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

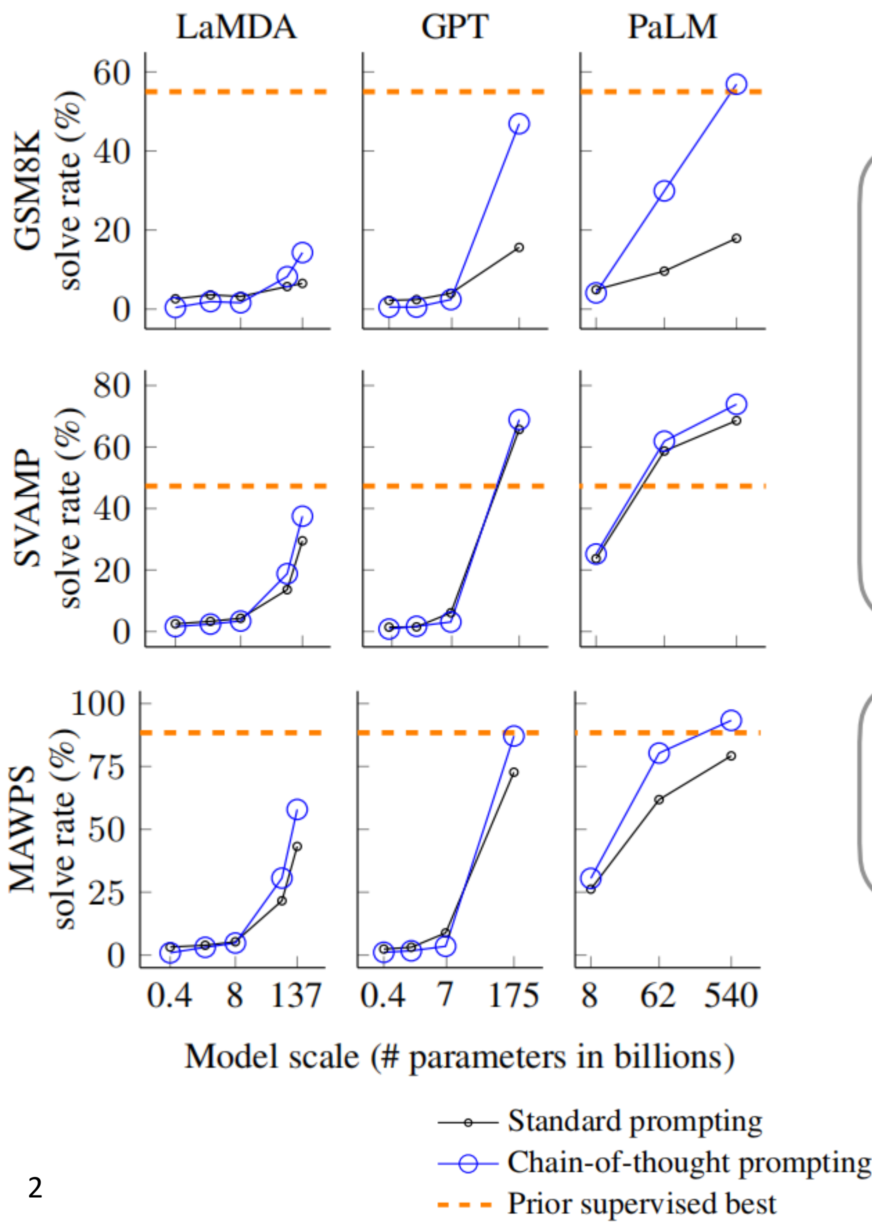
Jason Wei **Xuezhi Wang** **Dale Schuurmans** **Maarten Bosma**
Brian Ichter **Fei Xia** **Ed H. Chi** **Quoc V. Le** **Denny Zhou**

Google Research, Brain Team
{jasonwei,dennyzhou}@google.com

- Introduce "Chain of Thought" as a way of prompt engineering
- Published in Jan 2022
- Cited **10,505 times** (as of Feb 4, 2025)

  10,719 times (as of Feb 5, 2025)

Overview



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. ✓

THE UNIVERSITY OF ARIZONA

Key Idea in Chain of Thought (CoT) Prompting

Motivations

- ***(Interpretability)*** Arithmetic reasoning ability can benefit from natural language rationales, but costly to create large sets of high-quality rationales.
- ***(Prompt)*** LLMs offer prospect of in-context few-shot learning via prompting, which works poorly on tasks that require reasoning , and doesn't scale well.

CoT

- Input prompts instruct the model to break down complex problems into logical steps, mimicking human reasoning.
- The output generates intermediate reasoning steps explicitly during inference.
- Adapt large language models (LLMs) to complex reasoning tasks.

Previous Prompting

- Input prompts directly ask for the final answer without logical breakdowns.
- The intermediate reasoning steps based on learned distributions are implicit and hidden.
- Performs poorly on tasks requiring multi-step reasoning.

Methods of CoT Prompting

PROMPT FOR MATH WORD PROBLEMS

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?

A: There are 15 trees originally. Then there were 21 trees after some more were planted. So there must have been 21 - 15 = 6. The answer is 6.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

A: There are originally 3 cars. 2 more cars arrive. 3 + 2 = 5. The answer is 5.

Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?

A: Originally, Leah had 32 chocolates. Her sister had 42. So in total they had 32 + 42 = 74. After eating 35, they had 74 - 35 = 39. The answer is 39.

Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?

A: Jason started with 20 lollipops. Then he had 12 after giving some to Denny. So he gave Denny 20 - 12 = 8. The answer is 8.

Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?

A: Shawn started with 5 toys. If he got 2 toys each from his mom and dad, then that is 4 more toys. 5 + 4 = 9. The answer is 9.

Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?

A: There were originally 9 computers. For each of 4 days, 5 more computers were added. So 5 * 4 = 20 computers were added. 9 + 20 is 29. The answer is 29.

Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?

A: Michael started with 58 golf balls. After losing 23 on tuesday, he had 58 - 23 = 35. After losing 2 more, he had 35 - 2 = 33 golf balls. The answer is 33.

Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?

A: Olivia had 23 dollars. 5 bagels for 3 dollars each will be 5 x 3 = 15 dollars. So she has 23 - 15 dollars left. 23 - 15 is 8. The answer is 8.

- 8 exemplars are provided in sequence before the test question.
- Each exemplar includes: A question (task statement); a chain of reasoning: A natural language explanation detailing intermediate steps; a final answer.
- The exemplars are straight - forward and manually constructed without prompt engineering.
- Most of the datasets under the same task share the same exemplar sets.

Methods of CoT Prompting

Table 21: Few-shot exemplars for full chain of thought prompt for AQuA.

PROMPT FOR AQUA ALGEBRAIC WORD PROBLEMS

Q: John found that the average of 15 numbers is 40. If 10 is added to each number then the mean of the numbers is?
Answer Choices: (a) 50 (b) 45 (c) 65 (d) 78 (e) 64

A: If 10 is added to each number, then the mean of the numbers also increases by 10. So the new mean would be 50. The answer is (a).

Q: If a / b = 3/4 and 8a + 5b = 22,then find the value of a.
Answer Choices: (a) 1/2 (b) 3/2 (c) 5/2 (d) 4/2 (e) 7/2

A: If a / b = 3/4, then b = 4a / 3. So 8a + 5(4a / 3) = 22. This simplifies to 8a + 20a / 3 = 22, which means 44a / 3 = 22. So a is equal to 3/2. The answer is (b).

Q: A person is traveling at 20 km/hr and reached his destiny in 2.5 hr then find the distance?
Answer Choices: (a) 53 km (b) 55 km (c) 52 km (d) 60 km (e) 50 km

A: The distance that the person traveled would have been 20 km/hr * 2.5 hrs = 50 km. The answer is (e).

Q: How many keystrokes are needed to type the numbers from 1 to 500?
Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788

A: There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500. 9 + 90(2) + 401(3) = 1392. The answer is (b).

- 8 exemplars are provided in sequence before the test question.
- Each exemplar includes: A question (task statement); a chain of reasoning: A natural language explanation detailing intermediate steps; a final answer.
- The exemplars are straight - forward and manually constructed without prompt engineering.
- Most of the datasets under the same task share the same exemplar sets.

THE UNIVERSITY OF ARIZONA

Methods of CoT Prompting

PROMPT FOR STRATEGYQA

Q: Do hamsters provide food for any animals?

A: Hamsters are prey animals. Prey are food for predators. Thus, hamsters provide food for some animals. So the answer is yes.

Q: Could Brooke Shields succeed at University of Pennsylvania?

A: Brooke Shields went to Princeton University. Princeton University is about as academically rigorous as the University of Pennsylvania. Thus, Brooke Shields could also succeed at the University of Pennsylvania. So the answer is yes.

Q: Yes or no: Hydrogen's atomic number squared exceeds number of Spice Girls?

A: Hydrogen has an atomic number of 1. 1 squared is 1. There are 5 Spice Girls. Thus, Hydrogen's atomic number squared is less than 5. So the answer is no.

Q: Yes or no: Is it common to see frost during some college commencements?

A: College commencement ceremonies can happen in December, May, and June. December is in the winter, so there can be frost. Thus, there could be frost at some commencements. So the answer is yes.

Q: Yes or no: Could a llama birth twice during War in Vietnam (1945-46)?

A: The War in Vietnam was 6 months. The gestation period for a llama is 11 months, which is more than 6 months. Thus, a llama could not give birth twice during the War in Vietnam. So the answer is no.

Q: Yes or no: Would a pear sink in water?

A: The density of a pear is about $0.6g/cm^3$, which is less than water. Objects less dense than water float. Thus, a pear would float. So the answer is no.

- 8 exemplars are provided in sequence before the test question.
- Each exemplar includes: A question (task statement); a chain of reasoning: A natural language explanation detailing intermediate steps; a final answer.
- The exemplars are straight -forward and manually constructed without prompt engineering.
- Most of the datasets under the same task share the same exemplar sets.

Experiments Setup

Models

- **GPT-3** (350M, 1.3B, 6.7B, 175B)
- **LaMDA** (422M, 2B, 8B, 68B, 137B)
- **PaLM** (8B, 61B, 540B)
- **UL2** (20B)
- **Codex**

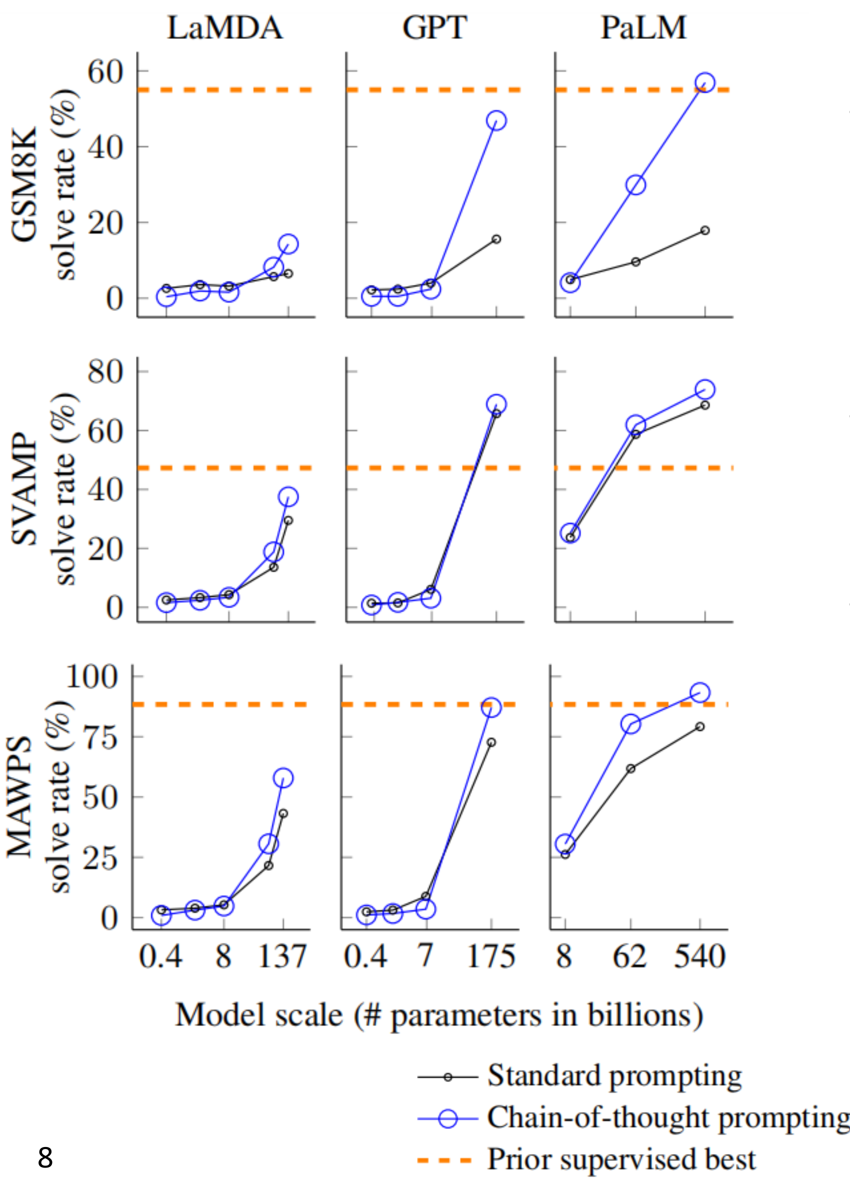
Prompts

- **Standard few-shot prompting**
  model is given in-context exemplars of input–output pairs
- **Chain-of-thought prompting**
  augment each exemplar in few-shot prompting with a chain of thought for an associated answer

Benchmarks

- **GSM8K**: benchmark of math word problems
- **SVAMP**: dataset of math word problems with varying structures
- **ASDiv:** dataset of diverse math word problems
- **AQuA:** dataset of algebraic word problems
- **MAWPS**

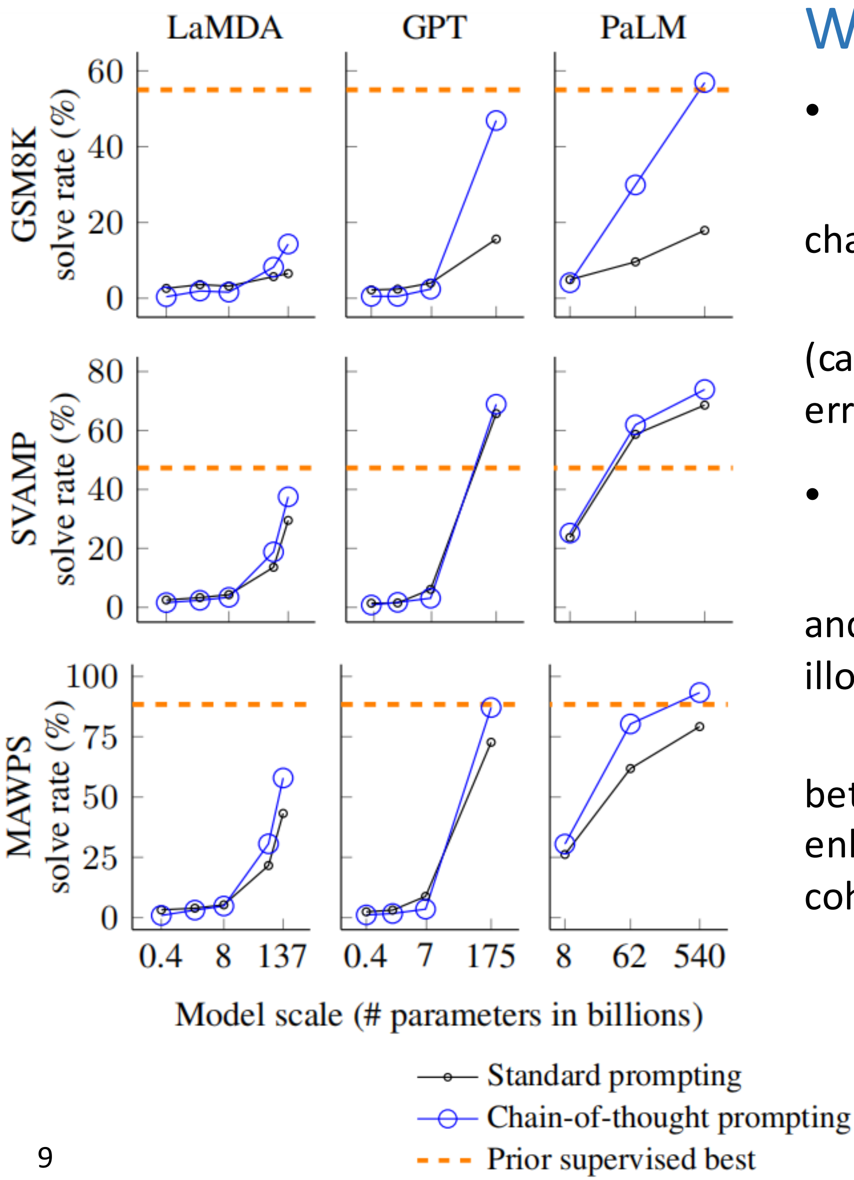
Results (1): Solving Math Problems



Observations:

- **Emergent ability:** CoT prompting does not positively impact performance for small models, and only yields performance gains when used with models of ∼100B parameters.
- **Greater Gains on Complex Tasks:** CoT prompting has larger performance gains for more complicated problems.
- **State of the art:** CoT prompting on large models (GPT-3 175B and PaLM 540B) compares favorably to previous best works based on task-specific finetuning.

Results (1): Solving Math Problems



Why?

- **Error analysis:**

  - 96% of correct answers had logically and valid chains; only 4% were coincidental.

  - Among incorrect answers: 46% had minor errors (calculator mistakes, missing steps); 54% had major errors (semantic misunderstand or incoherent reasoning).

- **Scaling Benefits:**

  - Small models fail at symbol mapping, arithmetic, and coherent reasoning, often producing fluent but illogical CoT outputs.

  - Larger models, with greater parameter capacity and better data utilization, capture complex patterns, enhance semantic understanding, and produce more coherent outputs, significantly improving reasoning tasks.

THE UNIVERSITY OF ARIZONA

Results (2): Ablation Study

Variations:

- **Equation only:** the model is prompted to output only a mathematical equation before giving the answer.
- **Variable compute only**: model is prompted to output a only sequence of dots (...) equal to the number of characters in the equation needed to solve the problem.
- **CoT after answer**: the CoT prompt is given after the answer, isolating whether the model actually depends on the produced CoT to give the final answer.

	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

THE UNIVERSITY OF ARIZONA

Results (2): Ablation Study

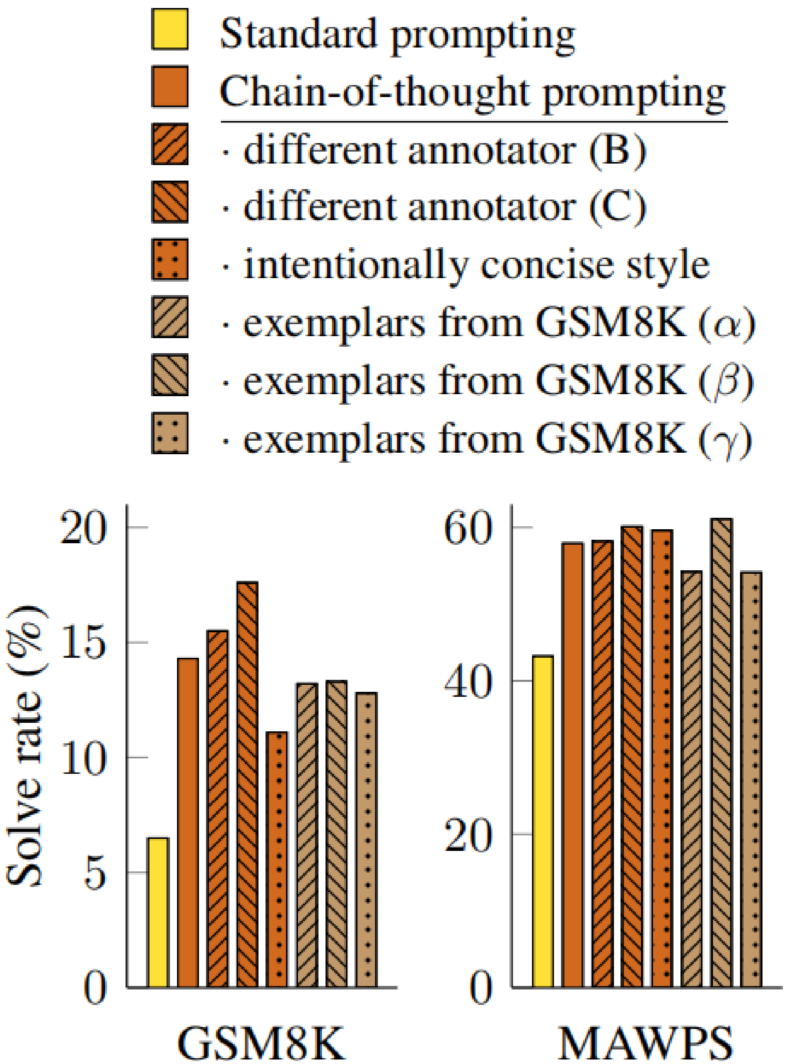
Observation:

- **Equation only:** Performs well on simpler datasets (e.g., SVAMP) but poorly on complex ones (e.g., GSM8K) → *Natural language reasoning is essential for semantically challenging problems.*
- **Variable compute only**: No significant improvement → *The success of CoT depends on reasoning steps, not computational effort.*
- **CoT after answer**: Similar to baseline performance → *Reasoning before the answer is critical for guiding logical thought processes.*

	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

THE UNIVERSITY OF ARIZONA

Results (3): Robustness



Sensitivity consideration:

Performance varies significantly based on exemplar design, with accuracy ranging from 54.3% to 93.4% depending on exemplar variations (Zhao et al., 2021).

Factors:

- **Different annotators:**. 3 annotators' own linguistic style without writing instructions.
- **Annotators without machine learning background:** GSM8K dataset, written by crowd compute workers.
- **Different exemplars**
- **Different order of exemplars**
- **Different number of exemplars**
- **Different language models**

THE UNIVERSITY OF ARIZONA

Results (3): Robustness

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?

A: There are 15 trees originally. Then there were 21 trees after some more were planted. So there must have been 21 - 15 = 6. The answer is 6.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

A: There are originally 3 cars. 2 more cars arrive. 3 + 2 = 5. The answer is 5.

Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?

A: Originally, Leah had 32 chocolates. Her sister had 42. So in total they had 32 + 42 = 74. After eating 35, they had 74 - 35 = 39. The answer is 39.

Annotator A

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?

A: There are 21 trees now and there are 15 trees in the beginning, so the workers plant 21 - 15 = 6 trees. The answer is 6.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

A: There are 3 cars in the beginning, 2 more arrive, so now there should be 3 + 2 = 5 cars. The answer is 5.

Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?

A: Leah had 32 chocolates and her sister had 42, in total they have 32 + 42 = 74 chocolates. After they ate 35, now there are 74 - 35 = 39 chocolates. The answer is 39.

Annotator B

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?

A: We start with 15 trees. Later we have 21 trees. The difference must be the number of trees they planted. So, they must have planted 21 - 15 = 6 trees. The answer is 6.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

A: There are 3 cars in the parking lot already. 2 more arrive. Now there are 3 + 2 = 5 cars. The answer is 5.

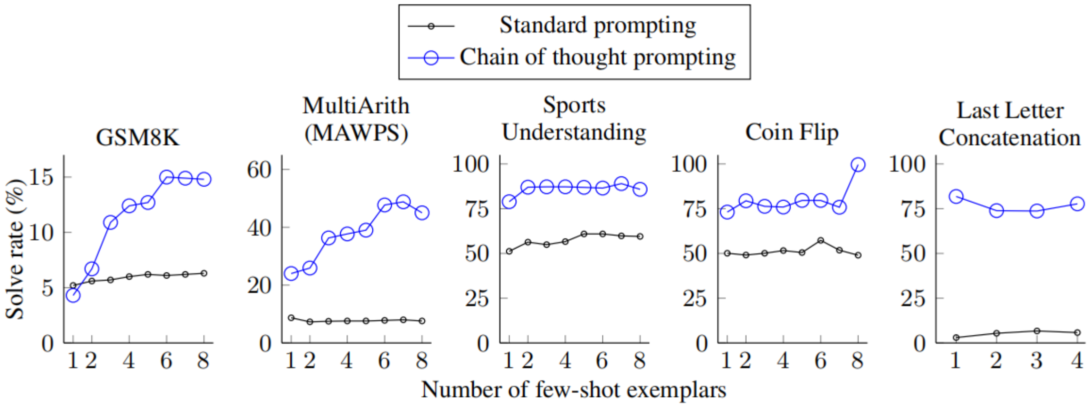
Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?

A: Leah had 32 chocolates and Leah's sister had 42. That means there were originally 32 + 42 = 74 chocolates. 35 have been eaten. So in total they still have 74 - 35 = 39 chocolates. The answer is 39.

Annotator C

THE UNIVERSITY OF ARIZONA

Results (3): Robustness



	GSM8K	SVAMP	ASDiv	MAWPS
Standard prompting	6.5 $\pm$0.4	29.5 $\pm$0.6	40.1 $\pm$0.6	43.2 $\pm$0.9
Chain of thought prompting	14.3 $\pm$0.4	36.7 $\pm$0.4	46.6 $\pm$0.7	57.9 $\pm$1.5
Robustness				
· different annotator (B)	15.5 $\pm$0.6	35.2 $\pm$0.4	46.5 $\pm$0.4	58.2 $\pm$1.0
· different annotator (C)	17.6 $\pm$1.0	37.5 $\pm$2.0	48.7 $\pm$0.7	60.1 $\pm$2.0
· intentionally concise style	11.1 $\pm$0.3	38.7 $\pm$0.8	48.0 $\pm$0.3	59.6 $\pm$0.7
· exemplars from GSM8K (α)	12.6 $\pm$0.6	32.8 $\pm$1.1	44.1 $\pm$0.9	53.9 $\pm$1.1
· exemplars from GSM8K (β)	12.7 $\pm$0.5	34.8 $\pm$1.1	46.9 $\pm$0.6	60.9 $\pm$0.8
· exemplars from GSM8K (γ)	12.6 $\pm$0.7	35.6 $\pm$0.5	44.4 $\pm$2.6	54.2 $\pm$4.7

Observations:

- **Robust Across Annotators:**

  - All annotations outperform standard prompting → *CoT is independent from linguistic style*.

  - Variance among annotations → *Well-designed prompts for specific tasks are critical.*

- **Robust Across Exemplars:**

  CoT prompting works well with different sets of exemplars.

- **Robust Across Order and Quantity:**

  Performance remains stable across changes in exemplar order and number.

THE UNIVERSITY OF ARIZONA

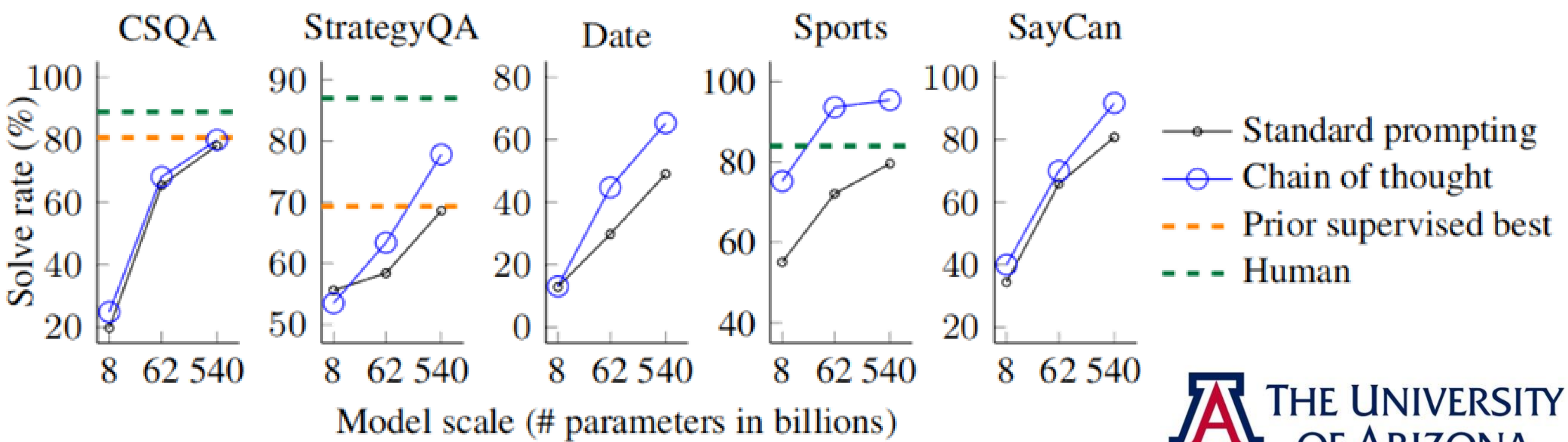
Results (4): Generalization

Commonsense Reasoning:

- **Benchmarks:**
  - **CSQA:** Requires prior knowledge for answering complex semantic questions.
  - **StrategyQA**: Multi-hop reasoning for strategy answers.
  - **BIG-bench(Date):** Inferring dates from context.
  - **BIG-bench(Sports):** Determining plausibility of sports-related sentences.
  - **SayCan**: Translating natural language instructions into robot actions.

- **Observations:**

Scaling up model size and using chain-of-thought prompting significantly improved performance, especially for large models like PaLM 540B.



Results (4): Generalization

Symbolic Reasoning:

- **Tasks:**

  - **Last letter concatenation:** concatenate the last letters of words in a name.

    e.g., "Amy Brown" → "yn"

  - Coin flip: Simulate flipping coins based on context.

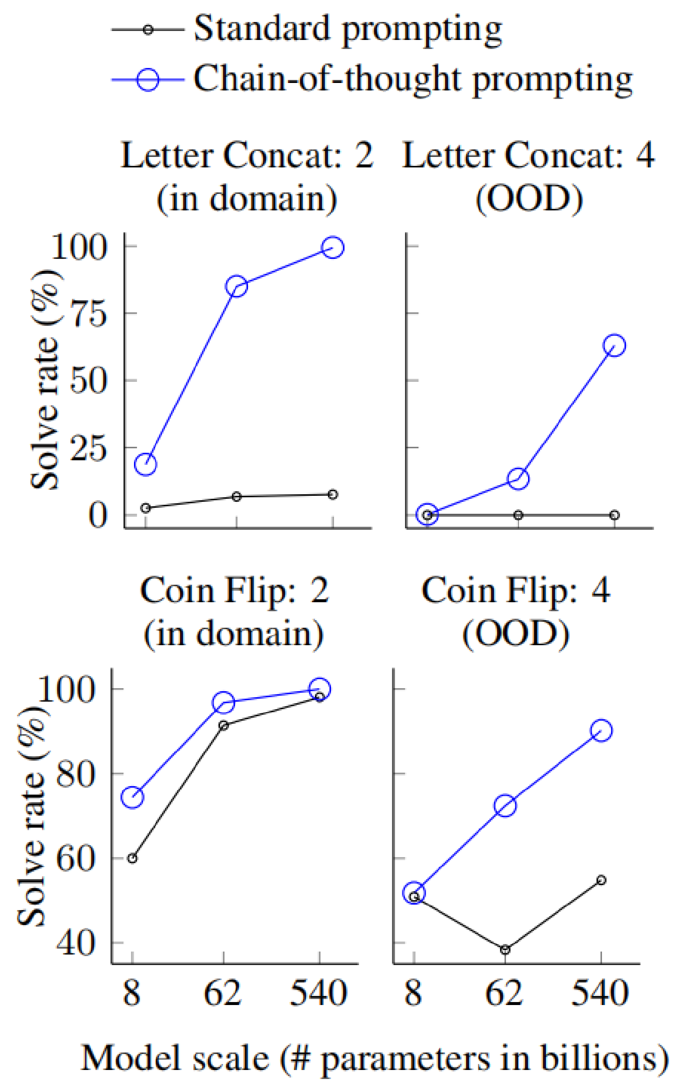
    e.g., "A coin is heads up. Phoebe flips the coin. Osvaldo does not flip the coin. Is the coin still heads up?"→ "no"

- **Observations:**

  - CoT prompting improves symbolic reasoning tasks, outperforming standard prompting significantly.

  - Enables length generalization, allowing models to handle inputs longer than those in few-shot exemplars.

  - Larger models benefit most from CoT, particularly in OOD tasks.



Discussions (1): How Does CoT Work?

- **Information Decomposition:** CoT breaks down complex problems into intermediate reasoning steps, aligning with the model's attention mechanism and simplifying the solution process.
- **Explicit Activation of Implicit Knowledge:** Large models learn implicit reasoning patterns during pretraining. CoT prompts reveal and utilize these patterns explicitly through natural language reasoning.
- **Natural Language as Universal Representation:** CoT relies on natural language reasoning, which may serve as an optimal format for guiding logical and semantic processes in models.

Discussions (2): What Remains Unanswered?

- **Mechanisms of CoT:** How does CoT influence attention and token dependency during inference? Does CoT activate new reasoning abilities or merely refine pre-existing ones?
- **Interpretability:** How does CoT align with human cognitive reasoning? Can CoT prompts be designed to ensure causal, interpretable reasoning chains?

Thank you

Yuxin Ren

2025 Feb 6

THE UNIVERSITY OF ARIZONA