Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, Denny Zhou

Google Research, Brain Team

Presented by:

Anny Dai, Yixin Chen

Feb 7, 2025
Introduction

**Problem:** scaling up large language model size alone has not proved sufficient for achieving high performance on challenging tasks such as arithmetic, commonsense, and symbolic reasoning

**Two Ideas:**

1. Rationale-augmented training and fine tuning methods: costly to create a large set of high quality rationales
2. Few shot prompting method: works poorly on tasks that require reasoning abilities, and often does not improve substantially with increasing language model scale

⇒ **Chain-of-thought prompting:** a combination of the two ideas

An approach where a sequence of intermediate natural language reasoning steps are generated, leading to the final output.
Introduction

Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.
Key Features:

**Decomposition**: Breaks down complex problems into manageable steps, allowing for targeted computation on each component.

**Interpretability**: Provides insight into how the model processes and arrives at an answer, offering a way to trace and debug the reasoning path.

**Applicability**: Useful across various domains including arithmetic, commonsense, and symbolic reasoning tasks.

**Implementation**: Can be activated in large pre-trained models through few-shot prompting with exemplars that demonstrate chain-of-thought reasoning.

---

**Figure 3**: Examples of (input, chain of thought, output) triples for arithmetic, commonsense, and symbolic reasoning benchmarks. Chains of thought are highlighted. Full prompts in Appendix G.
Arithmetic Reasoning

Experiment Setup

- **Benchmarks Overview:**
  - Variety of Math Problems: Includes GSM8K, SVAMP, ASDiv, AQuA, and MAWPS for a comprehensive assessment.

- **Prompting Approaches:**
  - Standard Few-shot Prompting vs. Chain-of-Thought Prompting

- **Language Models Tested:**
  - Range of Models: Includes GPT-3, LaMDA, PaLM, UL2 20B, and Codex, showcasing a spectrum from 350M to 540B parameters.

| Dataset  | $N$ | Example problem |
|----------|-----|-----------------|
| GSM8K    | 1,319 | Josh decides to try flipping a house. He buys a house for $80,000 and then puts in $50,000 in repairs. This increased the value of the house by 150%. How much profit did he make? |
| SVAMP    | 1,000 | Each pack of dvds costs 76 dollars. If there is a discount of 25 dollars on each pack. How much do you have to pay to buy each pack? |
| ASDiv    | 2,096 | Ellen has six more balls than Marin. Marin has nine balls. How many balls does Ellen have? |
| AQuA     | 254  | A car is being driven, in a straight line and at a uniform speed, towards the base of a vertical tower. The top of the tower is observed from the car and, in the process, it takes 10 minutes for the angle of elevation to change from 45° to 60°. After how much more time will this car reach the base of the tower? Answer Choices: (a) $5\sqrt{3} + 1$ (b) $6\sqrt{3} + \sqrt{2}$ (c) $7\sqrt{3} - 1$ (d) $8\sqrt{3} - 2$ (e) None of these |
| MAWPS: SingleOp | 562 | If there are 7 bottle caps in a box and Linda puts 7 more bottle caps inside, how many bottle caps are in the box? |
| MAWPS: SingleEq | 508 | Benny bought a soft drink for 2 dollars and 5 candy bars. He spent a total of 27 dollars. How much did each candy bar cost? |
| MAWPS: AddSub | 395 | There were 6 roses in the vase. Mary cut some roses from her flower garden. There are now 16 roses in the vase. How many roses did she cut? |
| MAWPS: MultiArith | 600 | The school cafeteria ordered 42 red apples and 7 green apples for students lunches. But, if only 9 students wanted fruit, how many extra did the cafeteria end up with? |
Arithmetic Reasoning

Key Findings

1. **Scale Matters**: The effectiveness of chain-of-thought prompting increases with the model size.

2. **Greater Gains on Complex Problems**: Chain-of-thought prompting significantly boosts performance on complex arithmetic problems, especially in larger models like GPT and PaLM.

3. **Surpassing Previous Benchmarks**: Using chain-of-thought prompting, large models like GPT-3 175B and PaLM 540B have exceeded previous state-of-the-art performances on several challenging benchmarks.

Figure 4: Chain-of-thought prompting enables large language models to solve challenging math problems. Notably, chain-of-thought reasoning is an emergent ability of increasing model scale. Prior best numbers are from Cobbe et al. (2021) for GSM8K, Jie et al. (2022) for SVAMP, and Lan et al. (2021) for MAWPS.
Arithmetic Reasoning

Ablation Study

- Mathematical equation?
  - -> Equation only
- Variable computation (i.e., intermediate tokens)?
  - -> Variable compute only
- Prompts allow the model to better access relevant knowledge acquired during pretraining
  - -> Chain of thought after answer

Figure 5: Ablation study for different variations of prompting using LaMDA 137B and PaLM 540B. Results for other datasets are given in Appendix Table 6 and Table 7.
Arithmetic Reasoning

Robustness

- Three different annotators
- An additional, more concise chain of thought following a specific style by Annotator A
- Three sets of eight exemplars randomly sampled from the GSM8K training set

Figure 6: Chain-of-thought prompting has variance for different prompt examples (as expected) but outperforms standard prompting for various annotators as well as for different exemplars.
Arithmetic Reasoning

Table 6: Ablation and robustness results for arithmetic reasoning datasets. Chain of thought generally outperforms ablations by a large amount. “Equation only” performs in between standard prompting and chain of thought prompting, as it allows for intermediate reasoning steps via equations but does not leverage natural language. Chain of thought prompting has variance (as expected) when used with prompts written by different annotators or when using other exemplars, but still outperforms standard prompting by a large margin. Standard deviation shown is for different order of few-shot prompting exemplars, with five different random seeds. Results here are shown for LaMDA 137B, as additional queries for GPT-3 and PaLM are both limited and expensive.

|                      | GSM8K  | SVAMP  | ASDiv  | MAWPS  |
|----------------------|--------|--------|--------|--------|
| Standard prompting   | 6.5 ±0.4 | 29.5 ±0.6 | 40.1 ±0.6 | 43.2 ±0.9 |
| Chain of thought prompting | 14.3 ±0.4 | 36.7 ±0.4 | 46.6 ±0.7 | 57.9 ±1.5 |
| Ablations            |        |        |        |        |
| • equation only      | 5.4 ±0.2 | 35.1 ±0.4 | 45.9 ±0.6 | 50.1 ±1.0 |
| • variable compute only | 6.4 ±0.3 | 28.0 ±0.6 | 39.4 ±0.4 | 41.3 ±1.1 |
| • reasoning after answer | 6.1 ±0.4 | 30.7 ±0.9 | 38.6 ±0.6 | 43.6 ±1.0 |
| Robustness           |        |        |        |        |
| • different annotator (B) | 15.5 ±0.6 | 35.2 ±0.4 | 46.5 ±0.4 | 58.2 ±1.0 |
| • different annotator (C) | 17.6 ±1.0 | 37.5 ±2.0 | 48.7 ±0.7 | 60.1 ±2.0 |
| • intentionally concise style | 11.1 ±0.3 | 38.7 ±0.8 | 48.0 ±0.3 | 59.6 ±0.7 |
| • exemplars from GSM8K (α) | 12.6 ±0.6 | 32.8 ±1.1 | 44.1 ±0.9 | 53.9 ±1.1 |
| • exemplars from GSM8K (β) | 12.7 ±0.5 | 34.8 ±1.1 | 46.9 ±0.6 | 60.9 ±0.8 |
| • exemplars from GSM8K (γ) | 12.6 ±0.7 | 35.6 ±0.5 | 44.4 ±2.6 | 54.2 ±4.7 |
Commonsense Reasoning

Experiment Setup

● Benchmarks
  ○ **CSQA**: Questions requiring deep commonsense knowledge about the world.
  ○ **StrategyQA**: Demands reasoning over multiple steps to derive answers.
  ○ **BIG-bench (Date and Sports Understanding)**: Tests abilities in context-specific date inference and plausibility assessments in sports contexts.
  ○ **SayCan**: Involves translating natural language instructions into robotic actions.

● Prompts
  ○ Similar approach as previous sections with manual composition of chain-of-thought prompts for few-shot learning.
  ○ Utilization of both predefined training examples and first-seen examples in evaluations to assess model generalization and reasoning capabilities.
Commonsense Reasoning

Figure 7: Chain-of-thought prompting also improves the commonsense reasoning abilities of language models. The language model shown here is PaLM. Prior best numbers are from the leaderboards of CSQA (Talmor et al., 2019) and StrategyQA (Geva et al., 2021) (single-model only, as of May 5, 2022). Additional results using various sizes of LaMDA, GPT-3, and PaLM are shown in Table 4.
Symbolic Reasoning

Experiment Setup

- **Specific Tasks Employed:**
  - **Last Letter Concatenation:** Models concatenate the last letters of names (e.g., "Amy Brown" to "yn").
  - **Coin Flip:** Models determine if a coin is heads up after a series of flips (e.g., after mixed actions by multiple people).

**Prompt for Last Letter Concatenation**
- **Q:** Take the last letters of the words in "Elon Musk" and concatenate them.
- **A:** The last letter of "Elon" is "n". The last letter of "Musk" is "k". Concatenating them is "nk". The answer is nk.

- **Q:** Take the last letters of the words in "Larry Page" and concatenate them.
- **A:** The last letter of "Larry" is "y". The last letter of "Page" is "e". Concatenating them is "ye". The answer is ye.

**Prompt for Coin Flip**
- **Q:** A coin is heads up. Ka flips the coin. Sherrie flips the coin. Is the coin still heads up?
- **A:** The coin was flipped by Ka and Sherrie. So the coin was flipped 2 times, which is an even number. The coin started heads up, so after an even number of flips, it will still be heads up. So the answer is yes.

- **Q:** A coin is heads up. Jamey flips the coin. Teresa flips the coin. Is the coin still heads up?
- **A:** The coin was flipped by Jamey and Teresa. So the coin was flipped 2 times, which is an even number. The coin started heads up, so after an even number of flips, it will still be heads up. So the answer is yes.
Symbolic Reasoning

Experiment Setup

- **Specific Tasks Employed:**
  - Last Letter Concatenation: Models concatenate the last letters of names (e.g., "Amy Brown" to "yn").
  - Coin Flip: Models determine if a coin is heads up after a series of flips (e.g., after mixed actions by multiple people).

- **Both in-domain and out-of-domain (OOD) test sets used:**
  - In-domain: Tasks with the same complexity as training examples.
  - OOD: Tasks with increased complexity compared to training examples.

Figure 8: Using chain-of-thought prompting facilitates generalization to longer sequences in two symbolic reasoning tasks.
Symbolic Reasoning

Table 5: Standard prompting versus chain of thought prompting enables length generalization to longer inference examples on two symbolic manipulation tasks.

| Model  | Last Letter Concatenation | Coin Flip (state tracking) |
|--------|---------------------------|-----------------------------|
|        | 2 standard | CoT standard | OOD: 3 standard | CoT standard | OOD: 4 standard | CoT standard | OOD: 3 standard | CoT standard | OOD: 4 standard |
| UL2 20B| 0.6 18.8   | 0.0 0.2      | 0.0 0.0         | 52.9 49.6    | 151.6 52.2      | 48.7 50.4    |
| LaMDA 420M | 0.3 1.6  | 0.0 0.0      | 0.0 0.0         | 52.9 49.6    | 150.0 50.5      | 49.5 49.1    |
| 2B     | 2.3 6.0    | 0.0 0.0      | 0.0 0.0         | 54.9 55.3    | 47.4 48.7       | 49.8 50.2    |
| 8B     | 1.5 11.5   | 0.0 0.0      | 0.0 0.0         | 52.9 55.5    | 48.2 49.6       | 51.2 50.6    |
| 68B    | 4.4 52.0   | 0.0 0.8      | 0.0 2.5         | 56.2 83.2    | 50.4 69.1       | 50.9 59.6    |
| 137B   | 5.8 77.5   | 0.0 34.4     | 0.0 13.5        | 49.0 99.6    | 50.7 91.0       | 49.1 74.5    |
| PaLM 8B| 2.6 18.8   | 0.0 0.0      | 0.0 0.2         | 60.0 74.4    | 47.3 57.1       | 50.9 51.8    |
| 62B    | 6.8 85.0   | 0.0 59.6     | 0.0 13.4        | 91.4 96.8    | 43.9 91.0       | 38.3 72.4    |
| 540B   | 7.6 99.4   | 0.2 94.8     | 0.0 63.0        | 98.1 100.0   | 49.3 98.6       | 54.8 90.2    |

UNIVERSITY OF TORONTO
Commonsense Reasoning
Symbolic Reasoning

Ablation & Robustness

Table 7: Ablation and robustness results for four datasets in commonsense and symbolic reasoning. Chain of thought generally outperforms ablations by a large amount. Chain of thought prompting has variance (as expected) when used with prompts written by different annotators or when using other exemplars, but still outperforms standard prompting by a large margin. Standard deviation shown is for different order of few-shot prompting exemplars, with five different random seeds. Results here are shown for LaMDA 137B, as additional queries for GPT-3 and PaLM are both limited and expensive. The exception is that we run SayCan using PaLM here, as the SayCan evaluation set is only 120 examples and therefore less expensive to run multiple times.

|                      | Commonsense |         | Symbolic |         |
|----------------------|-------------|---------|----------|---------|
|                      | Date        | Sports  | SayCan   | Concat  | Coin   |
| Standard prompting   | 21.5 ±0.6   | 59.5 ±3.0 | 80.8 ±1.8 | 5.8 ±0.6 | 49.0 ±2.1 |
| Chain of thought prompting | 26.8 ±2.1 | 85.8 ±1.8 | 91.7 ±1.4 | 77.5 ±3.8 | 99.6 ±0.3 |
| Ablations            |             |         |          |         |
| · variable compute only | 21.3 ±0.7  | 61.6 ±2.2 | 74.2 ±2.3 | 7.2 ±1.6 | 50.7 ±0.7 |
| · reasoning after answer | 20.9 ±1.0  | 63.0 ±2.0 | 83.3 ±0.6 | 0.0 ±0.0 | 50.2 ±0.5 |
| Robustness           |             |         |          |         |
| · different annotator (B) | 27.4 ±1.7  | 75.4 ±2.7 | 88.3 ±1.4 | 76.0 ±1.9 | 77.5 ±7.9 |
| · different annotator (C) | 25.5 ±2.5  | 81.1 ±3.6 | 85.0 ±1.8 | 68.1 ±2.2 | 71.4 ±11.1 |
Scope and Limitation

**Broad Applicability:** The research demonstrates the effectiveness of chain-of-thought prompting across a range of tasks, from commonsense and symbolic reasoning to arithmetic problem-solving. This approach significantly improves model performance, particularly in tasks that require multi-step reasoning or complex thought processes.

**Model Scalability:** The results indicate that chain-of-thought prompting benefits from increased model size, with larger models like PaLM and GPT-3 showing remarkable improvements in solving capabilities, suggesting scalability is a key factor in its success.
Scope and Limitation

**Nature of Reasoning**: Unclear if the neural network truly "reasons" like humans.

**Annotation Costs**: Manually augmenting exemplars is expensive for fine tuning.

**Accuracy of Reasoning Paths**: No guarantee of correct reasoning, leading to both correct and incorrect outputs, highlighting the need for improved factual accuracy in future work.

**Scalability Issues**: Effective chain-of-thought reasoning currently requires large models, which are costly; further research is needed to enable reasoning in smaller models.

**Number of Exemplars**: Increasing the number of few-shot exemplars does not consistently lead to performance gains, indicating a potential plateau in effectiveness beyond a certain point.

**Transferability of Gains**: Gains from chain-of-thought prompting do not always transfer perfectly among different models, which raises questions about how pre-training datasets and model architectures influence performance gains.
Code Notebook and Visual Demonstration

https://colab.research.google.com/drive/1rZp4EQOIaBf6qhygCfaYJ8BN6nCX-UW?usp=sharing
Conclusion

- **Methodology**: The paper explored chain-of-thought prompting as an effective method to enhance reasoning capabilities in language models.
- **Findings**: Experiments across arithmetic, symbolic, and commonsense reasoning reveal that chain-of-thought reasoning is an emergent property linked to model scale, enabling large models to excel in tasks with previously flat scaling curves.
- **Implications**: This advancement broadens the range of reasoning tasks manageable by language models, paving the way for further research into language-based reasoning methods.