Towards Reasoning in Large Language Models

Jie Huang    Kevin C.C. Chang
University of Illinois at Urbana-Champaign
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.
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.
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. ✗
Chain-of-Thought (CoT) Prompting

### Standard Prompting

**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. ✗

### Chain-of-Thought Prompting

**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. ✓
**GSM8K**

- Standard
- Chain of thought

Finetuned SOTA at the time

**StrategyQA**

- Standard
- Chain of thought

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)

Rationale Engineering

[Input] 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? [Rationale] Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. [Output] The answer is 11.
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]

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)

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)
Reasoning in Large Language Models

Techniques (§3)
- Fully Supervised Finetuning (§3.1)
- Prompting & In-Context Learning (§3.2)
- Hybrid Method (§3.3)
- End Task Performance (§4.1)
- Analysis on Reasoning (§4.2)

Evaluation & Analysis (§4)
- Chain of Thought and Its Variants (§3.2.1)
- Rationale Engineering (§3.2.2)
- Problem Decomposition (§3.2.3)
- Others (§3.2.4)
- Reasoning-Enhanced Training & Prompting (§3.3.1)
- Bootstrapping & Self-Improving (§3.3.2)

Findings & Implications (§5)

Reflection, Discussion & Future Directions (§6)

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

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. 14)” while Valmof...

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