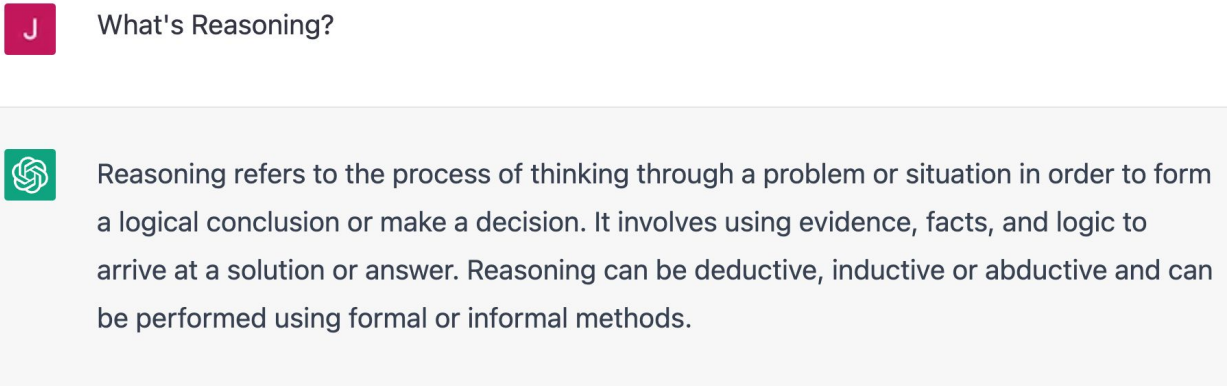
Humans vs *Machines*

- Humans can handle a lot of tasks, even with only a few examples. **[Few-shot Learning Ability]**
- Humans possess the ability to generalize beyond familiar scenarios. **[Out-of-distribution Robustness]**
- Humans are capable of providing explanations for their decisions or predictions, whereas machines (especially deep neural networks) are often considered "black box" with limited explainability. **[Explainability]**

Humans are able to reason, while reasoning is frequently considered missed in machines.

The summary was borrowed from the discussion with Denny Zhou.

What's Reasoning?



Reasoning refers to the process of thinking through a problem or situation in order to form a logical conclusion or make a decision. It involves using evidence, facts, and logic to arrive at a solution or answer. Reasoning can be ***deductive***, ***inductive*** or ***abductive*** and can be performed using ***formal*** or ***informal*** methods.

https://chat.openai.com/chat

Common Types of Reasoning

- **Deductive Reasoning:**
  - Premise: All mammals have kidneys.
  - Premise: All whales are mammals.
  - Conclusion: All whales have kidneys.
- **Inductive Reasoning:**
  - Observation: Every time we see a creature with wings, it is a bird.
  - Observation: We see a creature with wings.
  - Conclusion: The creature is likely to be a bird.
- *Abductive Reasoning, analogical reasoning, causal reasoning, probabilistic reasoning…*

Formal Reasoning vs *Informal* Reasoning

- **Formal Reasoning** is a systematic and logical process that follows a set of rules and principles, often used in mathematics and logic.

more structured and reliable

- **Informal Reasoning** is a less structured approach that relies on intuition, experience, and common sense to draw conclusions and solve problems, and is often used in everyday life.

more adaptable and open-ended

Reasoning in Language Models

usually no clear definition

focus on **informal deductive reasoning** — a widely used form in which the conclusion is guaranteed to be true as long as the premises are true

Reasoning in Large Language Models

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.

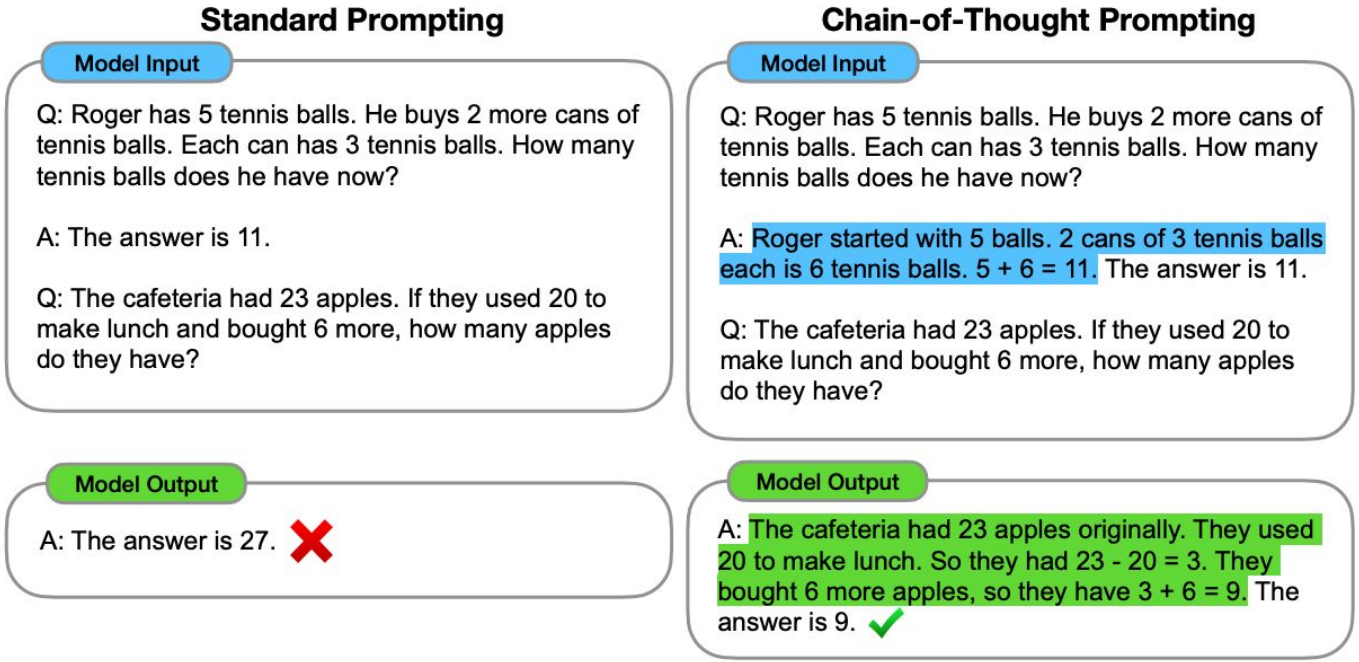
Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. 

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. NeurIPS 2022.

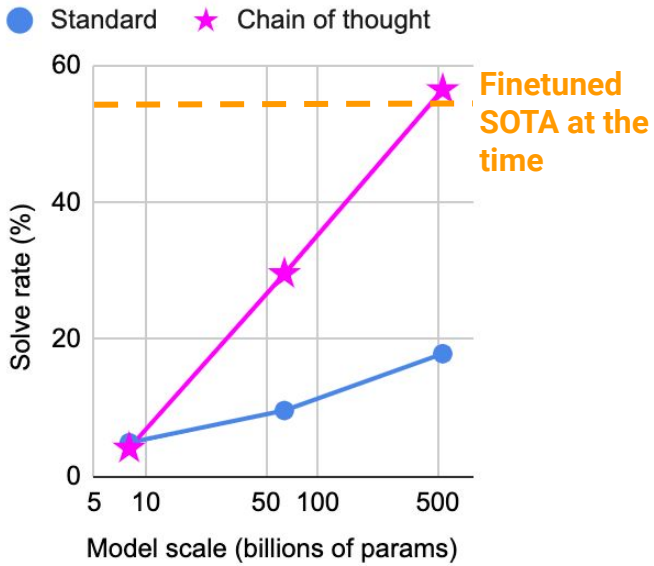
Chain-of-Thought (CoT) Prompting



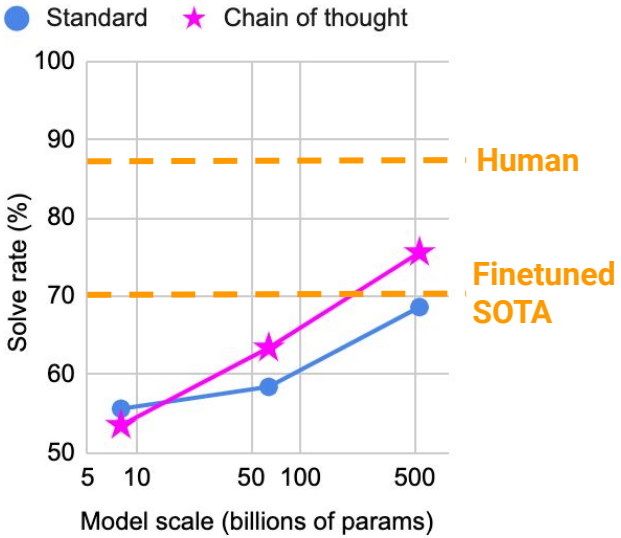
Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. NeurIPS 2022.

Chain-of-thought elicits “reasoning” of LLMs

GSM8K



StrategyQA



Real model output #1

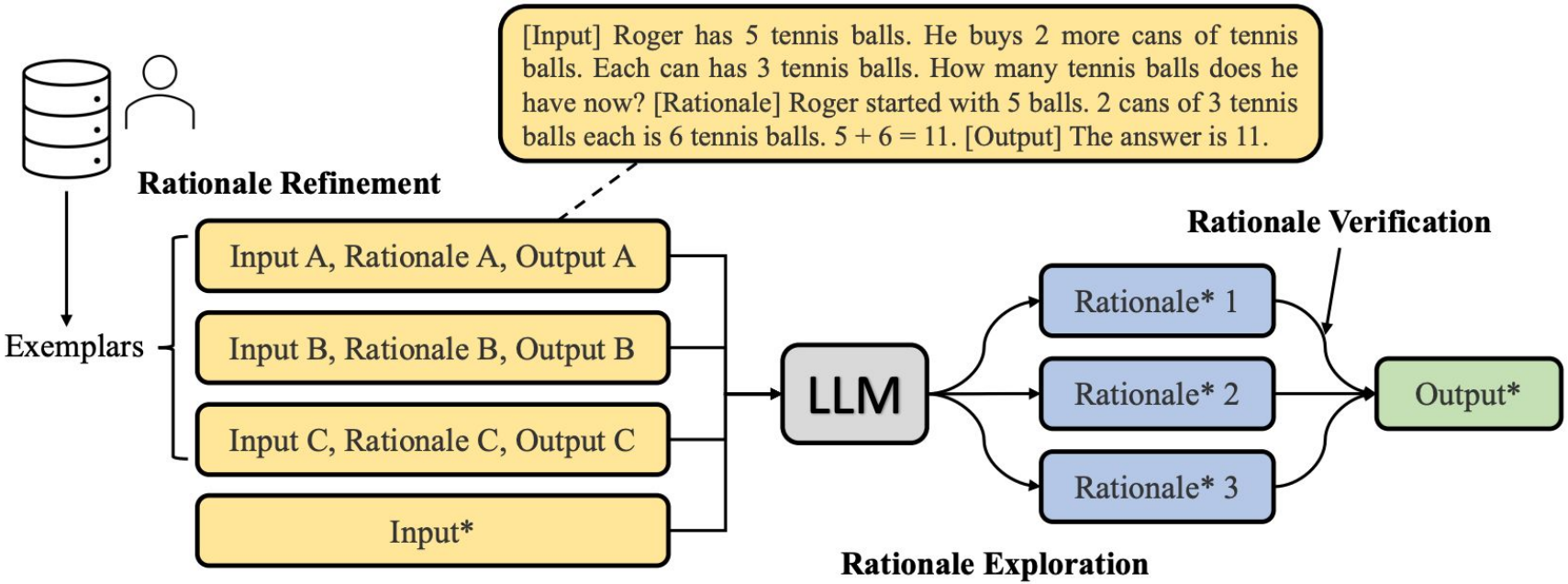
Q: Can you hide a basketball in a sand cat's ear?

A: A basketball is about 30 cm in diameter. A sand cat's ear is about 2 cm in diameter. Thus, a basketball would not fit in a sand cat's ear. So the answer is no.

(Copy from Jason’s slides)

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. NeurIPS 2022.

Rationale Engineering



Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. 2023.

Why Reasoning?

- **Few-shot Learning Ability⬆️**
- **Out-of-distribution Robustness⬆️**
- **Explainability⬆️**
  - Humans can handle a lot of tasks, even with only a few examples. **[Few-shot Learning Ability]**
  - Humans possess the ability to generalize beyond familiar scenarios. **[Out-of-distribution Robustness]**
  - Humans are capable of providing explanations for their decisions or predictions, whereas machines (especially deep neural networks) are often considered “black box” with limited explainability. **[Explainability]**

Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. 2023.

Are language models really able to reason?

There are several indications that LLMs are able to reason, including

- **high performance on various tasks requiring reasoning** (Suzgun et al., 2022)
- **the ability to reason step-by-step with chain-of-thought prompting** (Wei et al., 2022)
- **the reflection of human-like content effects on reasoning** (Dasgupta et al., 2022)

Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. 2023.

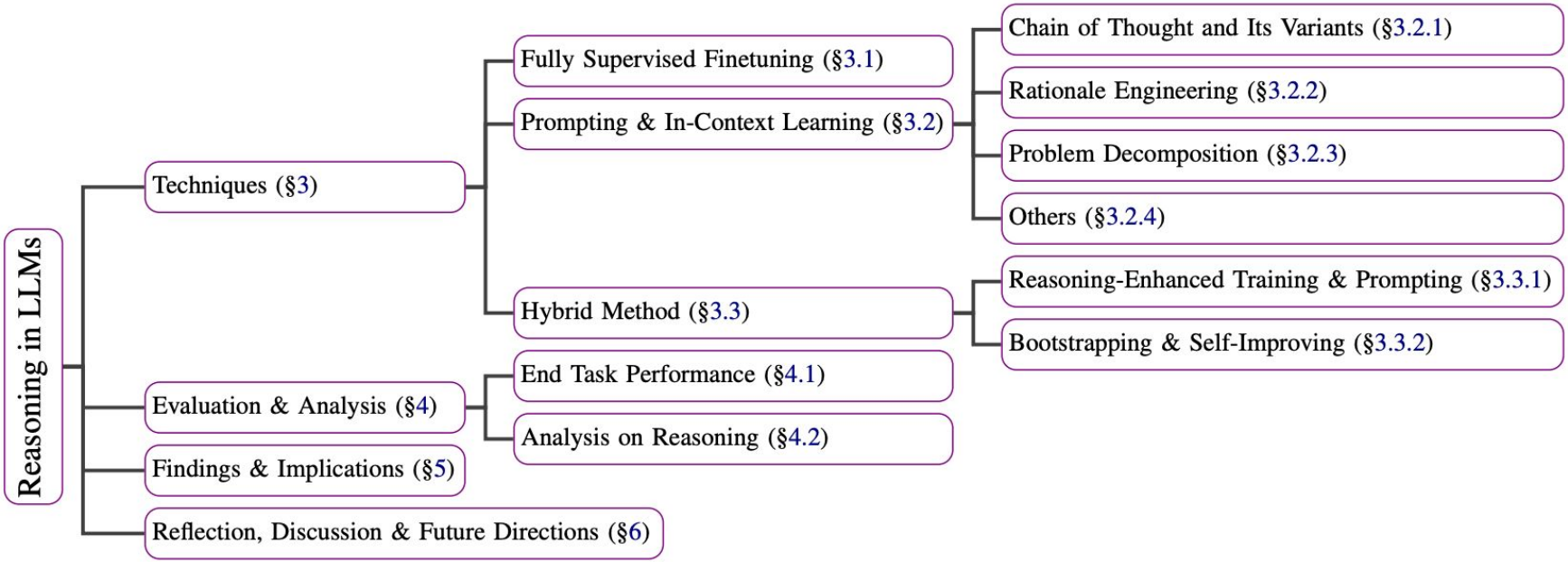
Are language models really able to reason?

There are several observations that suggest LLMs may not be capable of reasoning:

- **LLMs still struggle with tasks that require complex reasoning** (Valmeekam et al., 2022; Han et al., 2022; Ruis et al., 2022)
- **LLMs make mistakes in their reasoning**
- **The performance of LLMs has been found to be sensitive to the frequency of certain terms** (Razeghi et al., 2022; Jung et al., 2022)
- **Language models have been found to struggle with associating relevant information that they have memorized** (Huang et al., 2022)

Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. 2023.

Reasoning in Large Language Models



Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. 2023.

Email: jeffhj@illinois.edu
Twitter: @jefffhj

Towards Reasoning in Large Language Models: A Survey

Jie Huang **Kevin Chen-Chuan Chang**
Department of Computer Science, University of Illinois at Urbana-Champaign
{jeffhj, kcchang}@illinois.edu

Survey

Github

Abstract

Reasoning is a fundamental aspect of human intelligence that plays a crucial role in activities such as problem solving, decision making, and critical thinking. In recent years, large language models (LLMs) have made significant progress in natural language processing, and there is observation that these models may exhibit reasoning abilities when they are sufficiently large. However, it is not yet clear to what extent LLMs are capable of reasoning. This paper provides a comprehensive overview of the current state of knowledge on reasoning in LLMs, including techniques for improving and eliciting reasoning in these models, methods and benchmarks for evaluating reasoning abilities, findings and implications of previous research in this field, and suggestions on future directions. Our aim is to provide a detailed and up-to-date review of this topic and stimulate meaningful discussion and future work.[1]

they are large enough (Wei et al., 2022a). For example, by providing the models with "*chain of thoughts*", i.e., reasoning exemplars, or a simple prompt "*Let's think step by step*", these models are able to answer questions with explicit reasoning steps (Wei et al., 2022b; Kojima et al., 2022), e.g., "all whales are mammals, all mammals have kidneys; therefore, all whales have kidneys." This has sparked considerable interest in the community since reasoning ability is a hallmark of human intelligence that is frequently considered missed in current artificial intelligence systems (Marcus, 2020; Russin et al., 2020; Mitchell, 2021; Bommasani et al., 2021).

However, despite the strong performance of LLMs on certain reasoning tasks, it remains unclear whether LLMs are actually reasoning and to what extent they are capable of reasoning. For example, Kojima et al. (2022) claim that "LLMs are decent zero-shot reasoners (p. 1)", while Valmeekam

Thank Jason Wei and Denny Zhou for their valuable advice and constructive feedback on this work!